University of Gloucestershire
School of Business Computing and Social Sciences
Doctor of Philosophy

A Hybrid Approach to Intelligent Prediction of Medical Conditions

A Framework for Advancing Medical Diagnostics through Novel Hybrid Deep Learning Models

DenCeption and HyBoost for Enhanced Feature Extraction and Predictive Accuracy in Medical Image Analysis

Zainab Loukil

A thesis submitted to the University of Gloucestershire in accordance with the requirements of the degree of PhD in the School of Business Computing and Social Sciences.

Date: October 30, 2024

Word Count : 83,082

To my lovely husband Ahmed, thanks for lighting up my world, this is for you.
To my dear parents and their unconditional love and prayers.
To our precious baby on the way, you are already a source of inspiration and joy. This journey is as much for you as it is for us ...

Declaration of Authorship

I, Zainab Loukil, declare that the work in this thesis was carried out in accordance with the regulations of the University of Gloucestershire and is original except where indicated by a specific reference in the text. No part of this thesis has been submitted as part of any other academic award. The thesis has not been presented to any other educational institution in the United Kingdom or abroad.

Any views expressed in this thesis are those of the author and in no way represent those of the University of Gloucestershire.

Signed:

DOI: 10.46289/8XU8FE24

Date: 30/10/2024

Abstract

Medical image analysis is currently challenged by the need to achieve precision in diagnostic methods while maintaining broad applicability across diverse datasets. This challenge is intensified by the intricate details present in high-dimensional medical imaging data, affecting diagnostic effectiveness and patient care. Precise feature extraction is crucial in identifying patterns vital for medical diagnoses, yet current models often struggle with real-world variability, including diverse imaging conditions and patient demographics. This research advances the field by introducing a novel hybrid feature extraction framework, DenCeption, and a predictive model, HyBoost, which address these challenges through improved disease prediction accuracy, generalisability and adaptability. DenCeption, combining DenseNet-169 and Inception-V4 architectures, achieved a notable accuracy of 91.3%, surpassing existing models like DenseNet-121 (89.3%) on the BRATS MRI dataset, demonstrating its superior feature extraction capabilities. The hybrid feature extraction framework also proved adaptable across multiple datasets, including MRI and Retinal images, with accuracy reaching 98.9% in the Retinal dataset, significantly outperforming traditional methods. HyBoost, integrating multiple machine learning algorithms and leveraging patient demographic and physiological data, further enhances predictive accuracy. For instance, on the OCT dataset, HyBoost achieved an accuracy of 98.33%, with a sensitivity of 99.45%, outperforming existing models like XGBoost and AdaBoost. These improvements are supported by extensive testing across various datasets, Fundus, OCT, and X-ray, where HyBoost consistently demonstrated superior performance metrics, including low mean absolute error and high precision. Moreover, a new evaluation mechanism, involving a sophisticated performance measurement matrix (PMM), systematically selects the most optimal evaluation metrics, ensuring the robustness and clinical applicability of the models. This mechanism addresses the limitations of existing evaluation approaches, further enhancing the interpretability and reliability of the developed models. This research represents a significant advancement in medical image processing, setting new benchmarks in predictive medical imaging analytics. By systematically improving model per-

formance and integrating advanced machine learning and deep learning applications, this work revolutionises medical diagnostics, achieving high accuracy rates and robust disease prediction across multiple imaging modalities.

Acknowledgements

First and foremost, I extend my deepest gratitude to Allah for the countless blessings and the strength throughout my PhD journey. It is with His gracious support that I have been able to pursue and accomplish this significant academic milestone.

I am profoundly grateful to my supervisor, Dr. Qublai Khan Ali Mirza, for his invaluable guidance, patience, and expertise. Dr. Ali Mirza's mentorship has been a beacon of light, guiding me through the challenges of research with wisdom and encouragement. His unwavering support and insightful feedback have been instrumental in shaping my work.

I also wish to express my sincere appreciation to my second supervisor, Dr. William Sayers, for his constructive criticism and pivotal contributions to my research. Dr. Sayers' expertise and thoughtful advice have significantly enriched my academic and personal growth during this journey.

To my parents, who have always been my source of inspiration and support, I owe everything. Your unconditional love, sacrifices, and belief in my potential have been my stronghold. Your prayers and blessings have empowered me to strive for excellence in all my endeavors.

My deepest thanks go to my husband, whose love, understanding, and encouragement have been my constant companions. Your unwavering faith in me and your sacrifices have made this journey possible. Your partnership is my greatest treasure.

To my sister Sihem and my brother Mohammed Ali, thank you for your endless support, laughter, and care. Your presence and belief in my abilities have been a source of comfort and motivation. Sharing this journey with you has been one of my greatest joys.

To all of you, I am eternally grateful. This thesis is not only a reflection of my work but also a testament to the love and support that each of you has provided. Thank you for being part of my journey.

Publications

Journal Papers:

Loukil, Z., Mirza, Q.K.A., Sayers, W. and Awan, I., 2023. A Deep Learning based Scalable and Adaptive Feature Extraction Framework for Medical Images. Information Systems Frontiers, pp.1-27. https://doi.org/10.1007/s10796-023-10391-9.

Loukil, Z, Mirza, Q.K.A, and Sayers, W., 2024. Advancing Intelligent Medical Diagnostics with HyBoost: A Robust Predictive Framework for High-Dimensional Imaging Data. IEEE Access. Under Review.

Conference Papers:

Loukil, Z. and Salah, A.M., 2020, December. Toward Hybrid Deep Convolutional Neural Network Architectures For Medical Image Processing. In 2020 International Conference on Innovation and Intelligence for Informatics, Computing and Technologies (3ICT) (pp. 1-6). IEEE.

Loukil, Z., Mirza, Q.K.A. and Sayers, W., 2023, August. A Novel and Adaptive Evaluation Mechanism for Deep Learning Models in Medical Imaging and Disease Recognition. In 2023 10th International Conference on Future Internet of Things and Cloud (FiCloud) (pp. 270-277). IEEE.

Contents

Declaration of Authorship ii

Abstract iii

Acknowledgements v

Publications vii

List of Acronyms xxii

1 Introduction 1

1.1 Motivation 1

1.1.1 Shortcomings of Traditional Techniques in Case of Complex Medical Images 2

1.1.2 Transformative Impact of Deep Learning Approaches on Medical Imaging 3

1.1.3 Critical Role of Feature Extraction in Medical Image Analysis 6

1.1.4 Integration of Demographic and Physiological Data for Personalised Diagnostics 7

1.2 Problem Statement 8

1.3 Aim 9

1.4 Objectives 9

1.5 Contributions 10

1.6 Scope . . . 15
1.7 Thesis Structure . . . 16

2 Literature Review **18**

2.1 Introduction . . . 18
2.2 Research Background . . . 19
2.2.1 Neural Networks: A Focus on Convolutional Neural Networks . . . 19
2.2.2 ML Classifiers . . . 65
2.2.3 Medical Image Modalities . . . 69
2.2.4 Performance Evaluation Metrics . . . 72
2.3 Advancements in Learning-based Segmentation Techniques for Medical Image Processing . . . 78
2.4 Hybrid Models Versus Singular Approaches . . . 86
2.4.1 Hybrid Methodologies in Medical Image Analysis: Bridging Computational Techniques for Enhanced Performance . . . 86
2.4.2 Advancing Medical Diagnostics with Hybrid Computational Models . . . 91
2.5 Progression of Disease Detection and Classification Techniques . . . 117
2.5.1 The impact of Learning-based Approaches on DMO Disease Classification . . . 123
2.5.2 Advancements in Learning-based Approches for DR Detection . . . 134
2.5.3 Progression of Analysis Techniques for Pulmonology Related Conditions 142
2.6 Identified Challenges . . . 144
2.7 Datasets Overview and Selection Rationale . . . 148
2.7.1 MRI Dataset . . . 149
2.7.2 Retinal Dataset . . . 150
2.7.3 Fundus Dataset . . . 151
2.7.4 OCT Dataset . . . 151
2.7.5 X-ray Dataset . . . 152

2.7.6 Dataset Usage for Training, Testing, and Validation 152
2.8 Datasets Selection Strategy 153
2.8.1 Generalisation and Robustness 153
2.8.2 Comprehensive Evaluation 153
2.8.3 Enhancing Model Capabilities 154
2.8.4 Applicability in Multi-Disease Diagnosis 154
2.9 Cross-Validation Method 154
2.9.1 10-Fold Cross-Validation Overview 154
2.9.2 Application of 10-Fold Cross-Validation in Each Chapter 155
2.9.3 Consistency and Clarity in Cross-Validation 156
2.10 The Testbed 156
2.11 Conclusion 158

3 DenCeption: A Novel Hybrid Deep Learning Based Model 160
3.1 Introduction 160
3.2 Related Works: Deeper and Wider Hybrid CNN Architectures 163
3.3 Identified Challenges 169
3.4 Proposed Hybrid Deep Learning Model: DenCeption 170
3.5 Rationale Behind the Selection of the DenCeption Model 178
3.5.1 Addressing Identified Challenges 179
3.5.2 Selection Criteria for Hybridisation 179
3.5.3 Comparative Analysis and Experimental Validation 180
3.5.4 Alignment with Research Goals 180
3.5.5 Justification for Model Components 180
3.6 Conducted Experiments 181
3.7 Dataset 187
3.7.1 Benchmarking and Validation 187
3.7.2 Real-World Clinical Relevance 187

3.7.3 Complexity in Data Structure . . . 187
3.7.4 Strategic Use in Early Model Development . . . 188
3.7.5 Contribution to Model Robustness . . . 188
3.8 Results and Discussion . . . 188
3.8.1 Training and Testing Results: DenCeption Versus DenCeption Variants and Benchmarking Methods . . . 188
3.8.2 Validation Results: DenCeption Versus Variants and Benchmarking Methods . . . 198
3.9 Conclusion . . . 202

4 A Deep Learning based Scalable and Adaptive Feature Extraction Framework for Medical Images 205

4.1 Introduction . . . 205
4.2 Related Works . . . 209
4.3 Problems Identified . . . 215
4.4 Methodology of the Proposed Features Extraction Model . . . 216
4.4.1 Image Pre-Processing . . . 219
4.4.2 Image Segmentation and Associated Parameters . . . 221
4.4.3 Proposed Hybrid High-Level Features Extraction Model . . . 223
4.4.4 Deep Hidden Features Extraction Method . . . 234
4.4.5 Features Weighting . . . 235
4.4.6 Features Fusion . . . 237
4.5 Dataset . . . 239
4.5.1 BRATS MRI Dataset . . . 239
4.5.2 Retinal Dataset . . . 241
4.5.3 Combined Justification . . . 242
4.6 Conducted Experimentation and Research Evaluation Mechanism . . . 243
4.7 Results and Discussion . . . 245

4.7.1 Proposed Method Versus its Variants . . . 245
4.7.2 Benchmarking Methods . . . 256
4.8 Conclusion . . . 266

5 Advancing Intelligent Medical Diagnostics with HyBoost: A Robust Predictive Framework for High-Dimensional Imaging Data 268

5.1 Introduction . . . 268
5.2 Research Contributions . . . 273
5.3 Related Works . . . 274
5.3.1 Identified Challenges . . . 280
5.4 Proposed Prediction Framework . . . 282
5.4.1 Block 1: Image Pre-Processing . . . 282
5.4.2 Block 2: Features Extraction . . . 283
5.4.3 Block 3: Additional Features Incorporation . . . 284
5.4.4 Block 4: Prediction . . . 285
5.4.5 Block 5: Shapley Additive Explanation . . . 287
5.4.6 Block 6: Adaptive Performance Evaluation . . . 289
5.5 Rationale Behind the Selection of the HyBoost Model . . . 291
5.5.1 Addressing Identified Challenges in Disease Prediction . . . 293
5.5.2 Justification for XGBoost . . . 293
5.5.3 Justification for AdaBoost . . . 293
5.5.4 Integration of XGBoost and AdaBoost in HyBoost . . . 294
5.5.5 Practical Benefits of HyBoost . . . 294
5.5.6 Alignment with Research Goals . . . 294
5.6 Datasets . . . 295
5.6.1 Fundus Dataset as the Primary Dataset . . . 295
5.6.2 OCT and X-ray Datasets as Secondary Datasets: Selection Justification 296
5.6.3 Rationale for Using Different Datasets . . . 297

5.6.4 Visualisation of Additional Parameters Across Datasets: Focus on Fundus and OCT Datasets . . . 297
5.7 Experimentation Scenarios . . . 300
5.8 Features Extraction and Sample Testing . . . 304
5.8.1 Data Preparation: Image Pre-Processing Outcome . . . 304
5.8.2 Features Extraction and Selection . . . 305
5.9 Prediction Results and Discussion . . . 317
5.9.1 Baseline Scenario Results . . . 317
5.9.2 Enhanced Scenario Results: Validation on Selected Prediction Models . 338
5.9.3 Benchmarking: Comparative and Critical Discussion . . . 345
5.10 Conclusion . . . 364

6 Conclusion and Future Work **365**
6.1 Performance Overview of the Research Contributions . . . 366
6.2 Research Contributions Against Chosen Datasets . . . 369
6.2.1 Contribution 1: Design of a Novel DL-Based Hybrid Model (DenCeption) Against MRI Dataset . . . 370
6.2.2 Design of an Adaptive and Scalable Features Extraction Framework Against MRI and Retinal Datasets . . . 371
6.2.3 Design of a Novel Evaluation Mechanism for DL Models Against all Datasets . . . 372
6.2.4 Design of an Intelligent and Robust Predictive Framework Against Fundus, OCT and X-ray Datasets . . . 373
6.3 Limitations and Challenges . . . 374
6.4 Future Work . . . 375

A Appendices: Algorithms **379**
A.1 Algorithm 1 - Texture Features Extraction . . . 380
A.2 Algorithm 2 - Shape and Colour Features Extraction . . . 381

A.3 Algorithm 3 - Optimal Performance Evaluation Metrics 382
A.4 Algorithm 4 - Algorithm for HyBoost Hybrid Predictive Model 383

B Appendices: Code **384**

B.1 Pre-processing and Segmentation - Features Extraction Framework 384
B.2 High Level Features Extraction: Texture, Shape, Colour Features 386
B.3 Deep Hidden Features Extraction: DenCeption Model 390
B.4 Features Weighting 395
B.5 Features Fusion 395
B.6 Classification Block 397
B.7 Proposed Features Extraction Framework: Full Pipeline Execution 397
B.8 Image Pre-processing: Prediction Framework 398
B.9 Proposed HyBoost Predictive Model 400
B.10 Performance Measurement Matrix 404
B.11 SHAP Analysis 407

List of Tables

2.1 Neural Network Methods in Image Registration 22

2.2 Neural Network Methods in Image Segmentation and Edge Detection 23

2.3 Neural Network Methods in CAD 25

2.4 Automated Segmentation Techniques 27

2.5 Comparative Analysis of Deep Learning Architectures 28

2.6 Comparison of Traditional Versus Deep Learning Architecture for Automatic Localisation (El-Shafai, Abd El-Samie, Soliman, and Mostafa, 2024) 33

2.7 Generalised Paradigm for Multiple-Lesion Recognition (Jiang et al., 2023; Khan, Sohail, Zahoora, and Qureshi, 2020; McNeely-White, Beveridge, and Draper, 2020) 42

2.8 Multiple-lesion Recognition in Different Body Regions (Ananda et al., 2021; Cuevas-Rodriguez et al., 2023; Krizhevsky, Sutskever, and Hinton, 2012) . . . 43

2.9 Overview of Additional CNN Architectures: Advantages and Disadvantages . . 54

2.10 ML Models Working Principles and Mathematical Representation 65

2.11 Medical Images Characteristics: MRI, OCT, Fundus, and X-ray 71

2.12 Comparative Evaluation of Performance Metrics 76

2.13 Comparative Evaluation of Loss Metrics 77

2.14 Comparative Analysis of Hybrid Models: Advantages and Disadvantages . . . 104

2.15 ML and DL Applications in DMO Related Prediction Literature 124

2.16 Resnet-50 Outperformance in DR Classification Task (Athira and Nair, 2023) . 141

2.17 Summary of Datasets Used in Research 148

2.18 System Specifications, Operating System, and Drivers 157

3.1 ResNet Network: Advantages and Disadvantages (Xu, Fu, and Zhu, 2023) . . . 164
3.2 DenseNet Network: Advantages and Disadvantages (Huang, Liu, Van Der Maaten, and Weinberger, 2017b) 166
3.3 Inception Networks Family: Advantages and Disadvantages (Szegedy, Ioffe, Vanhoucke, and Alemi, 2017) 167
3.4 InA Filter Numbers Per HDB Block 174
3.5 InB Filter Numbers Per HDB Block 175
3.6 InC Filter Numbers Per HTB Block 177
3.7 RA Filter Numbers Per HTB Block 178
3.8 RB Filter Numbers Per HTB Block 178
3.9 Hybrid DenCeption Variants Architecture 183
3.10 Training Results of DenCeption Versus Other Variants and Benchmarking Methods 189
3.11 Validating Results with Growth Rate k=32 199
3.12 Validating Results with Growth Rate k=24 201

4.1 Examples of Existing Texture Features Extraction Methods 230
4.2 Examples of Existing Shape Features Extraction Methods 231
4.3 Block 1 Variants - HF Fusions 244
4.4 Individual Blocks Variants Testing using Performance Metrics: BRATS Dataset 246
4.5 Integration Block Variant Testing using Performance Metrics: HF-DHF Fusion using BRATS Dataset 247
4.6 Block3 Variants Evaluation using RASR for BRATS Dataset 249
4.7 Individual Blocks Variants Testing using Performance Metrics: Retinal Dataset 251
4.8 Integration Block Variant Testing using Performance Metrics: HF-DHF Fusion using Retinal Dataset 252
4.9 Block3 Variants Evaluation using RASR for Retinal Dataset 253

4.10 Comparison of Benchmarking Feature Extraction Methods and Their Corresponding Imaging Data Type Coverage 261
4.11 Benchmarking Results of BRATS Dataset Versus Proposed Methods 264
4.12 Benchmarking Results of Retinal Dataset Versus Proposed Methods 265

5.1 HyBoost Major Blocks Description 288
5.2 Hyperparameters Overview 303
5.3 Datasets Heatmap Observations 308
5.4 SOM Key Components (Miljković, 2017) 316
5.5 Fundus-DR Prediction Results without Fine-tuning 318
5.6 Fundus-DR Prediction Results with Fine-tuning 318
5.7 OCT-DMO Prediction Results without Fine-tuning 326
5.8 OCT-DMO Prediction Results with Fine-tuning 326
5.9 X-ray without Fine-tuning 334
5.10 X-ray with Fine-tuning 335
5.11 AdaBoost Performance Metrics: Demographic and Physiological Features Case 338
5.12 XGBoost Performance Metrics: Demographic and Physiological Features Case 341
5.13 HyBoost Performance Metrics: Demographic and Physiological Features Case 343
5.14 Benchmarking Results Using Fundus Dataset 347
5.15 Benchmarking Results for X-ray Dataset 353
5.16 Benchmarking Results Using OCT Dataset 358

6.1 Summary Table of Contributions Performance Against Chosen Datasets 378

List of Figures

1.1 Medical Image Classification Process using Deep Learning. 4

2.1 Artificial Neural Networks Learning Process. 20

2.2 Data Augmentation Process where TF is the Applied Transformation Function (Garcea, Serra, Lamberti, and Morra, 2023) 35

2.3 Transfer Learning from ImageNet (Mukhlif, Al-Khateeb, and Mohammed, 2023). 38

2.4 Medical Images Considered in This Research: (a) Chest X-ray (CXR) Image, (b) Brain MRI Scan, (c) OCT Scan, and (d) Fundus Image. 70

2.5 Confusion Matrix Representation. 73

2.6 CASF-Net Architecture Proposed in (Zheng, Liu, Feng, Xu, and Zhao, 2023). . 80

2.7 LAEDNet Encoder Structure Proposed in (Zhou et al., 2022) 82

2.8 ci-DMO Prediction Process Using OCT and Related Fundus Scans (Varadarajan et al., 2020). 130

3.1 Typical CNN Architecture for Medical Image Analysis (Yang, Lan, Gao, and Gao, 2020) 161

3.2 Original Inception-A module (Szegedy, Ioffe, Vanhoucke, and Alemi, 2017). . 171

3.3 Original Inception-B module (Szegedy, Ioffe, Vanhoucke, and Alemi, 2017). . . 171

3.4 Original Inception-C module (Szegedy, Ioffe, Vanhoucke, and Alemi, 2017). . . 172

3.5 Original Reduction A module (Szegedy, Ioffe, Vanhoucke, and Alemi, 2017). . 172

3.6 Original Reduction B module (Szegedy, Ioffe, Vanhoucke, and Alemi, 2017). . 173

3.7 Proposed Hybrid DenCeption Architecture 173

3.8 Hybrid Dense Block Architecture . . . 173
3.9 DenCeption InA Module . . . 174
3.10 DenCeption InB Module . . . 174
3.11 Hybrid Transition Block Architecture . . . 175
3.12 DenCeption InC Module . . . 176
3.13 DenCeption RA Module . . . 177
3.14 DenCeption RB Module . . . 177
3.15 Testing Results: Top-1 Error of DenCeption Variants . . . 195
3.16 Testing Results: Top-1 Error of DenCeption Versus Benchmarking Methods . . 195

4.1 DL Impact On Medical Applications . . . 207
4.2 Proposed Features Extraction Methodology Framework . . . 216
4.3 Pre-Processing Result on FLAIR Modality . . . 221
4.4 MRF-EPM Segmentation: Test Done on a Scan Sample from the BRATS Dataset 222
4.5 Image Dimensionality Reduction Through Convolutional Layer where (a) Feature Width, (b) Feature Height, (c) Number of Channels . . . 235
4.6 Critical Sample Testing of Block3 - BRATS Dataset . . . 249
4.7 Critical Sample Testing of Block3 - Retinal Dataset . . . 254
4.8 ROC Curve of Block3 Variants Testing: (a) BRATS Dataset, (b) Retinal Dataset 256

5.1 Power of ML and DL in Healthcare Data Handling (Rahmani et al., 2021). . . . 270
5.2 DR Detection Using Non Referable and Referable Fundus Imgaes: DL Vs Specialists (Noriega et al., 2021). . . . 271
5.3 OCT Scan Dimensional Representation: (A) Axial Scan, (B) Cross-sectional Scan, and (C) OCT Volume (Khan, Sohail, Zahoora, and Qureshi, 2020). . . . 271
5.4 Automated Approach for X-ray Analysis (Ait Nasser and Akhloufi, 2023). . . . 272
5.5 Proposed Prediction Framework. . . . 282
5.6 Image Pre-processing Steps. . . . 283
5.7 HyBoost Blocks: Training, Testing and Re-training Phases. . . . 286

5.8 Optimal Parameters Selection Block of HyBoost Model. 287
5.9 Proposed Performance Evaluation Framework 289
5.10 Three-Dimensional Representation of PMM Matrix 291
5.11 Conversion of Two-Dimensional PMM Matrix into Optimal Set of Performance Measurement Metrics Vector 291
5.12 Age Distribution by Diabetic Type for - (a): Fundus Images Affected by DR and (b): OCT Scans Affected by DMO. 298
5.13 CRT Distribution by Diabetic Type for – (a): Fundus Images Affected by DR and (b): OCT Scans Affected by DMO. 300
5.14 Experimentation Process. 302
5.15 Image Pre-Processing Outcome. 304
5.16 Correlation Heatmap for Fundus Dataset. The HF feature are from left to right (x-axis) and top to bottom (y-axis) as follows: Texture Contrast, Texture Energy, Texture Homogeneity, Entropy, Coarseness, Directionality, BGR, mean_values_std, Shape Area, Shape perimeter. 306
5.17 Correlation Heatmap for OCT Dataset. The HF feature are from left to right (x-axis) and top to bottom (y-axis) as follows: Texture Contrast, Texture Energy, Texture Homogeneity, Entropy, Coarseness, Directionality, BGR, mean_values_std, Shape Area, Shape perimeter. 306
5.18 Correlation Heatmap for X-ray Pneumonia Dataset. The HF feature are from left to right (x-axis) and top to bottom (y-axis) as follows: Texture Contrast, Texture Energy, Texture Homogeneity, Entropy, Coarseness, Directionality, BGR, mean_values_std, Shape Area, Shape perimeter. 307
5.19 Correlation Heatmap Generated DHF Features for: (a) Fundus DR Dataset, (b) OCT DMO Dataset, (c) X-ray Pneumonia Dataset 313
5.20 Fundus AdaBoost ROC (a) PR (c) Curves without Hyperparameters Fine-tuning and ROC (b) PR (d) Curves with Hyperparameters Fine-tuning. 322

5.21 Fundus XGBoost ROC (a) PR (c) Curves without Hyperparameters Fine-tuning and ROC (b) PR (d) Curves with Hyperparameters Fine-tuning. 323
5.22 Fundus HyBoost ROC (a) PR (c) Curves without Hyperparameters Fine-tuning and ROC (b) PR (d) Curves with Hyperparameters Fine-tuning. 325
5.23 OCT AdaBoost ROC (a) PR (c) Curves without Hyperparameters Fine-tuning and ROC (b) PR (d) Curves with Hyperparameters Fine-tuning. 330
5.24 OCT XGBoost ROC (a) PR (c) Curves without Hyperparameters Fine-tuning and ROC (b) PR (d) Curves with Hyperparameters Fine-tuning. 331
5.25 OCT HyBoost ROC (a) PR (c) Curves without Hyperparameters Fine-tuning and ROC (b) PR (d) Curves with Hyperparameters Fine-tuning. 332
5.26 Confusion Matrix of AdaBoost for – (a): Fundus Dataset, (b): OCT Dataset and (c): X-ray Dataset. 339
5.27 Confusion Matrix of XGBoost for – (a): Fundus Dataset, (b): OCT Dataset and (c): X-ray Dataset. 341
5.28 Confusion Matrix of HyBoost for (a): Fundus Dataset, (b): OCT Dataset and (c): X-ray Dataset. 343
5.29 SHAP Values Explainer for (a): (Alryalat et al., 2022), (b): (Shimpi and Shanmugam, 2023), (c): Proposed Method. 360

List of Acronyms

Acronym	Definition
AI	Artificial Intelligence
ANN	Artificial Neural Network
AE	Autoencoder
AUC	Area Under the Curve
AMD	Age-related Macular Degeneration
AdaBoost	Adaptive Boosting
AUC-PR	Area Under the Precision-Recall curve
Acc	Accuracy
AP	Average Precision
AD	Alzheimer Disease
BM	Boltzmann Machine
BNN	Bayesian Belief Network
BM3D	Block Matching and 3Ds filtering
BMU	Best Matching Unit
BGR	standard deviations of RGB colours
CNN	Convolutional Neural Network
CRT	Central Retinal Thickness
CAD	Computer-Aided Diagnosis
CT	Computed Tomography
CapsNet	Capsule Neural Network
CASF-Net	Cross-attention and Cross-scale Fusion Network
CMRI	Cardiac MRI Images
CHI-Net	Context Hierarchical Integrated Network

Continued on next page

Acronym	Definition
CRF-RNN	Conditional Random Field as a RNN
C-SVM	Cascade SVM
ConvNet	Convolutional Network
CE-MRI	Contrast-Enhanced MRI
CGAN	Conditional GAN
CBIR	Content-Based Image Retrieval
CADNet	Convolutional Attention-to-DMO Network
CFP	Colour Fundus Photographs
CN	Choroidal Nevus
CCNN	Concatenated CNN
CXR	Chest X-ray
CovidDWNet	Covid19 Detection Network based on FRB and DDC
CPU	Central Processing Unit
CB	Convolutional Block
Coa	Coarseness
CHKM	Colour histogram of K-mean
CCM	Colour Co-occurrence Matrix
ci-DMO	center-involved DMO
DenCeption	DenseNet-InCeption hybrid model
DBP	Diastolic Blood Pressure
DenseNet	Densely Connected Network
DNN	Deep Neural Network
DC-ELM	Deep Convolutional Extreme Learning Machine

Continued on next page

Acronym	Definition
DBM	Deep Boltzmann Machine
DBN	Deep Belief Network
DeepLab	Semantic segmentation architecture
DR	Diabetic Retinopathy
DCNN	Deep CNN
DMO	Diabetic Macular Oedema
DT	Decision Tree
DAL	Deep Active Learning
DDC	Dense Dilated Convolution
DSC	Dice Similarity Coefficient
DTL	Dimensional Transfer Learning
DLF	Deep Label Fusion
DDC	Depthwise Dilated Convolutions
DB	Dense Block
DHF	Deep Hidden Features
Dir	Directionality
DBPSP	Difference Between Pixels of Scan Pattern
DM	Diabetes Mellitus
ER-Net	Edge-Reinforced Neural Network
ERloss	Edge-Reinforced Optimisation Loss
EE-UNet	EfficientNet-based U-Net
EOG	Edge Operating Gradient
ELM	External Limiting Membrane
FFNN	Feed Forward Neural Networks

Continued on next page

Acronym	Definition
FCNN	Fully Connected Neural Network
FLOP	Floating-Point Operations
FN	False Negative
FSM	Feature Selection Module
FPN	Feature Pyramid Network
FRB	Feature reuse Residual Block
FM	Feaute Map
FCM	Fuzzy c-means clustering
GB	Gradient Boosting
GBDT	Gradient Boosting DTs
GBM	Gradient Boosting Machines
GD	Gestational Diabetes
GLCM	Grey Level Co-occurrence Matrix
GMRF	Gaussian Markov Random Field
GN-AlexNet	GoogleNet and AlexNet Network
GPU	Graphics Processing Unit
GRU	Gated Recurrent Units
GTR	Gross Total Resection
HAML	Hybrid Automated Medical Learning
HDL	Hybrid DL
HF	High-level features
HigherHRNet	Higher HRNet
HRCT	High-resolution computed tomography
HRNet	High-Resolution Network

Continued on next page

Acronym	Definition
HyBoost	Hybrid Boosting model
HAL-IA	Hybrid Active Learning framework using Interactive Annotation
HSI	Hyperspectral Image Classification
InA	Inception A Block
InB	Inception B Block
InC	Inception C Block
ILD	Interstitial Lung Disease
KNN	K-Nearest Neighbours
LAEDNet	Lightweight Attention Encoder–Decoder Network
LCE	Categorical Cross-Entropy
LR	Logistic Regression
LRSE	Lightweight Residual Squeese-and-Excitation
LSTM	Long Short-Term Memory
LWS-Net	Light Weight Stacking Network
MAE	Mean Absolute Error
MAGAN	Multi-scale Attention GAN
MBConv	Mobile Inverted Bottleneck Convolution
MAS	Multi-Atlas Segmentation
MedShift	Identification of Shift Data for Medical Image Dataset Curation
ML	Machine Learning

Continued on next page

Acronym	Definition
ML-FEC	Multi-Label Feature Extraction and Classification
MLP	Multi-Layer Perceptron
MPI	Message Passing Interface
MRI	Magnetic Resonance Imaging
MTANN	Multi-resolution Massive Training ANN
MSDNet	Multi-Scale DenseNet
MURA	MUsculoskeletal RAdiographs
mAP	mean Average Precision
MTDL	Multi-Task DL
MGMT	O6-methylguanine-DNA methyltransferase
MRF	Markov Random Field
MODY	Maturity Onset Diabetes of the Young
MK	Microbial Keratitis
NASNet	Neural Architecture Search Network
NB	Naive Bayes
NED	Neural Edge Detector
NFN	Network Followed Network
NInA	Without InA Block
NPDR	Non-Proliferative DR
NSCT	Non-Subsampled Contourlet Representation
NSST	Non-Subsampled Shealet Transform
NRA/B	without RA/B
OCT	Optical Coherence Tomography

Continued on next page

Acronym	Definition
OD	Optic Disc
ONH	Optic Nerve Head
PCA	Principal Component Analysis
PET	Positron Emission Tomography
PMM	Performance Measurement Matrix
PNN	Probabilistic Neural Network
PR	Precision-Recall
PSNR	Peak Signal-to-Noise Ratio
PSO	Particle Swarm Optimisation
PyramidalNet	Pyramidal Network
PDR	Proliferative DR
QNN	Quantizer Neural Network
QMPA	Quantum Marine Predator Algorithm
RBFNN	Radial Basis Function Neural Network
RV	Right Ventricle
RT	Radiation Therapy
RNN	Recurrent Neural Network
ResNet	Residual Network
ResNeXt	Alternative model based on the original ResNet design
RCNN	Region Based Convolutional Neural Network
RetinaNet	Feature Pyramid Based Network
RL	Reinforcement Learning
ReLU	Rectified Linear Unit

Continued on next page

Acronym	Definition
RGB	Red, Green, and Blue
RF	Random Forest
ROC	Receiver Operating Characteristic curve
REAM	Reverse Edge Attention Module
RBM	Restricted BM
RMSE	Root Mean Square Error
RMSProp	Root Mean Squared Propagation
RPN	Region Proposal Network
RU-Net	Residual U-Net
RA	Reduction A Block
RB	Reduction B Block
RoI	Region of Interest
SHAP	SHapley Additive exPlanations
SBP	Systolic Blood Pressure
SOM	Self-Organising Map
SCNN	Soft Cluster Neural Network
SVM	Support Vector Machine
SegNet	Deep Convolutional Encoder-Decoder Architecture for Image Segmentation
SSD	Single Shot Detector
SENet	Squeeze-and-Excitation based Network
SGE	Sun Grid Engine
Sen	Sensitivity
Spe	Specificity

Continued on next page

Acronym	Definition
SynthSeg	Synthetic Data Based Segmentation CNN network
SRP	Stacked Residual Pooling
SDL	Single DL
SI	Similarity Index
SPECT	Single-Photon Emission Computed Tomography
SSIM	Structural Similarity Index Measure
SENSE	Sensitivity Encoding
SVD-Net	fully-Supervised pre-training Network
SSL-AnoVAE	Self-Supervised Learning Based Anomaly Detection Framework
SSL	Self-Supervised Searning
STR	Subtotal Resection
SFS	Sequential Forward Selection
SMM	Second Moment Matrix
TF	Transformative Function
TL	Transfer Learning
TP	True Positive
TN	True Negative
TOF-MRA	Time-of-Flight Magnetic Resonance Angiograph
TDA	Topology Data Analysis
TB	Transition Block

Continued on next page

Acronym	Definition
T1Gd	Post-contrast T1-weighted
T2-FLAIR	T2 Fluid Attenuated Inversion Recovery
T2DM	Type 2 Diabetes Mellitus
T1DM	Type 1 Diabetes Mellitus
U-Net	Convolutional Networks for Biomedical Image Segmentation
UM	Uveal Melanoma
V-Net	Volumetric Convolutional Networks for Biomedical Image Segmentation
VGG	Visual Geometry Group
VDT	Variable Dimension Transform
VEGF	Vascular Endothelial Growth Factor
VBM	Voxel-Based Morphometry
WideResNet	Wide Residual Network
WTL	Weight TL
Xception	Extreme Inception Network
XGBoost	eXtreme Gradient Boosting
X-ray	X-Radiation
YOLO	You Look Only Once
2D	Two Dimensional/Dimension
3D	Three Dimensional/Dimension

Chapter 1

Introduction

1.1 Motivation

The field of medical diagnostics is at a critical juncture where traditional image processing techniques are increasingly inadequate to meet the demands of modern healthcare. The growing complexity and volume of medical imaging data expose the limitations of manual and conventional methods, which struggle with precision, scalability, and real-time adaptability. This research is motivated by the urgent need to address these challenges through the integration of advanced machine learning (ML) and deep learning (DL) methodologies. By leveraging the transformative power of these technologies, this research aims to revolutionise medical diagnostics, enhancing the accuracy and efficiency of disease detection and treatment planning. The research also seeks to bridge critical gaps in current approaches, such as the insufficient use of combined high-level and deep hidden features (denoted HF and DHF, respectively), and the often-overlooked potential of integrating demographic and physiological data for personalised healthcare. In tackling these multifaceted challenges, this work aspires to advance the field of medical imaging, ultimately improving patient outcomes and contributing to a more effective and trustworthy healthcare system.

1.1.1 Shortcomings of Traditional Techniques in Case of Complex Medical Images

The ever-growing complexity and volume of medical imaging data present a formidable challenge to traditional image processing techniques, which struggle to deliver the precision, scalability, and adaptability required for modern diagnostics. This research is driven by the urgent need to overcome these limitations through the development of advanced ML and DL methodologies. Using the power of these cutting-edge technologies, the aim of this work is to revolutionise the accuracy and efficiency of medical diagnostics, ultimately improving patient outcomes and advancing the field of healthcare.

Rapid advances in image processing have revolutionised various fields, with medical diagnostics standing out as a key beneficiary. In medical imaging, pre-processing tasks such as image enhancement, correction of brightness and geometric distortions, noise reduction, and edge delineation are critical for ensuring the accuracy of subsequent analysis. Such preparatory steps are indispensable for the success of subsequent analytical phases like segmentation, classification and prediction which is essential for both qualitative and quantitative evaluations of medical images. However, reliance on manual processing by medical professionals can lead to inaccuracies resulting from human limitations, such as fatigue or inconsistencies in level of expertise (Panayides et al., 2020). This underscores the imperative for automated segmentation methods to ensure both precision and dependability in medical diagnostics.

Although traditional image processing techniques have been fundamental in the advancement of this field, they often perform poorly when dealing with complex medical images. Challenges such as handling images under varied lighting conditions, deciphering ambiguous boundaries, and the need for extensive manual adjustments reveal the limitations of conventional methods (Razzak, Naz, and Zaib, 2018). Moreover, the lack of scalability and adaptability in traditional classification methods significantly impacts medical image processing by limiting the efficiency and effectiveness of diagnostic tasks (Panayides et al., 2020). In fact, these methods often struggle to handle diverse datasets with varying sizes, resolutions, and

complexities commonly found in medical imaging. This limitation obstructs the smooth integration of new data sources and restricts the scalability of classification models.

Another major problem is the inability to adapt to complex scenarios. In fact, the lack of adaptability in traditional methods can hinder their performance in real-time scenarios where reliable and accurate diagnoses are crucial. Without the ability to adapt to dynamic changes in imaging conditions or patient data, traditional models can face difficulties in providing timely and reliable results, which affects patient care and treatment decisions (Abdou, 2022). Additionally, difficulty in incorporating advanced features represents a major motivation to this research. In fact, scalability issues in traditional methods can make it challenging to incorporate advanced features or adapt to evolving medical imaging technologies. This limitation restricts the models' ability to leverage cutting-edge techniques for improved diagnostic accuracy and may result in unsatisfactory performance compared to more adaptable automated approaches such as ML and DL.

In light of these challenges, there is a compelling need to transition towards more advanced and automated solutions. The limitations of traditional methods in incorporating advanced features, adapting to evolving medical imaging technologies, and scaling effectively underscore the urgency of innovation in this space. This research is driven by the need to develop sophisticated methodologies that can overcome these intrinsic constraints, ultimately improving the adaptability, efficiency, and accuracy of medical image processing.

1.1.2 Transformative Impact of Deep Learning Approaches on Medical Imaging

ML and DL are pivotal subsets within the broad spectrum of Artificial Intelligence (AI), each playing a crucial role in the development and application of intelligent systems (Janiesch, Zschech, and Heinrich, 2021). ML is a subset of AI that focuses on the development of algorithms and statistical models that enable computers to perform specific tasks without using explicit instructions, relying instead on patterns and inference derived from data. It is the foun-

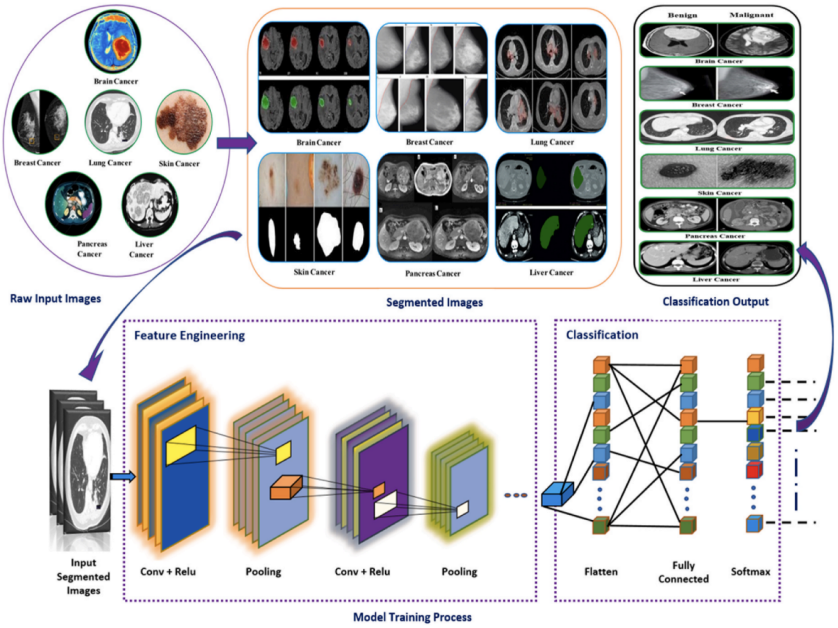


Figure 1.1: Medical Image Classification Process using Deep Learning.

dation for systems that can improve their performance on a task over time with more data. DL, a subset of ML, takes inspiration from the human brain's structure and function, employing artificial neural networks (ANNs) with many layers (hence "deep") to learn and make intelligent decisions. DL models are particularly adept at handling large amounts of data and excel at tasks such as medical image processing and disease detection, where they can automatically learn complex representations. Both ML and DL fall under the AI umbrella, a field dedicated to creating machines capable of performing tasks that would typically require human intelligence. Together, they represent the cutting edge of efforts to permeate machines with the ability to learn, reason, and adapt to new information or environments (Shinde and Shah, 2018).

Among DL architectures, convolutional neural networks (CNN) have revolutionised the field of medical imaging, marking a significant milestone in the application of AI to healthcare (Li, Liu, Yang, Peng, and Zhou, 2021b; Gu et al., 2018; O'shea and Nash, 2015) (Figure 1.1). CNNs have had a profound impact on medical imaging, enabling significant advancements in disease detection, diagnosis, and treatment planning. Their architectures excel at learning from image data without requiring manual feature extraction, thus creating highly accurate and efficient diagnostic tools (Albawi, Mohammed, and Al-Zawi, 2017; Yamashita, Nishio, Do,

and Togashi, 2018). Additionally, CNNs have been instrumental in enhancing the precision of diagnoses across a variety of modalities, including MRI, OCT, Fundus, and X-rays, facilitating early detection of diseases such as cancer, neurological disorders, eye related diseases, cardiovascular and pulmonology conditions (Li et al., 2014). The effectiveness of CNNs is largely attributed to their hierarchical feature extraction capability, which automatically learns representations from data. This has been further driven by the vast amount of data and advancements in hardware technology. The significant advancement in CNN performance is attributed to architectural innovations that focus on exploiting spatial and channel information, enhancing depth and width, and incorporating multi-path information processing (Anwar et al., 2018).

Despite their transformative impact on medical imaging and disease prediction, CNNs have yet to overcome several challenges that limit their full potential in healthcare applications. One of the primary issues is the requirement for large annotated datasets for training, which can be limited in the medical field due to privacy concerns, the labour-intensive nature of labelling, and the rarity of certain conditions (Tajbakhsh et al., 2016). Additionally, CNNs often struggle with generalising from one dataset to another, a phenomenon known as domain shift, which is particularly problematic in medical imaging where equipment and protocols vary widely across institutions. The interpretability of CNN decisions also remains a significant problem; the "black-box" nature of these models can impact trust and adoption by medical professionals who require understandable diagnostic rationales (Yadav and Jadhav, 2019). Furthermore, CNNs are susceptible to biases present in training data, which can lead to skewed or inaccurate predictions, especially for underrepresented groups in the data. Lastly, the computational complexity and resource requirements of CNNs pose challenges for deployment in resource-constrained settings, limiting accessibility to advanced diagnostic tools. Addressing these challenges is crucial for the broader acceptance and effective utilisation of CNNs in improving patient care and outcomes in the medical domain.

1.1.3 Critical Role of Feature Extraction in Medical Image Analysis

A critical aspect of medical image analysis is the feature extraction stage, which underpins the accuracy and effectiveness of disease diagnosis and prediction (Nixon and Aguado, 2019). This crucial phase involves identifying and isolating relevant information from complex medical images, transforming raw data into a structured format that can be effectively analysed by ML and DL models. Effective feature extraction not only enhances the model's ability to discern subtle patterns and anomalies indicative of specific medical conditions but also significantly reduces computational complexity by focusing on the most informative aspects of the data. By capturing the essential characteristics of medical images, this stage ensures that the subsequent analysis is both accurate and efficient, facilitating early detection and precise characterisation of diseases (Kumar and Bhatia, 2014).

Within the medical imaging pipeline, the distinction between HF and DHF features significantly influences the performance of intelligent models, including both ML and DL frameworks. HF features refer to the more abstract, global attributes of an image, such as overall shape and structure, which are often directly interpretable by humans. In contrast, DHF features delve into the finer, more granular details, capturing textures, edges, and intensity variations that may not be immediately perceptible to the human eye but are crucial for understanding the nuanced patterns characteristic of various medical conditions. The unique contribution of both feature types is indispensable for enhancing the accuracy and robustness of diagnostic models. HF features provide a macroscopic overview that helps in distinguishing between broadly different categories of abnormalities, while DHF features offer microscopic insights essential for detecting subtle anomalies that could be early indicators of disease (Lai and Deng, 2018). Together, these feature sets equip DL and ML models with a comprehensive understanding of medical images, enabling them to make more accurate predictions and diagnoses.

Despite the strength of this combination, a major shortcoming in current methodologies is the tendency to use these features in isolation instead of integrating them into a unified analytical framework. This separation limits the potential synergies that could be achieved by integrating the global contextual insights provided by HF features with the detailed precision

of DHF features. The lack of a cohesive approach in combining these feature types restricts the depth of analysis possible, potentially overlooking complex patterns that could be pivotal for diagnosis. Bridging this gap by developing techniques that effectively merge HF and DHF features could lead to a significant enhancement in the performance of intelligent diagnostic models, resulting a more holistic and nuanced analysis of medical images. Such advancements would not only enhance diagnostic accuracy but also contribute to the development of more sophisticated, interpretable, and clinically relevant AI tools in healthcare.

1.1.4 Integration of Demographic and Physiological Data for Personalised Diagnostics

The integration of demographic and physiological patient data into disease prediction models represents a critical yet often overlooked dimension in enhancing the accuracy and personalisation of healthcare diagnostics (Ganesan, Venkatesh, Rama, and Palani, 2010; Gichoya et al., 2022; Siuly and Zhang, 2016). Current state-of-the-art methods in medical prediction primarily focus on clinical imaging and genetic information, frequently neglecting the rich insights that demographic factors (such as age and sex) and specific physiological measurements (like Systolic Blood Pressure (SBP) and Diastolic Blood Pressure (DBP) for cardiovascular and pulmonary conditions, or Central Retinal Thickness (CRT) for eye diseases) can offer. These parameters are vital for a comprehensive understanding of a patient's health profile, as they can significantly influence disease emergence, progression, and severity. For instance, the risk factors and manifestations of many pulmonology conditions and eye-related diseases can vary markedly with age and sex, while physiological markers like blood pressure and CRT provide direct insights into the current state of a patient's health. By incorporating these demographic and physiological dimensions into prediction models, the approach moves toward more nuanced, accurate, and individualized disease prediction and management strategies. This holistic approach not only holds the promise of improving diagnostic precision but also paves the way for tailored treatment plans that account for the unique characteristics of each patient,

ultimately enhancing patient outcomes.

Medical image analysis is currently facing a significant challenge in order to achieve both precision and broader applicability. However, this is quite challenging because of the complex details found in high-dimensional medical imaging data. This issue goes beyond technical difficulties; it crucially affects the effectiveness and trustworthiness of diagnostic methods and patient care. The complex nature of high-dimensional data requires sophisticated analysis to identify implicit yet important patterns, which are vital for medical diagnosis. The current landscape is marked by models that, while often effective in controlled scenarios, underperform when confronted with the vast heterogeneity and the nuanced complexities presented by real-world data. The lack of robustness refers to these models' susceptibility to variations in imaging conditions, patient demographics, and pathological manifestations. Moreover, the lack of interpretability, which is a clinical necessity results the lack of clarity and integration of such models into the clinical workflow which represents the motivation of this research.

1.2 Problem Statement

The emergence of ML and DL technologies has indicated significant advancements in medical imaging, suggesting revolutionary improvements in diagnosing and predicting diseases with high accuracy. However, despite these advancements, several critical challenges persist, restricting the optimal utilisation of ML and DL in healthcare. These challenges include the difficulty of combining various algorithmic strategies, the complexities of extracting and assimilating DHF image features effectively, and the complications associated with incorporating patient-specific information into prediction models. Furthermore, the growing complexity of medical imaging data necessitates a more sophisticated and adaptable approach to feature extraction and model assessment, ensuring that these technologies can fully cater to the nuanced demands of contemporary medical diagnostics.

At the core of this research lies a crucial question:

How can the existing gaps in ML and DL applications for medical imaging be bridged

by incorporating hybrid algorithms in features extraction framework and improving prediction performance to accommodate the complex requirements of current medical diagnostic processes?

Addressing this question is essential to meeting the increasingly complex demands of modern medical diagnostics.

1.3 Aim

The aim of this study is to identify the current automated ML and DL models used in medical image processing by investigating the architecture of well used models in medical imaging, which helps to evaluate the performance of current ML/DL-based feature extraction framework and identify their limitations against the evolving complexity of diseases classification using medical images. This paves the way to design, develop, and evaluate an adaptive and scalable disease prediction framework which amalgamates novel hybrid DL-ML based models and reliable features extraction system to enhance the prediction performance of medical imaging data.

1.4 Objectives

Following objectives had to be fulfilled to achieve this aim:

1. Investigate the deployment of diverse ML and DL algorithms in medical image processing and delineate existing frontier gaps, paving the way for future innovations related to disease prediction in healthcare field.

2. Assess the differential impacts of hybrid ML and DL models across a spectrum of medical imaging modalities, aiming to comprehend their transformative impact and optimise their use in clinical diagnostics.

3. Design and develop a novel DL-based hybrid model to enhance classification performance of medical image processing compared to the existing state-of-the-art methods.

4. Explore the relationship between HF and DHF feature sets integration within medical imaging processing, aiming to significantly enhance the performance of ML and DL models across a variety of computational tasks including segmentation, classification, and prediction.

5. Design and develop an adaptive and scalable features extraction framework by integrating the novel designed hybrid DL model to combine an optimised set of HF and DHF features.

6. Evaluate how the inclusion of detailed demographic and physiological patient data can refine and enhance the performance disease prediction models, contributing to more targeted and effective medical outcomes.

7. Develop a robust PMM tailored for ML and DL models, with the specifics intent of advancing algorithmic sophistication and precision in medical imaging applications.

8. Build a sophisticated hybrid model for disease prediction that combines various algorithmic approaches, setting a new benchmarking predictive medical imaging analytics.

1.5 Contributions

The key innovation of this research is the design and implementation of an intelligent and adaptive framework for disease prediction using medical images. To achieve this, four main contributions have been made in this research:

- **Design of a novel DL-based hybrid model:** This work conducted a detailed examination of selected CNN architectures, including wider and deeper networks which have historically delivered impressive results. Addressing the need for an innovative approach, this research introduces a novel hybrid architecture, DenCeption, which merges the strengths of DenseNet-169 and Inception-V4. This combination aims to leverage the unique advantages of these networks to further enhance medical image processing. This

combination was backed with rigorous and thorough experimentation that involved both existing hybrid models as well as DenCeption's variants.

The development of the DenCeption model was driven by a critical need to overcome the limitations inherent in existing CNN architectures. While traditional models like DenseNet-169 and Inception-V4 have shown remarkable performance in medical image processing, they each have unique strengths and limitations. DenseNet is known for its efficiency in parameter usage and feature reuse, while Inception-V4 excels in capturing diverse features through its multi-scale processing capability. However, neither architecture alone fully addresses the challenges posed by the complexity of medical imaging data, such as the need for deep feature extraction and efficient handling of high-dimensional data.

The significance of DenCeption lies in its innovative combination of these two powerful architectures, merging their strengths to create a model that is more robust and versatile. This hybrid model was meticulously designed through a detailed examination of existing CNN architectures, identifying the specific elements of DenseNet-169 and Inception-V4 that could be synergised to enhance medical image processing. The selection of this combination was purposeful; it was the result of rigorous experimentation and comparative analysis, ensuring that the proposed model would offer a substantial improvement over existing methods.

DenCeption has demonstrated superior performance in medical image analysis, particularly in handling complex classification tasks. The extensive testing and validation against existing hybrid models and variants further confirmed its effectiveness, making it a valuable contribution to the field of medical diagnostics. By achieving higher accuracy and better feature extraction, DenCeption stands as a significant advancement, potentially revolutionising the way medical images are processed and analysed.

- **Design of an adaptive and scalable features extraction framework:** Towards tackling the drawback caused by the lack of features combination and enhancing the reliability of

features extraction methods, this work proposes a new hybrid features extraction framework that focuses on the fusion and optimal selection of HF and DHF. The scalability and reliability of the proposed method is achieved by the automated adjustment of the final optimal features based on real-time scenarios resulting an accurate and efficient medical images disease classification. The proposed framework has been tested on two different datasets to include BRATS and Retinal sets achieving outstanding results compared to benchmarking methods.

The motivation behind developing this hybrid feature extraction framework arose from the recognition that existing methods often failed to adequately combine HF features with DHF, leading to suboptimal classification performance. Traditional feature extraction approaches tended to focus either on superficial features or on complex deep features, without integrating the two in a meaningful way. This gap often resulted in unreliable and inconsistent outputs, especially in medical image analysis where both types of features are crucial for accurate diagnosis.

The proposed framework is significant because it addresses these shortcomings by introducing a scalable and adaptive method that dynamically adjusts the feature set based on complex data scenarios. This ensures that the most relevant features are selected and fused, optimising the classification process for diverse medical imaging tasks. The decision to focus on the fusion of HF and DHF was informed by a thorough analysis of their individual contributions to classification accuracy and the need to enhance their combined impact.

This framework has been proven to be highly effective through extensive testing on diverse datasets, where it achieved outstanding results compared to benchmarking methods. The scalability and adaptability of this approach make it particularly useful in real-world applications, ensuring consistent and accurate medical image classification across different imaging conditions and data complexities.

- **Design of a novel evaluation mechanism for DL models** : Recent advancements in

learning algorithms have led to their extensive use in various fields, including healthcare. However, selecting the most appropriate evaluation metrics for these algorithms remains a challenge. In this work, a novel evaluation mechanism that takes into account the problem domain and the specific application area is proposed. The mechanism involves randomly assigning weights to each metric for each dimension, applying a correlation operation to measure the degree to which these variables are related, and repeating this process for all stages. The resulting matrix is then used to calculate the final PMM vector that reflects the most optimal measurement metrics. The proposed mechanism provides a systematic and objective approach to selecting evaluation metrics for DL and ML algorithms, and can be applied to a wide range of applications. The work demonstrates the effectiveness of the mechanism using a case study on medical image processing applications.

In the rapidly evolving field of ML, particularly in healthcare applications, selecting the appropriate evaluation metrics is crucial yet challenging. The proliferation of DL models has led to a variety of evaluation metrics, but these are often selected without a systematic approach, potentially leading to biased or suboptimal evaluations. The need for a more structured evaluation process became apparent through the analysis of existing methods, which often failed to account for the specific requirements of different medical imaging tasks.

The novel evaluation mechanism introduced in this research is significant because it provides a systematic and objective approach to metric selection tailored to the problem domain. By incorporating a correlation-based approach and applying random weight assignments, the mechanism ensures that the most relevant metrics are identified and prioritised, leading to more accurate and meaningful evaluations of DL models.

The PMM mechanism is particularly useful in the context of medical image processing, where the consequences of misclassification are significant, and the complexity of the data requires careful consideration of multiple evaluation dimensions. This contribution

is valuable because it enhances the reliability and validity of DL model assessments, ultimately leading to better-informed decisions in model selection and deployment in clinical settings.

- **Design of an intelligent and robust predictive framework**: This work proposes an innovative prediction framework incorporating DenCeption, a scalable feature extraction framework, coupled with HyBoost, a novel hybrid predictive model. Designed to adapt to medical image characteristics, HyBoost enhances disease prediction by leveraging the strengths of various ML algorithms and incorporating vital patient demographic and physiological data. This approach significantly improves the model's performance, as confirmed by 10-fold cross-validation and SHapley Additive exPlanations (SHAP) explainability analysis. Significant performance enhancements have been resulted across various datasets including Fundus, OCT, and X-ray scans, indicating a substantial progression in predictive medical image analysis compared to existing models such as XGBoost, AdaBoost, U-Net, Inception-V4, YoLo-V7, EfficientNet-B5, VGG-16/19, and ResNet.

  The need for a robust predictive framework that can adapt to the specific characteristics of medical images is critical in advancing medical diagnostics. Traditional predictive models often fail to incorporate vital patient demographic and physiological data, which are essential for accurate disease prediction. Additionally, these models tend to rely on a narrow set of features, which limits their generalisability and effectiveness across different types of medical imaging data.

  The proposed predictive framework, which integrates DenCeption and the HyBoost model, is significant because it represents a comprehensive solution that enhances prediction accuracy by leveraging a wide range of ML algorithms and incorporating crucial patient data. The HyBoost model is a novel hybrid approach that combines the strengths of XGBoost and AdaBoost, two of the most powerful and widely used boosting algorithms in ML. XGBoost is known for its efficiency and scalability, offering robust performance

with large datasets and complex feature spaces. It is particularly effective in handling sparse data and has a built-in regularisation mechanism that reduces the risk of overfitting. On the other hand, AdaBoost excels in improving the performance of weak learners by focusing on the errors of previous models, making it highly effective in scenarios where the data is noisy or imbalanced. The combination of these two algorithms in the HyBoost model is justified by their complementary strengths: XGBoost's ability to handle diverse and complex data, and AdaBoost's proficiency in refining predictive accuracy through iterative learning. This hybrid approach enhances the overall robustness and reliability of the predictive framework, making it better suited to the intricate demands of medical image analysis.

The integration of XGBoost and AdaBoost within the HyBoost model significantly improves the framework's performance across various datasets. This combination has been shown to enhance the accuracy of disease prediction while maintaining the adaptability needed to apply the model to different types of medical imaging data. By leveraging the strengths of these boosting algorithms, the HyBoost model not only achieves higher predictive accuracy but also provides more reliable and consistent results, making it a valuable tool for clinicians. The SHAP explainability analysis further strengthens its utility by offering clear insights into the model's decision-making process, which is crucial for gaining trust and ensuring the practical adoption of the model in real-world clinical settings.

1.6 Scope

This research will encompass a systematic investigation of the various ML and DL algorithms currently utilised in medical image processing across different modalities such as MRI, OCT, Fundus, and X-rays. It will explore the integration of HF and DHF features within these images to boost the performance of related models. The scope includes the development of an innovative hybrid DL architecture for feature extraction and a versatile feature extraction framework.

Additionally, a robust PMM will be crafted to evaluate and enhance the algorithms involved in medical imaging. The research will focus on the integration of comprehensive patient data into predictive models to refine diagnostic accuracy. The study will be bound by the fields of current technological capabilities and data availability, with an emphasis on addressing the practical challenges of deploying these models in clinical settings. The deployment of the proposed features extraction and prediction frameworks is not part of the scope of this work. In addition, medical expert validation of the proposed solution is not part of the scope. Also, the validation of the proposed framework on real clinical data is not part of this research focus.

1.7 Thesis Structure

The remaining parts of the thesis are structured as follows:

- **Chapter 2: Literature Review**

  This chapter presents a background of the research followed by a thorough discussion on the evolution of medical image processing from traditional techniques to more automated approaches. The discussion on different processing methods that can be applied for more efficient training, features extraction and disease prediction is then presented. Subsequently, relevant recent studies with their benefits and drawbacks are discussed, which lays the foundation for the presented research.

- **Chapter 3: DenCeption : A Novel Hybrid Deep Learning Based Model**

  This chapter presents the design, implementation and validation details of the proposed DenCeption model. It starts by the examination of the DenCeption variants and followed by a thorough testing process. A benchmarking with existing hybrid DL-based models will be performed using a set of evaluation metrics. The conducted experiments will use MRI images for training, testing, and validation stages.

- **Chapter 4: A Deep Learning based Scalable and Adaptive Feature Extraction Framework for Medical Images**

This chapter presents the design and implementation of the proposed features extraction framework. It covers the integration of DenCeption model, presented in Chapter 3, within the proposed framework. The chapter conveys the proposed algorithm for HF and DHF features extraction and the impact of their fusion of the overall performance of the proposed framework. A benchmarking with existing features extraction methods will be conducted for validation purpose on various medical images to include Fundus and MRI.

- **Chapter 5: Advancing Intelligent Medical Diagnostics with HyBoost: A Robust Predictive Framework for High-Dimensional Imaging Data**

  This chapter presents the combination of this research contributions detailed in Chapter 3 and 4. It details the design and implementation of the proposed prediction framework by introducing the novel ML-based model, HyBoost. The chapter provides a thorough examination of the proposed framework through different medical image modalities to include OCT, Fundus and X-ray examining various set of diseases to include eye and pulmonology related conditions. This part of the research also conducts a benchmarking process to validate the proposed prediction framework against related works. This validation process will be powered by the proposed PMM evaluation mechanism.

- **Chapter 6: Conclusion and Future Work**

  The conclusion offers an analysis of the challenges addressed by the suggested frameworks and the underlying models, emphasising the advantages of the introduced solutions. It proceeds to outline the limitations of the proposed approach and strategies for their mitigation. Concluding, it explores potential improvements to the proposed framework and its applicability to a wider range of domains.

Chapter 2

Literature Review

2.1 Introduction

Image processing has seen rapid advancements, particularly, in medical diagnostic using diverse set of images. This field encompasses various pre-processing steps like image reconstruction, brightness and geometric transformations, noise filtering, and edge detection. Preprocessing is critical for subsequent steps like segmentation which is vital for qualitative and quantitative analysis. However, manual processing by physicians can introduce errors due to human factors such as fatigue and varying levels of experience, highlighting the need for automated segmentation for precision and reliability. Traditional image processing methods, though foundational, often face limitations with complex images, leading to challenges in adaptability and manual tuning requirements. Their performance in diverse and complex image scenarios, like varying lighting or ambiguous boundaries, is limited, necessitating the shift towards more advanced and automated techniques. This Literature review chapter will provide a thorough presentation and discussion about processing techniques, with particular focus on medical imaging building the ground for their applications in disease classification and prediction.

2.2 Research Background

The research landscape began witnessing a pivotal shift towards the exploration of automated image processing techniques, encompassing segmentation, classification, and detection, with a particular emphasis on enhancing the analysis and interpretation of medical images. In contemplating the future trajectory of the proposed method, several potential areas of improvement can be identified to maximise its performance and utility. Firstly, there's a compelling need for further optimisation designed to large datasets, ensuring that the algorithm retains its efficiency when confronted with an increased volume of data or images of higher resolution. Additionally, enhancing the algorithm's capacity for noise discrimination presents a significant avenue for development; refining its ability to accurately distinguish between genuine vessel structures and extraneous elements like noise or other retinal features, such as the optic disc (OD), could substantially reduce misdetection and elevate the overall precision. Moreover, adopting a more sophisticated classification process that overcomes the simplistic binary vessel/non-vessel categorisation could significantly improve the method's competence in managing complex scenarios and mitigating instances of under-segmentation. Lastly, the incorporation of additional image features and the leverage of advanced image processing techniques, including DL, stand as promising strategies to further augment the method's accuracy and robustness, paving the way for a more comprehensive and nuanced analysis of medical images.

2.2.1 Neural Networks: A Focus on Convolutional Neural Networks

The pivotal role that ANNs have come to play in the domain of medical image processing began with their early application in tasks such as image registration, segmentation, and edge detection (Jiang, Trundle, and Ren, 2010). These initial deployments marked a significant breakthrough, leading to the widespread adoption of ANNs in various image processing systems. The literature highlights the transformative impact of neural networks on medical imaging analysis, especially in the domains of medical image registration, segmentation, and edge detection. These foundational processes are crucial for effective content analysis and regional

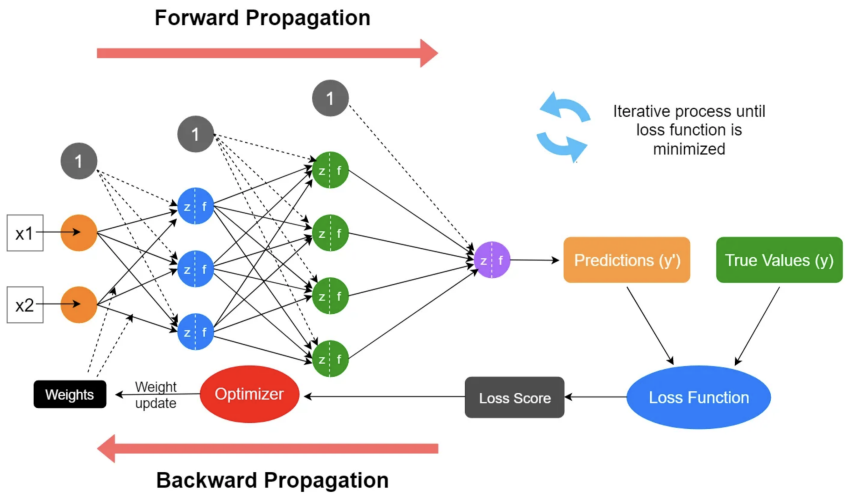


Figure 2.1: Artificial Neural Networks Learning Process.

inspection in medical imaging. The state-of-the-art works shed light on the potential enhancements that neural network applications can bring to medical image processing applications. These works elaborate on the adaptability and learning capabilities of neural networks, which can significantly refine diagnostic processes by optimising the relationship between inputs and outputs. This is particularly evident in visual inspection and visualisation tasks, where medical imaging serves as an indispensable tool.

The integration of neural networks in medical image processing represents a transformative shift towards more intelligent, adaptive, and precise healthcare solutions (Jiang, Trundle, and Ren, 2010). Their power lies in their ability to learn from data, improve their performance over time, and provide insights that might not be visible to the human eye (Figure 2.1). Their application in medical imaging has shown promising results in enhancing the accuracy of diagnoses, reducing the workload of healthcare professionals, and ultimately paving the way for personalised and preventive medicine. The continuous evolution of neural network techniques, coupled with increasing computational power and data availability, is set to further revolutionise the field of medical imaging and healthcare at large.

These ANN-based techniques have been applied for several purposes to include:

- **Image Registration:** Techniques for aligning images from different datasets or modalities are explored, with a focus on the use of Self-Organising Maps (SOMs) and other neu-

ral network-based methods. These techniques aim to establish correspondence between different images, accommodating for variations due to movements or changes over time.

- I**mage Segmentation and Edge Detection:** The paper discusses neural network approaches for segmenting medical images into regions with homogeneous properties and detecting edges of organs or tumours. Methods like the Hopfield network, Kohonen's competitive learning, and Multi-Layer Perceptron (MLP) are highlighted for their ability to classify medical images into content-consistent regions.

- **Computer-Aided Diagnosis (CAD):** Neural networks are extensively applied in CAD systems for detecting and diagnosing diseases from medical images. These systems illustrate various applications, including breast cancer detection from mammograms and lung disease identification, showcasing the utility of neural networks in enhancing diagnostic accuracy and reducing false positives.

These methods have been applied for several diseases such as breast cancer, lung cancer, general tumour detection, brain disorders (e.g. Alzheimer's disease (AD)). Tables 2.1, 2.2 and 2.3 provide a critical representation of ANN-based techniques applied on each of the aforementioned methods, respectively.

Table 2.1: Neural Network Methods in Image Registration

Method	**Description**	**Advantages**	**Disadvantages**
SOMs (Miljković, 2017)	Utilises the learning capability of SOMs to classify medical images into content-consistent regions, completing segmentations and edge detections.	- Preserves topological features of input data - Good for visualising complex datasets	- Training can be computationally intensive - May require fine-tuning for optimal results
Principal Component Analysis (PCA) with Neural Networks (Kurita, 2019)	Employs PCA for dimensionality reduction before feeding data into a neural network for registration.	- Reduces data complexity - Can improve computational efficiency	- PCA step may discard useful information - Sensitive to the scaling of input data
Multi-scale SOMs (Chen, Ashizawa, Yeo, Yanai, and Yean, 2021)	Utilises a multi-scale approach to handle different frequency components of the images for registration.	- Handles various spatial-frequency components effectively - Improves accuracy of registration	- More complex to implement - Higher computational cost

Table 2.1:: Neural Network Methods in Image Registration (Continued)

Method	**Description**	**Advantages**	**Disadvantages**
Competitive Learning, Self-Organizing, and Clustering (Li, Liu, Jiao, Chen, and Li, 2022b)	Designs neural networks for alternative solutions via competitive learning, self-organising, and clustering for image feature processing.	- Provides robust alignment solutions - Good for complex datasets with varying features	- Can be challenging to set up and train - May require extensive computational resources

Table 2.2: Neural Network Methods in Image Segmentation and Edge Detection

Method	**Description**	**Advantages**	**Disadvantages**
MLP (Desai and Shah, 2021)	Uses MLPs for the binary classification of pixels to identify boundaries and non-boundaries in medical images.	- Good for non-linearly separable data - Flexible and widely applicable	- Requires careful tuning of network architecture - Prone to overfitting
Fuzzy and Soft Competition Learning (Chouhan, Kaul, and Singh, 2019)	Implements competitive learning fused with soft competition and fuzzy c-means membership functions for segmentation.	- Reduces noise effects in medical images - Useful in MRI segmentation	- Complexity in implementation - May require extensive training data

Table 2.2: Neural Network Methods in Image Segmentation and Edge Detection (Continued)

Method	**Description**	**Advantages**	**Disadvantages**
Quantizer Neural Network (QNN) (Rokh, Azarpeyvand, and Khanteymoori, 2023)	A novel structure trained by genetic algorithms for segmentation, particularly of MRI and CT head images.	- Efficient classification performance - Requires fewer neurons and shorter training time	- Less interpretability of the network structure - Reliant on the quality of genetic algorithms
Multi-resolution Massive Training ANN (MTANN) (Tajbakhsh and Suzuki, 2018)	Utilises a multi-resolution approach to handle different frequency components for edge detection.	- Effectively removes noise and tiny details - Good for high-frequency component handling	- Computationally intensive - Complex architecture and training process
Neural Edge Detector (NED) (Su et al., 2021)	Employs a modified multi-layer neural network trained via supervised learning to extract contours from images.	- Good agreement with manual edge extraction - Efficient in contour extraction	- Sensitive to noise and low contrast - May require extensive training

Table 2.3: Neural Network Methods in CAD

Method	**Application**	**Advantages**	**Disadvantages**
Radial Basis Function Neural Network (RBFNN) (Montazer, Giveki, Karami, and Rastegar, 2018)	Applied for fast detection of masses in mammograms.	- Faster training compared to MLP - Effective in function approximation	- Sensitive to the choice of basis functions - May not scale well with large datasets
Soft Cluster Neural Network (SCNN) (Verma, McLeod, and Klevansky, 2009)	Employed for the classification of suspicious areas in digital mammograms.	- Increases generalisation ability - Depicts relationships between features and classifications effectively	- Performance dependent on image properties - May require fine-tuning for different imaging conditions
Probabilistic Neural Network (PNN) (Savchenko, 2019)	Utilised for the prediction of histological grade, hormone status, and axillary lymphatic spread in breast cancer patients.	- Fast training process - Effective in classification and pattern recognition	- Can be computationally expensive during testing phase - Sensitive to the smoothing parameter

While neural networks present a robust and adaptive framework for addressing a wide range of problems in medical image processing, certain fundamental challenges need to be acknowledged. Firstly, the complexity and interpretability of these models, especially those

involving DL, cause significant challenges. They often operate as 'black boxes,' with limited transparency in their decision-making processes, making it difficult for practitioners to interpret and fully trust their outputs. Secondly, the training of neural networks, particularly the DL variants, is resource-intensive. It requires significant computational power and time, making it less accessible for setups with limited resources. Lastly, the performance of neural networks is heavily contingent on the quality and quantity of the input data. They require large and well-curated datasets to function effectively. Inadequate, skewed, or poor-quality data can lead to models that do not generalise well and perform poorly on unseen data. Hence, while neural networks hold immense promise in revolutionising medical image processing, addressing these challenges is crucial for their effective and widespread adoption. The transition from traditional to automated segmentation techniques marks a significant advancement in medical image processing, particularly in radiation therapy (RT). This shift is driven by the need for rapid and accurate segmentation of medical images, which is critical for effective treatment planning and delivery. Traditional methods like manual delineation, while considered the gold standard, are time-consuming and suffer from intra- and inter-observer variations, limiting their efficiency and reliability (Sharp et al., 2014).

Traditional segmentation techniques such as thresholding, region-based methods, edge detection-based methods, and deformable models like geodesic active contours, have served as the foundation for medical image segmentation. These methods primarily rely on the analysis of image content, like voxel intensities and image gradients, to differentiate between distinct regions in medical images. However, they often lack the incorporation of prior knowledge about anatomical structures, which can be crucial for achieving high accuracy in complex medical scenarios. In contrast, automated segmentation methods introduce sophisticated approaches that leverage prior knowledge and ML to enhance segmentation accuracy. Techniques like atlas-based segmentation, multi-atlas segmentation, and model-based segmentation (using statistical shape models or statistical appearance models) have significantly improved the robustness and accuracy of segmentation. These methods benefit from prior information about the morphology of anatomical structures or the appearance of organs in different imaging modalities, providing a

more refined and anatomically consistent segmentation output (Table 2.4).

Table 2.4: Automated Segmentation Techniques

Technique	**Description**	**Advantages**	**Disadvantages**
Atlas-based Segmentation (Bach Cuadra, Duay, and Thiran, 2015)	Uses a reference image (atlas) with pre-segmented structures to guide segmentation.	Can provide anatomically viable results.	Performance heavily depends on the choice of the atlas.
Multi-atlas Segmentation (Sun, Zhang, and Zhang, 2019)	Combines multiple atlas registrations for robust segmentation.	Improved robustness and accuracy.	Computationally intensive and may produce topological errors.
Model-based Segmentation (Ecabert et al., 2008)	Utilises statistical models of shape and appearance to guide segmentation.	Provides anatomically correct segmentation.	Requires a comprehensive training set and may lack flexibility for atypical cases.

The importance of this transition cannot be understated, especially in the context of RT, where the precise delineation of target volumes and organs at risk directly impacts treatment effectiveness. Automated segmentation methods offer several advantages over traditional methods, as follows (Sharp et al., 2014):

- **Speed and Efficiency:** Automated methods significantly reduce the time required for segmentation, allowing for faster treatment planning and adaptation.
- **Consistency and Objectivity:** By reducing human intervention, automated methods offer more consistent and objective segmentation results, minimising the variability associated with manual delineation.

- **Integration of Multimodal Data:** Automated segmentation techniques can effectively integrate information from different imaging modalities (like CT, MRI, and PET), providing a more comprehensive understanding of the anatomical structures.

However, this transition also presents challenges, including the dependence on the quality and quantity of training data, the need for significant computational resources, and the complexity involved in interpreting the results of sophisticated models like DL networks.

The evolution of DL in the field of medical imaging is revolutionising the field, offering significant enhancements over traditional ML and pattern recognition methods (Razzak, Naz, and Zaib, 2018). Unlike conventional ML techniques like support vector machines (SVMs), neural networks, and k-nearest neighbours (KNN), which rely on expertly crafted features and struggle with raw image data, DL algorithms such as CNNs, recurrent neural networks (RNNs), and generative adversarial networks (GANs) have the ability to digest raw data and learn features autonomously. This attribute of DL dramatically reduces reliance on domain expertise and manual feature engineering, allowing for rapid learning and adaptation. DL's layered approach and ability to learn hierarchical feature representations make it exceptionally suited for dealing with the high-dimensional, variable-rich medical images derived from diverse platforms like CT and MRI scans. A critical assessment reveals that while DL is a powerful tool for image processing, it's not without challenges. Table 2.5 provides a comparative analysis of several existing DL architectures and their respective advantages and disadvantages (Razzak, Naz, and Zaib, 2018; Iqbal, N. Qureshi, Li, and Mahmood, 2023).

Table 2.5: Comparative Analysis of Deep Learning Architectures

Deep Learning Architecture	**Advantages**	**Disadvantages**
Deep Neural Network (DNN) (Liu et al., 2019)	Able to model complex non-linear relationships, used for classification and regression	Training process can be slow, requires a substantial amount of labelled data
		Continued on next page

Table 2.5: Comparative Analysis of Deep Learning Architectures (Continued)

Deep Learning Architecture	**Advantages**	**Disadvantages**
Deep Convolutional Extreme Learning Machine (DC-ELM) (Pang and Yang, 2016)	Computationally efficient, fast training	Optimisation of parameters can be challenging for large datasets
Deep Boltzmann Machine (DBM) (Duong, Luu, Quach, and Bui, 2019)	Top-down feedback integrates ambiguous data for robust inference	Initialisation makes the training process computationally expensive
Deep Belief Network (DBN) (Sohn, 2021)	Greedy layer-wise strategy and inference are feasible	Training can be complex due to error propagation to individual layers
Autoencoders (AEs) (Pratella, Ait-El-Mkadem Saadi, Bannwarth, Paquis-Fluckinger, and Bottini, 2021)	Can automatically compress data, useful in reducing dimensionality.	Compression may lead to loss of important information, leading to suboptimal reconstruction.
Deep AE (Chen and Guo, 2023)	Excellent for unsupervised learning and dimensionality reduction	Training can vanish, pre-training step required

Continued on next page

Table 2.5: Comparative Analysis of Deep Learning Architectures (Continued)

Deep Learning Architecture	**Advantages**	**Disadvantages**
Feed Forward Neural Networks (FFNN) (Arumugadevi and Seenivasagam, 2016)	Simplicity, suitable for basic tasks in neural networks.	Lack of feedback loops, limiting the ability to handle sequence data or temporal patterns.
Long-Short Term Memory (LSTM) (Mei, Li, Liu, Cai, and Du, 2021)	Solves vanishing/exploding gradient issue, capable of learning long-term dependencies.	More complex and computationally intensive than traditional RNNs.
Gated Recurrent Units (GRU) (Ikuta and Zhang, 2022)	Reduced complexity and computational load compared to LSTM, solves lengthy dependence in RNN.	Still computationally intensive, may not capture long-term dependencies as effectively as LSTM.
U-Net (Siddique, Paheding, Elkin, and Devabhaktuni, 2021)	Specialised for biomedical image segmentation, capable of precise localisation.	Requires substantial annotated data for training, may struggle with very large images.
V-Net (Liu, Pang, Jin, Liu, and Wang, 2022)	Specialised for 3D medical image segmentation, offering detailed anatomical insights.	Computational complexity is high due to 3D data processing, may require substantial memory.
SegNet (Zhang, Lu, Wu, Ni, and Wang, 2024)	Efficient for indoor scene understanding and pixel-wise segmentation.	May struggle with complex scenes or overlapping objects, requires good quality data.

Continued on next page

Table 2.5: Comparative Analysis of Deep Learning Architectures (Continued)

Deep Learning Architecture	**Advantages**	**Disadvantages**
You Look Only Once (YOLO) network (Chen et al., 2024)	Real-time object detection, faster processing by looking at the entire image only once.	Less accurate compared to two-stage detectors, may struggle with small objects.
DeepLab (Azad et al., 2022)	Effective for semantic segmentation with atrous convolution, handling objects at multiple scales.	Complexity increases with advanced versions, may demand more computational resources.
Bidirectional RNNs (Kim, An, Chikontwe, and Park, 2021)	Can process data in both forward and backward directions, capturing future context effectively.	More complex and computationally intensive, potentially overfitting in smaller datasets.
Highway Networks (Ha et al., 2020)	Facilitate training deeper networks by enabling information highways, improving optimisation.	Complexity in architecture and hyperparameter tuning can make them hard to implement.
Wide Residual Network (WideResNet) (Nakayama, Lu, Li, and Kamiya, 2020)	Increased feature map size per layer for more expressive models, faster training.	Higher computational cost than standard ResNet, may require more memory during training.
Pyramidal Network (PyramidalNet) (Duta, Liu, Zhu, and Shao, 2020)	Combines top-down and bottom-up approaches, capturing multiscale context effectively.	Complexity increases with depth and pyramid levels, may require substantial compute resources.

Continued on next page

Table 2.5: Comparative Analysis of Deep Learning Architectures (Continued)

Deep Learning Architecture	**Advantages**	**Disadvantages**
Xception (Carnagie, Prabowo, Budiana, and Singgih, 2022)	Uses depthwise separable convolutions, more efficient parameter usage, improved performance.	May not always outperform other architectures with similar parameter counts.
ResNeXt (Koné and Boulmane, 2018)	Simplified architecture with repeated building blocks, improving scalability and performance.	Similar to ResNet, complexity and resource requirements increase with model size.
SqueezeNet (Koonce, 2021)	Highly compact model, reduced parameter count without significant loss in accuracy.	May not capture as complex features as larger models, potentially lower accuracy.
Fully Connected Neural Network (FCNN) (Basha, Dubey, Pulabaigari, and Mukherjee, 2020)	Can handle input of random size, flexibility in architecture design.	May lack some spatial hierarchies due to absence of fully connected layers.
Fast Region Based Convolutional Neural Network (Fast-RCNN) (Siradjuddin and Muntasa, 2021)	Faster object detection by providing entire image to CNN, generating convolutional feature map.	Still involves region proposals, which can be computationally expensive compared to later models.

Continued on next page

Table 2.5: Comparative Analysis of Deep Learning Architectures (Continued)

Deep Learning Architecture	**Advantages**	**Disadvantages**
Mask RCNN (Wang et al., 2021)	Precise instance segmentation, extends Faster RCNN by adding a branch for predicting masks.	Complexity and computational requirements are high, particularly for large datasets.
RetinaNet (Miao et al., 2022)	Addresses class imbalance with focal loss, effective for dense and small object detection.	Complexity in balancing between speed and accuracy, requires careful tuning.
Boltzmann Machine (BM) (Jeyaraj and Nadar, 2019)	Can model complex distributions and capture high-order correlations between observable variables.	Training is computationally intensive and may require careful tuning to avoid local minima.
GAN (You et al., 2022)	Powerful for generating realistic samples, useful in data augmentation, image generation, and unsupervised learning.	Training can be unstable and challenging, requiring careful balance between generator and discriminator.

Table 2.6: Comparison of Traditional Versus Deep Learning Architecture for Automatic Localisation (El-Shafai et al., 2024)

Aspect	**Traditional Architecture**	**Deep Learning Architecture**
Feature Extraction	Relies on manually extracted, DHF features.	Automatically learns HF feature representations from the data.
		Continued on next page

Table 2.6: Comparison of Traditional Versus Deep Learning Architecture for Automatic Localisation (Continued)

Aspect	**Traditional Architecture**	**Deep Learning Architecture**
Performance	Often limited by the quality and depth of feature engineering.	Exhibits superior performance, especially in complex and high-dimensional data scenarios.
Computational Requirement	Generally, less computationally intensive.	Requires significant computational resources, particularly for training.
Learning Capability	Limited learning capability, often requiring manual feature selection.	Exhibits deep learning capabilities, capturing complex patterns in data.
Adaptability	Less adaptable to new, unseen data.	Highly adaptable, can generalise well to new, unseen data.
Implementation Complexity	Less complex models, easier to interpret.	More complex models, sometimes referred to as "black boxes".
Data Requirement	Less reliant on large datasets.	Typically requires large datasets to perform optimally.
Accuracy and Precision	May struggle with very complex tasks.	Tends to offer higher accuracy and precision, especially in complex tasks.

Despite these architectures offering sophisticated approaches to model complex data, the application of DL in medical image processing faces several challenges (Razzak, Naz, and Zaib, 2018):

- **Dataset Availability:** DL requires large datasets to establish classifier accuracy, and medical imaging datasets are particularly challenging to compile due to the need for expert annotation and the rarity of certain conditions.
- **Privacy and Legal Issues**: Sharing medical data is complicated due to strict privacy

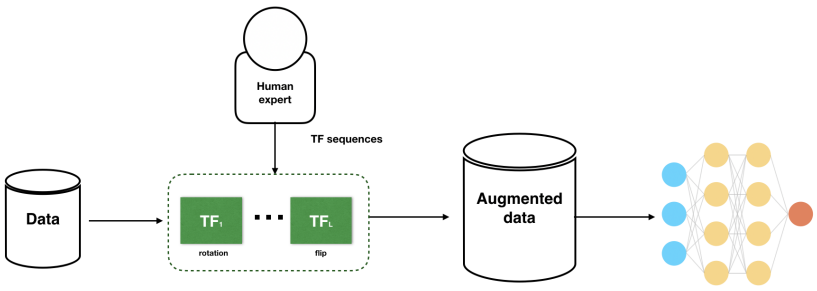


Figure 2.2: Data Augmentation Process where TF is the Applied Transformation Function (Garcea, Serra, Lamberti, and Morra, 2023)

laws, which impose massive restrictions on the way patient data can be used and disclosed.

- **Data Interoperability and Standards:** The lack of standardisation in medical data, arising from variations in hardware and sensor technologies, restricts the creation of universally applicable models.

The impact of DL on automatic localisation in medical images has been significantly positive, transforming the landscape of medical image analysis. DL models, particularly DNNs, have outperformed traditional ML methods in various complex computer vision tasks. The success of DNNs in object localisation, especially in medical images, has garnered attention due to their superior performance over conventional methods (Alaskar et al., 2022). Table 2.6 presents a comparative analysis of traditional techniques and DL architectures for different aspects of image processing for localisation purposes.

While traditional architectures have been fundamental in the initial stages of computer vision and image processing, the advent and advancement of DL architectures have significantly shifted the paradigm towards more efficient, accurate, and sophisticated methods of automatic localisation in medical images. Despite this, challenges like data availability, computational requirements, and the interpretability of DL models persist, shaping the future directions of research in this domain (Alaskar et al., 2022).

Through literature, the profound impact of DL on data augmentation is thoroughly reviewed, particularly within the field of medical imaging (Figure 2.2).

DL has been a significant factor in advancing the adoption of sophisticated data augmentation techniques, effectively addressing pressing challenges such as limited data availability, strict privacy concerns, and the substantial costs associated with data labelling. The scope of data augmentation strategies in DL extends across a broad spectrum, ranging from relatively simple transformations, including cropping, padding, and flipping, to the deployment of complex generative models. This expansion significantly enhances the diversity and volume of training data, enabling the enrichment of datasets without the need for additional data collection. Moreover, DL-driven data augmentation facilitates targeted class augmentation, such as the artificial generation of lesions, thereby providing invaluable support for underrepresented classes within medical datasets. This capability is especially critical in ensuring the robustness and efficacy of DL models in medical applications, where the balance and completeness of training data directly influence model performance and diagnostic accuracy (Garcea, Serra, Lamberti, and Morra, 2023).

The exploration into the strengths of specific DL architectures reveals a nuanced understanding of the way different data augmentation strategies can enhance model performance in medical imaging. Affine transformations, known for their ability to preserve lines and parallelism, emerge as a fundamental tool for geometric adjustments without compromising the integrity of image content. Their simplicity and ease of integration into DL pipelines make them a staple augmentation strategy. Erasing transformations, by selectively removing parts of images, strengthen model robustness against occlusions and mitigate dataset biases, prompting models to avoid oversimplified detection patterns.

Elastic transformations introduce local shape variations, proving invaluable in simulating real-world deformations such as those induced by breathing or patient movements, especially for augmenting organs or lesions. Pixel-level transformations adjust image attributes like brightness, contrast, and saturation to enhance model robustness across varying scanners and imaging protocols. GANs stand out for their ability to generate realistic images, significantly diversifying data and pushing models towards improved generalisation. They are particularly beneficial for controlled image generation tasks, such as synthesising lesions or enriching un-

derrepresented classes in datasets.

Feature mixing methods, which combine features or parts of different images to create new samples, and model-based methods that incorporate physically or biologically inspired models for image generation or modification, both contribute to model generalisation and robustness. The latter notably supports the synthesis of artificial lesions or the simulation of disease progression. Reconstruction-based methods, operating on raw scanner data, allow for the simulation of acquisition artifacts or the variation of acquisition angles, furthering model Acc and robustness. However, the literature also delves into the challenges and considerations fundamental in selecting and implementing these data augmentation strategies. The effectiveness of these approaches can vary significantly depending on the organ, pathology, and data modality in question, necessitating a designed approach for different medical imaging tasks. Despite the potential of data augmentation to substantially elevate model performance and generalisation, it's imperative to maintain a balance between complexity and practicality, ensuring that augmented data accurately reflects realistic scenarios (Garcea et al., 2023).

In addition to tackling the constitutive challenges in medical image analysis, the adoption of transfer and reinforcement learning (TL and RL, respectively) techniques has emerged as a transformative strategy, allowing DL models to leverage pre-trained knowledge and significantly enhance performance, especially in scenarios with limited annotated medical datasets (Atasever, Azginoglu, Terzi, and Terzi, 2023; Hu, Zhang, Matkovic, Liu, and Yang, 2023). TL has been instrumental in addressing the challenges of medical image analysis, especially in situations where labelled data is insufficient, costly, or time-consuming to acquire. The fundamental concept of TL is to transfer knowledge from a source domain (where abundant data is available) to a target domain (where data is limited) (Figure 2.3).

This approach is similar to a person leveraging their existing knowledge to learn a new, related task more efficiently. In medical image analysis, TL has been widely used due to the strict requirements for expert annotations and the often-limited availability of labelled medical images. Pre-trained models on large datasets, such as ImageNet, are adapted to medical tasks through techniques like weight initialisation and fine-tuning. In former case, a pre-trained

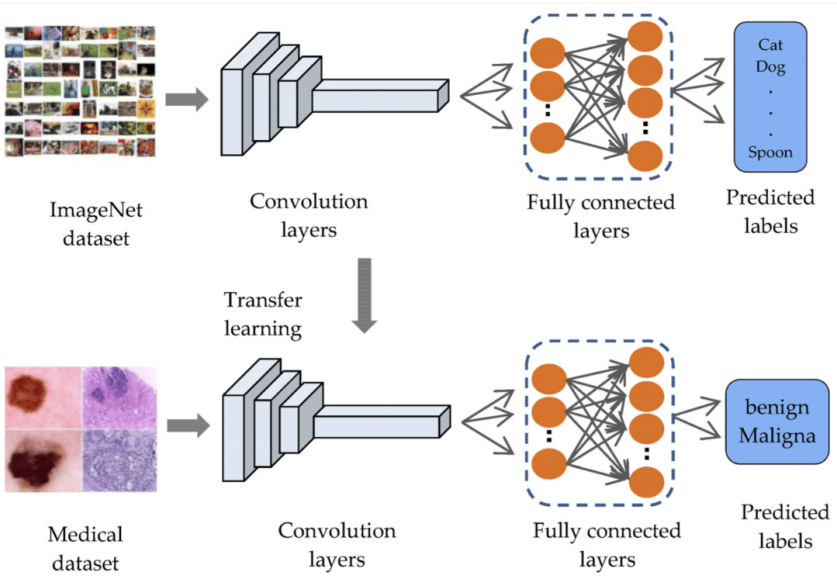


Figure 2.3: Transfer Learning from ImageNet (Mukhlif, Al-Khateeb, and Mohammed, 2023).

model's weights are used as a starting point and further updated with medical data. Fine-tuning involves adjusting the weights of certain layers while keeping others frozen, especially when the new task is similar to the pre-trained task but the dataset is relatively small. The impact of TL on medical image analysis has been significant, particularly in improving diagnostic accuracy, automating disease detection, and reducing the time and resources required for model training from scratch. For instance, TL has been employed successfully in various medical imaging tasks such as brain tumour segmentation, lung nodule detection, and breast cancer classification.

RL represents a dynamic and powerful approach in medical image analysis, offering the potential to significantly enhance decision-making processes and automate complex tasks by iteratively learning optimal strategies from interactions with the environment (Hu et al., 2023). The impact of RL on medical image analysis has been increasingly recognised for its potential in enhancing diagnostic accuracy and automating complex tasks. This comprehensive review delves into the integration of RL in various medical imaging tasks, highlighting its unique advantages and the wide range of its applications. Unlike traditional supervised and unsupervised learning models, RL thrives in environments where vast amounts of annotated data are limited or susceptible to bias, learning through interaction and exploiting past experiences.

This adaptability makes RL particularly valuable in medical settings, where data annotations can be resource-intensive and subject to human error. Medical image analysis tasks, including detection, segmentation, classification, and synthesis, have all benefited from RL's dynamic approach. From enhancing landmark and lesion detection to optimising segmentation processes and improving image classification with minimal training data, RL has demonstrated significant promise. Moreover, its role in medical image synthesis, particularly in semantic map generation and pixel value alteration, showcases its adaptability.

TL and RL are powerful paradigms in the field of medical image analysis, each with its unique advantages. However, they are not seamless, presenting several obstacles that can limit their effective implementation (Atasever et al., 2023; Hu et al., 2023). TL, while reducing the need for large, annotated datasets by leveraging knowledge from pre-trained models, faces the challenge of domain adaptation. When the source and target domains differ significantly, the transferred knowledge may not align well, leading to a phenomenon known as negative transfer. Ensuring that the model generalises well to the new domain without overfitting to the source domain's features requires careful fine-tuning and validation, a process that can be both time-consuming and computationally demanding (Atasever et al., 2023). RL, on the other hand, is fundamentally suited for environments where interaction and sequential decision-making are crucial. In medical image analysis, RL can navigate through sequential data, learning from the environment to make predictions or decisions. However, the complexity of defining the state, action, and reward space in medical contexts can be challenging. The high dimensionality of medical images and the implications involved in the medical decision-making process demand thorough design and tuning of the RL model. Furthermore, the training process for RL models is often computationally intensive and time-consuming, requiring a substantial number of trial-and-error interactions, which can be a significant bottleneck in time-sensitive medical settings (Hu et al., 2023). In both techniques, the interpretation and explainability of the model's decisions are crucial, especially in medical applications where decision-making needs to be transparent and justifiable. Ensuring that the models not only perform well but also provide insights into their decision process is a challenge that needs continuous attention.

In the field of medical image analysis, DL architectures have made significant progress across various disease types, each showcasing unique contributions along with fundamental limitations (Razzak, Naz, and Zaib, 2018; Mall et al., 2023). The most commonly used DL architectures for these applications vary based on the specific requirements of the disease diagnosis and the nature of the imaging techniques involved. For Diabetic Retinopathy (DR), automated detection of related diseases has been realised using Deep CNNs (DCNNs). Studies have demonstrated high sensitivity and specificity in classifying and detecting DR cases, particularly using datasets like EyePACS-1, Messidor-2, Kaggle fundus, DRIVE, and STARE. Additionally, in ophthalmology, architectures like the Visual Geometry Group 16 (VGG-16) model and the Network Followed Network (NFN+) model have been successful in tasks such as retinal vessel mapping and DR classification, achieving high area under the curve (AUC) scores. The ability of these networks to process high-resolution images and extract minute features is commendable. Nonetheless, the models may struggle with generalisability across different datasets and require extensive computational resources. For tumour detection in various body parts, DL methods have been utilised to process mammographic images, ultrasound images, and MRI scans. CNNs, along with SVMs for classification, have shown promising results. DL has also been applied in the analysis of MRI, Positron Emission Tomograp (PET) images, and functional MRI for the detection of AD's and Parkinson's diseases. CNNs, DBNs, and sparse AEs have been effectively used for feature extraction and classification. For lung diseases, U-Net and V-Net models have demonstrated efficacy in segmentation tasks, with high dice coefficient indices indicating precise delineation of lung regions. These models have also been pivotal during the COVID-19 pandemic for rapid diagnosis and segmentation of infections in lung CT scans. However, their performance heavily depends on the quality and size of the training data. Also, the time required for training and fine-tuning these complex models can be substantial.

While DL architectures continue to revolutionise medical image analysis, offering more automated, accurate, and faster diagnostics, several challenges persist. The reliance on large, annotated datasets for training, the computational cost of training and inference, and the need

for models that generalise well across diverse medical settings are some of the critical problems. Furthermore, the interpretability of these models remains a crucial aspect, especially in medical applications where understanding the model's decision-making process is vital for clinical acceptance and trust. Future research directions may focus on addressing these challenges, developing more robust and generalisable models, and enhancing the interpretability and transparency of DL systems in healthcare.

DL has revolutionised the field of medical image analysis, particularly in the recognition of multiple lesions from medical images (Jiang et al., 2023). This technology has brought significant advancements by enabling the analysis of complex image patterns and the identification of implicit differences between various lesion types, which are often challenging for human observers. DL models, particularly CNNs, have demonstrated remarkable performance in accurately classifying, detecting, and segmenting lesions in various organs such as the brain, skin, breast, lungs, and abdomen. These models have the ability to learn hierarchical feature representations from medical images, allowing them to capture both the local details and the global context of the lesions. As a result, DL-based approaches have shown great potential in improving diagnostic accuracy, reducing the workload of radiologists, and ultimately enhancing patient care by enabling early detection and treatment of diseases. Below are the tables (2.7 and 2.8) summarising the contribution of DL, the limitations, and the architectures/techniques applied for each of the following:

- Generalised paradigm for multiple-lesion recognition
- Multiple-lesion recognition in different body regions

Table 2.7: Generalised Paradigm for Multiple-Lesion Recognition (Jiang et al., 2023; Khan et al., 2020; McNeely-White, Beveridge, and Draper, 2020)

Aspect	**Contribution of Deep Learning**	**Limitations**	**Architectures/ Techniques Applied**
Classification	Improved accuracy in identifying specific diseases from medical images.	Requires large datasets for optimal performance; may struggle with highly imbalanced data.	CNNs, ResNet, VGG-16, VGG-19, Transfer Learning (TL) with fine-tuning
Detection	Ability to detect implicit and early-stage lesions, enhancing early diagnosis.	Challenges in dealing with varying image quality and artifacts.	Two-stage models (RCNN, Fast-RCNN), Single-stage models (YOLO, SSD), Attention Mechanisms
Segmentation	Precise segmentation of lesion areas, crucial for treatment planning and monitoring.	Requires high computational resources and can be sensitive to hyperparameter settings.	U-Net, V-Net, nnU-Net, Encoder-decoder architectures, Dice Loss for optimisation

Table 2.8: Multiple-lesion Recognition in Different Body Regions (Ananda et al., 2021; Cuevas-Rodriguez et al., 2023; Krizhevsky, Sutskever, and Hinton, 2012)

Disease	**Dataset**	**Contributions**	**Limitations**	**DL Architecture**
Brain Lesions	BRATS, CuRIOUS, HECKTOR, SLCN, ADNI, CAD-Dementia MRI	Identification of tumours. Widespread use in categorising conditions like AD and segmenting brain structures. Enhanced ability to differentiate between various types of brain lesions. Detection and classification of brain images	May require preprocessing to handle variations in image acquisition protocols.	CNN, DBN, Sparse AE, SVM, TL, Attention Mechanisms, Capsule Neural Network (CapsNets)

Continued on next page

Table 2.8: Multiple-lesion Recognition in Different Body Regions (Continued)

Disease	**Dataset**	**Contributions**	**Limitations**	**DL Architecture**
Ocular Lesions	EyePACS-1, Messidor-2	Detection and classification of DR cases. Improved classification of ocular diseases, such as Age-related Macular Degeneration (AMD) and Diabetic Macular Oedema (DMO). High AUC scores in tasks like retinal vessel mapping and DR classification.	Struggles with generalisability across datasets. Requires extensive computational resources. Sensitivity to image quality and the need for precise lesion segmentation.	VGG-16, NFN+, DCNN, CNN with multi-scale feature fusion, TL, Attention-CNN

Continued on next page

Table 2.8: Multiple-lesion Recognition in Different Body Regions (Continued)

Disease	**Dataset**	**Contributions**	**Limitations**	**DL Architecture**
Lung Lesions	LUNA16, ANODE09	High accuracy in segmentation. Essential during COVID-19 for rapid diagnosis and lung segmentation. Identification, description, and categorisation of tumours from CT imaging and radiography using CNNs. Unified techniques for identifying various illnesses with lung X-rays.	Requires large, quality datasets for training. Time-consuming training and fine-tuning. Need for large and diverse datasets to cover the spectrum of lung diseases.	V-Net, 2D-CNN, 3D-CNN, U-Net based segmentation, Multi-stage approaches.

The concept of using blocks of layers as structural units is also highlighted as an emerging direction (Khan et al., 2020; McNeely-White, Beveridge, and Draper, 2020). The survey categorises recent CNN architectures into seven distinct categories based on their innovative approaches: spatial exploitation, depth enhancement, multi-path processing, network width, feature-map exploitation, channel boosting, and attention mechanisms. The examination of CNNs underscores the shift towards more sophisticated architectural designs, moving beyond simple parameter optimisation to more complex structures that significantly enhance the network's ability to learn from data (Ananda et al., 2021; Cuevas-Rodriguez et al., 2023). Datasets of labelled images were relatively small, and object recognition in realistic settings was a chal-

lenge due to the variability in object presentations. Therefore, there was a need to leverage a large learning capacity provided by CNNs, combined with recently available large labelled datasets like ImageNet, to significantly improve performance on object recognition tasks. In this context, AlexNet, introduced by Alex Krizhevsky et al. in 2012, is one of the well-known CNNs that has significant breakthrough in the field of DL and computer vision where it became the foundational architecture of several subsequent CNN models (Krizhevsky, Sutskever, and Hinton, 2012). It significantly outperformed other models in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC-2012). AlexNet's architecture consists of 5 convolutional layers followed by 3 fully connected layers. It utilises Rectified Linear Unit (ReLU) as the activation function, deviating from the classic application of tanh and sigmoid functions commonly used at the time (Ananda et al., 2021). The use on ReLU nonlinearity improves training speeds by preventing vanishing gradient problems. Additionally, AlexNet employs dropout to avoid overfitting in its fully connected layers. It is also known for its implementation of data augmentation techniques such as image translations, horizontal reflections, and patch extractions. Towards reducing the size of the network and improving its ability to pick up on features, AlexNet uses overlapping max pooling. Despite its attempt to reducing overfitting and the usage of parallel splitting across graphics processing units (GPUs), AlexNet presented several drawbacks to include:

- Size and complexity where is relatively large and computationally intensive compared to traditional vision models, requiring significant GPU resources for training.
- Potential overfitting where despite the use of dropout, the network's complexity and depth mean it can still be vulnerable to overfitting without thorough oversight of regularisation and training data augmentation.
- Static Architecture where the latter is relatively fixed and does not employ modules or blocks that can be easily stacked or modified, limiting flexibility compared to other CNN modular-based designs.

These drawbacks led to the investigation of the impact of CNNs depth on the accuracy of

large-scale image recognition. Attempts to enhance these architectures have mainly focused on aspects like the size of the feature maps or the stride of convolutional layers. Towards filling a crucial gap by systematically exploring how increasing network depth, facilitated by small convolutional filters, VGG network has been introduced by Simonyan and Zisserman in 2014 (Simonyan and Zisserman, 2014). The primary attribute of VGG network is its genetic layout architecture's simplicity using fixed size of input RGB images of 224x224 passing it through 3x3 convolution filters. Similar to AlexNet, these are then followed by ReLU activation functions. These blocks are stacked increasing depth between 16 and 19 convolutional layers, namely VGG-16 and VGG-19 respectively. Each of these models performs max-pooling after a set of fully connected layers. VGG's design choice is pivotal, demonstrating that deeper networks could significantly improve upon prior architectures (Razzak, Naz, and Zaib, 2018; Bressem et al., 2020). Subsequently, VGG significantly improved the accuracy on the ImageNet dataset, showcasing the benefits of deeper networks (Yang et al., 2021; Bressem et al., 2020). Also, features learned by VGG networks have demonstrated a well transfer to other image recognition tasks, demonstrating the adaptability of the learned representations. Despite the deeper representation of VGG-19 compared to VGG-16 with the expectation to capture more complex features, the former could be resource-intensive and only provide marginal accuracy improvement which may not justify the increased complexity. That is, the computational intensity is a common theme for both VGG's versions where it limits their deployment on hardware with limited memory. Given their capacity and depth, both VGG networks could be vulnerable to overfitting, especially when trained on smaller datasets without adequate regularisation or data augmentation (Yang et al., 2021). Addressing the degradation problem observed when networks become deeper, where performance saturates and then degrades rapidly, has always been of great importance. Contrary to expectations, this degradation was not due to overfitting but because deeper networks are harder to optimise (Yang et al., 2021).

Authors in (He, Zhang, Ren, and Sun, 2016a) have introduced Residual Network (ResNet) in 2015. This novel architecture, through its residual learning framework, makes it easier to optimise the network and enable accuracy improvement from significantly increased depth. The

key innovation in ResNet is the introduction of "skip connections" or "shortcut connections" that allow gradients to flow through the network directly, addressing the vanishing gradient problem and enabling the training of networks with depths of up to hundreds or even thousands of layers. The main principle of those connections is to fit the input from preceding layer to the following layer without introducing any modification to the current input, which leads to have a deeper network. The core idea behind ResNets is to learn residual functions with reference to the layer inputs, instead of learning unreferenced functions. This approach allows the training of networks with depths of up to 152 layers, substantially deeper than conventional networks like VGG networks, yet maintaining lower complexity (He, Zhang, Ren, and Sun, 2016b). ResNet network utilises bottleneck designs for efficiency, with blocks containing 1x1, 3x3, and again 1x1 convolutions, where 1x1 convolutions reduce and then increase dimensions, leaving the 3x3 layer as a bottleneck.

Two major types of blocks are composing a ResNet network defined respectively as the identity and convolutional blocks, resulting two versions of skip connections including the identity and projection shortcuts. Identity connections bypass the original input to the addition operator of the current block in order to make sure that the following layer performance level will be at least the same as the previous layer without degradation. Projection shortcut, however, is a connection that performs a convolutional operation to ensure the volume size remain the same after each addition operator. The image volume size injected into a ResNet architecture has a noticeable impact on the skip connections. In case the input dimension is very small compared to the output image dimension, three solutions could be introduced to solve image size compatibility problem described as the following (Szegedy et al., 2015):

- Increasing the input dimension by performing all the skip connections as identity shortcuts mapping and setting a zero padding. In such case no extra parameters are required.
- Performing the projection shortcut mapping to only increase the input dimension and set the skip connections as an identity shortcut. As a result, extra few parameters are needed.
- Setting all the skip connections as projection shortcuts where a larger number of param-

eters is required, indeed.

The third solution proved a better efficiency compared to the other two solutions, based on experiments performed in (Huang, Liu, Van Der Maaten, and Weinberger, 2017b), where projection shortcuts ensured the volume size stability. ResNet network has been introduced in several versions differing mainly by the increase of the number of layers. Each of these versions present an ease of training due to skip connections helping with vanishing gradients, consistently strong performance, and adaptability in application beyond classification. However, a common drawback is that ResNet architecture is resource-intensive, particularly deeper models, diminishing outcomes on performance with increased depth, complexity in deployment on edge hardware.

Preventing overfitting has always been a target for several years. In fact, given the large number of parameters in deeper and wider networks, there is a significant risk of overfitting, especially when the number of labelled training examples is limited. Additionally, there has been a consistent need to accomplish computational efficiency. Towards tackling these issues, Inception network, originally introduced as GoogleNet by Google researchers in 2014, has been designed with the aim to approximate an optimal sparse structure with dense, computationally efficient components, allowing for increased network depth and width without a significant increase in computational requirements (Szegedy et al., 2015). Inception is well known for its deep and wider architecture that achieves high accuracy in image classification tasks with relatively efficient computation. This was achieved by introducing the inception module allowing for efficient computation by combining filters of different sizes to include 1x1, 3x3, and 5x5 convolutions within the same layer. Inception's foundational architecture applies max-pooling layers to reduce the spatial dimensions of the feature maps. The network also uses batch normalisation (BN), Root Mean Squared Propagation (RMSprop) optimiser, and employs a global average pooling layer at the end of the network instead of fully connected layers, which reduces the total number of parameters and helps to control overfitting. The original version of Inception network, known as Inception-V1, has introduced a level of implementation complexity due its novel inception module which did not exist in prior CNNs. In addition, this version

lacks in residual connections, which could have improved training for even deeper networks. Follow-up versions have been developed towards enhancing the foundational architecture of Inception-V1, namely, Inception-V2 and Inception-V3. Despite the improved architecture, the reduction of number of parameters and computational cost through factorisation as well as label smoothing implementation towards regularising the model and preventing overfitting on the training data, Inception-V2 and Inception-V3 faced several drawbacks. In fact, as improvements were added, the architectures of these versions became increasingly complex and harder to replicate. Moreover, the increase in accuracy began to require more complex engineering and hyperparameter tuning. Deeper networks, unlike wider networks such as Inception family, are suffering from several problems including:

- Hard detection of salient objects that have large size variation.
- Choosing the right size of kernel.
- Stacking convolutional layers to get a very deep network causes an overfitting and an expensive computation cost.

In order to overcome the aforementioned problems, Inception-V4 has been introduced performing with filters that have multiple sizing based on the information distribution: (1) larger kernel for globally distributed information, and (2) smaller kernel for locally distributed information. This leads to obtain a wider network instead of deeper one which represents the principle of inception module composing the Inception family networks. The drawbacks presented by the naïve Inception-V1, complex Inception-V2 and V3, it has been suggested to have uniform modules in order to boost the network performances. As a result, the stem modules have been modified and introduced three different versions of inception modules named A, B, and C respectively (Tan and Le, 2019). Inception-V3 has introduced a new module called the reduction module that aims to reduce the computation complexity by going wider instead of deeper and to avoid the loss of information that might be due to an excessive reduction of dimensions. However, its reduction blocks have not been explicitly implemented, unlike reduction blocks of Inception-V4. In fact, two main reduction blocks have been introduced

including reduction block A that reduces a 35x35 dimension into 17x17 dimension, and a reduction block B that reduces the resulted dimension into 8x8 size. The hybrid integration of Inception with ResNet network, resulting Inception-ResNet, has combined the benefits of inception modules with residual connections, leading to easier training of deeper networks. In addition, this combination has offered improvements in accuracy and efficiency, matching or surpassing other architectures of its time. Therefore, it allowed for scaling the network deeper without the degradation problem. However, this architecture became one of the most complex, combining two powerful ideas but at the cost of simplicity. Also, training and deploying these models require significant computational resources, especially for real-time applications.

DL models, including CNNs, struggled with the problem of vanishing gradients, making it challenging to train very deep networks. While architectures like ResNets introduced skip connections to mitigate this issue, a novel model introduced by Gao Huang et al. in 2017, took this concept further by ensuring maximum information and gradient flow between all layers in the network, named Densely Connected Convolutional Network (DenseNet) (Huang et al., 2017b). This was achieved by connecting each layer to every other layer directly in feed-forward mode, enhancing feature propagation and encouraging feature reuse, which in turn addressed the vanishing gradient problem more effectively and led to models that were both deep and efficient to train. For a DenseNet with L layers, there are L(L+1)/2 direct connections. Unlike traditional CNNs, where the input to each layer is only the output from the previous layer, in DenseNet, each layer receives the concatenated outputs from all preceding layers as its input. This dense connectivity pattern allows for substantial depth in the network with fewer parameters. DenseNets consist of multiple densely connected blocks, with layers within each block being directly connected to every other layer. Transition layers, which perform convolution and pooling operations, connect these dense blocks. This design allows for significant reductions in parameters through feature reuse while maintaining or improving model performance on various benchmarks. Similar to ResNet, DenseNet presented several versions varying on the number of layers, hence parameters, as follows:

- DenseNet-121: this version is designed to balance between efficiency and computational

requirement, which makes it suitabefor wider range of applications.

- DenseNet-169: this version presents a deeper architecture towards improving accuracy on complex datasets.
- DenseNet-201: this version provides a higher capacity and potentially better performance on very challenging visual recognition tasks.
- DenseNet-264: this version represents the most extensive version, aiming at expanding the limits of performance in various benchmarks.

Each DenseNet variant is characterised of a growth rate which critically influences the model's depth and complexity, effectively resulting the number of new features each layer contributes to the global feature map, thus balancing between model efficiency and its ability to represent complex features. Despite their significant contributions in enhancing deep CNNs architectures, DenseNet networks present a variety of disadvantages to include:

- Memory Consumption: The concatenation of feature maps from all preceding layers could lead to increased memory usage during training, which may be a constraint on hardware with limited resources.
- Computational Overhead: While parameter-efficient, the dense connections increase the computational complexity, especially as the network depth increases.
- Potential for Overfitting: Despite its regularising effect, the extensive reuse of features in some scenarios might lead to overfitting, particularly on smaller datasets without the consideration of relevant regularisation techniques.

Another version fundamentally composed of DenseNet architecture has been introduced by same authors in 2018, namely Multi-Scale DenseNet (MSDNet) (Huang et al., 2017a). MSDNet has been designed on the basis that it can adapt its computational resource usage dynamically, depending on the complexity of the input image. This adaptability makes it highly suitable for real-world applications where computational resources are a bottleneck. The complex

design combining multi-scale representation and dense connectivity could potentially introduce challenges in understanding, implementing, and optimising the network. The architecture of MSDNet introduces a risk of overfitting, especially when trained on limited data without appropriate regularisation techniques. Similar to DenseNet, while designed for efficient processing, the multi-scale and dense connectivity aspects of MSDNet network could negatively impact the computational overhead during the training phase.

In addition to the discussed well known CNN architectures, several convolution-based models have been introduced for image processing to include: SqueezeNet, MobileNet, ShuffleNet, EfficientNet, Extreme Inception Network (Xception), High-Resolution Network (HRNet), HigherHRNet, Neural Architecture Search Network (NASNet), NoisyStudent, ResNeXt, and Squeeze-and-Excitation based Network (SENet) (Bianco, Cadene, Celona, and Napoletano, 2018). Table 2.9 presents a brief overview of each architecture, its advantages and disadvantages.

Table 2.9: Overview of Additional CNN Architectures: Advantages and Disadvantages

CNN Network	**Architecture Overview**	**Advantages**	**Disadvantages**
SqueezeNet (Iandola et al., 2016)	• Utilises Fire modules consisting of a squeeze layer with 1x1 filters followed by an expand layer with a mix of 1x1 and 3x3 filters. • Begins with a single convolution layer, followed by eight Fire modules, concluding with a final convolution layer. • Employs delayed downsampling for accuracy improvement.	• Reduces communication overhead in distributed training. • Facilitates deployment on hardware with limited memory. • Maintains competitive accuracy with significantly reduced size.	• Complexity in Design where it requires critical trade-off considerations between model size, efficiency, and accuracy. • Limited capacity where parameter reduction might affect performance on more challenging tasks.

Continued on next page

Table 2.9 : Overview of Additional CNN Architectures: Advantages and Disadvantages (Continued)

CNN Network	**Architecture Overview**	**Advantages**	**Disadvantages**
MobileNet-V1 (Howard et al., 2017)	• Uses depth-wise separable convolutions	• Highly efficient • Suitable for mobile devices • Retains reasonable accuracy	• Lower accuracy compared to larger models • Trade-off between latency and accuracy
MobileNet-V2 (Howard et al., 2017)	• Replaced residual structure with linear bottlenecks • Uses shortcut connections	• Improved efficiency and accuracy compared to V1 • Better performance on various tasks without significant size increase	• Might underperform against recent architectures in speed-accuracy trade-off • Replaced residuals can be complex to implement
MobileNet-V3 (Howard et al., 2017)	• Combines architecture search techniques with NetAdapt algorithm • Includes squeeze-and-excitation blocks	• Further efficiency and accuracy improvements • Incorporates feature recalibration leading to higher performance	• More complex due to manual and architecture search components • Optimisation can be challenging

Continued on next page

Table 2.9 : Overview of Additional CNN Architectures: Advantages and Disadvantages (Continued)

CNN Network	**Architecture Overview**	**Advantages**	**Disadvantages**
ShuffleNet-V1 (Zhang, Zhou, Lin, and Sun, 2018)	• Introduced group convolutions and channel shuffle operations to reduce computational cost while maintaining accuracy.	• Highly efficient • Suitable for mobile and embedded devices. • Maintains competitive accuracy with fewer parameters.	• Might limit representational power due to group convolutions • Requiring critical tuning of hyperparameters.
ShuffleNet-V2 (Ma, Zhang, Zheng, and Sun, 2018)	• Focused on practical design for real-world applications • Optimising for direct metrics like speed and efficiency.	• Improved computational efficiency and simplified architecture without sacrificing performance. • Better utilisation of hardware acceleration.	• Enhancements require architectural changes, complicating migration from initial version. • Might still not match the accuracy of more complex models on certain tasks.

Continued on next page

Table 2.9 : Overview of Additional CNN Architectures: Advantages and Disadvantages (Continued)

CNN Network	**Architecture Overview**	**Advantages**	**Disadvantages**
EfficientNet-B0 to B7 (Tan and Le, 2019; Tan, Pang, and Le, 2020)	• Based on a baseline architecture (B0) scaled up to B7 using an aggregated coefficient. • Utilises mobile inverted bottleneck convolutions with squeeze-and-excitation optimisation.	• High efficiency with fewer parameters and floating-point operations (FLOPs). • Scalable across different computational budgets.	• Decrease in accuracy in case of larger models (B6, B7) • resource-intensive.
EfficientNet-V2 (Tan and Le, 2021)	• Incorporates Fused- Mobile Inverted Bottleneck Convolution (MB-Conv), progressive learning, and optimised scaling for depth, width, and resolution.	• Faster training and improved efficiency without sacrificing accuracy. • Models are more lightweight and require less computational power for both training and testing.	• More complex and resource-intensive then the original EfficientNet.

Continued on next page

Table 2.9 : Overview of Additional CNN Architectures: Advantages and Disadvantages (Continued)

CNN Network	**Architecture Overview**	**Advantages**	**Disadvantages**
Xception (Chollet, 2017)	• Comprises 36 convolutional layers organised into 14 modules around depthwise separable convolutions, which include depthwise convolutions followed by pointwise convolutions. • Uses linear residual connections around most modules.	• Outperforms Inception V3 on the ImageNet and significantly on larger datasets. • Easier to define and modify due to its linear stack of depthwise separable convolutions. • Achieves performance enhancement through more efficient use of parameters.	• May need optimisations at depthwise convolution operations to match Inception V3's speed. • Performance may heavily rely on optimisation configurations. • further tuning required for optimal results on various datasets.
HRNet-V1 (Wang et al., 2020)	• Starts with high-resolution stream • Adds lower-resolution streams with parallel connections.	• High accuracy for pose estimation and segmentation. • Adaptable across tasks	• Higher computational and memory demands

Continued on next page

Table 2.9 : Overview of Additional CNN Architectures: Advantages and Disadvantages (Continued)

CNN Network	**Architecture Overview**	**Advantages**	**Disadvantages**
HRNet-V2 (Wang et al., 2020)	• Similar to the initial version with modifications for object detection, improved cross-scale connections.	• Improved feature fusion for detection • Maintains task flexibility with enhancements.	• Increased model complexity • Higher training and testing time.
HRNet-W (Wang et al., 2020)	• Core HRNet ideas optimised for classification • reduces network width	• More efficient than original HRNet • Competitive performance for classification	• Trade-off between efficiency and peak accuracy, especially for high-resolution images.

Continued on next page

Table 2.9 : Overview of Additional CNN Architectures: Advantages and Disadvantages (Continued)

CNN Network	Architecture Overview	Advantages	Disadvantages
HigherHRNet (Cheng et al., 2020)	• HRNet Backbone: Utilises HRNet's high-resolution maintenance capabilities. • High-Resolution Feature Pyramid: Incorporates a deconvolution module for generating higher resolution feature maps. • Multi-Resolution Supervision: Applies different resolution targets during training to ensure effective scale variation handling. • Multi-Resolution Heatmap Aggregation: Aggregates heatmaps from different resolutions in testing for improved keypoint detection.	• Significantly enhances keypoint localisation accuracy for small figures. • Learns to produce heatmaps that account for person scale variations, improving pose estimation. • Outperforms existing bottom-up and some top-down methods on benchmarks without needing refinement techniques.	• Added modules introduce computational complexity. • The approach may present training and optimisation challenges, including potential overfitting on certain scales.

Continued on next page

Table 2.9 : Overview of Additional CNN Architectures: Advantages and Disadvantages (Continued)

CNN Network	**Architecture Overview**	**Advantages**	**Disadvantages**
NASNet-A, B, C (Qin and Wang, 2019)	• Variants discovered during neural architecture search • focusing on scalability and performance. • NASNet-A is the most performant model.	• High accuracy on benchmarks • Scalability across different computational requirements • Transferability to other tasks.	• Computationally intensive search process • Complex architecture.
NASNet-Mobile (Qin and Wang, 2019)	• A scaled-down version of NASNet-A • optimised for efficiency and performance on mobile devices.	• Designed for efficiency on mobile devices. • Competitive accuracy for its size.	• Sacrifices some accuracy for efficiency • Adapting and optimising for a wide range of devices can be challenging
NASNet-Large (Qin and Wang, 2019)	• A scaled-up version of NASNet-A • optimised for maximum accuracy and feature richness.	• Top accuracy on image classification tasks • Produces rich feature representations.	• High computational requirements • May not be suitable for real-time applications or on-device testing

Continued on next page

Table 2.9 : Overview of Additional CNN Architectures: Advantages and Disadvantages (Continued)

CNN Network	Architecture Overview	Advantages	Disadvantages
NoisyStudent (Xie, Luong, Hovy, and Le, 2020)	• Base Models: EfficientNet architectures used as starting points for teacher and initial student models. • Data Augmentation and Noise Injection: Variations include stochastic depth, and dropout. • Model Size Scaling: Application of NoisyStudent across different scales of EfficientNet (B0, B1, B2, etc.), demonstrating larger models benefit more from the approach.	• Achieves state-of-the-art results on benchmarks like ImageNet. • Makes models more robust, improving generalisation. • Effectively leverages abundant unlabelled data, making it a cost-effective strategy for improving model performance.	• Requires significant resources for training, pseudo-labelling, and iterative retraining. • Effectiveness depends on the quality of pseudo-labels generated by the teacher model. • The iterative training process adds complexity, requiring careful management of noise and augmentation strategies.

Continued on next page

Table 2.9 : Overview of Additional CNN Architectures: Advantages and Disadvantages (Continued)

CNN Network	**Architecture Overview**	**Advantages**	**Disadvantages**
ResNeXt (Xie, Girshick, Dollár, Tu, and He, 2017)	• Utilises residual blocks with grouped convolutions, featuring multiple parallel paths (or groups) within each block. • This structure is repeated throughout the network, creating a highly modular and scalable architecture. • Different versions include ResNeXt-50 (32x4d), ResNeXt-101 (32x4d), and ResNeXt-101 (64x4d), varying in depth and group configurations.	• Balances computational efficiency with high accuracy. • Modular design allows for easy adjustment of depth, width, and cardinality making it scalable. • Generally, outperforms ResNet models of similar complexity.	• More complex to implement than traditional architectures due to grouped convolutions. • Larger versions require significant computational power. • Increasing model size can lead to decreasing accuracy improvements.

Continued on next page

Table 2.9 : Overview of Additional CNN Architectures: Advantages and Disadvantages (Continued)

CNN Network	**Architecture Overview**	**Advantages**	**Disadvantages**
SENet network (Hu, Shen, and Sun, 2018)	• The core of SENet, the SE block, involves two operations: Squeeze, which aggregates feature maps across spatial dimensions to produce a channel descriptor, and Excitation, which captures channel-wise dependencies through a gating mechanism. • This recalibration process enhances models performance with minimal additional computational overhead.	• Enhances the network's ability to focus on relevant features, improving performance on various tasks. • SE blocks can be integrated into a wide range of existing CNN architectures, making SENet a flexible enhancement method. • Despite their effectiveness, SE blocks introduce minimal computational overhead.	• While conceptually straightforward, fine-tuning augmented models to achieve optimal performance can require additional resources. • The added parameters and complexity might lead to overfitting without proper regularisation. • The improvements may be less significant when applied to already highly optimised models.

2.2.2 ML Classifiers

When CNNs are integrated with ML classifiers, the potential for creating an efficient and accurate medical image processing framework is dramatically enhanced. CNNs, as mentioned above, known for their prowess in feature extraction directly from image data, combined with the robust classification capabilities of ML classifiers such as Random Forest (RF), Decision Tree (DT), eXtreme Gradient Boosting (XGBoost), and Adaptive Boosting (AdaBoost), could lead to a synergistic effect that can significantly improve the performance of medical image analysis. This fusion, in addition to its potential to leverage the DL strengths of CNNs in handling raw image data but also it capitalises on the nuanced decision-making and interpretability offered by traditional ML classifiers. Such a hybrid approach promises to enhance to current state-of-the-art methods in medical imaging, offering the potential for even more precise diagnoses, improved patient outcomes, and a deeper understanding of complex diseases.

Table 2.10: ML Models Working Principles and Mathematical Representation

ML Model	**Overview**	**Mathematical Representation**
DT (Charbuty and Abdulazeez, 2021)	• DT is a tree-like model that resembles a flowchart. • It includes internal nodes, branches, leaf nodes, and root node. • Each tree partitions data recursively based on attribute values, a process known as recursive partitioning. • The structure of a DT assists in visualising the decision-making process.	The information gain for a split on feature A is given by Equation 2.1, where: • I is the impurity measure, • D_p, D_{left}, D_{right} are the datasets of the parent, left child, and right child nodes, • N_p, N_{left}, N_{right} are the number of samples in each dataset.

Continued on next page

Table 2.10: ML Models Working Principles and Mathematical Representation (Continued)

ML Model	**Overview**	**Mathematical Representation**
RF (Dai, Bai, Sun, Huang, and Wang, 2018)	• RF uses multiple DTs for tasks like classification and regression. • It builds many DTs at training time. • For classification, it uses the mode of the classes predicted by the trees; for regression, it uses the mean prediction. • Overfitting Correction where RF addresses the tendency of DTs to overfit their training data. • Randomness in trees where each tree is built with a degree of randomness, either from different data samples or using different feature subsets for splits.	For a regression problem, the prediction of an RF is presented in Equation 2.2, where • $y_t(x)$ is the prediction of the t-th DT • T is the total number of trees. For a classification problem with classes C, the prediction is the class with the majority vote, as shown in Equation 2.3

Continued on next page

Table 2.10: ML Models Working Principles and Mathematical Representation (Continued)

ML Model	Overview	Mathematical Representation
XGBoost (Ramraj, Nagamalai, Pandian, and Vimala, 2016)	• XGBoost is an advanced form of gradient boosting, known for its efficient and scalable performance. • Parallel Processing: Utilises parallel tree boosting, improving speed and accuracy. • Versatile framework where it is compatible with Gradient Boosting DTs (GBDT) and Gradient Boosting Machines (GBM) methods. • Distributed Computing: Operates on distributed systems like Hadoop, Sun Grid Engine (SGE) grid, and Message Passing Interface (MPI). • Large-Scale Application: Capable of handling problems with billions of data points.	XGBoost involves the summing of predictions from N additive functions (trees), formulated as in Equation 2.4, where: • y^i is the predicted value for the i-th instance, • ϕ is the model, • f_k is a function representing an individual tree, • F is the space of all possible trees. The objective function to be minimised is presented in Equation 2.5, where: • l is a differentiable convex loss function that measures the difference between the predicted value y^i and the actual target y_i, • Ω is the regularisation term.

Continued on next page

Table 2.10: ML Models Working Principles and Mathematical Representation (Continued)

ML Model	Overview	Mathematical Representation
AdaBoost (Sevinç, 2022)	• AdaBoost can be combined with other learning algorithms to boost their performance. • Weak learners integration where it integrates outputs from multiple weak learners into a weighted sum for the final prediction. • Adjusts the influence of weak learners based on the accuracy of previous predictions. • It is particularly sensitive to noisy data and outliers. • AdaBoost is less prone to overfitting compared to other learning algorithms in certain situations.	The AdaBoost model representation is as formulated in Equation 2.6, where: • $h_t(x)$ is the weak learner obtained in the t-th iteration, • α_t is the weight assigned to $h_t(x)$, which is computed based on the error rate of h_t on the training data. The weights α are calculated as shown in Equation 2.7, where ε_t is the error rate of the classifier h_t.

The aforementioned ML classifiers have made substantial contributions to the field of medical image processing. These algorithms offer a range of benefits, including the ability to handle complex datasets with high dimensionality, provide interpretability in their decision-making processes, and achieve high accuracy rates in classification tasks. In medical image processing, these classifiers play a crucial role in identifying and categorising various pathologies, enhancing diagnostic precision, and facilitating timely and personalised treatment plans. Each algorithm utilises the collective power of simpler models to achieve greater predictive performance. RF leverages the diversity of multiple DTs to reduce overfitting. On the other hand, DT serves as the fundamental building block for more complex models. In addition, XGBoost optimises gradient boosting for speed and performance, and AdaBoost iteratively refines its learning from

the errors of previous models. Table 2.10 explains the mechanisms and mathematical underpinnings that empower these algorithms to excel in predictive tasks, where Equation 2.1, 2.2, 2.3, 2.4, 2.5, 2.6, 2.7 are defined as follows:

$$IG(D_p, A) = I(D_p) - \frac{N_{left}}{N_p} * I(D_{left}) - \frac{N_{right}}{N_p} * I(D_{right}) \tag{2.1}$$

$$y(x) = \frac{1}{T} \sum_{t=1}^{T} y_t(x) \tag{2.2}$$

$$y(x) = mode\{y_1(x), y_2(x), \ldots, y_T(x)\} \tag{2.3}$$

$$\hat{y}_i = \phi(x_i) = \sum_{k=1}^{N} f_k(x_i), f_k \in F \tag{2.4}$$

$$Obj(\Theta) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{N} \Omega(f_k) \tag{2.5}$$

$$F(x) = \sum_{t=1}^{T} \alpha_t * h_t(x) \tag{2.6}$$

$$\alpha_t = \frac{1}{2} ln(\frac{1 - \varepsilon_t}{\varepsilon_t}) \tag{2.7}$$

2.2.3 Medical Image Modalities

By efficiently analysing vast amounts of medical imaging data, such as MRI scans, OCT Fundus and X-rays, these ML classifiers help uncover critical insights into patient health, significantly contributing to advances in medical diagnostics and treatment strategies (Figure 2.4).

MRI scans are a sophisticated medical imaging technique used to capture high-resolution images of the inside of the human body without the use of ionising radiation, making them a safer alternative to X-rays and CT scans. These scans utilise strong magnetic fields and radio waves to generate detailed images of organs, soft tissues, bone structures, and other internal body parts. These images are crucial for diagnosing a wide range of conditions, from brain tumours and spinal cord injuries to musculoskeletal disorders and diseases affecting the heart and inter-

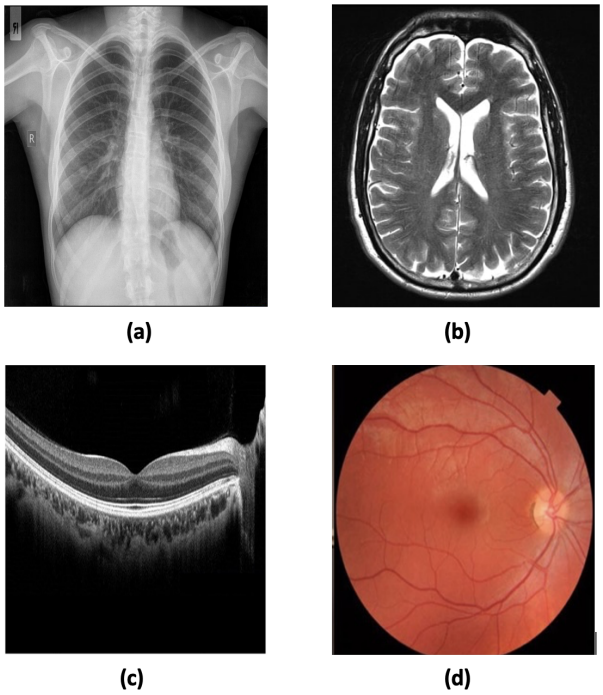


Figure 2.4: Medical Images Considered in This Research: (a) Chest X-ray (CXR) Image, (b) Brain MRI Scan, (c) OCT Scan, and (d) Fundus Image.

nal organs. OCT scans, on the other hand, offer a window into the eye's retina, capturing its layers with significant clarity to aid in the management of ocular diseases. Fundus photography complements this by providing detailed images of the eye's interior, crucial for tracking changes over time. Meanwhile, X-ray imaging pierces beyond the surface, revealing the hidden architecture of bones and dense tissues. Together, these imaging modalities constitute a set of investigative imaging modalities that enhance the understanding of various medical conditions, allowing for early detection and informed treatment decisions. These medical images will be the main focus in the various experiments conducted in this research. Table 2.11 summarises an overview of each medical imaging type.

Table 2.11: Medical Images Characteristics: MRI, OCT, Fundus, and X-ray

Medical image	**Characteristics**
MRI	• High-Resolution Interior Imaging: MRI scans produce detailed images of internal body structures, including soft tissues, organs, and bones, using magnetic fields and radio waves. • Non-Ionising Procedure: Unlike X-rays and CT scans, MRI does not use ionising radiation, making it safer for repeated use. • Soft Tissue Contrast: Exceptionally effective in distinguishing between different types of soft tissues, making it invaluable for diagnosing brain, spinal cord, joint, and muscle disorders. • Disease Detection and Monitoring: Utilised for diagnosing and monitoring various conditions, such as tumours, stroke, and degenerative diseases. • Guidance for Procedures: MRI guidance can be used for certain types of biopsies or for planning surgeries (Desikan et al., 2006).
OCT	• Non-invasive Imaging Test: OCT uses light waves to create cross-sectional images of the retina. • Retinal Examination: It visualises the distinct layers of the retina, the light-sensitive part of the eye. • Mapping and Measurement: Allows ophthalmologists to map the retina and measure its thickness accurately. • Disease Detection: These detailed measurements assist in detecting various eye diseases. • Treatment Guidance: Helps guide treatment for conditions such as age-related macular degeneration, diabetic eye disease, and glaucoma (Wieser, Biedermann, Klein, Eigenwillig, and Huber, 2010).

Continued on next page

Table 2.11: Medical Images Characteristics: MRI, OCT, Fundus, and X-ray (Continued)

Medical image	**Characteristics**
Fundus	• Interior Eye Imaging: Fundus photography captures images of the eye's interior, such as the retina, OD, macula, and posterior pole. • Equipment Used: Utilises a complex microscope combined with a high-resolution camera. • Diagnostic Tool: Aids in diagnosing various eye diseases. • Documentation: Helps in documenting the progression of eye conditions. • Monitoring: Used for ongoing monitoring of eye health (Bernardes, Serranho, and Lobo, 2011).
X-ray	• Non-Invasive Technique: X-ray imaging views the body's internal structures without surgery. • Radiography: Utilises X-rays to create images inside the body. • Bone Imaging: Especially effective for visualising bones, which absorb X-rays differently from soft tissues. • Fracture Identification: Commonly used to detect bone fractures. • Injury and Infection Assessment: Helps in identifying areas of injury or infection. • Foreign Object Location: Aids in locating foreign objects embedded in soft tissues (Huda and Abrahams, 2015).

2.2.4 Performance Evaluation Metrics

Evaluation metrics play a critical role in the development and validation of ML and DL applications, serving as essential tools for quantifying the performance of algorithms and models. These metrics provide a quantifiable measure of the performance of a model in prediction decisions when aligned with the actual outcomes, thereby facilitating the comparison of different

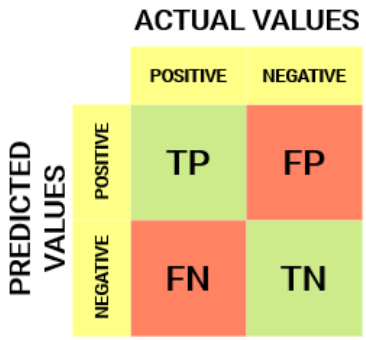


Figure 2.5: Confusion Matrix Representation.

models and guiding the selection of the most effective one for a given task. Common metrics such as accuracy, precision, recall, and the F1-score are widely used in classification tasks to evaluate the correctness and relevance of the predictions. For regression tasks, metrics like mean absolute error (MAE) are employed to assess the deviation of predicted values from the true values. More complex applications, especially those involving medical imaging or natural language processing, may require specialised metrics like the Area Under the Receiver Operating Characteristic curve (AUC-ROC) for evaluating model performance in a more nuanced manner.

In this work, the main focus is on the following set of key evaluation metrics to include:

- **Confusion Matrix:** is used to describe the performance of a classification model. As presented in Figure 2.5, confusion matrix summarises the number of correct and incorrect predictions with count values and is broken down by each class to include the following values:

  - True Positives (TP): The cases in which the class was positive and the model predicted positive.

  - True Negatives (TN): The cases in which the class was negative and the model predicted negative.

  - False Positives (FP): The cases in which the class was negative but the model predicted positive.

  - False Negatives (FN): The cases in which the class was positive but the model predicted

negative.

- **Accuracy (Acc):** widely used in ML and DL applications measuring the proportion of correctly classified examples over the total number of examples. Mathematically, Acc is represented as:

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.8}$$

- **Sensitivity (Sen):** Used in binary classification to measure the proportion oftrue positive examples that are correctly classified by the model. Mathematically, Sen is represented as:

$$Sen = \frac{TP}{TP + FN} \tag{2.9}$$

- **Specificity (Spe):** Used in binary classification to measure the proportion of TN examples that are correctly classified by the model. Mathematically, Spe is defined as follows:

$$Spe = \frac{TN}{TN + FP} \tag{2.10}$$

- **Precision-Recall (PR)** Used to measure the trade-off between precision and recall at different classification thresholds. Precision and Recall are represented as follows:

$$Precision = \frac{TP}{TP + FP} \tag{2.11}$$

$$Recall = \frac{TP}{TP + FN} \tag{2.12}$$

- **ROC curve:** Provides a visualisation of the trade-off between true positive and false positive rates (TPR and FPR, respectively). TPR and FPR are presented as follows:

$$TPR = \frac{TP}{TP + FN} \tag{2.13}$$

$$FPR = \frac{FP}{FP + TN} \tag{2.14}$$

- **F1-score:** is the harmonic mean of precision and recall, giving both metrics equal weight. It is useful when you need a balance between precision and recall. Mathematically, it is presented as follows:

$$F1 - score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{2.15}$$

- **AUC:** Used to measure the performance of a classifier by calculating the area under the ROC and PR curve. Mathematically, the AUC can be found using integral calculus by integrating the curve in the ROC or PR space:
  For AUC-ROC:

$$AUC - ROC = \int_0^1 TPR(FPR)d(FPR) \tag{2.16}$$

  For AUC-PR:

$$AUC - PR = \int_0^1 Precision(Recall)d(Recall) \tag{2.17}$$

To draw a clear line on how to choose the best set of evaluation metrics, it is important to understand the advantages and inconveniences of each parameter. Table 2.12 provides a summary comparison of aforementioned parameters.

Table 2.12: Comparative Evaluation of Performance Metrics

Measurement parameter	**Pros**	**Cons**
Acc	Easy to understand and calculate	Can be misleading with imbalanced datasets
Sen	Focuses on TPR	Doesn't account for TNs
Spe	Focuses on TNR	Doesn't account for TPs
PR	Good for imbalanced datasets	Doesn't account for TNs
F1-score	Accounts for both precision and recall	Can be biased towards either precision or recall
AUC score	Aggregates performance across all classification thresholds	Doesn't account for class distribution
ROC curve	Useful for imbalanced datasets, where it may be more important to optimise the TPR or FPR/FNR than the overall accuracy.	The choice of the classification threshold can have a significant impact on the shape of the curve, and the optimal threshold may be problematic.

In addition to the above, loss metrics are commonly used evaluation metrics for ML and DL algorithms. These metrics measure the difference between the predicted and actual values and provide an indication of how well the algorithm is performing. The following are some of the commonly used loss metrics for different types of problems (Simonyan and Zisserman, 2014; Ronneberger, Fischer, and Brox, 2015):

- **MAE:** used mainly for regression problems. It measures the average of the absolute differences between the predicted and actual values. MAE is presented as follows:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{2.18}$$

where n represents the number of observations, y_i denotes the actual values, and $\hat{y}_i$ represents the predicted values.

- **Categorical Cross-Entropy (LCE):** is used for multi-class classification problems. It measures the difference between the predicted probabilities and the actual class labels. The formula of LCE for a single example is as follows:

$$LCE(y,\hat{y}) = -\sum_{j} y_i * log(\hat{y}_j) \tag{2.19}$$

where:

- y is the binary indicator (0 or 1) if class label j is the correct classification for the observation,

- $\hat{y}$ is the predicted probability of the observation being of class j. For a dataset with N examples is as follows:

$$LCE = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j} y_{ij} log(\hat{y_{ij}}) \tag{2.20}$$

Each loss metric has its own advantages and disadvantages, depending on the problem domain and the characteristics of the data. Table 2.13 presents a comparative summary of loss metrics.

Table 2.13: Comparative Evaluation of Loss Metrics

Loss Metric	**Pros**	**Cons**
LCE	Encourages the model to output high confidence for the correct class and low confidence for incorrect classes	Not accurate with imbalanced data and sensitive to label noise and mislabelling

Continued on next page

Table 2.13: Comparative Evaluation of Loss Metrics (Continued)

Loss Metric	**Pros**	**Cons**
MAE	Less sensitive to outliers	Does not punish larger errors more severely than smaller errors

2.3 Advancements in Learning-based Segmentation Techniques for Medical Image Processing

The field of medical image processing has experienced significant advancements, driven by the incorporation of learning-based techniques. These innovations, originating from the domains of ML and DL, have transformed the way medical images are analysed, interpreted, and leveraged for diagnostic objectives. From CNNs adept at image classification and segmentation to RL and TL algorithms that enhance image reconstruction, the deployment of these advanced computational strategies has markedly increased the precision, efficiency, and dependability of medical diagnostics. These advancements not only enable early and accurate disease detection but also improve treatment planning and patient monitoring, indicating a new era of personalised medicine. As the exploration of ML and DL in medical image processing progresses, the potential for novel discoveries and innovations that further refine and improve patient care is extraordinarily promising.

In this context, authors in (Ngo, Lu, and Carneiro, 2017) introduced a novel approach for the automated segmentation of the left ventricle from cardiac cine MRI data using DBN network, addressing the challenge of large shape and appearance variations in the visual object of interest, especially when the annotated training set is small. This advantage is particularly noteworthy as it underscores the method's efficiency in leveraging DBNs alongside distance regularised level set methods to enhance segmentation precision. However, the complexity and considerable computational time required by this approach, coupled with its reliance on initial segmentation techniques, indicates areas ready for further optimisation and refinement.

The suggestion to incorporate 3D shape modelling and motion models presents a promising avenue to transcend the current slice-by-slice segmentation limitation, potentially offering more coherent and comprehensive segmentation.

Similarly, authors in (Bao, Zhu, and Li, 2023) introduced a hybrid-scale contextual fusion network for medical image segmentation. It incorporates a hybrid-scale embedding layer before the transformer to capture object information across multiple scales. The network utilises standard transformers and pooling transformers in the first two and last two skip connections, respectively, to model long-range dependencies and handle long input sequences. A dual-branch channel attention module was also proposed to focus on crucial channel features and conduct multi-level features fusion. This fusion scheme effectively captures richer context and detailed features, leading to efficient encoding and better segmentation performance. The study addresses the challenge of automatic segmentation of medical images, which is complex due to varying positions, sizes, and shapes of medical objects like organs and tumours. However, the computational intensity and the complexity of the network architecture, despite efforts to mitigate these through pooling transformers, highlight significant challenges in implementation and scalability. Moreover, the performance variability across different medical datasets or segmentation tasks underscores the necessity for model-specific tuning, which could limit the method's applicability in diverse clinical scenarios. It becomes evident that while both approaches push the boundaries of medical image segmentation through DL, they encounter common problems in computational complexity, which highlights the need for model adaptability across various segmentation tasks.

In the same context, authors in (Zheng, Liu, Feng, Xu, and Zhao, 2023) proposed a novel neural network architecture for medical image segmentation named Cross-attention and Cross-scale Fusion Network (CASF-Net) as shown in Figure 2.6. This network is designed to integrate both coarse and detailed feature representations by employing a dual-branch encoder network that models non-local dependencies and multi-scale contexts. The proposed cross-attention and cross-scale module within CASF-Net efficiently perform multi-scale information fusion, capable of exploring long-range contextual information.

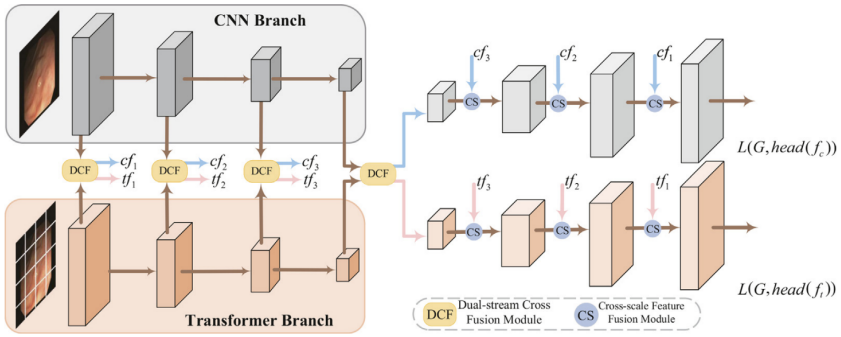


Figure 2.6: CASF-Net Architecture Proposed in (Zheng, Liu, Feng, Xu, and Zhao, 2023).

Similarly, authors in (Ammari, Mahmoudi, Hmida, Saouli, and Bedoui, 2023) introduced a novel Deep Active Learning (DAL) approach for right ventricle (RV) segmentation in cardiac MRI images (CMRI). The study targets the challenge of automatically segmenting the RV in CMRI images, a task traditionally done manually by radiologists, which is tedious and time-consuming. The RV's complex shape and the quality of CMRI images add to the segmentation difficulty (Ammari et al., 2023). The proposed approach was tested on images from public patients and custom subjects, resulting in an increase in the dice coefficient from 0.86 to 0.91, indicating better overlap between the predicted segmentation and the ground truth.

While both methodologies exhibit significant advancements in medical image segmentation (Zheng et al., 2023; Ammari et al., 2023), they are not without their potential drawbacks. CASF-Net's dual-branch architecture, integrating CNN and transformer features, may introduce increased model complexity and computational costs (Zheng et al., 2023). Such intricacy necessitates precise tuning of the cross-attention and fusion mechanisms across various medical imaging modalities, alongside attentive regularisation strategies to mitigate overfitting risks due to the model's complexity. Authors method in (Ammari et al., 2023), although benefiting from the DAL approach to efficiently utilise unlabelled data, might encounter challenges related to the substantial computational resources required for training and inference. The reliance on a considerable annotated dataset for initial model learning, despite efforts to minimise labelling workloads, could present practical constraints. Moreover, the method's strategies to counteract potential overfitting through data augmentation and uncertainty estimation highlight common concerns in deploying DL models in medical imaging. Comparatively, the CASF-Net's endeav-

our to maximise the joint advantages of CNN and transformer features reflects a compelling strategy to capture complex image details and contextual information (Zheng et al., 2023). This contrasts with the DAL approach's pragmatic use of unlabelled data to enhance model training efficiency and initial accuracy, demonstrating a focus on optimising data utilisation over architectural complexity (Ammari et al., 2023).

Under a similar premise, authors in (Billot et al., 2023) proposed SynthSeg, a neural network designed to segment brain scans across a wide range of contrasts and resolutions without the need for retraining or fine-tuning. It is trained on synthetic data generated from a generative model conditioned on segmentation. The approach uses domain randomisation, fully randomising the contrast and resolution of synthetic training data to achieve robust performance across varied target domains. On the other hand, authors in (Kunhimon, Shaker, Naseer, Khan, and Khan, 2023) introduced a learnable weight initialisation method for hybrid volumetric medical image segmentation models. This approach, designed for medical data, aims to utilise available medical training data to effectively learn contextual and structural indicators through self-supervised objectives. It integrates easily into any hybrid model without needing external training data. The method focuses on capturing the volumetric nature of medical data early in the training process, improving segmentation performance by inducing contextual indicators within the model.

SynthSeg's adaptability to new contrasts and resolutions without the need for retraining positions it as a highly adaptive tool in medical imaging analysis (Billot et al., 2023). The use of synthetic data for training on perfectly aligned ground truths, which can be automatically generated, along with its demonstrated robustness across a wide range of morphological variability, underscores its potential for widespread clinical application. However, potential limitations include its dependence on the generative model's assumptions and the comprehensive coverage of tissue tracings in training label maps to match all tissues present in test scans, with the effectiveness against various lesions or pathologies remaining an area for future exploration (Billot et al., 2023). Conversely, the method proposed in (Kunhimon et al., 2023) excels through data-dependent initialisation, learning weight initialisation from the training data itself.

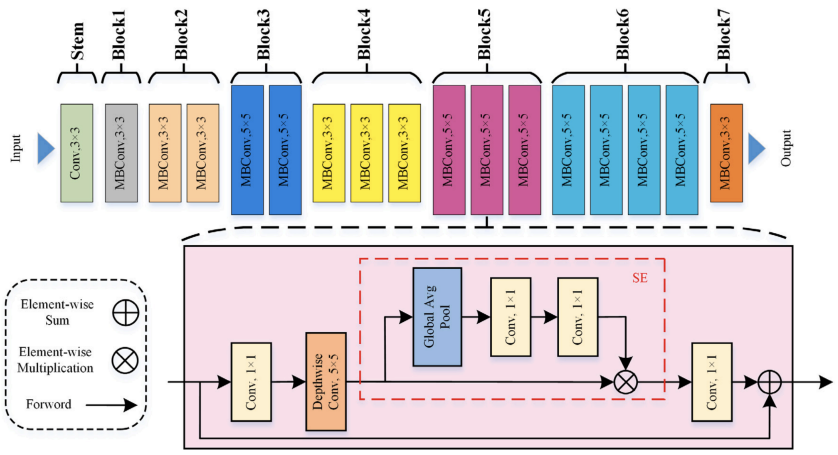


Figure 2.7: LAEDNet Encoder Structure Proposed in (Zhou et al., 2022)

This approach is particularly beneficial for medical imaging tasks characterised by small-scale datasets, as it doesn't require external data for achieving superior segmentation results. Its flexibility and the possibility of easy integration into various hybrid volumetric medical image segmentation models enhance its practical utility. The use of self-supervised learning objectives to exploit structural and contextual information further helps in effective model weight initialisation. On the other hand, the method's efficacy might be constrained by the specific characteristics of the training data, and its implementation complexity could cause challenges, especially in the fine-tuning of self-supervised learning objectives. Moreover, while validation on multi-organ and lung cancer segmentation tasks is robust, the broader applicability and performance across diverse medical imaging tasks require further investigation (Kunhimon et al., 2023). The network is designed to maintain a balance between segmentation accuracy and computational efficiency. LAEDNet is available in three model sizes: light (LAEDNet-S), medium (LAEDNet-M), and massive (LAEDNet-L), each using different versions of the EfficientNet backbone to accommodate to varying computational resource constraints (Zhou et al., 2022).

A research proposed in (Zhou et al., 2022) introduced a Lightweight Attention Encoder–Decoder Network (LAEDNet), novel and efficient encoder-decoder network for automatic ultrasound image segmentation. LAEDNet leverages a lightweight version of EfficientNet as the encoder (Figure 2.7) and integrates a Lightweight Residual Squeese-and-Excitation (LRSE) block in the decoder.

Similarly, authors in (Xie, Pan, Zhang, and An, 2022) proposed CHI-Net (Context Hierarchical Integrated Network), a CNN designed for medical image segmentation. CHI-Net is structured to address the challenges of low contrast, high similarity, and varying scales among different tissues in 2D medical images. It incorporates two primary modules: Dense Dilated Convolution (DDC) and Stacked Residual Pooling (SRP). The DDC module captures comprehensive complementary features at multiple scales, while the SRP module integrates encoder detail features through multiple effective field-of-views to generate more discriminative features. The network is designed to be flexible and adaptive for different medical image segmentation tasks (Xie et al., 2022). LAEDNet's lightweight architecture, not only accelerates the inference process but also ensures adaptability through its variable model sizes (e.i, small, medium, and large), addressing diverse computational needs and application scenarios (Zhou et al., 2022). The incorporation of the attention mechanism further refines segmentation accuracy, delivering smoother object contours. However, the complexity nature in its design, alongside the reliance on the diversity and quality of training datasets, presents challenges in terms of generalisability and implementation. Additionally, the potential for further optimisation in model size and computational speed suggests undiscovered efficiencies within its architecture. Conversely, CHI-Net proved its adaptability to scale and to object variations, through the contribution of its DDC module, and superior feature representation achieved by the SRP module (Xie et al., 2022). These innovations enable CHI-Net to excel in various medical image segmentation tasks, demonstrating robustness and superior performance against other segmentation methods. Nevertheless, the increased complexity and computational demands introduced by these modules could compromise its applicability in resource-limited settings or real-time applications. Moreover, its current limitation to two dimensional (2D) images causes a significant challenge for extending its application to three dimensional (3D) medical imaging, necessitating further research and development.

Another research introduced in (Xia et al., 2022) proposed Edge-Reinforced Neural Network (ER-Net), designed for segmenting vessel-like structures in 3D medical imaging modalities. ER-Net is an encoder–decoder architecture that incorporates a Reverse Edge Attention

Module (REAM) and an Edge-Reinforced Optimisation Loss (ERloss) to enhance the identification and preservation of spatial edge information. Additionally, a Feature Selection Module (FSM) is integrated to adaptively select discriminative features from both encoder and decoder, emphasising the weight of edge voxels and improving segmentation performance (Xia et al., 2022).

On the other hand, authors in (Wang, Li, and Cheng, 2023) introduced an extended EfficientNet-based U-Net architecture, named EE-UNet, for the automatic and accurate segmentation of the OD and optic cup (OC) in fundus images, which is crucial for clinical glaucoma screening. The method uses EfficientNet to extract features at various scales, incorporates a Conditional Random Field as a RNN (CRF-RNN) within the U-Net framework for end-to-end segmentation, and employs the ranger optimiser for better convergence. Additionally, a multi-label loss function is designed to balance the foreground and background pixels (Wang, Li, and Cheng, 2023).

Comparatively, ER-Net's emphasis on edge preservation and adaptive feature selection directly contrasts with EfficientNet/U-Net comprehensive approach, integrating advanced neural network architectures and optimisation strategies for improved feature extraction and segmentation refinement (Xia et al., 2022; Wang, Li, and Cheng, 2023). While both methods demonstrate superior segmentation capabilities, their respective disadvantages highlight critical areas for future research, particularly in simplifying network architectures, reducing computational demands, and enhancing model interpretability and generalisability (Xia et al., 2022; Wang, Li, and Cheng, 2023). Addressing these challenges will be paramount in advancing the field, potentially requiring a fusion of the innovative techniques presented in these works to develop more adaptive, efficient, and accessible segmentation tools for clinical applications.

On a similar pathway, a research proposed in (Xie et al., 2023a) introduced a deep adversarial co-training method for semi-supervised semantic segmentation of medical images, addressing the challenge of distribution shift between labelled and unlabelled data . The core idea is to enhance the model's robustness against distribution shifts by integrating adversarial training into the co-training process. This approach simulates distribution shift perturbations through

adversarial perturbations and applies them to challenge the supervised training phase towards enhancing the model's resilience. Co-training involves training two sub-models on disjoint subsets of the dataset to independently extract varied knowledge, enhancing overall performance by integrating insights from both views and reducing confirmation bias. The proposed method showed significant improvements on challenging medical datasets, achieving a Dice Similarity Coefficient (DSC) score of 87.37% with only 20% of labels on the ACDC dataset, comparable to using 100% of labels. On the SCGM dataset, which exhibits more pronounced distribution shift, the method achieved a DSC score of 78.65% with 6.5% of labels, outperforming the baseline by 10.30%. These results demonstrate the method's superior robustness against distribution shifts in medical imaging (Xie et al., 2023a).

On the other hand, Authors in (Messaoudi, Belaid, Salem, and Conze, 2023) introduced novel network architectures for 2D and 3D uni- and multi-modal medical image segmentation, leveraging the efficiency of pre-trained 2D classification networks on natural images. The key strategies include: (1) Weight TL (WTL) which is embedding a pre-trained 2D encoder into higher-dimensional U-Net architectures, and (2) Dimensional TL (DTL), which is expanding a 2D segmentation network into higher dimensions by extrapolating weights for use in 3D U-Net-like architectures. These methods are validated on various medical imaging modalities such as MRI, CT, and ultrasound, showcasing superior performance in challenges like CAMUS (for echocardiographic data) and CHAOS (for MRI and CT abdominal images) (Messaoudi et al., 2023). Authors focus on adversarial training and co-training aims to directly confront distribution shifts, enhancing model robustness and semi-supervised learning efficiency (Xie et al., 2023a). In contrast, the method's reliance on pre-trained networks and TL strategies aims to bypass the computational expensive requirements traditionally associated with training DL models from scratch, despite the fundamental risks of overfitting and the nuanced demands of medical image data (Messaoudi et al., 2023). It is evident that both approaches highlight the importance of innovative ML methodologies in improving segmentation outcomes, yet they also illuminate the necessity for careful consideration of each method's limitations and prerequisites (Xie et al., 2023a; Messaoudi et al., 2023). Addressing these challenges through

further research and development will be crucial for advancing the field, potentially requiring a synthesis of adversarial resilience and TL efficiency to create more adaptive, effective, and user-friendly segmentation tools for clinical and research applications (Zhu, Wang, Li, and Li, 2023; Shi, Lu, Yin, Zhong, and Yang, 2023; Vasudeva and Chandrashekara, 2023).

2.4 Hybrid Models Versus Singular Approaches

In recent years, the landscape of computational models, particularly in the field of medical image analysis, has witnessed a paradigm change towards the integration of hybrid methodologies. This section delves into the rationale behind the escalating interest in hybrid models as opposed to traditional singular approaches. Singular models, characterised by their reliance on a specific computational technique, have been the foundation of numerous advancements in automated image analysis and disease diagnosis. However, the complexity of medical data, coupled with the nuanced nature of various diseases, often overcomes the capacity of any single method to provide a comprehensive solution.

2.4.1 Hybrid Methodologies in Medical Image Analysis: Bridging Computational Techniques for Enhanced Performance

Hybrid models emerge as a sophisticated response to these limitations, combining the strengths of diverse computational strategies to enhance accuracy, robustness, and interpretability. By combining techniques such as DL, statistical analysis, and ML algorithms, hybrid models aim to offset the basic weaknesses of singular approaches. This section aims to investigate the comparative advantages of hybrid models, emphasising the way their multidimensional nature allows for a more nuanced understanding and processing of medical images. Through a review of related works and recent advancements, this section explores the transformative potential of hybrid models in overcoming challenges that single-method approaches face, such as data insufficiency, class imbalance, and the need for domain-specific adaptability. This overview not only highlights the superiority of hybrid models in certain contexts but also sheds light

on the evolving landscape of medical image analysis, where the synergy between different computational paradigms paves the way for ground-breaking innovations.

Hybrid DL models offer several advantages that address some of the critical challenges faced in data science and AI-driven fields, particularly in scenarios characterised by limited data availability (Gavrishchaka, Yang, Miao, Senyukova, et al., 2018). One of the primary benefits of these models is their improved performance in the face of data incompleteness. This is achieved by leveraging domain-expert knowledge alongside compact ensembles of complementary low-complexity models that are discovered through optimisation techniques. Such an approach enhances the model's tolerance to incomplete datasets, a common obstacle in many real-world applications. Furthermore, hybrid models excel in improving model accuracy (Gavrishchaka et al., 2018). By collaboratively combining the strengths of optimisation-based ensembles with DNNs, these models are capable of uncovering implicit patterns and non-linear mixed terms that might be overlooked by singular DNN-based or optimisation-based approaches. This fusion of methodologies allows for a significant increase in accuracy, making hybrid models particularly valuable in complex problem-solving scenarios where precision is paramount.

Operational simplicity is another notable advantage of hybrid models. The complexity involved in discovering and training optimal DNN architectures can be challenging, especially given the vast parameter spaces and architectural configurations possible (Gavrishchaka et al., 2018). Hybrid models address this challenge by simplifying the operational workflow. This is performed by reducing the problem's dimensionality through the initial use of optimisation-discovered components, streamlining the process of identifying and training the most effective DNNs for the task at hand. Additionally, the flexible incorporation of domain knowledge stands out as a critical benefit of hybrid models (Gavrishchaka et al., 2018). In areas with significant data limitations, the ability to effectively integrate and leverage existing domain expertise is invaluable. Hybrid models are designed to accommodate and utilise such knowledge, thereby enhancing the model's effectiveness and applicability in specialised fields. This aspect of hybrid models underscores their utility in addressing complex challenges where domain-specific

insights are crucial for achieving high levels of accuracy and performance.

Hybrid models bring together the strengths of deep automated techniques and traditional ML approaches, offering several advantages that enhance their performance, applicability, and scalability across a wide range of tasks (Bozkurt, 2022). One of the primary benefits of these hybrid models is their high accuracy, with DL components achieving accuracies as high as 96.81%. This level of performance significantly surpasses that of classical ML methods, making hybrid models especially valuable in applications where precision is paramount. Another key advantage is the robust feature extraction capability of DL models (Bozkurt, 2022). Unlike traditional methods that often require manual feature selection and extraction, DL models are capable of automatically learning complex feature representations directly from raw data. This ability not only simplifies the model development process but also ensures that the features used for classification or prediction are optimally representative of the underlying data patterns, thereby providing a more solid foundation for accurate decision-making. Moreover, hybrid models demonstrate improved generalisation to new, unseen data (Bozkurt, 2022). This characteristic is crucial for the practical deployment of models in real-world scenarios, where the ability to accurately predict or classify instances that were not present in the training dataset can significantly impact the effectiveness and reliability of the model. The enhanced generalisation capabilities of hybrid models stem from their sophisticated architecture, which combines DL's ability to model high-level abstractions with ML's efficiency in handling structured data. The adaptability and scalability of hybrid models stand out as significant advantages (Bozkurt, 2022). By integrating the strengths of ML and DL, these models offer flexible and scalable solutions that can be adapted to a wide variety of tasks beyond their initial application domain. Whether it's activity recognition, image classification, or predictive analytics, hybrid models can be adapted to meet the specific requirements of different applications, providing an adaptive toolset for tackling diverse challenges across industries and research domains.

A review paper on hybrid DL (HDL) models was reported in the literature for image classification emphasises the importance of transitioning from single DL (SDL) models to HDL models (Jena et al., 2021). This transition is crucial for enhancing performance by leveraging

the strengths of multiple DL architectures or combining DL with ML models. HDL models have demonstrated superior performance across various applications, particularly in medical and non-medical image processing, by integrating the best aspects of two or more SDL models or fusing DL with ML approaches. The importance of migrating to hybrid models is translated by the fact that HDL models offer improved stability and performance by combining the advantages of multiple SDL architectures or integrating DL with ML. Also, the transition to HDL models is driven by the need for more accurate and automated image classification solutions. The paper categorises HDLs into three main types: spatial, temporal, and spatial-temporal, based on the nature of the input data (images, videos, electronic time-series signals) (Jena et al., 2021). Examples include the Inception-ResNet model, which combines two SDL models (Inception and ResNet) for enhanced image classification. Applications range across medical imaging, hyperspectral image classification, emotion recognition from audio-visual data, human activity recognition, and time-series data analysis, showcasing the diverse utility of HDL models (Jena et al., 2021).

Hybrid models exhibit superior performance metrics compared to single models. This advantage is primarily due to the comprehensive feature extraction and classification capabilities essential in hybrid architectures. By integrating multiple learning approaches, HDL models can capture a wider range of data characteristics, from high-level abstractions to nuanced details that might be overlooked by single models. This capability enables hybrid models to achieve higher accuracy, sensitivity, and specificity in tasks such as image classification, object detection, and semantic segmentation. Furthermore, hybrid models offer enhanced flexibility and an expanded application scope. This flexibility allows researchers and practitioners to design adapted solutions that are specifically optimised for the complexities of various imaging tasks across different domains. Whether it's medical imaging, satellite imagery analysis, or automated quality inspection in manufacturing, hybrid models can be adapted to address the unique challenges of each task. This adaptability is facilitated by the models' ability to leverage both DL's powerful representation learning and the domain-specific insights that traditional ML models provide. Consequently, hybrid architectures are not only more adaptive but also

capable of expanding the limits of what can be achieved across a broad spectrum of imaging tasks, setting new standards for performance and applicability in the field.

The migration from solely using ML or DL models to adopting hybrid approaches is crucial for enhancing the accuracy and efficiency of medical diagnoses, particularly in cancer detection (Painuli, Bhardwaj, and köse, 2022). Hybrid models, which combine the strengths of various ML/DL techniques and sometimes incorporate traditional image processing methods, are shown to better address the complexities of medical image analysis. These models excel in handling diverse data types, extracting more nuanced features, and improving the interpretability of results, which are essential for early and accurate cancer diagnosis. Authors in (Painuli, Bhardwaj, and köse, 2022) discusses several hybrid models applied to the detection and classification of various cancers, including lung, breast, liver, pancreatic, and brain cancers, as well as skin cancer. These models often combine classical image processing techniques with advanced ML/DL algorithms to enhance feature extraction, segmentation, and classification accuracy. For example, the use of cascade SVM (C-SVM), CNNs with data augmentation techniques, and ensemble models incorporating different ML and DL architectures. The applications of these hybrid models span across different imaging modalities such as CT, MRI, PET, and dermoscopic images, demonstrating their reliability and effectiveness in improving diagnostic outcomes.

Hybrid DL models present a significant advancement in medical image processing, capitalising on the synergistic potential of combining various learning algorithms. These models stand out for their improved diagnostic accuracy, a direct result of leveraging the strengths fundamental in multiple approaches. The ability to integrate and synthesise diverse perspectives enables these models to achieve a level of precision that surpasses that of single-model systems. One of the critical advantages of hybrid models is their robustness to the essential variability and complexity of medical images. Traditional single-model approaches often struggle with the wide range of variations present in medical imaging data, from differences in imaging modalities to patient-specific characteristics. Hybrid models, however, can navigate these complexities more effectively, offering more reliable and consistent performance across a broad spectrum of imag-

ing scenarios. Efficient feature extraction is another benchmark of hybrid DL models. These systems excel at identifying and combining features from different data levels and aspects, such as spatial relationships, texture patterns, and contextual information. This capability is crucial for detecting implicit anomalies that may be indicative of early-stage diseases or conditions that are otherwise difficult to identify. Furthermore, the integration of various models within a hybrid framework helps mitigate the risk of overfitting. By drawing on diverse data representations and learning methodologies, hybrid models stimulate a more holistic understanding of the data. This approach not only enhances the model's accuracy on the training data but also improves its generalisation to unseen data. Consequently, hybrid DL models are not just more accurate and robust; they are also more adaptable and capable of delivering consistent performance across different patient populations and imaging conditions.

2.4.2 Advancing Medical Diagnostics with Hybrid Computational Models

The emergence of hybrid models as a solution to these challenges is underscored by their capacity to combine diverse data sources and modalities, thereby enhancing model generalisation and robustness. By incorporating domain knowledge, hybrid models can effectively navigate data imbalance and variability, improving their interpretability and reliability. Furthermore, hybrid approaches facilitate the combination of ML/DL techniques with traditional analysis methods, addressing complex medical data challenges, such as feature extraction and noise reduction. Importantly, hybrid models also offer pathways to enhance model explainability, stimulating greater acceptance and trust within the clinical community. Through these diverse advantages, hybrid models present a compelling framework for advancing diseases diagnosis, promising greater accuracy, adaptability, and clinical relevance in the face of the basic challenges caused by ML/DL techniques.

In this context, Yan presented a comprehensive review of the importance and advances in multi-task DL (MTDL) for medical image computing and analysis (Zhao, Wang, Che, Bao,

and Li, 2023). Their paper emphasises the critical shift from traditional single-task models to MTDL approaches due to the fundamental complexity and inter-connectivity of medical imaging tasks. MTDL models leverage the relationships between tasks to improve performance, generalisation, and computational efficiency. This paradigm shift is crucial for advancing medical image analysis, offering a more holistic and efficient way to handle multiple interrelated tasks simultaneously (Zhao et al., 2023). The review categorises MTDL network architectures into four main types to include cascaded, parallel, interacted, and hybrid, each suited for different task relationships and complexities. Notably, the hybrid architecture is highlighted for its ability to integrate the advantages of the other three, making it particularly suitable for complex task combinations in medical imaging. These architectures facilitate the joint learning of tasks such as segmentation, classification, and disease diagnosis, demonstrating MTDL's adaptability and potential for enhancing medical diagnostics (Zhao et al., 2023). The Hybrid MTDL models provide several benefits, including:

- **Efficiency in Learning:** By sharing common features across tasks, these models reduce redundancy and improve computational efficiency.
- **Improved Generalisation:** The models can leverage the noise patterns and task relationships to learn more generalised features that are robust to variations in the data.
- **Feature Prioritisation:** They can identify and prioritise important features across tasks, enhancing the model's focus and performance on critical aspects of the data.
- **Reduced Overfitting:** The integration of tasks introduces inductive biases, helping to mitigate overfitting compared to single-task models.

Authors in (Sun, Wang, and Tang, 2013) presented a hybrid DL model combining Convolutional Networks (ConvNets) and Restricted BMs (RBM) for face verification in uncontrolled environments (Sun, Wang, and Tang, 2013). This model is designed to learn relational visual features directly from raw pixels of paired face images. It employs multiple groups of ConvNets to extract local and global relational features which indicate identity similarities, followed by

a top-layer RBM that infers from these HF features. This structure aims to capture face similarities from various aspects and is fine-tuned jointly to optimise face verification performance. The research addresses the challenge of face verification given the significant intra-personal variations in faces due to changes in cause, illumination, expression, age, makeup, and occlusions, especially when the images are captured in wild, uncontrolled conditions (Sun, Wang, and Tang, 2013). Traditional methods typically lose critical relational information between face pairs during the feature extraction stage, which this model aims to preserve and utilise for more accurate verification. The hybrid ConvNet-RBM model demonstrates competitive performance on the LFW dataset.

Conversely, authors in (Bourouis, Alroobaea, Rubaiee, and Ahmed, 2020) proposed a hybrid solution for medical image segmentation, focusing on accurately identifying pathological regions in biomedical images, particularly for brain tumour segmentation. This hybrid framework integrates statistical-based, variational-based, and atlas-based techniques, aiming to leverage the strengths and mitigate the weaknesses of each approach. The proposed method consists of a pipeline framework with several steps to include: (1) pre-processing to improve image quality and remove noise, (2) classification using symmetry axis detection and SVM for learning, and (3) a refinement step employing a variational-based level set method for precise boundary detection of regions of interest. The challenge addressed is the precise analysis of medical images, such as segmentation, detection, and quantification of tumours and cancers, which is critical for numerous clinical applications (Bourouis et al., 2020). The complexity and immense volume of medical imaging data make it difficult to design effective segmentation algorithms. Additionally, manual delineation of specific regions is inconvenient, necessitating an automated and robust solution. The results showcase the effectiveness of the proposed hybrid framework through various metrics:

- **Similarity Index (SI):** Achieved an average of 80.9%, indicating a strong agreement with expert segmentation and demonstrating competitive performance compared to other methods.

- **Sensitivity and Specificity:** Reported high Sen (88.7%) and Spe (94.0%), showcasing the framework's ability to accurately detect tumour regions.
- **Comparative Analysis:** The framework's SI of 80.9% is compared with other approaches, showing its competitive edge, especially against methods with lower SI percentages.

Melanoma is a fatal form of skin cancer that can quickly spread to other parts of the body if not diagnosed early. Early detection significantly improves the chances of survival, making it crucial to develop automated diagnosis systems that can assist doctors and individuals in identifying potential melanoma lesions. The challenge lies in accurately distinguishing between benign and malignant skin lesions surrounded by various challenges, such as the variability in lesion appearance and the presence of noise in images. In this context, a research has introduced a hybrid method for melanoma skin cancer detection, combining predictions from three different models: a CNN and two classical ML classifiers (KNN and SVM) (Daghrir, Tlig, Bouchouicha, and Sayadi, 2020). These models are trained on features describing the borders, texture, and colour of skin lesions. The final decision is made using majority voting, where the classification result chosen by the majority of the models is taken as the final output. This approach aims to leverage the strengths of DL and classical ML to improve the accuracy of melanoma detection. The experiments conducted on a dataset from the ISIC (International Skin Imaging Collaboration) archive showed that:

- The CNN model alone achieved an Acc of 85.5%.
- The SVM classifier achieved an Acc of 71.8%.
- The KNN classifier had the lowest Acc of 57.3%.
- Combining the predictions of all three methods through majority voting significantly improved the performance, achieving an Acc of 88.4%.

Each of the three hybrid methods employs a fusion of technologies to improve upon the limitations of single-model systems, aiming for higher accuracy, robustness, and comprehensive data interpretation. ConvNets-RBM pushes the boundaries of direct feature learning, while

Statistical, variational, and atlas-based approach emphasises a cascading improvement in segmentation accuracy, and CNN-KNN-SVM showcases the strength of combining different classification techniques (Sun, Wang, and Tang, 2013; Bourouis et al., 2020; Daghrir et al., 2020). Despite these advances, each approach faces challenges in complexity, potential for overfitting, data dependency, and the necessity for careful integration and tuning of their composite parts. These factors highlight the trade-off between performance gains and increased operational demands based in hybrid computational methods.

Another research introduced a Hybrid Automated Medical Learning (HAML) framework, which is a sophisticated combination of distributed DL, multi-agent systems, and knowledge graphs (Belhadi, Djenouri, Diaz, Houssein, and Lin, 2022). HAML aims to efficiently and automatically learn from medical data, overcoming challenges like data heterogeneity and complex medical learning tasks. Each agent in the system is designed to learn patterns from medical data locally, while knowledge graphs facilitate the sharing of relevant patterns and ontologies among agents to enhance learning and communication. The research addresses the automation of medical learning to assist medical teams in making informed decisions. Automated medical learning faces challenges such as heterogeneity in medical data and the complexity of medical learning tasks, which can lead to inaccuracies in the learning process. HAML is proposed to mitigate these issues by utilising a combination of advanced intelligent approaches (Belhadi et al., 2022). HAML demonstrated superior performance in various case studies compared to the most up-to-date medical learning models, both in computational efficiency and the quality of solutions. In fact, in process mining, HAML achieved higher accuracy in detecting relevant patterns from event medical data. Also, in recognising patients' activities in a smart building context, HAML efficiently identified different activities. For medical image retrieval, HAML showed an impressive ability to find the most relevant medical images based on queries. The results indicate HAML's effectiveness in handling different types of medical data and tasks, outperforming existing approaches in terms of speed and accuracy.

Recent research has presented a hybrid DL method for multi-modal medical image fusion, aiming to address the issue of integrating information from different medical imaging modali-

ties (e.g., MRI, CT, and Single-photon emission computed tomography (SPECT)) into a single composite image (Li, Zhao, Lv, and Li, 2021a). This is crucial for enhancing diagnostic accuracy by providing comprehensive information in one image, thus helping medical professionals in making more informed decisions. The primary challenge addressed is the limitation of individual medical imaging techniques, which may only offer partial insights due to their specific focuses (e.g., anatomical versus functional imaging). The goal is to fuse images from multiple modalities to capitalise on the combined strengths and comprehensive details, overcoming the scattering of information across different images which could restrict diagnostic processes (Li et al., 2021a). The proposed method leverages DL to fuse multi-modal medical images into a single, comprehensive image that retains critical information from each modality. It aims to overcome deficiencies in existing fusion techniques by improving upon clarity, detail, and processing efficiency. The method involves pre-processing steps (noise removal, registration, standardisation), followed by DL-based fusion using a model trained on a diverse set of medical images. This approach ensures that the fusion process is not only aligned to the specific characteristics of medical imaging but also scalable and efficient for batch processing of multiple images (Li et al., 2021a). The paper reports improvements across various metrics used to evaluate image fusion quality, including Edge Operating Gradient (EOG), Root Mean Square Error (RMSE), Peak Signal-to-Noise Ratio (PSNR), Entropy, Structural Similarity Index Measure (SSIM), and others. These improvements indicate enhanced clarity, detail preservation, and overall image quality in the fused outputs compared to existing methods (Li et al., 2021a).

Authors in (Qaid et al., 2021) presented hybrid models that combine DL, TL, and ML techniques for the early detection and classification of COVID-19 from CXR images (Qaid et al., 2021). Specifically, it utilises CNNs and TL models (using VGG-16 and VGG-19 architectures) hybridised with powerful ML algorithms. These models are designed to distinguish COVID-19 cases from normal cases and other types of viral pneumonia, leveraging the feature extraction capabilities of CNNs and the pre-trained knowledge of TL models. The extracted features are then fed into various ML algorithms (e.g., SVM, RF) for classification. The challenge addressed is the early detection of COVID-19 to mitigate its spread and alleviate the pressure

on healthcare systems. Given the similarities between radiographic features of COVID-19 and other viral pneumonias, distinguishing between them is difficult, necessitating the use of AI to enhance accuracy and robustness in detection (Qaid et al., 2021). The proposed models achieved promising results across different configurations and data sets:

- Full accuracy in binary classification of COVID-19 versus normal cases and COVID-19 versus viral pneumonia cases using certain hybrid models.
- For multiclass classification (normal, viral pneumonia, COVID-19), Acc reached up to 97.8%.
- These models outperformed the baseline model, showing higher accuracy, precision, recall, and F1-score across various classification scenarios.

Time-consuming and expertise-reliant process of manually identifying regions of interest in lung high-resolution computed tomography (HRCT) images before applying DL algorithms for Interstitial Lung Disease (ILD) classification is a critical challenge. The diversity and complexity of ILD manifestations in HRCT images necessitate a robust and efficient classification approach to enhance healthcare outcomes for patients with ILD, a condition with a high risk of lung cancer. In this context, Pawar et al. proposed a novel two-stage hybrid approach utilising DL networks for the classification of ILD from HRCT images (Pawar and Talbar, 2022). This method aims to improve the accuracy and efficiency of ILD classification by automating the process, which traditionally relies on manual identification of the region of interest (ROI) in lung HRCT images. The paper reports considerable improvements in ILD classification performance due to the proposed method's stage-wise enhancement of DL algorithm performance. Specifically, the method achieves high accuracy in classifying six ILD classes (normal, emphysema, fibrosis, ground glass, micronodules, and consolidation), with precision, recall, F-score, and accuracy metrics demonstrating the effectiveness of the approach (Pawar and Talbar, 2022).

Accurate and automatic segmentation and classification of brain tumours from MRI scans is a challenging task complicated by the high spatial and structural variability of tumours. Manual segmentation of MRI data is time-consuming and prone to errors, which can adversely affect

patient outcomes. Early and accurate diagnosis is crucial for effective treatment planning and improving patient survival rates. To tackle this challenge, a research has proposed a hybrid deep TL model named GN-AlexNet, aimed at improving the classification of brain tumours into three types: pituitary, meningioma, and glioma (Samee et al., 2022). This model integrates the architecture of GoogleNet and AlexNet by removing five layers from GoogleNet and incorporating ten layers from AlexNet, enhancing feature extraction and automatic classification capabilities for BT tri-classification. The model was evaluated using a publicly available Contrast-Enhanced MRI (CE-MRI) dataset and demonstrated superior performance in accuracy and sensitivity compared to existing methods, including various TL techniques and ML/DL models. The GN-AlexNet model achieved remarkable performance metrics on the CE-MRI dataset, outperforming existing TL models to include VGG-16, AlexNet, SqueezNet, ResNet, MobileNet-V2, and ML/DL approaches. The model attained an Acc of 99.51% and a Sen of 98.90%, demonstrating its effectiveness in classifying brain tumours with high precision and reliability (Samee et al., 2022).

While these proposals exhibit high accuracy and robustness in their specific applications, they share a common drawback of complexity and a significant dependence on the quality and variety of the training data (Qaid et al., 2021; Pawar and Talbar, 2022; Samee et al., 2022). Model optimisation and computational efficiency are other areas of concern, which are essential when considering the practical implementation of these models.

The relatively poor generalisability of DCNN models to datasets with characteristics not well-represented in the training data, particularly in medical image segmentation tasks is a major challenge in image processing. Although DCNNs show high accuracy in segmenting anatomical structures, their performance often drops when applied to new datasets with different imaging conditions. MAS (Multi-Atlas Segmentation) methods, while less accurate in some cases, demonstrate better generalisation capabilities across diverse datasets. In light of this challenge, authors presented Deep Label Fusion (DLF), a novel hybrid method that integrates the strengths of DCNN and MAS for medical image segmentation (Xie et al., 2023b). This approach aims to leverage the high accuracy of DCNN in learning complex data represen-

tations and the robust generalisability of MAS to variations in image characteristics. The DLF method introduces an end-to-end pipeline with learnable weights, incorporating a weighted voting subnet for MAS and a fine-tuning subnet to correct residual errors, improving segmentation accuracy. DLF was evaluated on five datasets representing different anatomical structures and imaging modalities. The method achieved comparable accuracy to the state-of-the-art DCNN model, nnU-Net, on datasets similar to the training set. Notably, DLF outperformed nnU-Net in generalising to datasets with different characteristics, such as varying MRI field strengths or patient populations. Additionally, DLF consistently improved upon conventional MAS methods. The paper also introduces a modality augmentation strategy that enhances segmentation accuracy and interpretability in multimodal imaging scenarios (Xie et al., 2023b).

The variability in COVID-19 clinical presentations and outcomes, ranging from mild symptoms to severe complications requiring ICU admission or resulting in death is also a significant challenge. The goal is to support clinical decision-making by predicting the severity of patient outcomes based on CT images and clinical data, facilitating early intervention for those at higher risk. In light if this challenge, authors developed a hybrid ML/DL model to classify COVID-19 patients based on the severity of their condition, specifically distinguishing between those requiring ICU admission or facing death (ICU class) and those who do not (non-ICU class) (Chieregato et al., 2022). This classification was done using data from 558 patients admitted to a hospital in northern Italy during the early months of the COVID-19 pandemic. The hybrid model integrates a 3D CNN with CatBoost, a ML algorithm. The CNN serves as a feature extractor from baseline CT images, while the extracted features, along with laboratory and clinical data, are selected using the Boruta algorithm enhanced by SHAP values. The reduced feature set is then used to train a CatBoost classifier, achieving a probabilistic AUC of 0.949 on the holdout test set (Chieregato et al., 2022).

The need for an automated and accurate classification system for breast cancer histopathology to help clinical diagnosis and treatment planning is pivotal. The manual grading of cancer slides is time-consuming and requires expert knowledge, which is insufficient in many regions. The automated system aims to reduce diagnostic time and improve accuracy. In this context,

research in the literature proposed a novel approach for the automated classification of breast cancer histopathology slides into benign and malignant subtypes using a hybrid DL model that combines CNN and Long Short-Term Memory RNN (LSTM RNN) (Srikantamurthy, Rallabandi, Dudekula, Natarajan, and Park, 2023). This method, leveraging TL from ImageNet, was evaluated on the BreakHis dataset, comprising 2480 benign and 5429 malignant images across various magnifications. The proposed CNN-LSTM model utilises TL to classify four benign and four malignant breast cancer subtypes. It operates by extracting deep convolutional features using pre-trained CNN models (like ResNet50 and InceptionV3) from ImageNet, followed by an LSTM RNN model for classification. The model was trained and validated using various optimisers and configurations to achieve the best performance (Srikantamurthy et al., 2023). The hybrid CNN-LSTM model achieved the highest overall Acc of 99% for binary classification (benign vs. malignant) and 92.5% for multi-class classification (among subtypes of benign and malignant cancers). Among the optimisers tested (Adam, RMSProp, and SGD), Adam was found to be the most effective, producing the maximum accuracy with minimum model loss. The model outperformed existing CNN models such as VGG-16, ResNet50, and Inception in classifying breast histopathological images (Srikantamurthy et al., 2023).

One of the paramount challenges in medical image processing if the challenge of time-consuming MRI scanning procedures, which can affect patient comfort and introduce motion artifacts. By accelerating the MRI process through improved reconstruction of under-sampled images, the method aims to reduce scanning time, minimise patient stress, and decrease medical costs. In this scenario, authors introduced a novel DL framework for reconstructing MRI images from under-sampled k-space data, aiming to improve the accuracy of MRI reconstruction (Al-Haidri, Matveev, Al-Antari, and Zubkov, 2023). This framework leverages Conditional GANs (CGANs) with a U-Net architecture for the generator. Additionally, a unique hybrid loss function that considers both spatial and frequency domains is proposed to enhance the quality of the reconstructed images. This method is evaluated against traditional Sensitivity Encoding (SENSE) reconstruction and other DL approaches, focusing on the improvement of image quality metrics such as SSIM and PSNR (Al-Haidri et al., 2023). The proposed frame-

work demonstrated superior performance in reconstructing MRI images compared to traditional SENSE techniques, with improvements in PSNR by 6.84 and 9.57 for U-Net and CGAN models, respectively. SSIM metrics were comparable to those provided by SENSE, indicating that the reconstructed images maintain high fidelity to the original scans (Al-Haidri et al., 2023).

The automated, accurate detection and segmentation of cancerous regions in mammogram images, differentiating between benign and malignant cases represent a significant problem to help in timely and effective treatment decisions. For this purpose, Raaj developed a novel hybrid CNN architecture for classifying mammogram images into normal, benign, and malignant categories, aimed at improving the detection and segmentation of cancer regions in breast tissues (Raaj, 2023). This hybrid method incorporates a radon transform to convert spatial pixels into time–frequency variation images, a data augmentation module to enhance the dataset, and a mathematical morphological-based segmentation algorithm to precisely segment cancer pixels. The performance of this system is evaluated using the MIAS and Digital Database for DDSM datasets. The proposed architecture achieved impressive performance metrics (Raaj, 2023):

- DDSM Dataset: Sen of 97.91%, Spe of 97.83%, Acc of 98.44%, and Jaccard Index (JI) of 98.57%.
- MIAS Dataset: Se of 98%, Sp of 98.66%, Acc of 99.17%, and JI of 98.07%.

The high cost associated with collecting pixel-wise annotated data for medical image segmentation has become a paramount barrier. The goal is to achieve high accuracy segmentation labels with limited annotation effort, addressing issues like early stage, effective sample selection, and manual annotation workload. In this context, authors in (Li et al., 2023b) proposed a Hybrid Active Learning framework using Interactive Annotation (HAL-IA) for medical image segmentation, designed to reduce annotation costs by decreasing the number of annotated images required and simplifying the annotation process. This framework incorporates a novel hybrid sample selection strategy and an interactive annotation module, aiming to address these challenges.

Experimental results on four medical image datasets demonstrated the framework's effectiveness in achieving high-accuracy segmentation with less labelled data and fewer interactions. The HAL-IA framework outperforms other state-of-the-art methods by obtaining high-performance segmentation models with fewer labelled data and interactive clicks (Li et al., 2023b).

Challenges in content-based image retrieval (CBIR) for medical images have become critical, with special focus on the semantic gap between DHF visual features extracted by machines and high-level semantic understanding by humans. Vasudeva's research aims to improve the accuracy and efficiency of medical image classification and retrieval from large healthcare datasets, overcoming limitations of existing DL approaches that rely mainly on labelled data and lack of transparency (Vasudeva and Chandrashekara, 2023). For this purpose, they introduced a hybrid DL model combining CNN with LSTM networks, enhanced by feature extraction using the GLCM for medical image classification. This model aims to achieve higher classification accuracy and effective image retrieval by utilising additional layers in the CNN-LSTM architecture and improving retrieval performance with the Euclidean distance technique (Vasudeva and Chandrashekara, 2023). The hybrid model demonstrated superior performance compared to single ANN and CNN models, achieving a classification Acc of 99.4%. The precision, recall, and F1-score metrics also indicated improved accuracy for image classification on large healthcare datasets. These results underscore the model's effectiveness in extracting better medical image features and achieving higher classification accuracy (Vasudeva and Chandrashekara, 2023).

HAL-IA and CNN-LSTM-GLCM showcase innovative hybrid approaches aligned to specific challenges within medical imaging, emphasising efficiency and accuracy (Li et al., 2023b; Vasudeva and Chandrashekara, 2023). HAL-IA focuses on reducing annotation issues through active learning, beneficial for large-scale medical studies requiring extensive labelled data. CNN-LSTM-GLCM aims at enhancing classification and retrieval accuracy, showcasing the power of integrating DL with traditional feature extraction methods. Despite their advantages, both methods encounter challenges related to complexity, whether in implementation, compu-

tational demands, or optimisation. HAL-IA 's performance may vary depending on the imaging modality and complexity of the targets, while CNN-LSTM-GLCM 's sophisticated model architecture could cause challenges in training efficiency and application adaptability (Li et al., 2023b; Vasudeva and Chandrashekara, 2023). The advantages and limitations of the discussed and reviewed papers are summarised in Table 2.14.

Table 2.14: Comparative Analysis of Hybrid Models: Advantages and Disadvantages

Ref	**Hybrid Model**	**Advantages**	**Disadvantages**
(Sun, Wang, and Tang, 2013)	ConvNets-RBM	• Learns directly from raw data, potentially capturing more nuanced features. • Joint feature extraction maintains relational data between face pairs. • Multiple ConvNet groups increase robustness by capturing diverse similarities. • Unified architecture optimises feature extraction and recognition together. • Joint fine-tuning aligns the network closely with the verification task.	• The complexity of the model may demand high computational resources. • The potential for overfitting due to the model's high capacity. • Performance could be sensitive to the way well face images are aligned. • Learning from raw pixels could miss out on proven hand-crafted features' benefits unless optimisation is flawless.

Continued on next page

Table 2.14: Comparative Analysis of Hybrid Models: Advantages and Disadvantages (Continued)

Ref	**Hybrid Model**	**Advantages**	**Disadvantages**
(Bourouis et al., 2020)	Statistical, variational, and atlas- based	• Integrates statistical, variational, and atlas-based methodologies for segmentation accuracy. • Effective step-by-step initialisation leads to stable and accurate segmentation. • Competitive performance metrics indicate a strong match with ground truth.	• Relies heavily on the accuracy of pre-processing registration. • Applicability currently limited to specific tumour types. • Acknowledged need for improvement in registration algorithms and a more robust speed function for segmentation.

Continued on next page

Table 2.14: Comparative Analysis of Hybrid Models: Advantages and Disadvantages (Continued)

Ref	Hybrid Model	Advantages	Disadvantages
(Daghrir et al., 2020)	CNN-KNN-SVM	• Combines CNN, SVM, and KNN to outperform individual model accuracy. • Gains from CNN feature learning robustness and SVM/KNN classification effectiveness. • Can comprehensively analyse various lesion characteristics due to its adaptive approach.	• Managing three separate models increases computational complexity. • Performance is tied to the quantity and diversity of training data. • Requires precise calibration to ensure effective integration of model outputs without introducing bias.

Continued on next page

Table 2.14: Comparative Analysis of Hybrid Models: Advantages and Disadvantages (Continued)

Ref	Hybrid Model	Advantages	Disadvantages
(Belhadi et al., 2022)	HAML	• Efficient learning from heterogeneous medical data using a combination of distributed DL and multi-agent systems. • Improved communication through knowledge graphs, enhancing pattern sharing and ontology alignment. • Scalable architecture suitable for large-scale medical data analysis. • Adaptability to various medical learning tasks, increasing its adaptability.	• System complexity may limit the ease of use and maintenance due to the integration of multiple intelligent components. • Performance heavily reliant on the quality and diversity of the input data. • Optimisation challenges and potential computational overhead due to the need for tuning hyperparameters, particularly with evolutionary computation integration.

Continued on next page

Table 2.14: Comparative Analysis of Hybrid Models: Advantages and Disadvantages (Continued)

Ref	Hybrid Model	Advantages	Disadvantages
(Li et al., 2021a)	Multi Model DBM	• Enhanced clarity and detail in the output of fused medical images. • Efficient batch processing capabilities align with the demands of medical diagnosis. • Applicable to various multi-modal medical image fusion types, broadening its diagnostic utility.	• Some information loss during the fusion process, suggesting a need for model and parameter optimisation. • Success is dependent on high-quality, diverse training data, necessitating extensive, well-labelled datasets for the best results.

Continued on next page

Table 2.14: Comparative Analysis of Hybrid Models: Advantages and Disadvantages (Continued)

Ref	**Hybrid Model**	**Advantages**	**Disadvantages**
(Qaid et al., 2021)	CNN-VGG-16/VGG-19	• High accuracy in differentiating COVID-19 from other conditions. • Robust feature extraction from X-ray images using CNNs and TL. • Adaptive design that incorporates various ML algorithms. • Demonstrated generalisability across different datasets.	• Complexity due to the combination of multiple learning techniques. • Performance mainly reliant on data quality and diversity. • Time-consuming model optimisation and hyperparameter tuning.

Continued on next page

Table 2.14: Comparative Analysis of Hybrid Models: Advantages and Disadvantages (Continued)

Ref	Hybrid Model	Advantages	Disadvantages
(Pawar and Talbar, 2022)	GAN-ResNet50	• Automated process that reduces the need for manual regions of interest extraction. • Increased classification accuracy through accurate lung segmentation. • Efficient handling of whole HRCT images.	• Additional complexity from a two-stage process. • Dependency on the initial lung segmentation quality. • Potential challenges in generalising to other ILD types not included in the study.
(Samee et al., 2022)	GN-AlexNet	• Exceptional classification accuracy and sensitivity for brain tumour identification. • Streamlined process with automatic feature extraction and classification. • Architectural adaptability suggesting potential for broader medical imaging applications.	• Complex model due to dual DL architectures. • High computational demands for training and potentially larger datasets. • Uncertainty about performance generalisation across varied and more extensive datasets.

Continued on next page

Table 2.14: Comparative Analysis of Hybrid Models: Advantages and Disadvantages (Continued)

Ref	Hybrid Model	Advantages	Disadvantages
(Xie et al., 2023b)	DCNN-MAS	• Generalisability: Excels in adapting to diverse datasets, surpassing DCNN-only models in varied conditions. • Accuracy: Matches or exceeds state-of-the-art DCNN performances, especially under variable conditions. • Multimodal Data Utilisation: Employs innovative augmentation strategies for effective multimodal integration. • Flexibility: Capable of handling different segmentation tasks with variable dataset sizes.	• Complexity: The combination of DCNN and MAS introduces a more complex segmentation process. • Computational Demand: Requires significant computational power for the end-to-end learnable pipeline, including fine-tuning. • Registration Dependence: Performance depends on the quality of atlas-to-target image registration, limiting flexibility in certain scenarios.

Continued on next page

Table 2.14: Comparative Analysis of Hybrid Models: Advantages and Disadvantages (Continued)

Ref	Hybrid Model	Advantages	Disadvantages
(Chieregato et al., 2022)	CNN-CatBoost	• Comprehensive Assessment: Incorporates both imaging and non-imaging data for a detailed analysis. • Interpretability: Provides insights at global and patient-specific levels, crucial for clinical decisions. • Predictive Accuracy: Offers high accuracy, helping in early intervention strategies for at-risk individuals.	• Limited Generalisability: The model's applicability may be restricted due to single-center data reliance and the specific pandemic timeframe. • Small Dataset Concerns: The relatively modest dataset size could undermine model robustness and prediction variability. • Outcome Definition Limitations: The model's focus on ICU admission or death may not translate across different medical protocols or institutions.

Continued on next page

Table 2.14: Comparative Analysis of Hybrid Models: Advantages and Disadvantages (Continued)

Ref	**Hybrid Model**	**Advantages**	**Disadvantages**
(Srikantamurthy et al., 2023)	LSTM-RNN	• Reduces feature extraction necessity via TL. • Effectively manages imbalanced classes with data augmentation. • Shows superior accuracy in varied classification tasks. • Adaptive to different cancer types and diseases.	• High computational cost and complexity. • Limited to certain magnification levels without adaptability. • Lacks interpretability for clinical decision-making.

Continued on next page

Table 2.14: Comparative Analysis of Hybrid Models: Advantages and Disadvantages (Continued)

Ref	Hybrid Model	Advantages	Disadvantages
(Al-Haidri et al., 2023)	CGAN-U-Net	• Improves image quality significantly. • Simplifies the MRI reconstruction process. • Adaptable to various MRI applications.	• Increased computational complexity due to CGANs. • Effectiveness proven on a specific dataset, questioning broader applicability. • Unclear comparative effectiveness against diverse MRI techniques.
(Raaj, 2023)	Hybrid-CNN	• High performance in detection and segmentation metrics. • Uses image transformation and augmentation for better detection. • Efficient segmentation with a mathematical algorithm.	• Computational complexity due to the multi-step process. • Performance tested on specific datasets, requiring broader validation. • Segmentation may overlook external cancer pixels, indicating potential for improvement.

Continued on next page

Table 2.14: Comparative Analysis of Hybrid Models: Advantages and Disadvantages (Continued)

Ref	Hybrid Model	Advantages	Disadvantages
(Li et al., 2023b)	Hybrid Active Learning framework using Interactive Annotation (HAL-IA)	• Reduces the need for extensive manual annotation, lowering costs and effort. • Selects samples optimally to enhance model learning. • Offers an interactive module for quick, accurate annotations. • Addresses the early stage issue effectively with a progressive strategy.	• Implementation complexity due to multiple components. • May not perform uniformly across all medical imaging types, especially with complex shapes. • Initial sample selection strategy lacks optimisation, suggesting room for improvement with advanced techniques.

Continued on next page

Table 2.14: Comparative Analysis of Hybrid Models: Advantages and Disadvantages (Continued)

Ref	**Hybrid Model**	**Advantages**	**Disadvantages**
(Vasudeva and Chandrashekara, 2023)	CNN-LSTM Grey Level Co-occurrence Matrix (GLCM)	• Achieves high accuracy in classification using a CNN-LSTM model enhanced by GLCM features. • Extracts robust features, leveraging both DL and texture analysis. • Enhances image retrieval with the Euclidean Distance Technique, suitable for large datasets.	• Potentially high computational demand due to the complex model architecture. • Complexity in training and optimisation might require substantial data and resources. • Implementational challenges could arise from the integration of diverse techniques, complicating real-world applications.

These innovative hybrid methods have addressed medical imaging challenges, each with unique strengths like improved accuracy, enhanced image quality, and efficient segmentation (Srikantamurthy et al., 2023; Al-Haidri et al., 2023; Raaj, 2023). However, they share common disadvantages such as computational complexity and the necessity for validation across broader datasets.

2.5 Progression of Disease Detection and Classification Techniques

The progression of disease detection and classification techniques has been marked by significant advancements over recent years, driven by continuous technological advancements and research breakthroughs. This evolution has been characterised by a shift from traditional diagnostic methods towards more sophisticated, accurate, and faster automated systems. The integration of ML and DL models has played a pivotal role, offering unprecedented precision in identifying and classifying a wide array of diseases across various medical imaging modalities. These cutting-edge approaches not only facilitate early detection but also contribute to a deeper understanding of disease mechanisms, leading to more effective and personalised treatment options. Delving into this section, an exploration of the milestones achieved in this dynamic field will be critically discussed, underscoring the transformative impact of these techniques on the landscape of medical diagnostics and patient care.

In this context, a research proposed in (Mendonca and Campilho, 2006) presented an algorithm for the automated detection of the retinal vascular network, a significant evolution step from traditional manual image processing to automated techniques. The algorithm combines differential filters for centreline extraction and morphological operators for vessel segment filling. The approach considers various intensity and morphological properties of vascular structures, such as linearity, connectivity, and width. The proposed method underscores the transition towards automation, significantly enhancing the efficiency and accuracy of medical image processing. The method's adaptability, drawing from both local and global image features, ensures robust performance across a wide spectrum of cases. However, the complexity introduced by its multi-phase algorithm raises concerns regarding computational resource demands and processing time, particularly with large datasets. Moreover, the potential for misdetections and the dependency on accurate initial centreline detection highlight critical areas where inaccuracies can substantially impact the overall effectiveness. On the other hand, authors in (Deng et al., 2023) introduced an automated CT pancreas segmentation approach specifi-

cally for acute pancreatitis patients. This method combines a novel object detection approach namely Region Proposal Network (RPN) and U-Net for segmentation. The research focuses on the precise localisation and segmentation of the inflamed pancreas, demonstrating a specialised approach for acute pancreatitis. The use of RPN detectors and U-Net for segmentation effectively reduces background interference, enhancing focus on target regions. This method's specificity for acute pancreatitis showcases its designed utility in complex medical scenarios. Nevertheless, the two-stage process and reliance on RPN for initial detection introduce potential inefficiencies, with the risk of false positives in adjacent slices and the acknowledged room for performance improvement suggesting areas for future refinement.

MedShift, a pipeline designed to automatically identify and evaluate the significance of shift data in external medical datasets without requiring data sharing between internal and external organisations, proposed in (Guo, Gichoya, Trivedi, Purkayastha, and Banerjee, 2023). It utilises unsupervised anomaly detectors to understand the internal data distribution and identify significant deviations in external datasets. The pipeline then clusters these deviations and uses a multi-class classifier, trained on internal domain data, to assess the impact of removing the identified shift data on classification performance (Guo et al., 2023). Additionally, a data quality metric is proposed to quantify the dissimilarity between internal and external datasets. The efficacy of the method is validated using musculoskeletal radiographs (MURA) and CXR datasets from multiple sources. On the other hand, the paper in (Trombini, Solarna, Moser, and Dellepiane, 2023) introduced an unsupervised, graph-based image segmentation method, specifically designed to partition a digital image into homogeneous regions based on a user-defined, application-specific goal. This goal-oriented approach is innovative in that it doesn't just segment an image into parts but does so with a particular objective in mind, making the regions more meaningful for subsequent analysis (Trombini et al., 2023).

MedShift offers automatic detection and evaluation of shift data to improve AI model performance across various data sources. Its flexibility and quantitative analysis capabilities stand out as major strengths, enhancing dataset curation efforts with objective metrics. However, its approach to handling multi-class problems introduces computational challenges and extends

training times (Guo et al., 2023). The method's effectiveness across tasks beyond classification remains to be fully validated, with dataset-specific adaptations potentially required. The reliance on the performance of anomaly detectors further underscores a dependency that could influence the overall efficacy of the pipeline. The work in (Trombini et al., 2023), on the other hand, focuses on a goal-oriented, unsupervised approach to segmentation, leveraging both local and global image properties through a comprehensive framework. This method's flexibility and robustness across imaging domains highlight its adaptability. Yet, the complexity involved in configuring outcome-based propositions and the potential for over-segmentation present notable drawbacks (Trombini et al., 2023). Furthermore, the computation time and sensitivity to initiate placement may cause challenges in practical applications, affecting the method's utility and efficiency in diverse imaging contexts.

Lung cancer staging, crucial for treatment decisions, is currently a time-consuming and costly process, requiring expert analysis of clinical and imaging data. In this context, a research proposed in (Fotopoulos, Filos, Xinou, and Chouvarda, 2023) aimed to support and automate this process, making it faster, less expensive, and possibly more accurate, thereby facilitating better patient management and treatment planning. The study outlines a method for automating the classification of lung cancer stages using multi-positional radiomics and ML. The specific focus is on classifying lung cancer stages I and II (low severity) versus stages III and IV (high severity) based on CT images and radiomics features from both tumour and lung volumes (Fotopoulos et al., 2023). The proposed method achieved an AP of 77.5% and Recall of 78.7%. Nevertheless, the model's performance is limited by the size of the training set and the simplification of cancer staging into a binary classification, which may reduce the complexity of the disease stages. Moreover, the reliance on data quality and standardisation underscores potential variability in the model's accuracy (Fotopoulos et al., 2023). Similarly, the Gamma-based CNN method excels in delivering high accuracy and robustness in histopathology image analysis through an ensemble model that leverages adaptive weights for enhanced prediction capability (Majumdar, Pramanik, and Sarkar, 2023). This method demonstrates its effectiveness across multiple datasets, underscoring its generalisability. However, it faces chal-

lenges related to computational intensity, the complexity of model choice, and only marginal improvement over the best base learner, which may affect its scalability and practical implementation in resource-constrained environments (Majumdar, Pramanik, and Sarkar, 2023).

The limited size of annotated training datasets in 3D medical imaging results in the limitation of the development of robust 3D CNNs. Traditional methods often rely on self-supervised learning, which may not result semantically discriminative representations due to the lack of large-scale annotated data. In this context, authors in (Zhang, Li, Zhou, Ma, and Yu, 2023) proposed a fully-supervised pre-training framework, termed SVD-Net, which addresses the issue of data insufficiency in 3D medical imaging by leveraging large-scale 2D natural image datasets. The method involves a variable dimension transform (VDT) that reformulates 2D natural images to simulate 3D data, which enables the use of semantic supervision from the 2D domain to train 3D CNNs. The learned 3D representations can then be transferred to various medical imaging tasks. Towards tackling a similar issue, another study presented a Multi-scale Attention GAN (MAGAN) designed for medical image enhancement, specifically designed for unpaired images (Zhong, Ding, Chen, Wang, and Yu, 2023). MAGAN innovatively incorporates multi-scale information fusion in feature extraction through a feature pyramid network (FPN) and emphasises key image regions using attention mechanisms. It also addresses the enhancement process comprehensively by optimising for uniform illumination distribution, texture details, deep semantic features, and smoothness in the enhanced images. The approach leverages two generators and two discriminators in a GAN setup to achieve these goals. SVD-Net stands out for its novel approach of leveraging semantic supervision from 2D image datasets to pre-train 3D CNNs, addressing the insufficiency of annotated 3D medical images (Zhang et al., 2023). This strategy not only enhances model convergence and accuracy but also reduces the need for extensive annotated medical data, setting new performance benchmarks. However, the potential for domain shift and the method's dependence on the size and diversity of the 2D dataset used for pre-training, alongside the significant computational resources required, cause notable disadvantages (Zhang et al., 2023). MAGAN, through the integration of multi-scale information and attention mechanisms, focuses on optimising key

image areas while maintaining essential details, suitable for unpaired images (Zhong et al., 2023). This adaptability makes it a practical solution for enhancing medical images, improving downstream segmentation tasks. Despite these strengths, the complexity of the MAGAN architecture could lead to higher computational costs. Additionally, the method faces challenges in accurately differentiating specific image features and relies on unpaired images, which might limit learning capacity compared to supervised methods. The need for further research to address these limitations and streamline the model for faster processing without loss of quality is acknowledged (Zhong et al., 2023).

A research presented in (Kutan, KUTBAY, and ALGIN, 2023) focused on cerebrovascular vessel segmentation using DL approaches for Time-of-Flight Magnetic Resonance Angiographs (TOF-MRAs). This research is significant due to the impact of cerebrovascular diseases as a leading cause of death and disability worldwide. Accurate segmentation of cerebral vessels is crucial for early disease diagnosis and surgical planning. The study involves two main stages: (1) creating a labelled dataset through Hessian-based filters and image processing algorithms, and (2) comparing the performance of state-of-the-art DL architectures (U-Net, ResUNet, ResUNet++, TransUNet) in vessel segmentation where ResUNet++ achieved the highest performance, with a mean IoU score of 91.6%, outperforming other tested architectures.

Similarly, authors in (Yousaf, Iqbal, Fatima, Kousar, and Rahim, 2023) introduced a CNN based integrated model for the simultaneous detection and classification of two brain diseases: tumours and Ischemic stroke. This model is an advancement of the encoder-decoder architecture based on U-NET, enhanced to incorporate feature maps from one encoder block fused with the output of a subsequent encoder block. This approach aims to maintain low-level, detailed information and distinguish overlapping features during the encoding process, in addition to utilising U-NET skip connections. The proposed model demonstrated exceptional performance on a challenging combined medical dataset, achieving an average Acc of 99.56%, Spe of 99.99%, precision of 99.59%, and an F1-score of 99.57%.

Conversely, a research in (Li et al., 2023c) proposed a novel unsupervised anomaly de-

tection framework named SSL-AnoVAE, which incorporates a self-supervised learning (SSL) module into the anomaly detection process. This SSL module is designed to provide more precise semantic features (e.g., texture, structure, colour-related features) as prior information for better image reconstruction. The uniqueness of SSL-AnoVAE lies in its flexibility and universality, allowing application across different image modalities by adjusting the free-labels from image transformations to extract feature information with various semantic meanings (Li et al., 2023c). SL-AnoVAE achieves an Acc of 93.34%, Spe of 94.01%, and Sen of 92.30%. The method introduced in (Kutan, KUTBAY, and ALGIN, 2023) significantly reduces manual annotation efforts and shows promise for clinical applications in cerebrovascular disease diagnosis. However, its effectiveness depends on the quality of labelled data and requires considerable computational resources, raising concerns about biases and the practicality of deploying such complex models in resource-limited settings (Kutan, KUTBAY, and ALGIN, 2023). On the other hand, the enhanced U-NET model, with new skip connections, exhibits high clinical applicability through its performance metrics (Yousaf et al., 2023). Nonetheless, its generalisability may be constrained by reliance on specific datasets and untested applicability to a broader range of brain diseases or imaging modalities. Additionally, the method's focus on two-class classification and training on limited datasets could affect its robustness and generalisation capabilities (Yousaf et al., 2023). The proposed approach in (Li et al., 2023c) capitalises on self-supervised learning through the SSL-AnoVAE framework for improved anomaly detection and staging in retinal diseases. Its universal adaptability and insights into optimising unsupervised anomaly detection (UAD) methods underscore its potential clinical value. However, challenges with mathematical interpretability and detecting diseases with minimal structural or colour changes in retinal images suggest limitations in its applicability across different medical imaging tasks, particularly those less reliant on structure and colour information (Li et al., 2023c).

2.5.1 The impact of Learning-based Approaches on DMO Disease Classification

The field of ophthalmology has undergone a profound transformation, attributed to the integration of cutting-edge technologies, particularly DL and ML. Eye diseases, which encompass a wide spectrum of conditions affecting vision and ocular health, have garnered specific attention in this era of AI-driven healthcare innovation. DL and ML techniques are making significant strides in the early diagnosis, monitoring, and treatment of these ocular disorders, ultimately improving patient outcomes and quality of life. Table 2.15 summarises the discussed literature to include study objective, data source/sample, DMO related disease findings, and research gaps.

Table 2.15: ML and DL Applications in DMO Related Prediction Literature

Ref	Year of Publication	Study Objective	Data Source/ Sample	DMO Related Disease Findings	Research Gaps
(Varadarajan et al., 2020)	2020	Predict center-involved DMO directly from Fundus photographs.	• Thailand dataset • Eye PACS-DMO dataset in the US.	• Model outperformed retinal specialists in detecting center-involved DMO (ci-DMO). • Prediction of intraretinal and subretinal fluid. • AUC-ROC of 0.89, 85% Sen at 80% Spe.	• Dataset diversity • Data standardisation

Continued on next page

Table 2.15 ML and DL Applications in DMO Related Prediction Literature (Continued)

Ref	Year of Publication	Study Objective	Data Source/ Sample	DMO Related Disease Findings	Research Gaps
(Xu et al., 2022)	2022	Predict visual acuity outcomes in DMO patients following anti Vascular Endothelial Growth Factor (anti-VEGF) therapy using a GAN algorithm.	Retrospective review of DMO patients' records who underwent anti-VEGF therapy.	AI-based prediction displays the therapeutic effects of different treatment drugs on DMO patients. The model produces high-resolution, near-realistic OCT images.	• Inclusion criteria • Model generalisability • Information loss • Alternative model exploration • Experimental details • Sample size

Continued on next page

Table 2.15 ML and DL Applications in DMO Related Prediction Literature (Continued)

Ref	Year of Publication	Study Objective	Data Source/ Sample	DMO Related Disease Findings	Research Gaps
(Zhang et al., 2022)	2022	Predict visual acuity outcomes in diabetic patients post anti-VEGF therapy using an ensemble model of regression algorithms.	281 patients with clinical and OCT image-based features dataset.	The ensemble model of LR and RF models had the best predictive performance for visual acuity outcomes with Mean Average Errors (MAEs) between 0.137-0.153 for acuity and 0.164-0.169 for acuity variance.	• Generalisability • Validation necessity

Continued on next page

Table 2.15 ML and DL Applications in DMO Related Prediction Literature (Continued)

Ref	Year of Publication	Study Objective	Data Source/ Sample	DMO Related Disease Findings	Research Gaps
(Rasti et al., 2020)	2020	Predict the response of DMO patients to anti-VEGF treatment using a novel DL model named CADNet.	127 patients' pre-treatment OCT scans.	CADNet, with incorporation of attention mechanisms and feature selection.	• Dataset diversity • Threshold exploration • Model interpretability • External validation
(Chen, Chiu, Chen, Woung, and Lo, 2018)	2018	Present visual acuity outcomes in DMO patients at different timelines.	DMO dataset from DRCR.net, USA	Utilised multiple clinical variables for prediction with an MLP model; reported high correlation coefficients.	• Model comparison justification • Generalisability • Interpretability

Continued on next page

Table 2.15 ML and DL Applications in DMO Related Prediction Literature (Continued)

Ref	Year of Publication	Study Objective	Data Source/ Sample	DMO Related Disease Findings	Research Gaps
(Rajesh, Raajini, Sagayam, and Dang, 2020)	2020	Propose a ML-based model combining statistical methods with SVM and KNN for DMO detection.	DMO dataset with specific medical conditions.	DMO detection, achieving high performance.	• Methodology clarity • Generalisability • Interpretability
(Kumar and Gupta, 2023)	2023	Develop a DL-based model for binary classification of normal and eye disease images.	Kaggle dataset	ResNet50 and Xception models achieved the highest validation accuracies.	• Data augmentation specificity • Hyperparameter tuning details • Cross-validation • Generalisability

Continued on next page

Table 2.15 ML and DL Applications in DMO Related Prediction Literature (Continued)

Ref	Year of Publication	Study Objective	Data Source/ Sample	DMO Related Disease Findings	Research Gaps
(Mishra and Singh, 2022)	2022	Apply DL-based approaches for classifying OCT eye scans	Kaggle	External Limiting Membrane (ELM) Segmentation	• Benchmarking with current methods • Interpretability • Generalisability • Explanation of techniques and feature extraction
(Li et al., 2022a)	2022	Develop a framework for DR and DMO classification using Fundus images	8739 Fundus images	DR and DMO Classification achieving high performance for classification tasks. Outperformed ophthalmologists and state-of-the-art methods.	• Dataset biases and quality • Data heterogeneity impact • Generalisability • Validation

One of the main challenges identified in the literature is the detection of ci-DMO. This is caused by the human evaluation of Fundus photographs which has several limitations. Authors

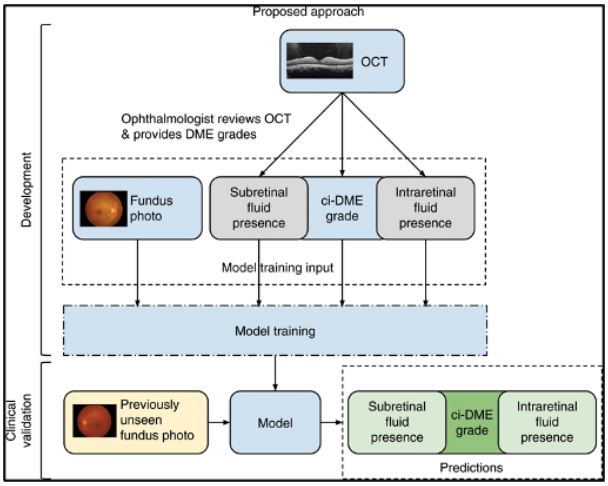


Figure 2.8: ci-DMO Prediction Process Using OCT and Related Fundus Scans (Varadarajan et al., 2020).

in (Varadarajan et al., 2020) proposed leveraging DL to predict ci-DMO directly from Fundus images, aiming to improve accuracy and cost-effectiveness (Varadarajan et al., 2020) (Figure 2.8).

This research used two independent datasets, one from Thailand and another from Eye PACS-DMO in the US for training and validation of the DL model. Their model outperformed retinal specialists in detecting ci-DMO demonstrating higher sensitivity and specificity. Their model also predicted the presence of intraretinal and subretinal fluid. The proposed framework covers the generalisability criterion presented by the test of a secondary dataset. Their experiments proved that the use of larger training datasets helps in further enhancing the model Acc where the model outperformed human evaluation by achieving a ROC-AUC equal to 89% corresponding to 85% Sen at 80% Spe against only 45% Spe (Varadarajan et al., 2020). The features used in this work also prove the importance of the region around the fovea in predicting ci-DMO from Fundus images. While the research used two datasets, both are relatively small and specific to certain populations which results a limited diversity in used datasets. In addition, data standardisation is a key gap in this work where the criteria for ci-DMO diagnosis and inclusion/exclusion varied between datasets. Interpretability is also one of the key gaps this work. In fact, while the study identifies the relevant region for predictions, further research could delve into the interpretability of the model, explaining why certain features are crucial for diagnosis.

The prediction of visual acuity outcomes in DMO patients following anti-VEGF therapy is a challenging theme covered by authors in (Xu et al., 2022). The paper retrospectively reviewed records of DMO patients who underwent of anti-VEGF therapy to include both A-scans and B-scans. The authors employed pix2pixHD GAN algorithm to generate post-therapeutic OCT images from pre-therapeutic ones. The model demonstrated a good accuracy level where the it objectively displays the therapeutic effects of different treatment drugs for DMO patients. Despite the high ability of their proposed framework to produce high reduction and near-realistic images (which are essential for prediction), authors' proposed methodology still suffers from multiple key gaps as follows:

- Restrictive inclusion criteria in the choice of the used dataset which may limit the diversity of the patient population.
- Lack of model generalisability where the framework did not include more complex clinical scenarios.
- Loss of important information due to some pre-processing stages undertaken by authors such as resizing.
- Lack of alternative models that could potentially perform better than the selected GAN. This would add additional insights into the suitability of the final framework.
- Lack of detailed experimentation outcomes.
- Sample size considered which has a direct impact on the model's predictive performance.

Predicting visual acuity outcomes in patients with diabetes after anti-VEGF therapy was also a recurrent theme in (Zhang et al., 2022). Authors proposed research involved 281 patients, where the used dataset included 18 features involving both clinical and OCT image-based features. Six regression algorithms were tested, with the ensemble model of Logistic Regression (LR) and RF achieving the best predictive performance for visual acuity (mean average error (MAE) of 0.137-0.153) and visual acuity variance (MAE of 0.164-0.169). Despite

the high accuracy presented by their model, it lacks in generalisability where it was only tested on relatively small dataset. In addition, while the paper mentions feature selection, it would be valuable to provide more details on the criteria or methods used to select the 18 features and justify why certain features were chosen over others. Also, towards validating the predictive models, external validation on an independent dataset would be beneficial. This will boost the robustness and generalisability approval of their proposed framework.

Predicting the response of patients with DMO to anti-VEGF treatment using pre-treatment OCT scans was also a complicated challenge addressed by to authors in (Rasti et al., 2020). The authors proposed a novel DL model named CADNet (Convolutional Attention-to-DMO Network) to predict treatment response. CADNet incorporates attention mechanisms and features selection techniques. To assess their model's generalisability, authors employed cross-validation. The study compared CADNet's performance with other baseline models, including traditional ML algorithms and popular DL architectures like VGG-16, ResNet50, InceptionV3, and Xception. CADNet outperformed these models. While the study's dataset consists of 127 patients, it would be beneficial to have a larger and more diverse dataset to enhance the model's generalisability. In addition, the study uses a fixed threshold of -10% to classify patients as responsive and non-responsive. Different thresholds may have varied clinical implication, and the choice of threshold should be further explored. DL models, including CADNet, often lack interpretability. Understanding which features or patterns the model uses for predictions is crucial in a clinical setting.

Another approach of applying ML-based models under the umbrella of predicting visual acuity outcomes was introduced in (Chen et al., 2018). Authors proposed a new model to predict visual acuity outcomes at different timelines using DMO dataset. The study employed MLP with a backpropagation learning rule. Resonating with the earlier referenced studies, generalisability remains a problem in this work. The study used several clinical variables for prediction, which is comprehensive. However, the rationale for including specific variables should be better explained. Additionally, the study did not incorporate certain factors, such as adverse effects, which could be relevant in real-world clinical decision-making. The choice of an MLP

neural network for prediction is reasonable. However, the paper lacks details on the model's architecture, hyperparameter tuning, and validation techniques. The paper reported high correlation coefficients for the MLP models (0.79), which is promising. However, it would be beneficial to provide additional metrics such as MAE to assess the model's predictive accuracy more comprehensively. Details about the validation techniques used for assessing the model's generalisation ability are missing. In fact, it is essential to describe how the dataset was split into training, validation, and testing sets and how cross-validation was performed.

By combining different statistical methods alongside ML techniques such as SVM and KNN, authors in (Rajesh et al., 2020) proposed a ML-based model using DMO dataset with specific medical conditions. The suggested hybrid approach enables a comprehensive analysis of epistasis in the context of DMO detection, achieving 99% recall, 73.88% precision, 99.99% Acc, and 84.61% F1-score. The complexity of the proposed methodology is quite challenging due to lack of explanation of used methods as well as used parameters. The paper compares the proposed statistical method with SVM and KNN working solely. However, a broader comparison with other state-of-the-art epistasis detection methods or an explanation of why these two ML algorithms were chosen for comparison would add depth to the analysis. Generalisability and interpretability are also problematic in this work.

Towards performing binary classification task, authors in (Kumar and Gupta, 2023) proposed a novel DL-based model for classifying normal and eye disease related images using Kaggle dataset. Their DL related experimentation involved multiple architectures to include CNN, deep CNN, AlexNet2, Xception, InceptionV3, DenseNet121, and ResNet50. Performance wise, the latter achieved the highest validation Acc of 98.9% followed by 98.4% Acc for Xception. While data augmentation techniques were employed to address the scarcity of training samples, the study does not specify the degree of augmentation on the specific parameters used, which can impact model's performance. The study also lacks details regarding hyperparameter tuning for each model. Cross-validation is a crucial step in assessing model generalisability, however, the study does not mention whether cross-validation was performed or how the dataset was split for training and validation. Interpretability, equally, is one of the

key gaps of the proposed framework. In fact, DL models particularly CNNs, are often considered as "black boxes", which makes it challenging to understand the decision-making process of these models in a medical context. Similarly, authors in (Mishra and Singh, 2022) applied DL-based approaches to classify normal and DMO affected eye scans (OCT). The paper lacks of an explanation as well as justification of the chosen technique for all stages to include processing, features extraction, and choice of evaluation metrics and their clinical relevance. The paper mentions that CNNs with varying numbers of convolutional layers were evaluated, but it also lacks detailed information on the architecture, hyperparameter tuning, and rationale behind choosing the final model. Despite the use of data from different sources, the proposed framework does not address how the model performs on data from different sources or if it is robust to variations in image quality.

2.5.2 Advancements in Learning-based Approches for DR Detection

DR is a prevalent microvascular complication of diabetes, leading to progressive vision impairment and, in severe cases, blindness. Early detection and timely intervention are crucial for managing the progression of DR. Traditional diagnostic methods involve manual examination of Fundus photographs by ophthalmologists, a process that can be time-consuming and subject to inter-observer variability. Recent advancements in ML and DL have showed in a new era for DR prediction through Fundus images. Automated algorithms can analyse vast datasets of Fundus photographs, identifying implicit pathological changes often overlooked in manual examinations. These ML and DL models not only enhance diagnostic accuracy but also facilitate a quicker and more scalable screening process. The expanding literature on this topic highlights the transformative potential of these technologies in revolutionising DR screening and management.

Studies presented in the literature in (Gargeya and Leng, 2017, Gulshan et al., 2016) delved into the development and validation of DL algorithms for the automated detection of DR from retinal Fundus images. The study proposed in (Gargeya and Leng, 2017) focused on an AI model trained using over 75,000 public images and presents a robust performance with an AUC

of 0.97. The efficacy of this model in identifying DR is noteworthy, suggesting a considerable potential for AI-driven DR screening on a global scale. On the other hand, the work in (Gulshan et al., 2016) introduced a deep CNN that has been trained on a substantially larger dataset of 128,175 retinal images. Notably, its performance metrics, particularly the AUC which exceeded 0.99 for both datasets, are impressive. The algorithm's performance was assessed using two separate datasets with high grading consistency by at least seven board-certified ophthalmologists. The algorithm displays adaptability, with the flexibility to balance between sensitivity and specificity based on different operational needs, making it potentially a very precise tool for referrable DR screening in clinical settings. In another dimension, the study in (Li et al., 2019) showcased the application of deep TL using the Inception-v3 architecture. This model was trained on 19,233 images from 5278 patients, and following a 10-fold cross-validation strategy, it displayed an Acc of 93.49% with an AUC of 0.9905. The model's performance, especially in distinguishing between cases requiring referral and those not, was comparable to human experts, indicating its reliability in automating DR detection and recommendations. While each algorithm showcases impressive performance metrics, the choice of dataset sizes and methodologies differ. Authors in (Gargeya and Leng, 2017) and (Gulshan et al., 2016) emphasise the breadth of their datasets, however, authors in (Li et al., 2019) leverage the power of TL. Additionally, the provision of visual heatmaps in (Gargeya and Leng, 2017) study can be a unique aid for clinicians. However, despite these advancements, there's an implicit call for further research in all studies, pointing towards a need for real-world implementation and exploration of the algorithms practical implications.

In the field of DR prediction, recent research has delved deeply into devising automated models that leverage the capabilities of advanced neural networks to identify DR features from medical imagery. The research proposed in (Tsiknakis et al., 2021) delves into creating an optimal model for detecting DR characterised by symptoms like augmented blood vessels, fluid leaks, exudates, haemorrhages, and micro-aneurysms. Emphasising the pivotal role of contemporary medical imagery in diagnosing DR, it also points out evaluation challenges. The study introduces an automated knowledge model to discern DR features via various neural networks,

namely Back Propagation Neural Network (BPNN), DNN, and CNN. The foundational BPNN model exhibited the least accuracy due to its singular hidden layer, contrasting the superior performance of DL structures such as DNN and CNN. These advanced models could locate DR features and ascertain their severity. A fundamental challenge was setting precise thresholds for each feature class, a challenge surmounted using the weighted Fuzzy C-means algorithm. Evaluating their efficacy on 1000 retinopathy images, the DNN model outperformed the BPNN in accuracy and efficiency, while the CNN, backed by the VGGnet model, produced a 72.5% Acc rate on a training set but waved slightly during tests on 300 images. Salient discoveries encompassed DNN's paramount accuracy outperforming both DNN and CNN, the central processing unit (CPU) imposed limitations affecting CNN's training time, and DNN's overarching effectiveness and precision, especially against DNN.

Leveraging the existing drive of ML advancements in DR detection, authors in (Mushtaq and Siddiqui, 2021) employed DenseNet-169 to detect DR in its early stages. The approach classifies Fundus images based on severity levels: No DR, Mild, Moderate, Severe, and Proliferative DR (PDR), using datasets from DR Detection 2015 and APTOS 2019 Blindness Detection. The DenseNet-169 model achieved a training Acc of 95% and a validation Acc of 90%. When compared to other models like SVM, DT, KNN, and a regression model, the proposed model achieved the highest Acc of 90%.

Yet another noteworthy contribution in this arena was a study in (Zhang et al., 2021) which focused on devising an intelligent diagnostic method for severe DR using colour Fundus images from a Kaggle dataset of 88,702 photos. Utilising the Inception-V3 classification algorithm, it was observed that images with a resolution of 896x896 pixels surpassed the performance of 299x299 pixel images in metrics like harmonic mean, AUC, sensitivity, and specificity. Notably, prediction errors predominantly surfaced in the moderate Non-PDR (NPDR) grade, especially in detecting the IRMA lesions. Despite its merits, the study identified the challenge of data imbalance in public datasets and underscored the need for well-balanced data to improve the model's accuracy. The findings underscore AI's promise in enhancing DR diagnostic precision, even though with a call for continued refinements.

These papers collectively underscore the expanding potential of advanced neural network models in DR detection. While traditional DR diagnostic methods are resource-intensive and often vulnerable, the integration of ML showcases promising results. However, each study, while innovative, also sheds light on the challenges and limitations faced. From data imbalance to the need for precise thresholds and the computational challenges imposed by hardware constraints, the research illuminates the outstanding challenges of automating DR detection. It is evident that while AI and ML present revolutionary possibilities in DR diagnosis, there's a persistent need for continued refinements, dataset optimisation, and computational advancements to realise their full potential.

Switching the focal point to glaucomatous retinal images, paper proposed in (Thanki, 2023) unveils a system underpinned by DNN and ML paradigms. Rigorous evaluations were conducted using performance metrics such as confusion matrix andtrue positives. The model, built on key training parameters, was benchmarked against datasets from the DRISTHI-GS and ORIGA archives. Leveraging the SqueezeNet model's capabilities, the paper illustrated that LR emerged as the leading classifier, surpassing its counterparts in accuracy and precision metrics. A study published in (Akella and Kumar, 2023), introduced a computerised system for analysing and assessing DR using retinal Fundus photographs. The research utilises the YOLO-V3 DL model to identify and categorise DR into five stages: normal, mild, moderate, severe, and proliferative, using colour Fundus images. The model exhibited high precision and sensitivity, with the mean average precision (mAP) measured for DR lesion detection. The findings suggest that the proposed model outperforms existing ones in accuracy and execution time, effectively distinguishing all DR stages. Another paper focused on the automatic detection of DR, delving into DR detection and its different stages using DL with Fundus images (Sunkari et al., 2024). Using a real-time hospital dataset, the study introduced a ResNet-18 architecture paired with a Swish function, achieving an Acc of 93.51%, Sen of 93.42%, precision of 93.77%, and an F1-score of 93.59%. The research concludes by comparing the effectiveness of different models like Simple CNN, VGGNet-16, MobileNet-V2, and ResNet, finding the ResNet-18 with Swish to be the most efficient for DR detection.

Contrastingly, authors in (Surya, Kashyap, Nadig, and Raman, 2023) aimed to create a predictive model for diagnosing DR in the Indian population, using systemic data and excluding Fundus photography. The study utilised ML on datasets from a population-based cross-sectional study with 1425 subjects, including known and newly diagnosed diabetes cases. Five ML algorithms were tested: RF, LR, SVM, ANN, and DT. Data were split in two experimental methods: an 80-20 percentage split and a three-way split (60% training, 30% validation, 10% test). 10-fold cross-validation was also applied. Performance was assessed using the ROC curve, AUC, accuracy, sensitivity, and specificity. The RF classifier stood out, achieving the highest performance with AUC values of 91% in the percentage split, 86% in the three-way split, and 90% in cross-validation. Given its superior performance, the RF classifier is recommended for targeted DR screening in the broader population.

Collectively, these studies underscore the evolving landscape of DR detection techniques. While the research in (Thanki, 2023) investigations underscore the potential of DL models in analysing retinal images, authors research uniquely diverges, focusing on systemic data, reflecting the diversity in approach (Wahab Sait, 2023). The YOLO V3's performance in (Akella and Kumar, 2023) might be commendable for its speed and accuracy, but the integration of Swish function in ResNet-18 indicates an innovative approach to improve detection capabilities (Sunkari et al., 2024). The other study in (Surya et al., 2023), however, emphasises the importance of expanding detection strategies beyond imagery, especially in resource-limited settings. As the prevalence of DR rises, such diverse methodologies ensure a broader, more inclusive approach to its early detection and management.

Several other studies have delved into this area, capitalising on the power of ML and DL techniques to interpret DR datasets for accurate disease prediction. One research emphasised the overlooked aspect of data pre-processing and dimensionality reduction, which is instrumental in generating unbiased results (Usman, Saheed, Ignace, and Nsang, 2023). In this venture, colour Fundus Photographs (CFPs) underwent a meticulous data pre-processing, followed by leveraging PCA for feature extraction. Building on this foundation, a DL Multi-Label Feature Extraction and Classification (ML-FEC) model was conceived. This model taps into the poten-

tial of pre-trained CNN structures. Furthermore, a strategic deployment of TL integrated with three renowned CNN architectures: ResNet50, ResNet152, and SqueezeNet1. Adjustments catering to lesion detection and classification provided promising outcomes, with ResNet152 taking the lead in Acc, reaching 94.40%. Such findings point towards the model's potential in real-world clinical setups, improving DR screening programs. Another study, inspired by the successes of DL in diagnosing various conditions, paved a way to design enhanced neural networks targeting the precise detection and categorisation of DR across five distinct stages (Khanna, Singh, Thawkar, and Goyal, 2023). The methodology encompassed the formulation of three novel CNNs. The first was an original creation, the second integrated the effectiveness of five elite networks, while the third synergised CNN with the capabilities of CNN-LSTM structures. Their performances were juxtaposed with twenty-one globally acknowledged image nets. Comprehensive evaluations manifested a peak in accuracy, sensitivity, and AUC score, suggesting that the proposed networks not only performed well often overpass many existing models. Such advancements can be instrumental in identifying retinal complications in diabetic patients, streamlining diagnosis, and fortifying preventive measures against vision deterioration.

Further expanding the horizon of eye-related disorders, another paper brought to light the risks of Uveal melanoma (UM) and choroidal nevus (CN) (Shakeri et al., 2023). While DR predominantly affects individuals aged between 20-65, UM, a grave intraocular cancer, primarily threatens those aged between 50-80. Early identification of UM can significantly limit associated mortality risks. To this end, a method integrating TL with a CNN was introduced, aiming to detect UM and enrich diagnostic interpretations (Shakeri et al., 2023). However, the complex nature of DL models can often obscure prediction comprehension. Addressing this challenge, the SHAP methodology was adopted. It strategically highlights the regions of an eye image that predominantly influence DR and CN predictions. The outcomes were promising, with SHAP analysis serving as a beacon, elucidating the underlying reasons behind classifications and offering a profound understanding of prediction dynamics.

Based on the above discussion, these studies underscore the immense potential and versatil-

ity of ML and DL techniques in revolutionising ocular diagnostics. While all three researches advocate the cause of early detection and intervention, their methodologies and focal points vary. The research in (Usman et al., 2023) accentuates the importance of data pre-processing and feature extraction, while the method proposed in (Khanna et al., 2023) delves into the innovation of CNN architectures. The study proposed in (Shakeri et al., 2023) uniquely integrates TL with SHAP to demystify predictions. Each study adds a unique dimension, and when viewed in conjunction, they offer a comprehensive insight into the future of DR detection and other ocular disorders. The consolidation of their findings can serve as a robust platform, pushing the boundaries of ocular diagnostics and paving the way for accurate and timely interventions.

A research published in (Keerthana, Tejasree, Rao, Kumar, and Yalla, 2023) examines the efficacy of various ML algorithms in detecting DR. The study utilises large sets of retinal images from multiple databases including Kaggle, Messidor, and IEMRC. The research employs an ensemble of ML classifiers applied to features extracted from retinal image processing to determine the disease's presence. The aim is early detection of DR to prevent severe vision loss. The study incorporates both supervised algorithms, such as SVM, ResNet, DensNet, Naive Bayes (NB), KNN, and neural networks, as well as unsupervised algorithms like k-means clustering, hierarchical clustering, and Markov chains for classification. The most accurate results were achieved by ResNet and DenseNet with an Acc rate of 96.22%.

Traditional DR diagnosis using Fundus images is challenging and time-consuming, as it requires expert professionals to detect minute features. Hence, an automated method to diagnose DR can be beneficial for early detection. A study proposed in (Athira and Nair, 2023) classifies DR into three stages: No D (No DR), NPDR, and Proliferative DR (PDR). Although DL algorithms are popular for classification tasks, many have low accuracy for early DR stages. The research introduces an algorithm employing advanced image processing, automatic hyperparameter tuning, and neural network training, emphasising the sensitive features for improved prediction. The algorithm's effectiveness was tested against modified versions of Resnet50, VGG-16, Mobilenetv2, Inceptionv3, and InceptionResnetv2. The values obtained are consoli-

dated and displayed in Table 2.16. The Resnet50-based network showed the best performance for both tasks.

Table 2.16: Resnet-50 Outperformance in DR Classification Task (Athira and Nair, 2023)

Evaluation Metric	**Resnet-50**	**VGG-16**	**Mobilenet-V2**	**Inception-ResnetV2**	**Inception-V3**
Classification Accuracy	0.947	0.861	0.858	0.870	0.852
Kappa Score	0.884	0.713	0.739	0.929	0.724
Detection Accuracy	0.998	0.942	0.94	0.982	0.949

Authors in (Syed and MA, 2023) proposed an automated system to detect, segment, and classify the microvascular complication of type 2 diabetes mellitus, DR, using the EyePACS dataset. The research introduces an RU-Net (Residual U-Net) for segmentation and a CCNN (Concatenated CNN) for DR's multi-class classification. The suggested classification approach achieved accuracies of 98.81% for benchmark data and 96.83% for real-time data, proving its potential to aid doctors in efficiently and accurately detecting and classifying DR.

Reflecting on the literature, the increasing reliance on advanced ML and DL methods in detecting and classifying DR using retinal images. From using ensemble classifiers and attention mechanisms to incorporating hybrid optimisation algorithms, the literature demonstrates rapid advancements in automated DR diagnosis. Notably, the ResNet-based approaches appear to outperform other methods, suggesting its prominence in the field. However, while many of these methods have shown high accuracy rates, practical implementation in real-world scenarios, scalability, and cost-effectiveness are aspects that need further exploration. It's also essential to evaluate the model's adaptability to different datasets and real-world conditions. The literature opens doors to more streamlined and efficient prediction of DR, which can significantly impact patient care and management. However, continuous validation, especially with larger and more diverse datasets, is imperative to ensure the reliability and generalisability

of these models in clinical settings.

2.5.3 Progression of Analysis Techniques for Pulmonology Related Conditions

ML and DL have equally emerged as revolutionary tools in the analysis of X-ray images. Over recent years, numerous studies have been conducted to leverage these techniques for the diagnosis of various disorders. Particularly, their application in predicting pneumonia from X-ray images has garnered significant attention. Traditional diagnostic methods often depend on radiologists' expertise, making them susceptible to human error and subjective variability. In contrast, ML and DL models offer a more standardised and efficient approach, with the potential to identify complex patterns not easily discernible by the human eye. The growing body of literature in this field underscores the promising potential of these models in enhancing the accuracy and speed of pneumonia detection.

In recent years, there's been a noticeable advancemetns in the application of DL models for detecting and classifying pneumonia and other related lung diseases using CXR. The research suggested in (Sharma and Guleria, 2023a) utilised a model based on VGG-16 to classify pneumonia from CXR datasets, showcasing an enhanced performance compared to other models with an Acc ranging between 92.15% to 95.4%. However, while CXRs serve as a vital diagnostic tool, interpreting them accurately requires skilled radiologists, which leads to challenges such as limited expert availability and high costs. A study proposed in (Yi, Tang, Tian, Liu, and Wu, 2023) introduced a DCNN designed to overcome these challenges. Their model was adept at categorising CXR images as normal or indicative of pneumonia, demonstrating superior efficiency in disease identification. Lung afflictions like pneumonia, COVID-19, and Tuberculosis have exhibited significant overlaps, necessitating holistic diagnostic approaches.

Recognising this need, authors in (Ahmed et al., 2023b) introduced a model capable of detecting these diseases simultaneously, displaying commendable accuracy across various datasets. Their innovative approach acknowledges that a patient testing negative for one disease might

still suffer from another. Performance evaluation using several public datasets from Kaggle generated significant results: an overall Acc of 98.72%, with recall scores of 99.66% for pneumonia, 99.35% for No-findings, 98.10% for Tuberculosis, and 96.27% for COVID-19. When subjected to unseen data from the same augmented dataset, the model surpassed previous research in terms of accuracy and other evaluation metrics.

With the pandemic's rapid spread, tools like CovidDWNet, proposed in (Celik, 2023), have become critical. This DL architecture, adept at promptly detecting COVID-19 using CT and X-ray images, showcased impressive accuracy metrics, emphasising its rapid prediction capability. The proposed architecture utilises feature reuse residual block (FRB) and depthwise dilated convolutions (DDC) units, enhancing feature acquisition from the images. Further integration with the Gradient boosting (GB) algorithm led to the CovidDWNet+GB model, which increased performance by roughly 7% in CT images and 3-4% in X-ray images. This combined architecture achieved an outstanding 99.84% and 100% Acc on binary class CT datasets and demonstrated 96.81% Acc on multi-class X-ray images and 96.32% on combined X-ray and CT datasets. Notably, CovidDWNet+GB can process thousands of images within seconds, highlighting its rapid prediction capability.

Given the importance of accurate CXR examinations, authors in (Ahmed, Nuwagira, Torlak, and Coskunuzer, 2023a) introduced a novel technique utilising topological data analysis (TDA) to extract unique patterns, making it exceptionally efficient even with small datasets. Furthermore, AI tools like the LWSNet, introduced by Lasker, have been pivotal in distinguishing between various lung diseases, exhibiting heightened accuracy through its unique DL structure (Lasker, Ghosh, Obaidullah, Chakraborty, and Roy, 2023). The proposed framework uses a simplified version of typical DL models paired with a lightweight CNN model, making it suitable for resource-limited devices. Testing on three separate public datasets and their combined version revealed the highest classification Acc of 98.54% with the quadruple stack. Comparative analyses showed that using LWSNet improved the average accuracy between individual to quadruple models by up to 2.88% across the datasets.

2.6 Identified Challenges

The landscape of diseases diagnosis leveraging ML and DL techniques is infested with serious challenges, which are pivotal in understanding the evolving interest in hybrid model approaches. These challenges not only highlight the limitations of singular ML/DL methodologies but also emphasise the necessity for innovative solutions that hybrid models present, as follows:

- **Data Requirements and Accessibility:** The efficacy of ML/DL models is heavily highly dependent on the availability of extensive datasets for training. However, acquiring such datasets is often restricted by privacy concerns and restrictive data sharing policies, which limit the scope of potential training material.
- **Insufficiency of Fully Labelled Datasets:** The lack of comprehensively labelled datasets further compounds the difficulty in training and validating models effectively, as full and accurate annotations are crucial for the development of reliable diagnostic tools.
- **Data Imbalance:** The prevalence of imbalanced datasets, where certain conditions are over-represented compared to others, introduces biases into the models, skewing predictions and undermining their clinical utility.
- **Input Data Variability:** The heterogeneity in data modalities, capturing techniques, and the conditions under which data is collected introduces variability that can significantly challenge the training process and the model's ability to generalise.
- **Feature Extraction and Selection:** The complexity of medical images and the refinement of disease manifestations necessitate sophisticated feature extraction and selection methods to ensure model accuracy, a task that remains challenging due to the complex nature of medical data.
- **Generalisation Across Modalities:** The specificity of models to particular data modalities restricts their applicability across diverse clinical settings, demanding flexible solutions that can adapt to various types of medical imaging.

- **Computational Demands:** Balancing high diagnostic accuracy with computational efficiency is a persistent challenge, impacting the feasibility of deploying ML/DL models in real-world clinical environments.
- **TL Limitations:** Although TL represents a promising avenue for overcoming dataset limitations, the distinct nature of medical data compared to the data used in pre-training often results in suboptimal model generalisation.
- **Noise and Variability in Imaging:** The presence of noise and the variability, fundamental in lesion appearances, necessitate advanced pre-processing and analysis strategies, complicating the diagnostic process.
- **Explainability and Clinical Trust:** The ambiguous nature of many ML/DL models causes a significant barrier to their clinical adoption, as the medical community prioritises understanding the decision-making process behind diagnostic predictions.

Hybrid DL models, while advancing the frontiers of medical image analysis and other complex applications, introduce certain limitations and challenges:

- **Complex Integration:** designing these advanced models necessitates a seamless integration of diverse ML techniques and architectures, elevating the complexity of development and necessitating more intensive computational power and expert intervention.
- **Data Intensiveness:** Despite their ability to operate on reduced data dimensions, these models still demand substantial, high-quality datasets to train effectively, particularly for the DL components.
- **Reliance on Expert Knowledge:** The efficacy of hybrid models can be significantly improved by incorporating domain expertise. However, acquiring and integrating this expertise remains a challenging task, often limited by availability and the complexities involved in translation to computational models.
- **Operational Complexity:** The implementation of these models can be resource-intensive, posing a potential bottleneck in environments where real-time analysis is paramount.

- **Data Dependency:** The success of DL components within hybrid models is deeply connected to the quality and wide range of the training data. In scenarios with limited or skewed datasets, the model's performance might not align with expectations.
- **Overfitting Risks:** Despite various strategies to mitigate overfitting, the sophisticated nature of DL frameworks within hybrid models can make them more susceptible to this issue compared to their simpler counterparts.
- **Interpretability Challenges:** Complex models, particularly DL-based ones, can obscure the interpretability of the decision-making process, presenting a challenge in settings where understanding and trust in the model's reasoning are crucial.
- **Data Labelling and Quality:**The need for well-labelled and comprehensive datasets is a persistent challenge, directly influencing the training and generalisation capabilities of these models.
- **Balancing Multiple Objectives:** In multi-task hybrid models, achieving the ideal balance between different objectives and avoiding task dominance is a sensitive task, critical for the overall performance of the model.
- **Explainability and Clinical Acceptance:** The 'black box' nature of some components within hybrid models can impact their adoption in clinical practice. Enhancing the interpretability and explainability of these models is essential for bridging the gap between technological advancements and their practical utility in healthcare settings.

Traversing through recent related studies, it becomes evident that DL's influence on these diseases diagnosis and prediction is both profound and promising. However, the literature also underscores the challenges that demand further attention. The increasing reliance on advanced ML and DL methods in detecting and classifying multiple diseases using these different medical images underscores the transformative impact of these technologies in healthcare. These methods not only enhance diagnostic accuracy but also facilitate a more nuanced understanding of disease mechanisms, leading to more effective and personalised treatment strategies. From

using ensemble classifiers and attention mechanisms to incorporating hybrid optimisation algorithms, existing solutions demonstrates rapid advancements in automated disease prediction. Notably, DL and ML approaches appear to outperform classic methods, suggesting their prominence in the field. However, while many studies, such as those presented, achieve high accuracy rates, it's essential to critically assess the broader implications and limitations of these models. First of all, while accuracy is undoubtedly an essential metric, it does not always provide a complete picture of a model's effectiveness. High accuracy rates can sometimes mask model biases, especially if the datasets employed aren't sufficiently diverse. Overfitting on specific datasets might lead to poor real-world performance, especially given the vast diversity of potential patient data. It's also crucial to assess models based on metrics like sensitivity, specificity, and precision, which might offer a more nuanced understanding of the model's true diagnostic capabilities.

Another point of conflict is the reproducibility and scalability of these models. The majority of the discussed models are tested on selected datasets and under specific conditions. There's an underlying assumption that such models, once deployed, will function with the same efficiency in diverse medical settings, which might not always be the case. The real-world medical landscape is saturated with variations in equipment, patient demographics, and image quality. A model that performs exceptionally well in a controlled research setting might underperform when exposed to this variability.

Furthermore, a singular focus on technical metrics might overshadow other essential aspects of medical diagnostics, such as interpretability and clinical relevance. While many models produce high accuracy rates, their lack of interpretability can obstructs their clinical adoption. Medical professionals often require an understanding of how a particular diagnosis was derived, especially when making critical health decisions. A black-box model, despite its accuracy, might remain underutilised if it fails to provide insights into its decision-making process.

Lastly, while the introduction of novel architectures and techniques is commendable, the sustainability and computational feasibility of such models must be examined. Not every medical facility, especially those in resource-limited regions, can afford the computational power

required by some of the advanced DL models. The push for innovation should be balanced with a drive for accessibility, ensuring that these advancements benefit a broad spectrum of healthcare settings.

2.7 Datasets Overview and Selection Rationale

In this research, multiple datasets were selected and utilised across different chapters to evaluate and validate the proposed models and frameworks. The choice of datasets, including MRI, OCT, Fundus, and X-ray images, was driven by the need to cover a broad spectrum of medical imaging modalities that are critical in various diagnostic tasks. Each dataset was carefully selected based on its relevance to the study's objectives, its availability, and its capacity to provide diverse and complex data for training and testing the models. This section provides a detailed description of each dataset, their sizes, and the rationale behind their selection. Additionally, the section will outline how the data was split between training, validation, and testing to ensure robust model evaluation. Table 2.17 summarises the key details of the datasets used in this research.

Table 2.17: Summary of Datasets Used in Research

Dataset	**Imaging Modality**	**Number of Images**	**Used in Chapters**	**Additional Data**	**Rationale for Selection**
BRATS MRI	MRI	8,000	3, 4	Age, Survival days, Resection status	Benchmark in brain tumour analysis; robust testing of model adaptability.

Continued on next page

Table 2.17: Summary of Datasets Used in Research (Continued)

Dataset	**Imaging Modality**	**Number of Images**	**Used in Chapters**	**Additional Data**	**Rationale for Selection**
Retinal	Fundus	1,000	4	Age, Scanning history, Disease stage	Complexity and detailed labelling; critical for retinal disease detection.
Fundus Images	Fundus	18,615	5	Age, Gender, SBP, DBP, Diabetic type, CRT	Balanced dataset with demographic data; ideal for DR analysis.
OCT Scans	OCT	25,197	5	Age, SBP, DBP, Diabetic type, CRT	Key for ophthalmology, particularly DMO; large dataset for robust testing.
X-ray Scans	X-ray	5,467	5	Age, SBP, DBP	Fundamental for pulmonology; critical for testing disease prediction models.

2.7.1 MRI Dataset

The BRATS MRI dataset, obtained as part of the RSNA-ASNR-MICCAI Brain Tumor Segmentation Challenge 2021 (Baid et al., 2021), was selected for its comprehensive imaging data in brain tumour analysis, namely O6-methylguanine-DNA methyltransferase (MGMT) which is significant in determining the treatment response in glioblastoma patients, making accurate image processing critical for reliable analysis and prediction. Despite being simple and un-

labelled, this dataset was chosen due to its extensive use in the medical imaging community, particularly for benchmarking segmentation and classification models in brain tumour studies.

- **Size**: The dataset comprises 2,000 cases, equivalent to 8,000 MRI scans, each available as NIfTii files (.nii.gz) across four different modalities: Native (T1), Post-contrast T1-weighted (T1Gd), T2-weighted (T2), and T2 Fluid Attenuated Inversion Recovery (T2-FLAIR).

- **Usage**: The BRATS dataset was utilised in Chapters 3 and 4 to test and validate the proposed DenCeption model and the feature extraction framework. The inclusion of patient-specific information such as age, survival days, and resection status further enhanced the evaluation of the proposed models.

- **Rationale**: The BRATS dataset was selected due to its challenging nature in brain tumour segmentation and its widespread acceptance as a benchmark in the field. Its simplicity, despite being unlabelled, allowed for testing the robustness of the proposed models under varied clinical imaging protocols.

2.7.2 Retinal Dataset

The Retinal dataset, consisting of 1000 Fundus scans, was chosen for its complexity and the detailed labelling provided by medical experts (Linchundan, 2019). This dataset includes both normal and abnormal Fundus images, offering a rich source of data for evaluating models designed for retinal disease detection, particularly DR disease.

- **Size**: The dataset contains 1000 Fundus coloured scans available as .jpeg files, categorised into normal and abnormal (presence of DR) scans.

- **Usage**: This dataset was primarily used in Chapter 4 to validate the proposed feature extraction framework, focusing on its ability to distinguish between different stages of retinal diseases.

- **Rationale**: The Retinal dataset was selected for its relevance to ophthalmology, a field where accurate image analysis is crucial for early detection and management of diseases such as DR. The dataset's complexity and the additional parameters (patient age, scanning history, and disease stage) made it ideal for testing the scalability and reliability of the proposed frameworks.

2.7.3 Fundus Dataset

The Fundus dataset, obtained from the DR Kaggle competition, provides a balanced set of images for normal and DR-affected eyes (Tanlikesmath, 2019). It is particularly valuable for studying the impact of cardiovascular health and diabetic conditions on retinal health.

- **Size**: The dataset includes 18,615 Fundus images, evenly split between normal and DR-affected cases.

- **Usage**: This dataset was utilised in Chapter 5 to explore the integration of demographic and physiological features into the predictive models, enhancing the overall accuracy and interpretability of the results.

- **Rationale:** The Fundus dataset was chosen for its comprehensive representation of DR cases across a wide age range and its inclusion of relevant physiological data. This made it an excellent candidate for developing models that integrate visual data with patient-specific health information.

2.7.4 OCT Dataset

The OCT dataset, also from Kaggle, focuses on DMO and provides a substantial number of scans to test the robustness of the proposed models in detecting retinal diseases (Mooney, 2018b).

- **Size**: The dataset contains 25,197 OCT scans, with 11,599 scans showing DMO and 13,598 normal scans.

- **Usage**: In Chapter 5, this dataset was critical for testing the proposed predictive models, particularly in the context of ophthalmology where OCT scans are a key diagnostic tool.
- **Rationale:**OCT is a critical imaging modality in ophthalmology, particularly for conditions like DMO. The dataset's size and focus on a specific retinal condition made it an ideal choice for evaluating the proposed models in a real-world scenario.

2.7.5 X-ray Dataset

The X-ray dataset, focusing on pulmonary conditions such as pneumonia, was selected to extend the scope of the research into pulmonology (Mooney, 2018a). This dataset allowed for the testing of models designed to detect anomalies in chest X-rays, a common and critical diagnostic tool in medicine.

- **Size**: The dataset includes 5,467 X-ray scans, with 3,883 scans showing pneumonia and 1,584 normal scans.
- **Usage**: This dataset was used in Chapter 5 to evaluate the proposed predictive frameworks' effectiveness in detecting pulmonary diseases, integrating additional health data to improve prediction accuracy.
- **Rationale**: X-ray imaging is a fundamental diagnostic tool for pulmonary conditions. The dataset's focus on pneumonia, combined with its demographic and health information, provided a robust platform for testing the generalisability and accuracy of the proposed models.

2.7.6 Dataset Usage for Training, Testing, and Validation

For each dataset, the data was split into training, testing, and validation sets to evaluate the models effectively. Specifically, 60% of the data was used for training, 30% for testing, and 10% for validation. This split ensured that the models were trained on a sufficient amount of data while still providing a robust evaluation on unseen data, with an additional validation set to

fine-tune model parameters. The diverse nature of the datasets, along with the varied imaging modalities, allowed for comprehensive testing of the models across different medical fields, ensuring their adaptability and generalisability.

2.8 Datasets Selection Strategy

The decision to use a variety of datasets across different diseases and imaging modalities was intentional and strategic. The overarching goal of this research is to develop robust, adaptable, and generalisable models capable of performing well across a range of medical imaging tasks, rather than being tailored to a single disease or imaging modality.

2.8.1 Generalisation and Robustness

By using multiple datasets, each representing different diseases (e.g., brain tumours, retinal diseases, pulmonary conditions), the research aimed to test the adaptability and scalability of the proposed models. Focusing on a single disease dataset might limit the scope of the research and the applicability of the models. The use of diverse datasets ensures that the models are not overfitted to a specific disease or imaging modality but can generalise well across different medical conditions.

2.8.2 Comprehensive Evaluation

The research sought to address a broad spectrum of challenges in medical imaging, from segmentation and classification to disease prediction, across various medical fields such as neurology, ophthalmology, and pulmonology. The inclusion of datasets like BRATS MRI, Retinal, Fundus, OCT, and X-ray scans allowed for a comprehensive evaluation of the models under different scenarios, thus demonstrating their versatility and effectiveness in real-world applications.

2.8.3 Enhancing Model Capabilities

Each dataset contributes unique characteristics that challenge the models in different ways. For example, MRI scans present challenges in terms of 3D data processing, while Fundus, OCT, and X-ray scans require detailed analysis of retinal and lung structures respectively. By testing the models on these varied datasets, the research was able to refine and enhance the models' capabilities, ensuring they are equipped to handle the diverse demands of medical diagnostics.

2.8.4 Applicability in Multi-Disease Diagnosis

In clinical settings, it is common for patients to present with multiple conditions that may require different imaging modalities for diagnosis. Developing models that can process and analyse different types of medical images ensures their applicability in multi-disease diagnosis, which is increasingly relevant in modern healthcare.

2.9 Cross-Validation Method

A comprehensive 10-fold cross-validation strategy was implemented across all datasets used in this research. This approach ensures that each dataset is trained, tested, and validated in a consistent and robust manner, thereby enhancing the reliability and generalisability of the proposed models. The following sections outline the cross-validation process applied in each chapter and across all datasets, along with the rationale behind this approach.

2.9.1 10-Fold Cross-Validation Overview

10-fold cross-validation is a well-established method in ML for assessing model performance. In this approach, the dataset is randomly partitioned into 10 equal-sized subsamples. Of the 10 subsamples, a single subsample is retained as the validation data for testing the model, while the remaining 9 subsamples are used as training data. This process is repeated 10 times, with each of the 10 subsamples used exactly once as the validation data. The results from the 10

folds are then averaged to produce a single estimation of the model's performance.

2.9.2 Application of 10-Fold Cross-Validation in Each Chapter

Chapter 3: BRATS MRI Dataset

In Chapter 3, where the BRATS MRI dataset is used to train and validate the proposed DenCeption model, 10-fold cross-validation was applied to ensure the model's robustness in handling complex medical imaging data. Given the nature of the BRATS dataset, which includes 8,000 MRI scans across different modalities (T1, T1Gd, T2, T2-FLAIR), this method allowed for a thorough evaluation of the model's performance in tumour detection and segmentation tasks. Each fold provided insights into how well the model generalises across different subsets of the data, ensuring that the results are not biased by a particular data partition.

Chapter 4: Retinal and BRATS MRI Datasets

For Chapter 4, where both the Retinal dataset and BRATS MRI dataset are used, the same 10-fold cross-validation strategy was employed. The Retinal dataset, consisting of over 1000 Fundus images, was subjected to this rigorous validation process to assess the accuracy and consistency of the feature extraction framework. This dataset, being more complex and labelled, required careful cross-validation to ensure that the model performs well across varying image types and patient conditions. The repeated validation over 10 folds provided a reliable measure of the model's ability to classify normal versus abnormal retinal conditions, such as DR.

Chapter 5: Fundus, OCT, and X-ray Datasets

In Chapter 5, where multiple datasets are used, including Fundus, OCT, and X-ray scans, 10-fold cross-validation was applied consistently across all datasets. This was crucial to ensure that the predictive models, including the proposed HyBoost model, were thoroughly evaluated across different imaging modalities and patient demographics. For the Fundus dataset (18,615

images), OCT dataset (25,197 scans), and X-ray dataset (5,467 scans), the cross-validation process provided a comprehensive assessment of the model's adaptability and accuracy in predicting various conditions like DMO and pneumonia. By maintaining consistency in the validation process across different datasets, this ensures that the comparisons made between the models' performances were fair and reliable.

2.9.3 Consistency and Clarity in Cross-Validation

The application of 10-fold cross-validation across all datasets and chapters was critical for maintaining consistency and clarity in the evaluation of the proposed models. This approach not only validated the models' performance on unseen data but also ensured that the models were not overfitting to any particular dataset. The results from the cross-validation provided a robust measure of each model's ability to generalise across different types of medical images and conditions, thereby confirming the models' applicability in diverse clinical scenarios.

2.10 The Testbed

To process and evaluate the proposed models, University of Gloucestershire Research AI Server is used. The operating system setup is based on the Ubuntu Linux distribution with its latest Long Term Support release. NVIDIA and CUDA drivers have been installed to utilise the GPUs available in the system. The latest versions available for the RTX 2080Ti series cards have been used. The software environment considered is Python and MATLAB, in addition to the servers' tools. The programming environment used to build the models consists of Python programming language with TensorFlow as the CNN modelling framework. The latest release of Anaconda distribution of Python with all its supporting packages have been used. The only updates to the base Anaconda distribution consist of the TensorFlow and the OpenCV image processing libraries. MATLAB installation is required to provide a runtime environment for some of the tools in development. In addition to the tools required to build the models, two packages have been provided that are required in order to serve the models as tools for the

use by the server. Docker and K8S allow for containerisation of software, allowing tools and applications to run without a view or access to any components of the system. Table 2.18 summarises the considered testbed. This setup will be applicable for all conducted experiments in this work.

Table 2.18: System Specifications, Operating System, and Drivers

	Specifications	**Operating System**	**Drivers**
Hardware	1x ASUS ESC8000 2x Intel Xeon Gold 5218 16x 32GB DDR4 2666Mhz ECC RDIMM (512GB Total) 8x GPU RTX 2080TI 11GB Blower 1x 2TB Intel 760p M.2 PCIe NVME 8x 2TB Seagate Exos 7E2000 ST2000NX0433 10GbE SFP+ Network Adapter 1x PIKE II 3108-8i-16PD/1G	Ubuntu 18.04.3 LTS (kernel: 5.0.0-32-generic x86_64)	Nvidia Graphics Drivers 430.50 CUDA 10.1.
	Python environment	**MATLAB 2019b**	**Servers tools**

Continued on next page

Table 2.18: System Specifications, Operating System, and Drivers (Continued)

	Specifications	**Operating System**	**Drivers**
Software	Anaconda 2019.03 Conda 4.7.10 Python 3.7.4 Tensorflow 2.0.0 Opencv 4.1.1.26 keras (version 2.31) tensorflow-gpu (v1.2.1) matplotlib (v3.1.1) opencv (v4.1.2) flask (v1.2.1) scikit-learn (v0.21.3) scipy (v1.3.1) numpy (v1.17.3) h5py (v.2.10.0) pytorch (v1.3.0)	Deep learning toolbox Parallel computation toolbox Computer Vision toolbox Signal Processing toolbox Statistics and Machine Learning toolbox	Docker 19.03.2 Microk8s v1.16.2

2.11 Conclusion

Following an extensive exploration of traditional and automated methodologies in medical image processing, this literature review underscores the gradual shift from conventional techniques to more sophisticated, learning-based methods. By delving into the impact of traditional methods on image processing performance and the significant advancements brought about by automated and hybrid models, this thesis has illuminated the evolving landscape of medical image analysis. Notably, the comparative analysis of novel DL approaches and the exploration of hybrid models versus singular strategies highlight a critical path toward enhancing medical

diagnostics. Identifying gaps in reliability, efficiency, interpretability, scalability, and adaptability, this research is positioned to contribute significantly to the field. The forthcoming chapters are thoroughly designed to bridge these gaps. The first chapter will introduce a novel hybrid DL model, marking a significant advancement in medical image processing. Subsequent chapters will disclose an adaptive and scalable feature extraction framework and prediction method, designed to address the nuances of different types of medical imaging data, including MRI, Fundus, OCT, and X-ray images. This comprehensive approach not only endorses the integration of cutting-edge computational techniques but also aims to significantly enhance disease prediction performance. Through this holistic methodology, the thesis stands as an innovative effort to navigate the complexities of medical image analysis, setting a new benchmark for future research in the domain. In the following chapter, a design and implementation of a novel hybrid DL based model will be presented.

Chapter 3

DenCeption: A Novel Hybrid Deep Learning Based Model

3.1 Introduction

The domain of AI has witnessed significant growth, particularly in enhancing and innovating systems and technologies. Among its branches, DL is distinguished for its exceptional capabilities in processing and analysing images, showcasing significant advancements in the field of medical imaging. This area is crucial for advancing various applications related to medical diagnosis, treatment planning, and prognostic evaluations. CNNs are at the forefront of DL techniques employed for medical image analysis, with various architectures demonstrating considerable success in this domain. This chapter conducts a detailed examination of selected CNN architectures, including wider and deeper networks which have historically delivered impressive results. Addressing the need for an innovative approach, this work introduces a novel hybrid architecture, DenCeption, which merges the strengths of DenseNet-169 and Inception-V4. This combination aims to leverage the unique advantages of these networks to further enhance medical image processing.

CNNs have been applied to enhance diagnostic accuracy and operational efficiency in medical procedures, such as targeting tumour/cancer detection through analysis tools. Several re-

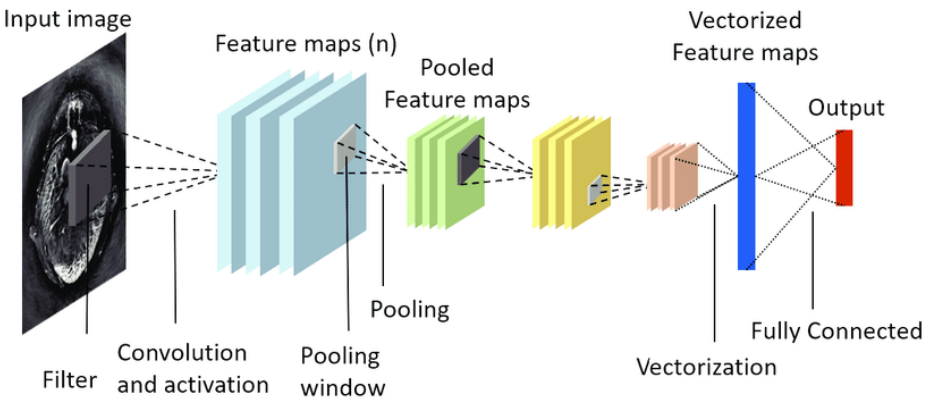


Figure 3.1: Typical CNN Architecture for Medical Image Analysis (Yang, Lan, Gao, and Gao, 2020)

searches explored the application of CNNs across various diagnostic images such as MRI, CT, aiming to significantly reduce the time consumed during image processing phases (Arena, Basile, Bucolo, and Fortuna, 2003; Mishra and Rahul, 2021). These researches focused on the fast and efficient analysis of these images to assist in identifying inherited mutations disposing individuals to various diseases and improving real-time processing capabilities in the health-care field. Through detailed studies, including the development of three-dimensional CNN models via simulation and their application in radiosurgery, as an example, several researches demonstrated the potential of CNNs to speed up image processing operations crucially (Arena et al., 2003; Zhao, Li, Rahaman, and Xu, 2022). Their applications highlight their ability to efficiently detect diseases, such as quickly isolate tumour regions, reconstruct cerebral ventricles for surgical planning, ...etc. The findings suggest a promising future for CNNs in advancing medical diagnosis and treatment (Bankman, 2008; Yang et al., 2021). Figure 3.1 shows a typical CNN architecture for medical image analysis.

CNNs, with their ability to automatically learn features from a large dataset of images, offer a more effective and rapid solution for medical image analysis. This includes tasks like segmentation, classification, and disease detection, which are crucial for accurate diagnosis and treatment planning. CNNs are well known for their high performance in various real-world applications beyond medical imaging, like speech and text recognition, due to their deep architectural design that mimics the human brain's neuron connectivity (Alzubaidi et al., 2021; Razzak,

Naz, and Zaib, 2018). This architecture allows CNNs to learn multiple levels of abstraction and representation from data, making them particularly suited for complex image analysis tasks in the healthcare sector. The outstanding performance showed by CNNs has led to their wide use in case of complex applications such as computer vision, segmentation, object-detection, video processing, natural language processing, speech recognition, and other several tasks particularly medical image processing. Their multilayer neural network architectures are designed to recognise visual patterns from pixels composing the image by minimising its pre- processing. There are four main categories composing deep CNNs to include: spatial exploitation (e.g., AlexNet), depth (e.g., Inception-V4), and multi-connection based CNNs (e.g., DenseNet and ResNet) underscoring their significance in expanding the limits of their potential contributions in medical image processing and analysis (Khan et al., 2020; Girdhar, Sinha, and Gupta, 2023).

The contribution of this work is to design and implement a hybrid DL model that combines the advantages of different CNN based architectures to enhance the existing state-of-the-art robustness and accuracy in image processing, with a particular focus on medical imaging. The chapter is composed as follows:

- Section 2 provides a discussion about the integration of these CNNs, with particular focus on DenseNet, ResNet, and Inception family, in hybrid architectures. The section also delves into a critical discussion about related works.
- Section 3 summarises the identified gaps and challenges in the literature.
- Section 4 focuses on a detailed presentation of the proposed hybrid DL model, DenCeption. This section will serve as the demonstration of the different hybrid blocks composing the proposed novel model.
- Section 5 presents the rationale behind the selection of the proposed DenCeption model.
- Section 6 covers the conducted experiments followed by an introduction to the used dataset in Section 7.

- Section 8 presents the outcomes of the performed experiments and a critical discussion of the obtained results at the testing and validation stages. The critical comparison will involve DenCeption variants as well as benchmarking methods.
- The chapter concludes with Section 9.

3.2 Related Works: Deeper and Wider Hybrid CNN Architectures

Given the critical evaluation performed on these different well known CNN architectures, the scope of this work is to focus on the outperforming models with a compromise of accuracy and computational demands. Therefore, DenseNet, ResNet, and Inception family networks will be the primary emphasis in this chapter. ResNet and DenseNet networks are characterised by a deeper architecture considering the number of: connections, layers, nodes, and parameters which leads to a more complicated networks (Pleiss et al., 2017). Unlike ResNet and DenseNet that stack convolutional blocks together to get a deeper network and a better performance, Inception family present an engineered architecture characterised by its complexity in obtaining higher speed and better accuracy (Emara, Afify, Ismail, and Hassanien, 2019). Tables 3.1, 3.2, and 3.3 present a critical review of the different network's versions of ResNet, DenseNet, and Inception family.

Table 3.1: ResNet Network: Advantages and Disadvantages (Xu, Fu, and Zhu, 2023)

ResNet Version	**Advantages**	**Disadvantages**
ResNet-18	• Fast processing and training due to fewer layers. • Suitable for applications with real-time requirements and less complex tasks. • Lower computational and memory requirements make it ideal for real-world deployment.	• Lower capacity may result in reduced accuracy on highly complex datasets compared to deeper versions.
ResNet-34	• Better accuracy than ResNet-18 for more complex tasks. • Maintains a balance between computational efficiency and model capacity. • Better processing and training time performance.	• Underperforms in case of complex datasets and tasks compared to deeper ResNet versions.
ResNet-50	• Uses bottleneck blocks for improved efficiency and depth. • Higher capacity and accuracy on complex tasks compared to ResNet-18 and ResNet-34. • Ensures a good balance of performance and computational demand.	• Higher computational requirements than ResNet-18 and ResNet-34. • May require more critical tuning and regularisation to avoid overfitting.

Continued on next page

Table 3.1: ResNet Network: Advantages and Disadvantages (Continued)

ResNet Version	**Advantages**	**Disadvantages**
ResNet-101	• Deeper architecture leads to improved accuracy and feature extraction capabilities. • Suitable for highly complex datasets and tasks. • Achieves better accuracy.	• Significantly increased computational and memory requirements compared to shallower versions. • Longer training and processing times, making it less suitable for real-time applications.
ResNet-152	• The deepest standard ResNet model, offering the highest capacity and potential accuracy. • Demonstrates state-of-the-art performance on various benchmarks and complex tasks.	• Very high computational cost and memory usage, challenging to deploy on resource-constrained platforms. • Diminishing performance improvements relative to the increase in depth and computational resources.

Table 3.2: DenseNet Network: Advantages and Disadvantages (Huang et al., 2017b)

DenseNet Version	**Advantages**	**Disadvantages**
DenseNet-121	• Offers a good balance between model complexity and computational efficiency, making it adaptable. • Suitable for environments with moderate computational resources.	• While efficient, it may not capture as complex features as its deeper counterparts, which might limit its performance on more complex tasks.
DenseNet-169	• Improved accuracy on complex datasets. • Maintains an equilibrium between network depth and required computations.	• Increased depth is resource-intensive compared to DenseNet-121.
DenseNet-201	• High capacity offering better performance on challenging visual recognition tasks. • Aims to maximise accuracy for demanding datasets.	• Very resource-intensive.
DenseNet-264	• Deepest version of DenseNet. • Best suited for benchmarking and research applications where performance is paramount.	• Highly demanding in computational resources. • Less practical for deployment in limited-resource scenarios. • Training and processing times are considerably longer due to its size.

Table 3.3: Inception Networks Family: Advantages and Disadvantages (Szegedy et al., 2017)

Inception Version	**Advantages**	**Disadvantages**
Inception-V1 (GoogLeNet)	• Significantly increases network depth and width without a substantial increase in computational cost. • Utilises 1x1 convolutions to reduce dimensionality and computational cost.	• Less efficient in handling high-resolution inputs compared to later versions. • Complexity of its architecture makes it harder to understand and modify.
Inception-V2	• Introduced BN. • Improved training speed and model performance. • Made architectural improvements to inception modules for increased efficiency.	• Still faced challenges with model complexity and understanding the interactions between different layers and blocks.
Inception-V3	• Further refined the inception modules with factorised convolutions to reduce the number of parameters. • Introduced RMSProp optimiser, label smoothing, and updated factorisation ideas for convolution. • Achieved significantly higher accuracy on ImageNet and other benchmarks.	• Computationally intensive. • The complexity of the architecture increased, requiring more resources for training.

Continued on next page

Table 3.3: Inception Networks Family: Advantages and Disadvantages (Continued)

Inception Version	**Advantages**	**Disadvantages**
Inception-V4	• Combined the strengths of Inception architecture with residual connections. • Significantly improved training speed and accuracy. • Adaptable to be integrated in hybrid models.	• High computational and memory requirements.

The significant contribution these architectures have offered to the image processing field has led to the high interest in combining their advantages towards getting better performance as well as less computational costs as a common drawback. Several researches have introduced hybrid combinations of these architectures. In fact, authors in (Yasashvini, Panjanathan, Graceline, and Jani Anbarasi, 2022) proposed a hybrid CNN architecture with ResNet and DenseNet to improve DR classification. These hybrids aim to leverage DL and TL for more accurate disease stage analysis. The advantages include enhanced feature extraction and classification, utilising large unlabelled datasets, and improved accuracy. The hybrid CNN with DenseNet has achieved the highest Acc of 96.22%. The disadvantages involve substantial computational resources for training as well as the complexity of integrating and optimising hybrid models.

Authors in (Alotaibi and Alotaibi, 2020) proposed a hybrid CNN architecture combining ResNet-152 and Inception-V1 models for Hyperspectral Image Classification (HSI), addressing the challenge of high feature dimensions with limited labelled samples in HSI classification. This hybrid architecture achieved notable accuracies: 95.31% on the Pavia University dataset, 99.02% on the Pavia Centre scenes dataset, 95.33% on the Salinas dataset, and 90.57% on the Indian Pines dataset. The advantage of this approach is the improved HSI classification performance by combining the strengths of both ResNet-152 and Inception-V1 architecture.

However, a potential disadvantage could involve increased model complexity and computational requirements due to the combination the high complexity of ResNet-152.

Another hybrid model has been proposed in (Zhang and Feng, 2019) using both Inception-V4 and DenseNet-121. Authors introduced their Inception-DenseNet architecture that embeds Inception-like blocks within a DenseNet framework. This architecture incorporates hybrid activation operations differing from previous Inception blocks, aiming for more flexible responses to object semantic regions. The Inception-DenseNet architecture demonstrates competitive or superior classification Acc with average of 88% on various image datasets compared to existing models like DenseNets and ResNets, with fewer trainable parameters. The key advantages of the proposed hybrid model include: (1) a reduction in the number of trainable parameters compared to original DenseNet, (2) an efficient training process due to dense connections, and (3) an enhanced nonlinear mapping and feature diversity through hybrid activation functions and multi-branch structure. In the following section, a proposal of a novel hybrid DL architecture will be introduced highlighting the hybrid structure of the resulted network.

3.3 Identified Challenges

Despite DL's profound capabilities, its models face several inherent challenges that can restrict their efficiency and effectiveness. Understanding these challenges is crucial for the development of more advanced and efficient models. The challenges faced by current DL models covered in the discussed literature include:

- **Size and Complexity:** Large, computationally intensive models requiring significant GPU resources compared to traditional vision models.
- **Potential Overfitting:** Vulnerability to overfitting despite dropout techniques, emphasising the need for rigorous regularisation and data augmentation.
- **Static Architecture:** Limited adaptability due to fixed structures, lacking the flexibility of modular designs.

- **Skip Connection Challenges:** Complexity in managing input dimensions and parameter efficiency across different skip connection strategies.
- **Detection of Salient Objects:** Difficulty in detecting objects with large size variation.
- **Deep Network Challenges:** Overfitting and high computational costs associated with very deep networks.
- **Memory Consumption and Computational Overhead:** Increased memory usage and computational complexity, particularly in models with dense connections.

Addressing these challenges is pivotal for advancing the field of DL in computer vision, necessitating innovative solutions that enhance model adaptability, efficiency, and effectiveness.

3.4 Proposed Hybrid Deep Learning Model: DenCeption

This work proposes an adaptive hybrid CNN based model that combines DenseNet-169 and Inception-V4. The motivation of this proposal is the idea of combining the advantages of DenseNet-169 and Inception-V4 (as per Tables 3.2 and 3.3) to result DenCeption. The particular focus of DenCeption is on reducing the complex medical image size during processing by keeping relevant features. The contribution of the proposed model is to potentially reduce the number of DHF features, hence reducing the number of trainable parameters. The major uniqueness of the DenCeption is the modification of both the original dense block and transition block of DenseNet-169.

The idea is to construct a new dense block (DB), namely hybrid dense block (HDB) composed of convolutional and inception modules. This will show the effect of the concatenation operation of each convolution on the output of inception modules. The new densely connections will be translated by the dense connectivity between all inception modules within the HDBs by conserving the initial dense connections between convolutional blocks (CBs). Towards minimising the size of the medical images while being processed, an integration of reduction A (RA) and reduction B (RB) blocks into the transition block (TB) takes place, constructing the

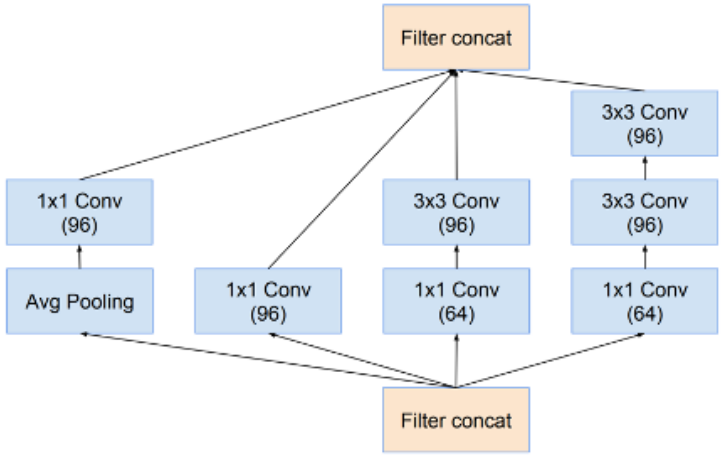


Figure 3.2: Original Inception-A module (Szegedy, Ioffe, Vanhoucke, and Alemi, 2017).

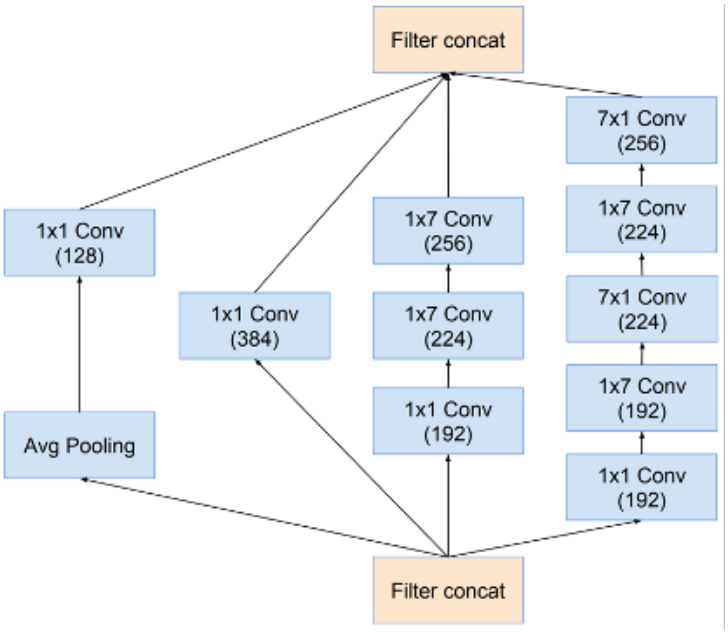


Figure 3.3: Original Inception-B module (Szegedy, Ioffe, Vanhoucke, and Alemi, 2017).

proposed hybrid transition block (HTB). As a result, reduction modules (RA) and (RB) will be densely linked to inception modules A (InA) and B (InB), respectively. A single Inception C (InC) module will be part of HTBs as well. Figures 3.2, 3.3, 3.4, 3.5 and 3.6 present the structures of the original InA, InB, InC, RA, and RB modules respectively (Szegedy et al., 2017). Figure 3.7 illustrates the proposed hybrid network.

DenCeption network consists of a convolutional layer linked to a max pooling layer, followed by alternation of four HDB blocks and three HTB blocks. Each HDB produces a set of features resulted from highly dense CB blocks linking different internal components including

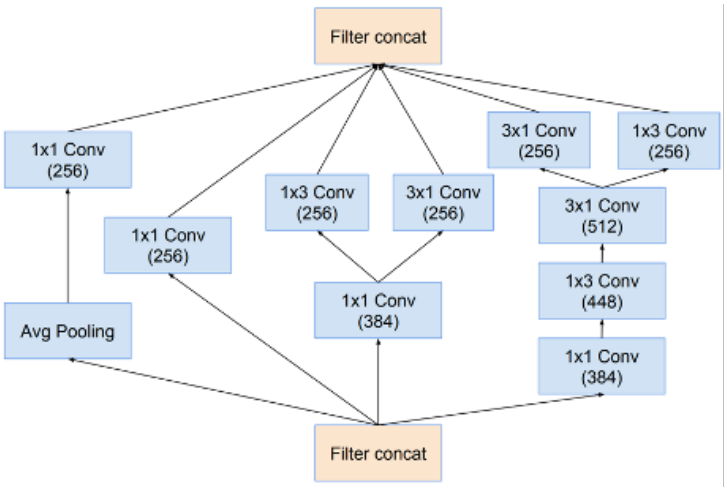


Figure 3.4: Original Inception-C module (Szegedy, Ioffe, Vanhoucke, and Alemi, 2017).

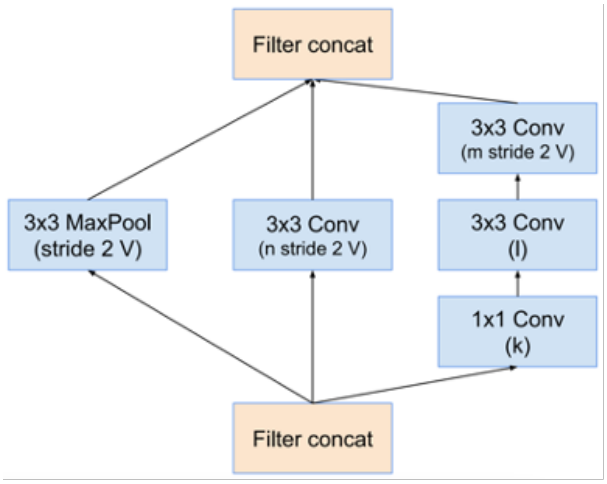


Figure 3.5: Original Reduction A module (Szegedy, Ioffe, Vanhoucke, and Alemi, 2017).

BN layer, Convolutional layer, ReLU activation function, inception modules to include InA and InB. Figure 3.8 illustrates the HDB composition.

The alternative integration of InA and InB increase the dense connections opting to minimise the total channels by that it means reducing the sample representation and used for key-points rearrangements along with 3*3 convolutional layer. The final outcome of $(n+1)^{th}$ HDB is $N*N*M*(1+\alpha)$, where N is the size of the of the feature map (FM) and M is the total number of channels considered, and $\alpha < 2$. Each of the proposed HDBs applies different filtering process in the inception modules InA and InB which is difference from the original inception modules as showed in Figure 3.9 and 3.10.

The hybrid InA, InB modules are presented in Figure 3.9 and 3.10 where n, l, k, t, and j are

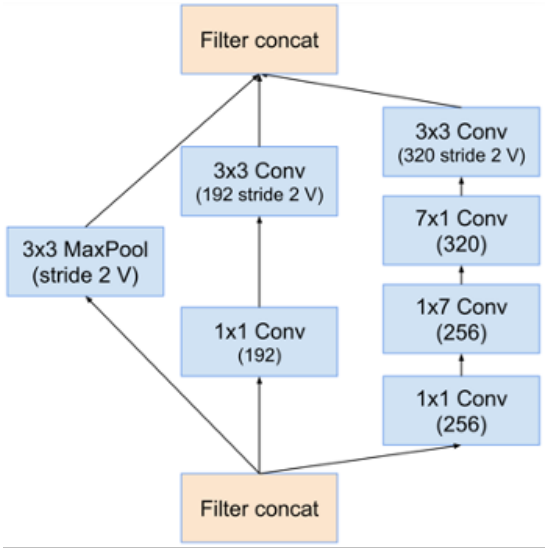


Figure 3.6: Original Reduction B module (Szegedy, Ioffe, Vanhoucke, and Alemi, 2017).

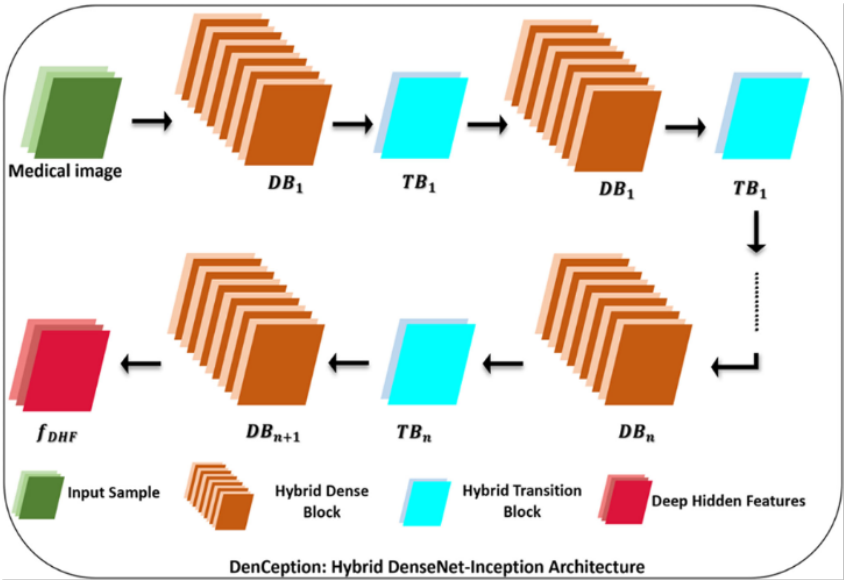


Figure 3.7: Proposed Hybrid DenCeption Architecture

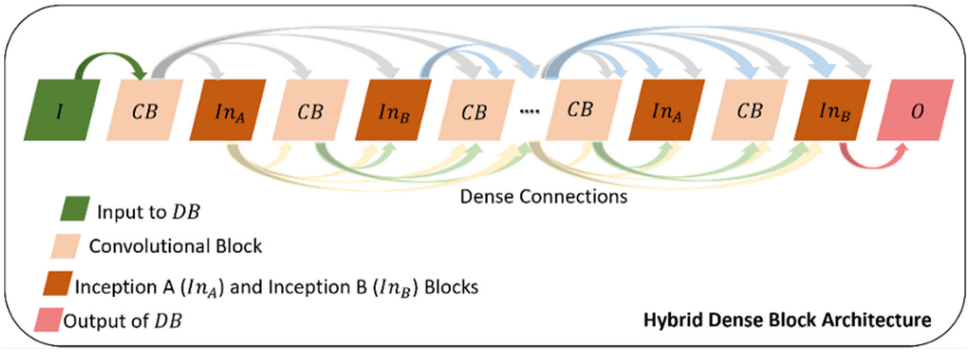


Figure 3.8: Hybrid Dense Block Architecture

the number of filters defined in Tables 3.4 and 3.5.

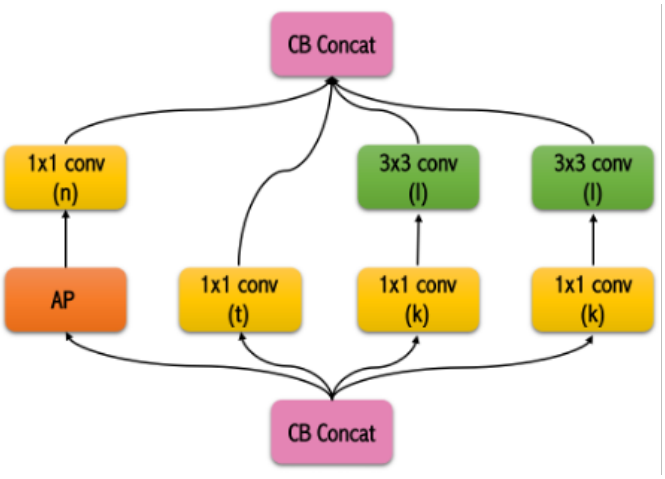


Figure 3.9: DenCeption InA Module

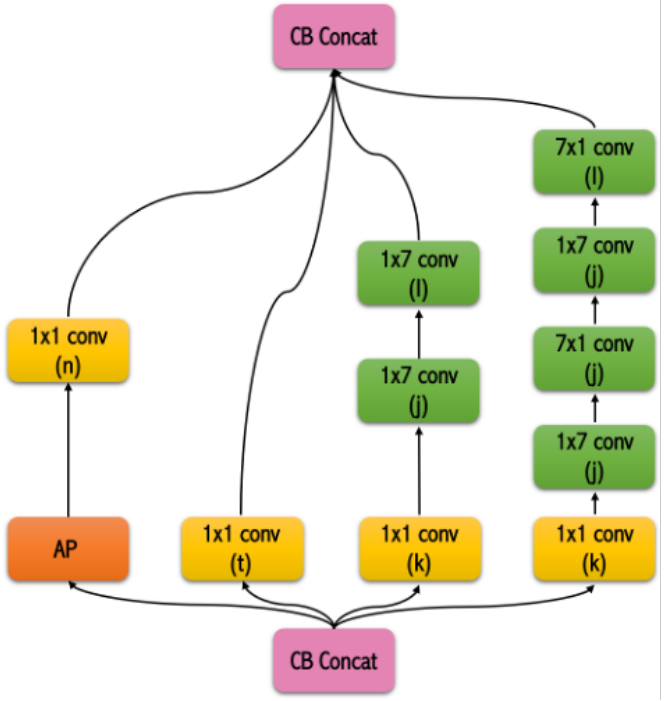


Figure 3.10: DenCeption InB Module

Table 3.4: InA Filter Numbers Per HDB Block

HDB Block	**n**	**l**	**t**	**k**
1	24	48	8	24

Continued on next page

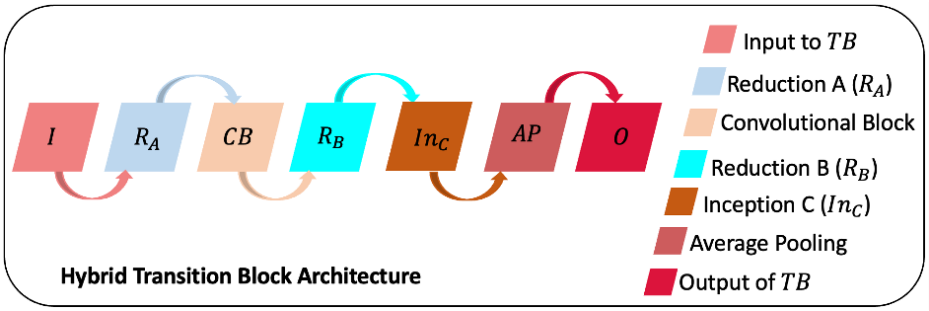


Figure 3.11: Hybrid Transition Block Architecture

Table 3.4: InA Filter Numbers Per HDB Block (Continued)

HDB Block	**n**	**l**	**t**	**k**
2	128	128	96	64
3	256	256	64	128
4	256	256	256	128

Table 3.5: InB Filter Numbers Per HDB Block

HDB Block	**n**	**l**	**t**	**k**	**j**
1	24	48	8	24	48
2	128	128	96	64	96
3	256	256	64	128	192
4	256	256	256	128	192

By increasing the dense connections composing the HDB, the FMs extracted at each n^{th} layer shows an increase as well. Therefore, the HTB is taking over the outcome and reduces the extracted FMs dimension from the

$$(n1)^{th}$$

HDB. Figure 3.11 presents the composition of the HTB block.

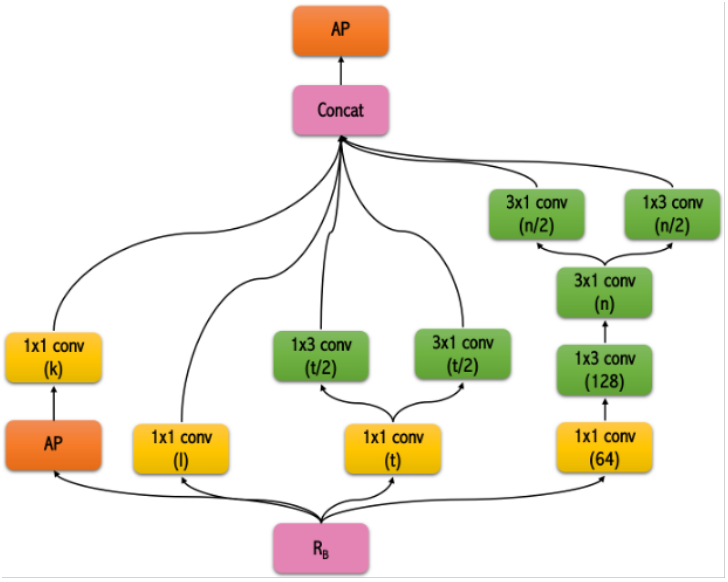


Figure 3.12: DenCeption InC Module

The presence of reduction blocks, inherited from Inception architecture, including RA and RB had a fundamental value in: (1) improving the FMs representation, (2) reducing the FMs dimension, and (3) emphasising on keeping the regions of interest's DHF features (loss rate is very low). The key value of integrating Inception C (InC) module is that it plays a crucial role in the proposed architecture by introducing specific operations and convolutions adapted to address critical aspects of feature representation. Its operation through the HTB block contributes to the overall performance of the hybrid model by providing additional flexibility in feature extraction and processing as well as maintaining relevant information that might be lost due to transition blocks. InC module represents the link between the resulted output of (RB) modules to the average pooling operation block. The RA module applies a stride equal to 2, followed by a 1x1 CB block, linked to RB which applies a stride equal to 1 and increases the number of filters towards enhancing the deep extraction of FMs. Afterwards InC module applies an increase number of filters to help balance the size of the output as well as the number of channels per operation followed by an average pooling layer which further applies a stride equal to 2 and maintains a constant number of filters as resulted from InC connection. Figures 3.12, 3.13, and 3.14 present the hybrid InC, RA, and RB modules where n, t, l, k, and m are the number of filters defined in Tables 3.6, 3.7, and 3.8.

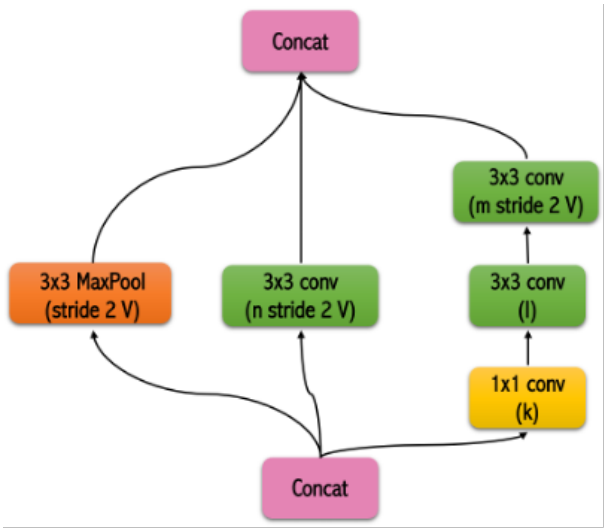


Figure 3.13: DenCeption RA Module

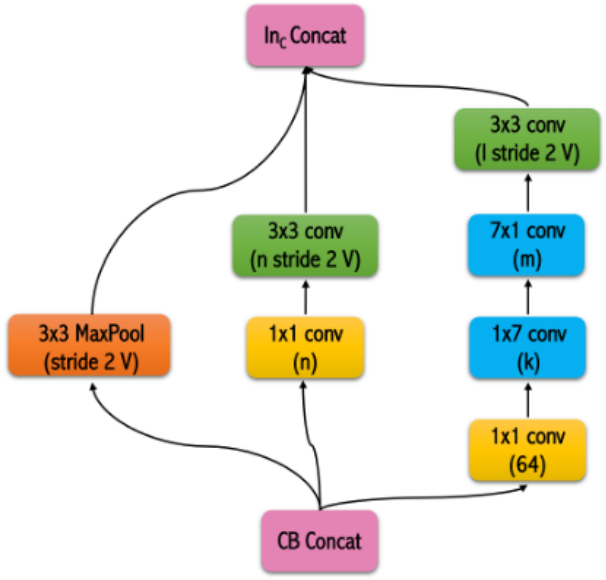


Figure 3.14: DenCeption RB Module

Table 3.6: InC Filter Numbers Per HTB Block

HTB Block	**n**	**t**	**l**	**k**
1	256	192	8	24
2	512	256	32	32

Continued on next page

Table 3.6: continued: InC Filter Numbers Per HTB Block (Continued)

HTB Block	**n**	**t**	**l**	**k**
3	512	256	128	128

Table 3.7: RA Filter Numbers Per HTB Block

HTB Block	**n**	**l**	**m**	**k**
1	64	48	64	24
2	224	128	256	64
3	320	256	512	128

Table 3.8: RB Filter Numbers Per HTB Block

HTB Block	**n**	**l**	**m**	**k**
1	128	128	128	96
2	256	256	256	128
3	320	512	512	256

3.5 Rationale Behind the Selection of the DenCeption Model

The selection of the DenCeption model as the proposed hybrid DL architecture was driven by a systematic and methodical approach, aimed at addressing specific challenges in medical image processing while leveraging the strengths of existing CNN architectures. The decision to combine DenseNet-169 and Inception-V4 into the DenCeption model was carefully considered; it was based on a thorough analysis of the advantages and limitations of various DL architectures, with the goal of optimising performance for complex medical image analysis tasks.

3.5.1 Addressing Identified Challenges

The primary motivation for developing the DenCeption model arose from the limitations observed in both traditional and existing DL models. Traditional CNNs, while effective in certain scenarios, often struggle with maintaining a balance between model complexity, computational efficiency, and accuracy, particularly in the context of high-dimensional medical imaging data. DenseNet and Inception architectures, each with their own unique strengths, were identified as potential solutions to these challenges. DenseNet-169's ability to enhance feature reuse and reduce the number of parameters, coupled with Inception-V4's multi-scale feature extraction and enhanced computational efficiency, made them perfect choices for integration.

3.5.2 Selection Criteria for Hybridisation

The hybridisation of DenseNet-169 and Inception-V4 into DenCeption was guided by several key criteria:

- **Feature Extraction and Efficiency**: DenseNet's densely connected layers facilitate efficient feature reuse, which is crucial for reducing the computational burden without compromising accuracy. Inception-V4, on the other hand, excels in extracting multi-scale features, which is particularly beneficial for medical images that often contain intricate details at various scales.

- **Model Adaptability**: The need for a model that can adapt to varying complexities in medical images was paramount. By combining the strengths of DenseNet-169 and Inception-V4, DenCeption was designed to be adaptable, allowing it to handle a wide range of medical imaging tasks with varying levels of difficulty.

- **Computational Constraints**: The selection process also considered the computational resources typically available in medical settings. While both DenseNet and Inception architectures are known for their computational demands, the hybrid model was optimised to strike a balance between performance and resource efficiency, ensuring its feasibility for practical deployment.

3.5.3 Comparative Analysis and Experimental Validation

Prior to finalising the DenCeption architecture, a comparative analysis was conducted, involving extensive experimentation with different CNN architectures. These experiments focused on evaluating the performance of various combinations of DenseNet and inception modules in terms of accuracy, computational efficiency, and scalability. The results of these experiments clearly indicated that the hybrid DenCeption model outperformed other configurations, particularly in tasks requiring precise feature extraction and classification in medical images. Details of the experiments conducted and the results obtained will be presented in the coming sections.

3.5.4 Alignment with Research Goals

The overarching goal of this research is to advance the state-of-the-art in medical image processing by developing a robust, efficient, and highly accurate DL model. DenCeption was selected because it directly addresses the identified gaps in existing methods, offering a novel approach that combines the best aspects of DenseNet and Inception architectures. This hybrid model not only meets the technical requirements but also aligns with the practical needs of the healthcare industry, where accuracy, efficiency, and adaptability are critical.

3.5.5 Justification for Model Components

Each component of the DenCeption model was carefully chosen to enhance its overall functionality. The HDB and HTB were designed to optimise feature extraction and dimensionality reduction, respectively, ensuring that the model remains both effective and efficient. The integration of inception modules within these blocks further contributes to the model's ability to handle complex and diverse medical imaging data, providing a comprehensive solution to the challenges identified in the literature.

3.6 Conducted Experiments

Towards testing the proposed DenCeption performance against the aforementioned hybrid and singular models on medical images, a classification-based task will be performed. The set of experiments conducted in this chapter will involve the test of the current proposal with different versions of DenCeption to include:

- DenCeption with no hybrid transition block, namely DenCeption-HDB, where the growth rate k=32 for all BN, ReLU and conv layers as per the conventional structure of the classic TB.
- DenCeption with hybrid dense block without InB Inception block, namely Denception-HDB-NInB.
- DenCeption with hybrid dense block without InA Inception block, namely Denception-HDB-NInA.
- DenCeption with no hybrid dense block, namely DenCeption-HTB.
- DenCeption with hybrid transition block without RB reduction block, namely Denception-HTB-NRB.
- DenCeption with hybrid transition block without RA reduction block, namely Denception-HTB-NRA.
- DenCeption with hybrid transition block without InC reduction block, namely Denception-HTB-NInC.
- DenCeption with both hybrid blocks using the foundation of DenseNet-121, namely DenCeption-121.
- DenCeption with both hybrid blocks using the foundation of DenseNet-201, namely DenCeption-201.

- DenCeption with both hybrid blocks using the foundation of DenseNet-161, namely DenCeption-161.

Table 3.9 presents the layers composition of each of the above versions. The comparison will also involve the reviewed state-of-the-art models to include ResNet-DenseNet (Yasashvini et al., 2022), ResNet-Inception (Alotaibi and Alotaibi, 2020), and Inception-DenseNet-121 (Zhang and Feng, 2019). The labelled MRI dataset is split into three main sets: 60% training, 30% testing, and 10% validation. To test the performance of the proposed framework the following metrics are considered: Acc, Sen, Spe, precision, F1-score, and MAE.

Table 3.9: Hybrid DenCeption Variants Architecture

Hybrid Models	HDB	HDB with no InB	HDB with no InA	HTB	HTB with no RB	HTB with no RA	HTB with no InC	Classic DB	Classic TB
	$\begin{bmatrix} 1\times 1CB \\ InA \\ 3\times 3CB \\ InB \end{bmatrix}$	$\begin{bmatrix} 1\times 1CB \\ InA \\ 3\times 3CB \end{bmatrix}$	$\begin{bmatrix} 1\times 1CB \\ 3\times 3CB \\ InB \end{bmatrix}$	$\begin{bmatrix} RA \\ 1\times 1CB \\ RB \\ InC \\ 2\times 2AP \end{bmatrix}$	$\begin{bmatrix} 1\times 1CB \\ RB \\ InC \\ 2\times 2AP \end{bmatrix}$	$\begin{bmatrix} RA \\ 1\times 1CB \\ InC \\ 2\times 2AP \end{bmatrix}$	$\begin{bmatrix} RA \\ 1\times 1CB \\ RB \\ 2\times 2AP \end{bmatrix}$	$\begin{bmatrix} 1\times 1CB \\ 3\times 3CB \end{bmatrix}$	$\begin{bmatrix} 1\times 1CB \\ 2\times 2AP \end{bmatrix}$
DenCeption-HDB	$\times \begin{bmatrix} 6 \\ 12 \\ 32 \\ 32 \end{bmatrix}$								$\times \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$
DenCeption-HDB-NInB		$\times \begin{bmatrix} 6 \\ 12 \\ 32 \\ 32 \end{bmatrix}$							$\times \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$

Continued on next page

Table 3.9: Hybrid DenCeption Variants Architecture (Continued)

Hybrid Models	HDB	HDB with no InB	HDB with no InA	HTB	HTB with no RB	HTB with no RA	HTB with no InC	Classic DB	Classic TB
DenCeption-HDB-NInA			$\times \begin{bmatrix} 6 \\ 12 \\ 32 \\ 32 \end{bmatrix}$						$\times \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$
DenCeption-HTB				$\times \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$				$\times \begin{bmatrix} 6 \\ 12 \\ 32 \\ 32 \end{bmatrix}$	
DenCeption-HTB-NRB					$\times \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$			$\times \begin{bmatrix} 6 \\ 12 \\ 32 \\ 32 \end{bmatrix}$	

Continued on next page

Table 3.9: Hybrid DenCeption Variants Architecture (Continued)

Hybrid Models	HDB	HDB with no InB	HDB with no InA	HTB	HTB with no RB	HTB with no RA	HTB with no InC	Classic DB	Classic TB
DenCeption-HTB-NRA						$\times \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$		$\times \begin{bmatrix} 6 \\ 12 \\ 32 \\ 32 \end{bmatrix}$	
DenCeption-HTB-NInC							$\times \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$	$\times \begin{bmatrix} 6 \\ 12 \\ 32 \\ 32 \end{bmatrix}$	
DenCeption-121	$\times \begin{bmatrix} 6 \\ 12 \\ 24 \\ 16 \end{bmatrix}$			$\times \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$					

Continued on next page

Table 3.9: Hybrid DenCeption Variants Architecture (Continued)

Hybrid Models	HDB	HDB with no InB	HDB with no InA	HTB	HTB with no RB	HTB with no RA	HTB with no InC	Classic DB	Classic TB
DenCeption-201	$\times \begin{bmatrix} 6 \\ 12 \\ 48 \\ 32 \end{bmatrix}$			$\times \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$					
DenCeption-161	$\times \begin{bmatrix} 6 \\ 12 \\ 36 \\ 24 \end{bmatrix}$			$\times \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$					
DenCeption	$\times \begin{bmatrix} 6 \\ 12 \\ 32 \\ 32 \end{bmatrix}$			$\times \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$					

3.7 Dataset

The decision to use the BRATS MRI dataset, despite its simplicity and the fact that it is unlabelled, was based on several strategic considerations that align with the research objectives. Below is a detailed justification addressing why the BRATS data was chosen and how it contributes to the overall goals of the research.

3.7.1 Benchmarking and Validation

The BRATS dataset is widely recognised in the medical imaging community, particularly in the field of brain tumour analysis. It has been used extensively as a benchmark for developing and testing segmentation and classification models. Despite being unlabelled, its established use in competitions like the RSNA-ASNR-MICCAI Brain Tumor Segmentation Challenge makes it a valuable resource for validating the performance of new models against industry standards.

3.7.2 Real-World Clinical Relevance

The BRATS dataset, although unlabelled, closely mirrors real-world clinical scenarios where MRI scans are often presented without pre-annotations. This reflects the reality that in many clinical settings, models need to perform accurately without the benefit of fully labelled datasets. The use of BRATS data, therefore, provides an opportunity to test the model's robustness and ability to handle real-world challenges, such as working with raw, unlabelled data.

3.7.3 Complexity in Data Structure

While the BRATS dataset is unlabelled, it is far from simplistic in terms of data structure. The dataset includes multiple modalities (T1, T1Gd, T2, T2-FLAIR) acquired from different clinical protocols and scanners. This complexity in data acquisition and the inherent variability in the images provide a challenging environment for testing the adaptability and performance of the proposed models. The dataset's multi-modality nature demands a model that can generalise

across different types of MRI scans, which is critical for ensuring the model's applicability in diverse clinical scenarios.

3.7.4 Strategic Use in Early Model Development

The BRATS dataset was strategically used in the early stages of model development (as detailed in Chapters 3 and 4) to fine-tune the feature extraction and segmentation capabilities of the proposed DenCeption model. The simplicity of the dataset in terms of labelling allows the research to focus on improving the core functionalities of the model, such as handling multi-modal MRI data, before moving on to more complex, fully labelled datasets.

3.7.5 Contribution to Model Robustness

The inclusion of BRATS data, despite its simplicity, contributes to enhancing the robustness of the proposed models. By testing the models on unlabelled data, the research ensures that the models are not overly reliant on labelled training data and can perform effectively even when such annotations are unavailable. This is particularly important for developing models that can be deployed in varied clinical settings, where access to fully labelled data may be limited.

3.8 Results and Discussion

3.8.1 Training and Testing Results: DenCeption Versus DenCeption Variants and Benchmarking Methods

The training results showcase DenCeption outperformed other models with the highest Acc of 91.3% (Table 3.10). Following it, DenCeption-201, DenCeption-161, and DenCeption-121 all presented a noticeable performance with accuracies around 89%. ResNet-Inception model has the lowest Acc of 73.4% (Alotaibi and Alotaibi, 2020). DenCeption-201 has the highest Sen (90%) compared to the other variants, indicating its ability to correctly identify positive cases, yet less than DenCeption's Sen (93%). DenCeption-HTB-NInC has the lowest

Sen (64%), suggesting it may miss a significant number of positive cases. DenCeption has also showed a precision of 93.7%, showing its strength in correctly identifying negative cases. The models with no InA and InB models in the HDB block, tend to have lower Spe, suggesting a difficulty in accurately classifying negative cases. As per the table, precision is highest for DenCeption (94%), indicating a high proportion of TP identifications out of all positive identifications. However, ResNet-Inception shows the lowest precision (70.3%), indicating more FPs. DenCeption has the highest F1-score of 93.4%, demonstrating excellent overall performance. Conversely, DenCeption-HTB-NInC has the lowest F1-score, reinforcing that it's the weakest performer compared to all other variants and benchmarking methods. DenCeption has the lowest MAE of 0.2, aligning with its high accuracy. Higher MAE in other models like DenCeption-HTB-NInC (0.88) and ResNet-Inception (0.84) indicates more significant errors in prediction.

Table 3.10: Training Results of DenCeption Versus Other Variants and Benchmarking Methods

Methods	**Acc (%)**	**Sen (%)**	**Spe (%)**	**Precision (%)**	**F1-score (%)**	**MAE**
ResNet-DenseNet (Yasashvini et al., 2022)	88.3	88	87	79	83.2	0.6
ResNet-InCeption (Alotaibi and Alotaibi, 2020)	73.4	74	73	70.3	72.1	0.84

Continued on next page

Table 3.10: Training Results of DenCeption Versus Other Variants and Benchmarking Methods (Continued)

Methods	**Acc (%)**	**Sen (%)**	**Spe (%)**	**Precision (%)**	**F1-score (%)**	**MAE**
InCeption-DenseNet-121 (Zhang and Feng, 2019)	88.5	87	89.2	83	84.9	0.61
DenCeption-HDB	88.1	80	78	81	80.4	0.58
DenCeption-HDB-NInB	83	78	79.6	80	78.9	0.72
DenCeption-HDB-NInA	77	69	70	72	70.4	0.8
DenCeption-HTB	76.4	68	68.1	70	68.9	0.84
DenCeption-HTB-NRB	76	68.3	67.4	69	68.6	0.85
DenCeption-HTB-NRA	76.2	67	68	73	69.8	0.83
DenCeption-HTB-NInC	74	64	62.3	66	64.9	0.88
DenCeption-121	89.3	89	88.4	90	89.4	0.4

Continued on next page

Table 3.10: Training Results of DenCeption Versus Other Variants and Benchmarking Methods (Continued)

Methods	**Acc (%)**	**Sen (%)**	**Spe (%)**	**Precision (%)**	**F1-score (%)**	**MAE**
DenCeption-201	89.7	90	89.5	91	90.3	0.38
DenCeption-161	89.5	89.6	89	90.3	89.8	0.4
DenCeption	91.3	93	93.7	94	93.4	0.2

These results indicate that both ResNet-DenseNet (Yasashvini et al., 2022) and InCeption-DenseNet-121 (Zhang and Feng, 2019) models perform comparably in terms of accuracy and specificity, but InCeption-DenseNet-121 slightly outperforms in F1-score and has a marginally lower MAE, indicating more consistent performance. The variations of DenCeption show a wide range of performance, with the 'HDB' and 'HTB' variants showing weaker results across metrics compared to '121', '201', and '161'. This indicates that certain configurations of DenCeption are superior to others, and fine-tuning the architecture is critical. Models without hybrid inception modules to include 'NInA', 'NInB', and 'NInC' tend to perform worse across all metrics than their counterparts, highlighting the importance of these components for model accuracy and reliability.

In critically analysing the DenCeption model in relation to its benchmarks and variants, it is essential to consider various factors such as overall performance metrics, model complexity, and practical applicability.

DenCeption Versus Benchmarks

Compared to these benchmarks, DenCeption generally outperforms ResNet-DenseNet (Yasashvini et al., 2022) and ResNet-Inception (Alotaibi and Alotaibi, 2020)significantly in terms of accuracy, precision, and F1-score. The ResNet-Inception model, with the stochastic depth, exhibits

the lowest performance among the benchmarks, which may suggest that while stochastic depth can potentially help in training deeper networks, it does not necessarily translate to higher performance in all cases. ResNet-DenseNet, while reasonably competitive, still underperform compared to DenCeption model's performance. On the other hand, InCeption-DenseNet-121 model closely outperform the DenCeption variants in accuracy, but still underperforms compared to the best-performing DenCeption models (Zhang and Feng, 2019). Its relatively high Sen and Spe indicate it is a robust model, but DenCeption's higher precision suggests it has better discriminative power for positive class identification.

DenCeption Versus its Variants

DenCeption-HDB variants show a drop in performance compared to the top DenCeption models. The hybrid HDB blocks, while innovative, may introduce complexity without proportional increase in performance, as indicated by their lower Sen and Spe scores. This suggests a potential overfitting issue or that the model architectures might not be capturing the features as effectively as the main DenCeption model. Conversely, in case of DenCeption-HTB variants, the different versions of the transition block seem to negatively impact the performance metrics, particularly for those missing RA, RB, or InC hybrid modules. This is possibly due to an over-reduction in feature space or ineffective feature translation from one block to another. The increased MAE across these variants also points to less accurate predictions. These results were quite different in case on DenCeption-121, DenCeption-201, and DenCeption-161. In fact, these models demonstrate a high level of performance across all metrics, suggesting that the specific configurations of hybrid HDB and HTB blocks in these architectures are well-suited to the classification task.

There is not a direct correlation between model complexity (number of parameters) and performance. While some DenCeption variants have fewer parameters but manage to achieve high accuracy and other performance metrics, some with more parameters do not perform as expected. This highlights the importance of architecture optimisation over highly increasing model depth or width. This proves the outstanding performance of DenCeption that, although

with small number of parameters compared to other models including those from its variants or from benchmarking models. The DenCeption model, with its inception, dense block, and transition block combination, showed an exceptional capability to capture a richer feature representation than its benchmarks and some of its variants. This is evident in its superior performance, particularly in precision and F1-score, which are critical for class imbalance scenarios often found in real-world datasets. Some variants present a potential of overfitting risks, where models perform well on the training data but may not generalise as well on unseen data. This is suggested by the increased MAE and lower Sen/Spe scores. Given the observations from the training outcomes, it became evident that removing hybrid Inception components including InA, InB, and InC tends to degrade performance, reaffirming the value of these modules in the DenCeption architecture for effective feature extraction and representation. The DenCeption model demonstrates high performance and is especially dominant in its precision and accuracy metrics. Its distinguished HDB and HTB combination proves that not all modifications lead to improvements, as seen with the HDB only and HTB only variants.

The results presented in Table 3.10 show a close range of accuracies among the DenCeption variants. While these results may initially appear similar, a deeper analysis reveals the nuances that differentiate these models, particularly when considering other performance metrics beyond Acc, Sen, Spe, precision, F1-score, and MAE, as follows:

- **Acc Analysis:** The slight difference in accuracy between DenCeption and its variants, such as DenCeption-201 and DenCeption-161, can be attributed to the architectural modifications aimed at balancing depth and complexity. DenCeption-201, for instance, includes a deeper structure compared to DenCeption-121, which may allow for more intricate feature extraction, particularly in complex imaging tasks. However, this increased depth also introduces a higher computational cost and potential for overfitting, which might slightly affect its generalisation capability, explaining the minor differences in accuracy.

- **Sen and Spe Analysis:** Sen and Spe are critical metrics in medical image analysis, par-

ticularly in scenarios where the cost of false negatives or false positives is high. The DenCeption model's Sen (93%) and Spe (93.7%) outperform those of the variants, suggesting that the original model is slightly better at correctly identifying both positive and negative cases. However, DenCeption-201 and DenCeption-161 are very close in these metrics, indicating that while they are slightly less accurate overall, they still maintain a strong balance between identifying true positives and true negatives.

- **Precision and F1-Score Analysis:** The F1-score and precision metrics also highlight the robustness of the DenCeption model. With a precision of 94% and an F1-score of 93.4%, DenCeption demonstrates superior performance in maintaining a high proportion of correctly identified positive results out of all positive results identified. The variants, particularly DenCeption-201 and DenCeption-161, maintain high precision and F1-score close to 90%, indicating that they are still highly effective but slightly less balanced than the original DenCeption model in this regard.

- **MAE Analysis:** The MAE metric further underscores the performance differences. DenCeption has the lowest MAE of 0.2, which reflects its high reliability and lower error rate in prediction tasks. The variants exhibit slightly higher MAE values, with DenCeption-201 and DenCeption-161 showing MAEs of 0.38 and 0.4 respectively. While these differences are not large, they do suggest that DenCeption is better optimised for minimising errors in predictions, making it more reliable across different scenarios.

While the accuracy differences among DenCeption and its variants are not substantial, the additional metrics provide a more complete picture. The original DenCeption model maintains a slight edge in overall performance, particularly in Sen, Spe, and MAE. The variants, while performing admirably and offering nearly comparable accuracy, show minor differences that could influence their suitability for specific medical imaging tasks where either computational efficiency or a particular balance of sensitivity and specificity is prioritised.

Towards better understanding the performance of the proposed hybrid model during the training and testing phase against other variants (DenCeption-121, DenCeption-201, and DenCeption-

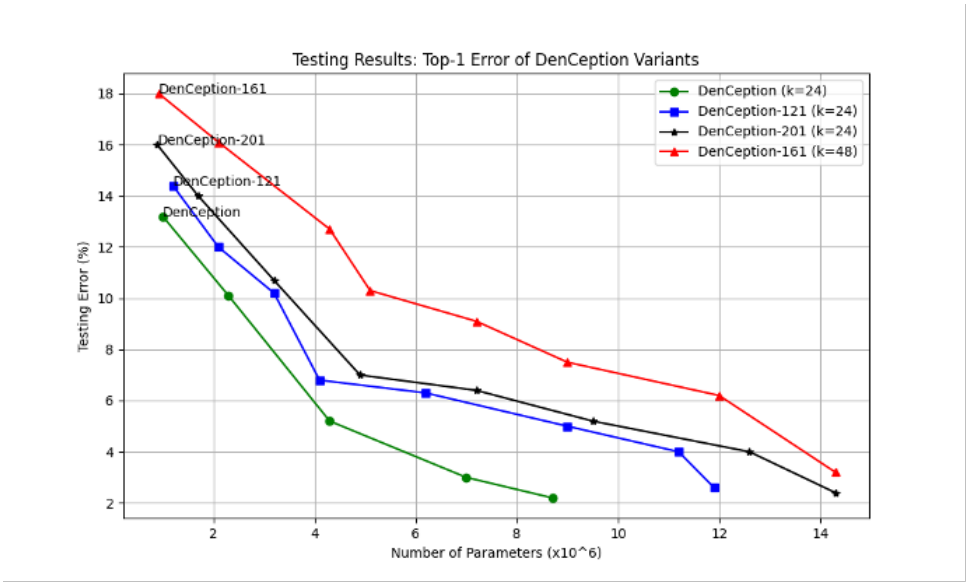


Figure 3.15: Testing Results: Top-1 Error of DenCeption Variants

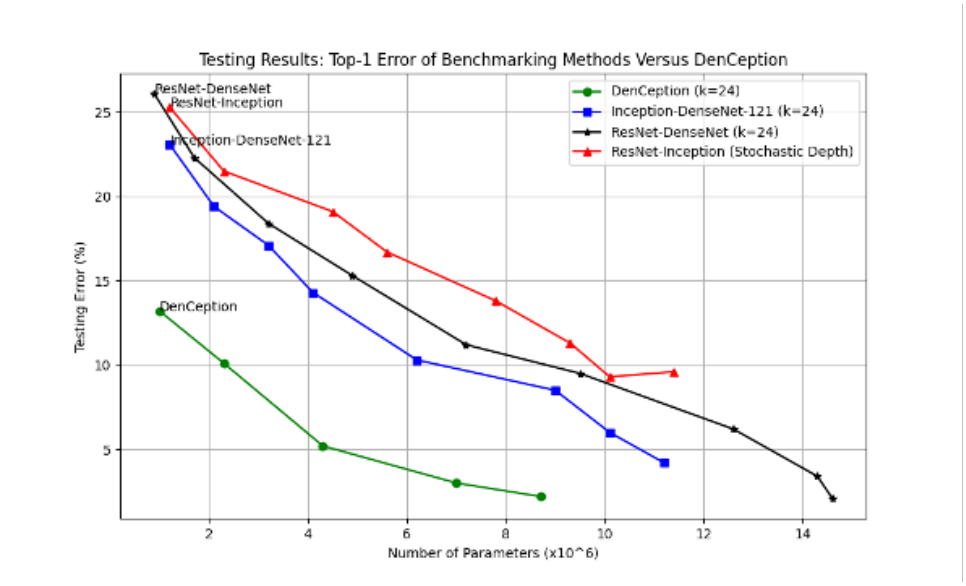


Figure 3.16: Testing Results: Top-1 Error of DenCeption Versus Benchmarking Methods

161) as well as existing benchmarking methods, relationship between Top-1 error versus the variation of the number of parameters per model has been conducted as shown in Figures 3.15 and 3.16, respectively.

Figure 3.15 reveals that, as the number of parameters increases, the Top-1 error for all DenCeption variants decreases, which suggests that having more parameters helps to reduce the classification error rate of the model. In particular, DenCeption's behaviour in decreas-

ing the Top-1 error while increasing the number of parameters is common, where additional parameters often correlate with a network's capacity to learn more complex features from the data. Similarly, the DenCeption-161 (k=48) variant starts with a higher error rate compared to the others but demonstrates a significant drop as parameters increase, ending up with the lowest Top-1 error among all the variants. This significant drop-in error rate suggests that this model is highly capable of learning from additional parameters, possibly due to a more effective network design that leverages the increased growth rate. However, it is worth noting that a compromise decision regarding the number of parameters and the efficiency of the model is critical as it impacts the adaptability and scalability of the model as well as its computational efficiency. On the other hand, DenCeption-201 (k=24) variant maintains a consistent advantage over the DenCeption-121 (k=24) when comparing models with the same growth rate (k). In fact, DenCeption-201 shows a consistently lower Top-1 error across all parameter sizes, indicating a better capacity to generalise or learn features. Its performance improvement is more pronounced at lower parameter counts, highlighting efficient use of the model's capacity. In contrast, DenCeption-121 initially starts with a higher error than the baseline but shows significant improvement with an increase in parameters. The model shows a stabilisation in performance improvement beyond a certain parameter count, suggesting a balance in learning capacity.

The proposed DenCeption (k=24) variant shows a less rapid descent in error rate compared to the more complex models. By observing the resulted graph, it looks that the variant DenCeption-161 (k=48) achieves a perfect balance between the number of parameters and performance, as it achieves the lowest error rate in the end. However, this has a drawback regarding the additional parameters, which might entail a more complex model that may require more computational resources. As per Figure 3.15, incremental improvements become less significant where increases in parameters lead to minor improvements in error rates, particularly in case of DenCeption-121 and DenCeption-201. This suggests that simply adding more parameters may not always result in significant performance enhancement and leads to consider the trade-off between complexity and performance. Additionally, the consideration of overfitting

issue is crucial in evaluating these models. In fact, deeper models with more parameters, like DenCeption-161 (k=48), may be at risk of overfitting. Therefore, it is crucial to validate these models on a separate validation set to ensure that their generalisability to unseen data.

Compared to other benchmarking methods, DenCeption maintains its high performance by showing the lowest Top-1 error rate across the number of parameters, indicating a robust model with high efficiency in parameter utilisation as shown in Figure 3.16. The graph also proves that a significant reduction in error rate with an increase in the number of parameters, is a highlight of an effective architecture for learning from data. Inception-DenseNet-121 demonstrates improved performance over the other models, except for DenCeption, as parameters increase (Zhang and Feng, 2019). There's a significant decrease in error rate with the first increase in parameters, which then starts to stabilise, resulting a drawback with additional parameters. ResNet-DenseNet achieved higher Top-1 error rates compared to DenCeption and Inception-DenseNet-121 (Yasashvini et al., 2022). However, despite an initial significant decrease, the enhancement in performance reaches a saturation point with an increase in parameters, indicating that ResNet-DenseNet may be less efficient at utilising additional parameters beyond a certain point. ResNet-Inception (Stochastic Depth), on the other hand, has a unique performance curve, starting with the highest error and showing a significant improvement as the number of parameters increases (Alotaibi and Alotaibi, 2020). This indicates that the use of stochastic depth might contribute to the efficient training of deep networks, allowing the model to eventually outperform ResNet-DenseNet, yet to outperform the proposed DenCeption. This proves that, while all models improve with more parameters, DenCeption achieves this more effectively, indicating it is well suited for environments where computational resources are limited, yet presenting highest Acc (91.3%), Sen (93%), Spe (93.7%), precision (94%), F1-score (93.4%), and lowest MAE (0.2).

In the following, as part of the validation stage, an evaluation of the unseen data will be conducted. This experiment enables to assess the generalisability of the proposed DenCeption model. This helps to avoid overfitting problem. By analysing the Top-1 and Top-5 errors, it enables the evaluation of the model's precision and its ability to distinguish the most likely

class and its top five likely classes, respectively. The analysis of the depth and the numbers of parameters per model allows a better understanding of the computational complexity, scalability and generalisability of these models. This will be considered as a validation process to the proposed DenCeption model. The results are presented in Tables 3.11 and 3.12 considering two growth rates with values of k=32 and k=24 respectively.

3.8.2 Validation Results: DenCeption Versus Variants and Benchmarking Methods

As shown in Table 3.11, as the depth and the number of parameters increase, there is an improvement in Top-1 and Top-5 error rates. However, this pattern is not consistent as some models with fewer parameters, such as DenCeption-121, outperform more complex models like DenCeption-HDB. DenCeption outperforms the other variants as well as the benchmarking methods with the lowest Top-1 and Top-5 error rates equal to 23.4 and 5.87 respectively. This performance is followed closely by the DenCeption-201 model. It suggests that the architecture of DenCeption is more effective in extracting and generalising features from the data, due to optimised layer configurations and parameter utilisation. The ResNet-Inception model with stochastic depth shows relatively higher error rates. This indicates that stochastic depth does not consistently provide benefits across different network architectures, especially when comparing against networks like DenCeption that potentially have more optimised pathways for feature propagation. The HDB and HTB variants show varied results, with HDB coupled with classic TB generally performing better than HTB coupled with classic DB. Removing inception modules from HDB (DenCeption-HDB-NInA and DenCeption-HDB-NInB) impacts performance negatively, signifying the importance of these modules in the learning ability of the network.

Table 3.11: Validating Results with Growth Rate k=32

Methods	**Top-1 error (%)**	**Top-5 error (%)**	**Depth**	**Number of parameters** ($x10^6$)
ResNet-DenseNet (k=32) (Yasashvini et al., 2022)	26.87	7.12	110	11.7
ResNet-InCeption (stochastic depth) (Alotaibi and Alotaibi, 2020)	28.32	8.93	110	11.9
InCeption-DenseNet-121 (k=32) (Zhang and Feng, 2019)	26.44	6.43	100	12
DenCeption-HDB (k=32)	27.44	8.1	150	15.4
DenCeption-HDB-NInB (k=32)	27.08	7.89	100	9.3
DenCeption-HDB-NInA (k=32)	29.7	9.02	100	9.7
DenCeption-HTB (k=32)	30.8	9.93	110	10.3
DenCeption-HTB-NRB (k=32)	31.27	10.16	90	9.5

Continued on next page

Table 3.11: Validating Results with Growth Rate k=32 (Continued)

Methods	**Top-1 error (%)**	**Top-5 error (%)**	**Depth**	**Number of parameters** ($x10^6$)
DenCeption-HTB-NRA (k=32)	31.03	10.07	90	9.6
DenCeption-HTB-NInC (k=32)	32.9	10.43	100	10.4
DenCeption-121 (k=32)	26.6	7.03	190	7.3
DenCeption-201 (k=32)	25.3	6.42	270	12
DenCeption-161 (k=32)	26.3	6.91	290	17.3
DenCeption (k=32)	23.4	5.87	230	8.17

By decreasing the growth rate, DenCeption with k=24 presents the best model across the variants and benchmarking methods with significantly lower error rates, despite having fewer parameters and moderate depth compared to other models. This proves the highly effective use of its architectural features and parameters, resulting in strong generalisation. DenCeption-121, DenCeption-201, and DenCeption-161, on the other hand, demonstrated good parameter efficiency, achieving low error rates with a smaller number of parameters compared to benchmarking methods. In fact, when compared with the ResNet-DenseNet (Yasashvini et al., 2022) and ResNet-Inception (Alotaibi and Alotaibi, 2020) models, the DenCeption variants show clear superiority, indicating that the hybrid DenseNet-Inception architecture, especially when fine-tuned (as in DenCeption variants), can outperform more traditional hybrid approaches. Hence the outperformance of DenCeption when compared with Inception-DenseNet-121.

Table 3.12: Validating Results with Growth Rate k=24

Methods	**Top-1 error (%)**	**Top-5 error (%)**	**Depth**	**Number of parameters** ($x10^6$)
ResNet-DenseNet (k=24) (Yasashvini et al., 2022)	21.3	4.22	90	10.4
ResNet-InCeption (stochastic depth) (Alotaibi and Alotaibi, 2020)	23.5	5.66	110	11.9
InCeption-DenseNet-121 (k=24) (Zhang and Feng, 2019)	20.1	4.03	70	11
DenCeption-HDB (k=24)	19.30	3.07	110	13.4
DenCeption-HDB-NInB (k=24)	20.06	4	90	6.1
DenCeption-HDB-NInA (k=24)	22.4	4.55	90	6.3
DenCeption-HTB (k=24)	22.7	4.7	60	7.5
DenCeption-HTB-NRB (k=24)	23	4.2	70	6.2

Continued on next page

Table 3.12: Validating Results with Growth Rate k=24 (Continued)

Methods	**Top-1 error (%)**	**Top-5 error (%)**	**Depth**	**Number of parameters** ($x10^6$)
DenCeption-HTB-NRA (k=24)	21.9	3.45	40	6.3
DenCeption-HTB-NInC (k=24)	24.2	5.7	60	6.9
DenCeption-121 (k=24)	18.3	2.87	150	5.4
DenCeption-201 (k=24)	17.5	2.3	200	8
DenCeption-161 (k=24)	18.04	2.5	250	10
DenCeption (k=24)	15.1	1.73	190	3.6

3.9 Conclusion

The comprehensive analysis of the DenCeption model, as evidenced by the provided training and testing results, Confirms its exceptional performance across various evaluation metrics. Notably, the DenCeption model achieves a novel standard of Acc achieving 91.3%, setting a new benchmark for excellence among the evaluated models. On a similar pathway, DenCeption-201, DenCeption-161, and DenCeption-121 variants demonstrate notable performance with accuracies around 89%. The ResNet-Inception model (Alotaibi and Alotaibi, 2020), however, underperformed with the least Acc of 73.4%. The DenCeption-201 variant excels in Sen, with a rate of 90%, showcasing its adaptability at accurately identifying positive instances, although slightly lower than DenCeption's highest Sen of 93%. Contrarily, the

DenCeption-HTB-NInC variant displays the lowest Sen at 64%, which shows the importance of integrating InC module within HTB increasing the recognition of a considerable number of positive cases. The precision rate of DenCeption reaching 94% underscores its efficiency in accurately identifying negative instances. The variants of InA and InB modules within the HDB block showed reduced Spe, reflecting a challenge in accurately classifying negative instances. Furthermore, DenCeption achieves the highest F1-score of 93.4%, illustrating its exceptional overall performance, while the DenCeption-HTB-NInC variant presenting the lowest performance further prove the importance of InC module alongside RA and RB reduction modules. The minimal MAE of 0.2 aligned with DenCeption's high accuracy is reflected in higher MAE values in other models, such as DenCeption-HTB-NInC (0.88) and ResNet-Inception (0.84), indicating significant prediction errors. The proposed DenCeption model's exceptional performance, as showed in the training results, suggests a significant advancement over the ResNet-DenseNet (Yasashvini et al., 2022) and InCeption-DenseNet-121 models in terms of accuracy and specificity. The diverse performance range of DenCeption variants highlights the critical impact of precise architecture fine-tuning. Variant models lacking hybrid inception modules, namely as 'DenCeption-HDB-NInA', "DenCeption-HDB-NInB', and "DenCeption-HTB-NInC', face performance degradation across all metrics, underscoring the indispensable role of these modules in the DenCeption architecture for effective feature extraction and representation. Transitioning from training and testing to validation stages is a critical step in affirming the DenCeption model's efficacy. The validation phase evaluates the model's generalisability and robustness, using unseen data to mitigate overfitting issues. Through an in-depth analysis of Top-1 and Top-5 errors, alongside depth and number of parameters per model, the validation process examines the computational complexity, scalability, and generalisability of the proposed model, DenCeption. This phase was pivotal in validating the DenCeption model's performance, ensuring its applicability and reliability in practical scenarios. DenCeption's performance, particularly against its variants and benchmarking methods, highlights a future path for research and development in DL models, addressing complex classification tasks with enhanced efficiency and effectiveness.

In this chapter, DenCeption has been introduced as an innovative DL model, paving the way for for a detailed discussion in the next chapter on the essential role of both HF and DHF feature extractions in enhancing the learning process. Moreover, the chapter will explore how DenCeption integrates in parallel with other techniques within the proposed advanced feature extraction framework, marking a significant advancement in the field.

Chapter 4

A Deep Learning based Scalable and Adaptive Feature Extraction Framework for Medical Images

4.1 Introduction

In the previous chapter, the development and validation of DenCeption has been detailed, a cutting-edge DL model designed to advance computational accuracy and efficiency. This chapter builds on the groundwork established by DenCeption, turning the attention to an in-depth examination of feature extraction techniques. By exploring both HF and DHF feature extractions, this work aims to highlight their crucial roles in enhancing the learning process. Furthermore, this chapter will show how integrating DenCeption within the newly proposed feature extraction framework not only improves the capabilities of state-of-the-art methods but also sets the stage for significant breakthroughs in the field of ML/DL and image analysis.

Medical image processing is a challenging step towards the efficiency enhancement of disease detection and diagnosis. The analysis of medical images has been considered as challenging and time-consuming task, particularly, for doctors and specialists. Improving the early diagnosis of a medical disease presents a serious challenge. To cope with this problem, medical

field is being in a massive progress to improve existing physiological analysis methods as well as medical machines for early disease detection and prediction. This topic has gained a great importance in medical innovative research, as a result it becomes an inner area for researchers including different specialities such as doctors and data scientists to use medical images in several applications.

One of the interesting stages in image processing is the medical image content-based retrieval. The complicated composition of medical images makes the information extraction a challenging step. Features extraction represents an important stage towards providing relevant image content based to result efficient medical application, for example, disease detection, medical analysis, as well as disease prediction. Each medical application is reflected by a focus area, namely, RoI, which contains most of the needed features to accurately accomplishing the target task, e.g., classification. In recent years, AI, approaches particularly DL, have evolved significantly due to the improvement in the processing capacity of computers and the accumulation of big data (Arel, Rose, and Karnowski, 2010). DL proved a strong ability to identify meaningful relationships in raw data, which justifies its application to support diagnosing, treating, and predicting outcomes in many medical situations. DL approaches, have already demonstrated their proficiency, surpassing human performance in medical applications particularly in diagnosing and predicting disease progression. DL proved a strong ability to identify meaningful relationships in raw data, which justifies its application to support diagnosing, treating, and predicting outcomes in many medical situations. DL is transforming the practice of medicine; it is helping doctors diagnose patients more accurately, make predictions about the patient's future health, and recommend better treatments (Ravì et al., 2016; Litjens et al., 2017). DL approaches present key-methods for several medical applications including decision making, disease stage tracking, disease detection, disease diagnosis and analysis. DL networks have shown a high sensitivity and accuracy for the detection of several diseases including breast cancer (Yala, Lehman, Schuster, Portnoi, and Barzilay, 2019), brain tumour (Zhao et al., 2018), DMO (Tang et al., 2021)...etc. Its application, particularly for features extraction, has contributed to managing the progression of these diseases by enhancing early detection. Figure 4.1

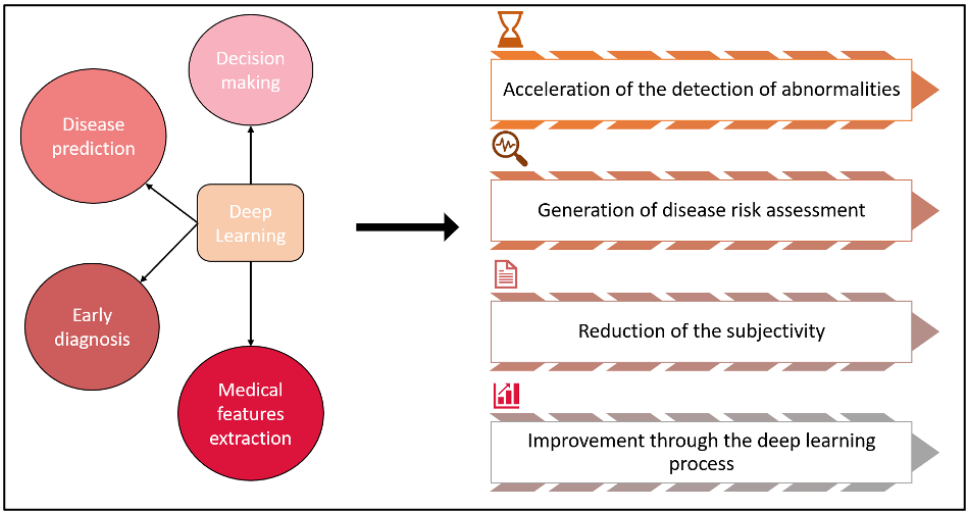


Figure 4.1: DL Impact On Medical Applications

presents a diagram summarising the impact of DL on medical applications.

The application of DL approaches has also increased the scalability and reliability of features extraction methods. As aforementioned, CNNs represent one of the most used architectures in features extraction (Razzak, Naz, and Zaib, 2018), in addition to MLP (Lai and Deng, 2018). The implication of CNN's architectures at this processing stage has shown a great improvement in the outcome of the classification and prediction tasks which represents a challenging area in medical imaging. Despite the integration of DL in medical image processing, traditional features extraction methods have also been applied concurrently. Particularly, salient and semantic features are one of the important extracted features in medical images (Gao, Ma, Liu, Liu, and Zhang, 2021; Conghua, Yuqing, and Jinyi, 2006). These features have been used in several applications such as images fusion, and image content-based retrieval.

A proposed research in (Gao et al., 2021) introduced a medical images fusion method based on salient feature extraction using Particle Swarm Optimisation (PSO) algorithm and the fuzzy logic. The suggested salient features extraction method is based on the non-subsampled shealet transform (NSST), where the latter helps into reducing the computational complexity of the proposed approach. The image fusion process is based mainly on the extraction of low and high frequency sub-bands features through the fuzzy logic and uses the PSO algorithm for optimisation. The proposed method has been tested on eight pairs of grey-scale and five pairs of

colour multimodal medical images.The amount of the testing set is considered very low in order to validate the suggested method. Subsequently, this limits the scalability of its application in real-time scenarios.

Semantic features have also been applied by authors in (Conghua, Yuqing, and Jinyi, 2006). Their method is based on the space density function, where they enhanced the original method which used Bayesian Belief network (BNN) (Peng and Long, 2001; Conghua, Song, Zhu, and Wang, 2005). The main idea is to transform the medical images from grey-scale to density function space. Their method has been tested on 400 pieces of images covering head, chest, abdomen and limbs of human bodies. The outcome precision of their method reflects a good image retrieve performance achieving 88.8%. One of the drawbacks presented by this method is the non-consideration of coloured dataset and the limited number of validation images. Comparable to (Conghua, Yuqing, and Jinyi, 2006), this leads to a potential scalability problem. In addition, based on (Gao et al., 2021; Conghua, Yuqing, and Jinyi, 2006), salient features extraction is mainly dependent on labelled datasets, which decrease its reliability and responsiveness in case of unlabelled input samples. That is, these irregular features are not efficient in such scenarios. Therefore, in this work, the main focus is on the regular features categorised as HF and DHF features.

Many classical and recent methods have been proposed to solve the features extraction step. Some of these methods consider single feature usage, such as texture, some others contemplate a combination of different feature levels. In this paper, a new features extraction framework is proposed. The first contribution of the method is to select the optimal features combination according to the input dataset. The fusion involves two main types of features including HF and DHF features in case of neural network application. The second contribution of the proposed extraction tool is the integration of the proposed automated hybrid deep network, DenCeption, for DHF features extraction. The massive enhancement of the resulted classification is dedicated to the resulted optimal features fusion. The structure of the chapter is as follows:

- Section 2 covers the related features extraction works.

- Section 3 highlights the identified problems.
- Section 4 presents the proposed features extraction methodology. A detailed explanation of the different types of features considered in this work will be covered, in addition to the features weighting and fusion stages.
- Section 5 presents the used datasets.
- Section 6 covers the conducted experimentation and the proposed research evaluation mechanism.
- Section 7 provides a critical evaluation of the obtained results for the different datasets.
- The chapter ends with Section 8.

4.2 Related Works

Medical image features extraction presents an important step towards resulting highly accurate analysis related, for instance, to disease detection, classification, and prediction. Extracting reflective features reinforces the efficiency rate of these particular applications. Features are categorised as two main types: HF and DHF features (Chowdhary and Acharjya, 2020). HF features, in particular, include texture, shape, and colour features. These features represent the fundamental factor that can be extracted from medical images (Mutlag, Ali, Aydam, and Taher, 2020). DHF features cover the low-level characteristics of a medical image. These include hidden information reflecting important analysis and leading to enhancing the diagnosis reliability (Jeyakumar and Kanagaraj, 2019). In this context, several proposed features extraction frameworks have been applied to address both HF and DHF extraction issues. However, these related works still outline some drawbacks in terms of the deployment (Huerga et al., 2021; Hazarika, Maji, Sur, Paul, and Kandar, 2021; Tsai, Zhang, Hung, and Min, 2017; Rundo et al., 2019; Rundo et al., 2021; Kavya and Padmaja, 2017; Xiao, Liang, Guan, and Hassanien, 2013; Zewail and Hag-ElSafi, 2017; Liu and Shi, 2011; Mingqiang, Kidiyo, Joseph, et al., 2008). HF

and DHF features, in particular, are considered the main point of interest in several features extraction methods. Multiple challenges have been highlighted in the literature to include:

- Testing models using different dataset sizes and complexities,
- Using different types of datasets to convey coloured and grey-scale based images,
- Potential of validating models using real-case scenarios,
- Testing features extraction systems responsiveness to multiple cases, and
- The lack of sufficient extracted features in some particular scenarios.

The size of datasets applied for features extraction experimentation has an impact on the complexity of the framework outcomes. Hence, considering different dataset sizes is of great importance in experiments validation. In fact, Tahira et al. evaluated their DL-based method over challenging datasets, namely, APTOS-2019 and IDRiD (Nazir et al., 2021). Both datasets have different sizes and complexities, hence the difference in the validation performances of the same model. Similar impact has been highlighted in the content-based image retrieval system proposed by Lin et al. in (Lin, Chen, and Chan, 2009). The use of different sets of data, covering multiple aspects, proved the importance of such consideration in features extraction-based models to achieve 99.2% Acc, 72.7% AP, and 50% average recall. That said, this factor has not been considered in several proposed features extraction frameworks. Despite the use of complex datasets, these methods lack the dataset experiments validation which, as a result, impact their reliability.

In this context, a supervised SVM based features extraction model has been suggested by Xiao et al. in (Xiao et al., 2017), providing a good model performance validation. Similar approach has also been proposed by authors in (Janakasudha and Jayashree, 2020), however, different datasets have been used. Considering the same extraction method, both works proved different validation performances which, as a result, stresses the importance of considering size and complexity of datasets when it comes to building a reliable model (Janakasudha and

Jayashree, 2020; Xiao et al., 2017). Moreover, considering the aforementioned factor will potentially add a scalability factor to the resulted model.

Medical images can be presented in different morphological manners. These multiple representations could also impact the final outcome of the features extraction framework. Considering the latter, types of medical images including both colour and grey-scale images, have significant effect on the processing stage. Features, to include HF and DHF, vary from one medical image type to another. In fact, colour-based images are a source of colour feature which is lacking in grey-scale based images.

Texture and shape features, on the other hand, can be extracted from both image types, however, some of these researches ignore the importance of coloured based datasets, instead focusing mainly on grey-scale medical images (Huerga et al., 2021; Tsai et al., 2017; Rundo et al., 2019; Rundo et al., 2021; Xiao et al., 2013; Zewail and Hag-ElSafi, 2017; Janakasudha and Jayashree, 2020; Xiao et al., 2017; Altaf, Anwar, Gul, Majeed, and Majid, 2017; Howarth and Rüger, 2004; Dara, Tumma, Eluri, and Kancharla, 2018; Madusanka, Choi, So, and Choi, 2019). This can be justified by the cost of considering coloured medical images, however, this has a drop impact on the reliability and scalability of these methods. In fact, as proposed in (Liu and Shi, 2011), the consideration of two types of datasets gives the model a free-space to interpret HF and DHF features; thus, removing the interpretability as a major challenge. This, however, was not the case for exiting features extraction models that focused mainly on texture and shape features extraction (Madusanka et al., 2019; Janakasudha and Jayashree, 2020; Xiao et al., 2017). Despite achieving interesting results in terms of accuracy, sensitivity, and specificity, these methods lack of the consideration of colour-based medical images dataset, hence, the non-scalability of their proposed models. DHF-based features extraction frameworks also have been, in multiple instances, part of the above challenges particularly when it comes to processing medical images such as CT and MRI scans proposed respectively by (Dara et al., 2018; Liu, Liu, and Zhu, 2020; Janakasudha and Jayashree, 2020). However, despite the consideration of colour-based medical images dataset, DHF extraction can also lack the importance of features that can be extracted through grey-scale based datasets (Nazir et al.,

2021). Hence, its lack of reliability and scalability as per the above.

The potential use of features extraction frameworks on real-case scenarios also makes several proposed methods under the question of their responsiveness, reliability and scalability towards particular testing experiments (Altaf et al., 2017; Howarth and Rüger, 2004; Liu, Liu, and Zhu, 2020). The consideration of multiple inquires helps in evaluating the consistency of the proposed model. Authors in (Altaf et al., 2017), for instance, considered multiple techniques combinations. However, no experimental setup has been in place to cover multiple scenarios, hence, the lack of sufficient features extraction. Similar experimentation approach has been considered by research proposed in (Howarth and Rüger, 2004), which limited their evaluation mechanism leading to a drawback in considering multiple features combinations, to include HF and DHF, hence the importance of features fusion.

Texture features have been the interest point of several features extraction frameworks proposed in the literature. In particular, several methods have been applied for texture features extraction. Grey-Level Co-occurrence Matrix (GLCM), has been widely used for texture features extraction (Hazarika et al., 2021). GLCM has demonstrated high efficiency in extracting discriminative features. Authors in (Tsai et al., 2017) proposed a GPU based features extraction from MRI images (grey-scale) with the objective of accelerating processing time metric and reducing its complexity. Based-on RoIs localised in the medical image, a set of Haralick features are derived from GLCM including: auto-correlation, dissimilarities, variance, entropy...etc. Despite the high level of efficiency obtained by the suggested method, the work lacks in identifying the complexity of the used dataset which limits the potential of benchmarking their proposed method.

In the same context, a research presented (Rundo et al., 2019) in proposed a new GPU-powered texture features extraction method based on the full dynamics of grey-scale levels . The tested dataset is composed of MRI and CT scans with no specification of related dataset size. This raises doubts about the scalability of these methods and their reliability in real-case scenarios. Recently, a CUDA-powered method for texture features extraction method has been proposed to cover unsupervised analysis of medical images, particularly CT scans

(Rundo et al., 2021). The suggested method relies on the mixture between GLCM and SOM, namely CHASM. The proposed method showed high performances in terms of responsiveness over-passing the pre-suggested methods (Tsai et al., 2017; Rundo et al., 2019). In addition, the proposed approach is based mainly on unsupervised extraction which covers the case of unlabelled dataset. A drawback presented by CHASM of the disregard of coloured dataset which can be a challenging problem when it comes to its possible application as a first or second clinical line tool. Texture features have also been extracted for medical disease detection, for instance, Glaucoma. Authors in (Kavya and Padmaja, 2017) proposed a new framework for Glaucoma detection using texture features extraction. In addition to GLCM, Gaussian Markov Random Field (GMRF) has been applied for texture extraction. The combination GLCM-GMRF reinforced the output result of the final classification task Acc to reach 86%. Despite the high performance of the proposed model, it did not cover the colour features of the used OCT dataset, which in turn limits the method's generalisability. Furthermore, utilising only 50 images for validation is deemed inadequate for thoroughly verifying the effectiveness of the proposed method.

Shape features are also one of the most useful HF features to extract relevant information from medical images, for instance, tumour shape in case of MRI images and Optic Nerve Head (ONH) in case of OCT images (Kavya and Padmaja, 2017; Xiao et al., 2013; Zewail and Hag-ElSafi, 2017; Liu and Shi, 2011; Mingqiang, Kidiyo, Joseph, et al., 2008). A proposed research has considered deformation-based features to construct a more accurate anatomical meaning from the images to represent the brain tumour proposed by Xiao et al. in (Xiao et al., 2013). Their work is based on the use of MRI image, particularly the lateral ventricular part of the brain towards extracting the deformation of the shape features. The method consists of retrieving the lateral ventricular shape, then the estimation of its deformation and finally transforming it into an actual representative feature. One of the advantages of this method is the use of the supervised and unsupervised methods including KNN and conventional Fuzzy c-means clustering (FCM), respectively. Their classification results shown a high Sen of 95.3% in case of supervised method (KNN), and 81.9% in case of unsupervised method (FCM). However, the

drawbacks of their method are that: (1) the exclusion of other features (e.g., texture) in order to improve the classification outcome, (2) the lack of covering colour-based dataset, and (3) the lack of scalability due to a limitation in validating the suggested method with very few cases (i.e., 15 cases).

Shape features have been also considered as key-features of proposed method by authors in (Zewail and Hag-ElSafi, 2017). This sparse contourlet-based extraction approach is composed of mainly two methods including Second Moment Matrix (SMM), and non-subsampled Contourlet representation (NSCT). The combination NSCT-SMM is based mainly on non-maximum suppression and thresholding after the generation of shape features strength. The outcome of the proposed method showed a high Acc of 78.91% and a low MAE to achieve 21.43%. Combining HF features presents a supporting factor for medical applications by providing extra relevant information about the target RoIs (Mutlag et al., 2020; Hazarika et al., 2021). Texture and shape features fusion has been considered in several works, particularly in (Nazir et al., 2021). In fact, authors proposed a new implementation of features extraction from medical images consisting of the extraction of three features levels: (1) key-points, (2) contours, and (3) textures, storing them into a feature vector and highlighting them on the original medical image. The suggested implementation has been tested on three different datasets including MRI, Iris and Bones achieving a classification Acc of 90% which is higher than the aforementioned proposed approaches. That is, features fusion represents a key-stage towards enhancing image content-based retrieval. One of the disadvantages of the method is, again, the scalability and reliability by the omission of different datasets and lack of validation experiments, as well as the lack of extraction of colour features. The latter represents a key-element in several medical applications that are colour-based.

Colour feature-based retrieve is of interest to many recent researches following the advanced engineering technologies in enhancing the medical image scans quality as well as integrating coloured options in image analysis and diagnosis, particularly in case of OCT scanss (Lin, Chen, and Chan, 2009). Traditional features extraction methods, including those to extract HF features, are still facing challenges extracting DHF due to their classical composition.

AI techniques, particularly DL, have shown an interesting enhancement of classification and prediction applications. Dara et al., proposed a DL-based deep features extraction method using CNN (Dara et al., 2018). The latter has been tested along with DBN and MLP on an MRI dataset composed of 69 subjects. CNN presented the most accurate network with 99% Acc. Despite considering the high accuracy, the suggested framework lacks of the reliability factor because of the small number of input sample which might risk causing a thrashing problem, and suffers from achieving scalable factor due to the disregard of unlabelled data. Similar approach has been adopted by (Nazir et al., 2021). The proposed method is based on CNN architecture, particularly DenseNet-100. The model accurately extracts hidden features and results an outstanding classification performance applied on OCT dataset for DMO detection. Despite the existence of colour features, the method eliminated the latter and focused mainly on DHF features. Subsequently, integrating colour features might have increased the final classification outcome. The complementarity of DHF and HF features represents the main contribution of this work. In the following section, a highlight of the identified problems is presented.

4.3 Problems Identified

The main drawbacks identified in existing features extraction methods include responsiveness, scalability, and reliability. Several approaches achieved high accuracy as an initial evaluation of the proposed framework, however many requirements have not been met so far. An efficient features extraction method consists of the consideration of every aspect that can be retrieved from the medical images in order to reflect a certain RoI. This includes texture, shape, and colour (in case of coloured dataset). In addition, hidden features that can be extracted through DL approaches represent additional important features towards having a complimentary feature set. The elimination of one of these features could result in affecting the final efficiency of the medical application. The use of small set of medical images represents an inefficient way to validate a proposed framework. In fact, features provided by a small set of images limits the generalisability of the evaluated method. Add to that, several problems can occur including

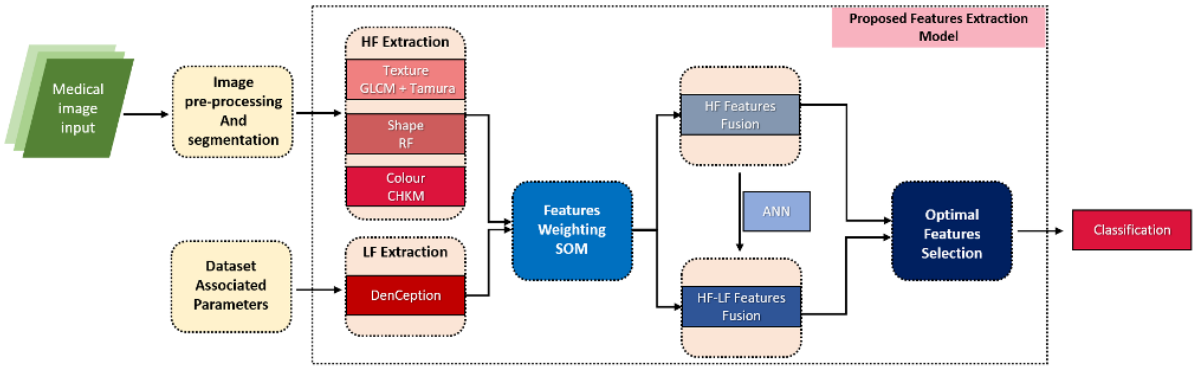


Figure 4.2: Proposed Features Extraction Methodology Framework

the overfitting where the DL model is not capable of successfully classifying the data when it becomes higher than what it has been trained on (small dataset), and under-fitting where the DL network's ability is limited in terms of finding the accurate relationship between the dataset used and the input samples. Thus, the non-scalability of the designed feature extraction model. In turn, this effects its reliability and initial efficient functionality. In this chapter, the addressed problems are as follows:

- Unautomated methods for medical image-based features extraction
- Non-scalability of existing approaches
- The lack of the use of HF and DHF features in a unique framework

4.4 Methodology of the Proposed Features Extraction Model

Feature extraction from image data is a crucial step, particularly, its significant application in case of medical images is considered challenging. The variety and deepness of extracted features represent the key-points in achieving high classification and prediction performances. The methodology presented in this research focuses mainly on the extraction and fusion of HF and DHF features. HF features extraction is based on segmented images whereas DHF features are derived from associated parameters provided along with the input dataset. The proposed features extraction framework is illustrated in Figure 4.2.

The figure illustrates the comprehensive and systematic approach of the proposed features extraction methodology framework. This framework is pivotal in processing and analysing medical images to extract the most relevant features that are critical for accurate classification and diagnosis. Below is a breakdown of each component of the framework to provide an overall understanding of its operation and significance:

- **Medical Image Input and Pre-Processing:** The process begins with the acquisition of medical images, which are the primary data source for this framework. These images undergo pre-processing and segmentation, a crucial step that prepares the raw image data for feature extraction. The pre-processing involves operations such as noise reduction, contrast enhancement, and image normalisation, ensuring that the image data is clean and standardised, which enhances the reliability of the subsequent analysis. Segmentation is employed to isolate ROIs within the images, focusing the feature extraction process on the most relevant areas that are indicative of disease.

- **HF and DHF Feature Extraction:** The framework distinguishes between two types of features: HF and DHF features. HF features capture the fine details and textures within the image. The framework utilises techniques like GLCM and Tamura for texture analysis, RF (Random Forest) for shape feature extraction, and CHKM (Color Histogram of K-Means) for colour features. These features are crucial for identifying subtle variations in image data, which may correspond to different stages or types of diseases. DHF features are extracted using the DenCeption model, a hybrid DL architecture that leverages the strengths of DenseNet-169 and Inception-V4. DHF features represent broader, more abstract patterns in the image data, capturing the overall structure and larger-scale anomalies. This dual approach ensures that both detailed local features and global structural information are considered in the analysis.

- **Dataset Associated Parameters:** In addition to the image data, the framework integrates associated patient parameters (such as age, survival days, and resection status) into the feature extraction process. These parameters provide essential context, enabling the

framework to tailor feature extraction based on patient-specific characteristics, thereby improving the relevance and accuracy of the extracted features.

- **Features Weighting and Fusion:** After HF and DHF features are extracted, the framework proceeds to the weighting and fusion stages. The SOM technique is used to assign weights to the extracted features. This step is crucial for prioritising features based on their importance in the classification task. The weighted features are then fused, ensuring that the most relevant features are emphasised in the final feature set. Afterwards, the framework performs two levels of feature fusion. The first level fuses HF features alone, which consolidates the detailed local information. The second level involves fusing both HF and DHF features, creating a comprehensive feature set that captures both fine-grained details and overarching patterns. This dual fusion approach ensures that the classification model has access to a rich and balanced set of features, enhancing its ability to make accurate predictions.

- **Optimal Features Selection:** Following the fusion process, the framework employs ANN to perform optimal feature selection. The ANN is trained to identify the most informative features from the fused set, discarding redundant or less relevant features. This step is vital for reducing the dimensionality of the data, which not only improves the efficiency of the classification model but also enhances its generalisation capability.

- **Classification:** The final step in the framework is the classification of the processed data. With the optimal set of features selected, the classification model can now accurately distinguish between different classes, such as healthy versus diseased states. The robustness of this framework ensures that the classification model is well-equipped to handle a wide variety of medical imaging data, making it adaptable to different diagnostic tasks.

The proposed framework's blocks, as illustrated in Figure 4.2, will be detailed in the following subsections, providing an in-depth explanation of each component and its role in the overall feature extraction process.

4.4.1 Image Pre-Processing

Image pre-processing represents a major and essential step towards improving features extraction by eliminating unwanted noise and irrelevant regions located in the medical image. The proposed image pre-processing model consists of the following principles:

- Ground truth extraction for data training and testing stage
- Images denoising using block matching and 3Ds filtering (BM3D) method.
- Bias field correction using N4 bias field correction method.

These steps are pivotal in order to enhance the quality of the image considered, from algorithmic perspective, as a matrix of pixels/intensities. It leads into the elimination of non-essential areas that contain unwanted signals which results image quality degradation. In fact, the medical image ground truth is important for the validation of the RoIs segmentation. Denoising of medical scanned images such as OCT, MRI, CT, ...etc is also an important stage towards enhancing the outcome of medical applications including, detection, analysis, and prediction. Subsequently, denoising stage generates clean images with high signal-to-noise ratio as well as high spatial resolution. In this denoising model, block-matching and BM3D method is used to denoise the input samples (Zhao, Hoffman, McNitt-Gray, and Ruan, 2019). Main steps used in BM3D are grouping, 3-dimensional discrete wavelet transformation and wavelet shrinkage. BM3D can remove the noise easily by eliminating it from the group of similar patches. The principle of denoising is to remove the additive noise and invert the blurring at the same time (Kaur, Singh, and Kaur, 2018). This method is called Wiener filter. The latter determines the optimal trade-off between the inverse filtering and the noise smoothing. The N4 bias field correction algorithm is a popular method for correcting low-frequency intensity and the non-uniformity present in the medical image data, known as a bias or gain field.

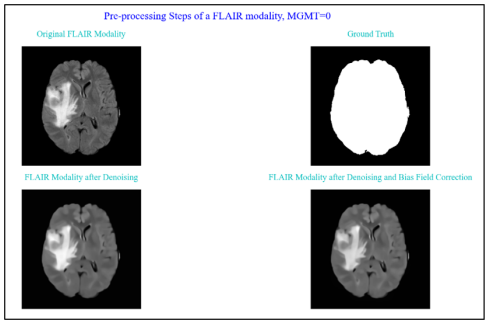
The main purpose of this stage is to ensure that the mask image and the main input image occupy the same physical space to ensure pixel to pixel correspondence. All these steps are complementary towards producing a high-quality input sample that can be effectively processed

and analysed. Figure 4.3 presents an example of image pre-processing applied on MRI image. The figure illustrates the pre-processing steps applied to the FLAIR modality in MRI scans for two subjects: one with a negative MGMT status (Figure 4.3.a) and another with a positive MGMT status (Figure 4.3.b).

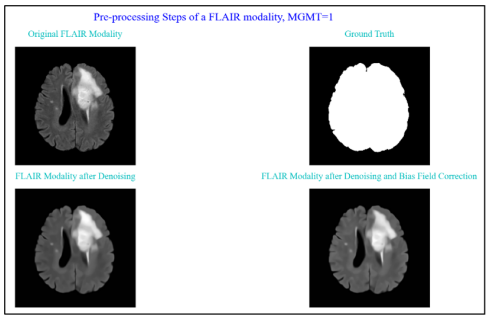
- **Subject with Negative MGMT:** In Figure 4.3.a, the pre-processing steps for a subject with a negative MGMT status are shown. The process begins with the original FLAIR modality, which often contains noise and potential artifacts. The first step involves denoising, which significantly enhances the clarity of the image by reducing noise while preserving essential features. The denoised image is then subjected to bias field correction, which compensates for intensity inhomogeneities that could otherwise distort the analysis. The corrected image shows improved contrast and uniformity, facilitating more accurate feature extraction and subsequent analysis. The ground truth image is also provided for comparison, representing the expected outcome or reference for validating the pre-processing steps.

- **Subject with Positive MGMT:** Similarly, Figure 4.3.b depicts the pre-processing steps for a subject with a positive MGMT status. The original FLAIR modality is first denoised to remove unwanted noise, enhancing the visibility of critical features such as tumour regions. The next step is bias field correction, which further refines the image by addressing any uneven intensities that may obscure the accurate interpretation of the data. This process ensures that the final pre-processed image is optimised for subsequent analysis, particularly in distinguishing between healthy and diseased tissue. The ground truth image is shown alongside the pre-processed images, serving as a benchmark for evaluating the effectiveness of the pre-processing steps.

The pre-processing of FLAIR modalities is a crucial step in medical image analysis, especially in the context of brain tumour detection and characterisation. The figure highlights how denoising and bias field correction are applied to improve image quality and ensure that the images are suitable for further analysis. By enhancing the clarity and consistency of the images,



(a) Subject with Negative MGMT



(b) Subject with Positive MGMT

Figure 4.3: Pre-Processing Result on FLAIR Modality

these pre-processing steps help in achieving more reliable and accurate diagnoses, particularly when distinguishing between different MGMT statuses. The inclusion of both negative and positive MGMT cases underscores the framework's adaptability and effectiveness across varied clinical scenarios.

4.4.2 Image Segmentation and Associated Parameters

Image segmentation characteristics represent the lower level of image characteristics including pixel intensities, RoI, bounding, edges...etc. It is defined by semantic image segmentation through extraction of RoIs of the input samples. Medical image segmentation represents a challenging step due to its deformable characteristics. The aim of semantic segmentation is to partition the image into multiple segments in order to simplify the representation of which makes it more significant and easier to process. The focus of this work is mainly on unsupervised segmentation algorithms. For instance, Markov Random Field (MRF) presents one of the well-used unsupervised segmentation algorithms, in addition to Expectation-Maximization (EPM)

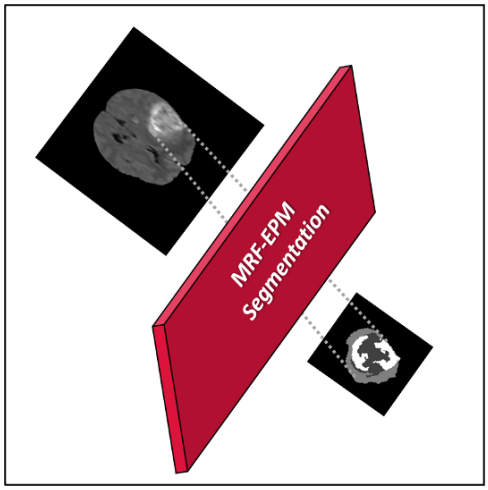


Figure 4.4: MRF-EPM Segmentation: Test Done on a Scan Sample from the BRATS Dataset

algorithm. The combination MRF-EPM iterates the posteriori probabilities and distributions of labelling in case there are no possibilities of the construction of an estimate segmentation model, i.e., no predefined classes. The segmentation process starts with randomly estimating the model parameters, then computing the conditional probabilities of a label given a random image region using naïve Bayes technique. The conditional probabilities are defined as follows (Equation 4.1):

$$P(\frac{\lambda}{r_i}) = \frac{P(\frac{r_i}{\lambda})P(\lambda)}{\sum\limits_{\lambda \in L} P(\frac{r_i}{\lambda})P(\lambda)} \tag{4.1}$$

where L represents the set of possible labels, is the given label, and ri is the region of features. Finally, MRF-EPM iterative algorithm uses the output of proceeding step in order to calculate the priori estimate of a given label, $\lambda \in L$. The computation involves a hidden estimate of the number of labels (β), knowing that the actual number of total labels is unknown. The priori estimate is defined as the following (Equation 4.2):

$$P(\lambda) = \frac{\sum\limits_{\lambda \in L} P(\frac{\lambda}{r_i})}{\mid \beta \mid} \tag{4.2}$$

Figure 4.4 shows the result of MRF-EPM algorithm applied on the MRI image presented in Figure 4.2. The method successfully segments the RoIs in the original image. Resulted segmented images will be used as the input data for the feature extraction model.

4.4.3 Proposed Hybrid High-Level Features Extraction Model

HF features are mainly defined by the features that can be interpreted by human brain. This is presented in the form of spectral features including texture, shape, and colour.

Texture Features Extraction

Texture features are based on the collection of image regions. It generally refers to a specific region within the image. Referred RoIs provide other important features such as shape and colour which will be discussed in the subsequent sections. Texture features extraction methods are generally divided into two main categories based on: (1) spatial relationship between regions, and (2) primitive attributes. The former includes (i) primitive region types presented as numbers and (ii) spatial organisation covering functional, structural, and statistical features. Primitive attributes texture features category focuses mainly on (i) grey-scale and (ii) geometrical attributes. The latter covers the shape, area, ...etc, whereas, grey-scale attributes enclose average and extremum. Texture features generally highlight discriminative features that represent key-features in disease detection and prediction applications. The focus of the proposed texture features extraction method is based on statistical features as the following:

- First- and second- order features including: contrast, entropy, angular second moment, and homogeneity.
- Additional features to include Coarseness and directionality.

Texture features reflect changes that might happen in the medical image due to disease detection and progression which in turn affects the pixels intensities. Several methods can be applied to extract texture features, for instance, GMRF (Kavya and Padmaja, 2017), SOM (Rundo et al., 2021), GLCM (Rundo et al., 2019; Kavya and Padmaja, 2017; Altaf et al., 2017), and Tamura (Mutlag et al., 2020; Umamaheswari, Bhavani, and Sikamani, 2018) approaches. Multiple performance parameters of the aforementioned methods have been reported in the literature to include classification accuracy, processing speed-up and other parameters. Despite

of demonstrating a valuable speed-up performance of 0.3 times, SOM method requires additional parallel computing platform that allows it to use certain types of GPUs which represents a limitation in case of resource-limited setup environments. GLCM, on the other hand, showed an independent processing speed-up to reach 19.5 times due to its processing optimisation and image handling. It also overpassed the classification Acc of GMRF to achieve 86%. By considering the optimal pixel direction and orientation, Tamura's features application showed quite an interesting classification Acc to reach 96% and 3.43% of the mAP retrieval. As per the above, in this study a combination of GLCM and Tamura is proposed.

First and Second Order Texture Features: GLCM Technique

GLCM technique is based mainly on pixels intensity and related changes. The major advantage of GLCM is that the co-occurring groups of pixels are spatially linked in multiple directions by referencing to two different factors including distance and angular second moment relationships. GLCM also highlights the busy texture regions defined by a very rapid changes of one-pixel intensity compared to its neighbours. Thus, it results a high intensity alteration of the related special frequencies. The GLCM algorithm first quantises the segmented input by specifying the value of each pixel intensity. The quantisation is specified based on a range of grey-scale included in the range of [2:256]. Second, it creates the co-occurrence matrix sized (n*n), where n presents the number of levels used in the quantisation step. The creation of co-occurrence matrix ($GLCM_f$) is based on the calculation of the number of occurrences of a pixel (p), located at (i,j) coordinates, in a pre-defined iterative window that covers the surrounding pixels. The steps are detailed as follows:

- Set p the sample considered for calculation.
- Set S the group of neighbour pixels surrounding p. The selected group is done under a centred window having as length, and height values in [3:999] interval.
- Each element (i,j) of GLCM matrix, based on S, is defined as (Eq 4.3):

$$GLCM(i,j) = occ(i,j) \tag{4.3}$$

where i,j are the i^th and j^th pixels intensities $\in$ [0:n-1], and occ() is the function representing the time of occurrence of i,j in S based on multiple direction and distance relationships. Which means (Eq 4.4):

$$GLCM(i,j) = \sum_k occ(i,j) \tag{4.4}$$

where k is the total occurrence of (i,j) in the centred window.

- Construct the symmetrical matrix of GLCM and add it to the co-occurrence matrix itself (Eq 4.5):

$$GLCM_f = GLCM + GLCM_s \tag{4.5}$$

where $GLCM_s$ is the symmetric matrix and $GLCM_f$ is the final co-occurrence matrix which results the following equation (Eq 4.6):

$$GLCM_f(i,j) = \sum_k occ(i,j) + occ(j,i) = 2\sum_k occ(i,j) \tag{4.6}$$

where occ(i,j) is the number of occurrences of intensity i as a reference in relationship with j, and occ(j,i) is the number of occurrences of intensity j as a reference in relationship with i.

- Normalisation of $GLCM_f$ (Eq 4.7):

$$GLCM_f(i,j) = \frac{\sum_k occ(i,j)}{M} \tag{4.7}$$

where M is the number of total elements, $M > 0$.

- Calculate first and second order texture features as the following:
  - Angular Second Moment (ASM): ASM is known also as Energy feature, denoted

f_ASM, it is defined as the squared elements of $GLCM_f$, as follows (Eq 4.8):

$$f_{ASM} = \sum_{j\in n}\sum_{j\in n} GLCM_f(i,j)^2 \tag{4.8}$$

– Entropy (E): It is determined as the quantification of randomness to be employed in distinguishing the texture of the segmented input sample, as follows (Eq 4.9):

$$f_E = -\sum_{i\in n}\sum_{j\in n} GLCM_f(i,j) * log(GLCM_f(i,j)) \tag{4.9}$$

– Contrast (C): Contrast is defined as the value of density contrast reference pixels and surrounding pixels, as follows (Eq 4.10):

$$f_C = \sum_{j\in n}\sum_{j\in n} (i,j)^2 GLCM_f(i,j) \tag{4.10}$$

where $GLCM_f$ (i,j) equals to pixel at the (i,j) location.

– Homogeneity (H): H is defined by approximately measure the $GLCM_f$ elements distribution to $GLCM_f$ diagonal (Eq 4.11):

$$f_H = \sum_{j\in n}\sum_{j\in n} \frac{GLCM_f(i,j)}{1+\mid j-i \mid} \tag{4.11}$$

Additional Texture Features

Tamura is also one of the well-used quantitative texture features extraction methods. It is based mainly on human visual perception and it represents an immense potential in image representation. Tamura provides a set of texture features including: roundness, directionality, line-likeness, regularity, coarseness, as well as contrast texture features. Ideally, Tamura's texture features present complementary features to those extracted through GLCM approach. The main drawbacks let to combining GLCM and Tamura are as follows:

- GLCM is a sparse matrix, containing many zero elements, which causes an increase in computational time and resource (Kaur, Singh, and Kaur, 2018; Baid et al., 2021).

- Tamura performs inefficiently in case of generic (non-homogeneous) images.

The proposed feature extraction approach includes Coarseness (Coa) and Directionality (Dir) as additional features. Tamura's discriminative features are defined in the following:

- Coarseness: Coa is defined by iteratively find the largest size in which the tissue is present through different patterns at multiple scales. The granularity measurement is done by calculating, for each pixel (i,j), six averages for a window of size $2^Z * 2^Z$, where Z $\in$[0:5], surrounding the pixel defined as follows (Eq 4.12):

$$Coa_z = \sum_{k=i-2^{Z-1}-1}^{i+2^{Z-1}} \sum_{t=j-2^{Z-1}-1}^{j+2^{Z-1}} \frac{pix(k,t)}{2^{2Z}} \tag{4.12}$$

  where pix(k,t) is the pixel intensity at location (k,t). Iteratively, at each pixel, calculation of non-overlapping neighbours defined by the absolute difference $A_Z(i,j)$ in both relationships: Vertically (V) and Horizontally (H) as follows (Eq 4.13a and 4.13b):

$$A_{Z,V}(i,j) = \mid Coa(Z,V)(i,j+2^{Z-1}) - Coa(Z,V)(i,j-2^{Z-1}) \mid \tag{4.13a}$$

$$A_{H,V}(i,j) = \mid Coa_{H,V}(i+2^{Z-1},j) - Coa_{Z,H}(i-2^{Z-1},j) \mid \tag{4.13b}$$

  Finally, considering either direction (V or H), calculation of the value of Z is processed in order to maximise $A_{Z,V}(i,j)$ or $A_{Z,H}(i,j)$, respectively. The function is defined as follows (Eq 4.14):

$$S_{Z,BEST}(i,j) = 2^Z \tag{4.14}$$

  resulting the final coarseness feature equation (Eq 4.15):

$$f_{Coa_Z} = \frac{Coa(i,j)}{S_{Z,BEST}(i,j)} \tag{4.15}$$

- Directionality Dir is defined by devolving the existence of any directional pattern in an image by measuring the overall degree of directivity (vertically, horizontally, or diago-

nally). This feature reflects the consistency of the region being processed. Dir consists in calculating the edge histogram (H_Dir). Dir texture feature is defined as follows (Eq 4.16):

$$f_{Dir} = 1 - N * N * m * \sum_{k=1}^{m} \sum_{\theta \in \psi_k} (\theta - \theta_k)^2 * H_{Dir}(\theta) \tag{4.16}$$

where:

- N: normalisation factor
- θ: quantisation angular position constructed by counting the edges of pixels with associated angles directions.
- m: number of peaks
- ψ_k: angles window associated to the k^{th} peak.

The remaining Tamura texture features are of importance but not considered in this method. The texture features extraction is implemented in Algorithm 1 and comprise two main steps:

- Step 1: Calculation of GLCM f matrix based on the occurrences (occ()) of pixels at a location (i,j) in the surrounding window S.
- Step 2: Calculation of Tamura texture features based on the best pixel direction and orientation at location (i,j) in the surrounding window S.

Algorithm 1: Texture Features Extraction

Data: RoIs (Region of Interests)

Result: $f_{ASM}, f_E, f_C, f_H, f_{Coa_{Z,BEST}}, f_{Dir}$

$n \Leftarrow$ number of levels

$M \Leftarrow n^2 > 0$

for $i \in \{0,\ldots,n\}$ **do**

  for $j \in \{0,\ldots,n\}$ **do**

    for $k \in \{0,\ldots,M\}$ **do**

      $GLCM_f(i,j) \Leftarrow \frac{2}{M}\sum occ(i,j)$

      $f_{ASM} \Leftarrow \sum\sum GLCM_f(i,j)^2$

      $f_E \Leftarrow -\sum\sum GLCM_f(i,j) * \log(GLCM_f(i,j))$

      $f_C \Leftarrow \sum\sum (i,j)^2 * GLCM_f(i,j)$

      $f_H \Leftarrow \sum\sum \frac{GLCM_f(i,j)}{1+|j-i|}$

$pix(i,j) \Leftarrow$ intensity value of the pixel at location (i,j)

$S_Z \Leftarrow 2^{2Z}$ where $Z \in [0:5]$

$N \Leftarrow$ normalisation factor

$\theta \Leftarrow$ quantisation angular position

$m \Leftarrow$ number of peaks

$\psi_k \Leftarrow$ angles window associated with the k^{th} peak.

$H_{Dir} \Leftarrow$ edge histogram

$M_w \Leftarrow$ measurement window

for $i,j \in \{0,\ldots,n\}$ **do**

  for $k = i - 2^{Z-1} - 1$ *to* $i + 2^{Z-1}$ **do**

    for $k = j - 2^{Z-1} - 1$ *to* $j + 2^{Z-1}$ **do**

      $Coa_Z \Leftarrow \sum \frac{pix(i,j)}{M_w}$

$A_{Z,V}(i,j) \Leftarrow \mid Coa_{Z,V}(i, j+2^{Z-1}) - Coa_{Z,V}(i, j-2^{Z-1}) \mid$

$A_{Z,H}(i,j) \Leftarrow \mid Coa_{Z,H}(i+2^{Z-1}, j) - Coa_{Z,H}(i-2^{Z-1}, j) \mid$

if $A_{Z,V}(i,j) > A_{H,V}(i,j)$ **then**

  $S_{Z,BEST}(i,j) \Leftarrow S_{Z,V}$

  $f_{Coa_{Z,V}} \Leftarrow \frac{Coa_{Z,V}}{S_{Z,BEST}}$

else

  $S_{Z,BEST}(i,j) \Leftarrow S_{Z,H}$

  $f_{Coa_{Z,H}} \Leftarrow \frac{Coa_{Z,H}}{S_{Z,BEST}}$

for $k = 1$ *to* m **do**

  for *each angle* $\theta \in \psi_k$ **do**

    $\rho \Leftarrow \sum(\theta - \theta_k)^2 * H_{Dir}(\theta)$

$f_{Dir} \Leftarrow 1 - N \cdot m \cdot \rho$

The hybrid composition of the proposed texture features extraction involves multiple interpretation levels of the input image which gives the system a better understanding of the image composition at different RoIs. GLCM-Tamura combination is considered as a booster to the whole feature extraction framework by: (1) speeding-up the processing time, (2) optimising the use of computational resources, and (3) increasing the efficiency of the final classification aiming to overpass the performance of existing methods as shown in Table 4.1.

Table 4.1: Examples of Existing Texture Features Extraction Methods

Methods	**Classification Accuracy**	**Processing Speed-up**	**Mean Average Precision Retrieval**
SOM (Rundo et al., 2021)	-	10.03 times	-
GLCM (Rundo et al., 2019)	-	19.5 times	-
GLCM-MRF (Kavya and Padmaja, 2017)	86%	-	-
GLCM (Altaf et al., 2017)	79.8%	-	-
Tamura (Mutlag et al., 2020)	96%	-	-
Tamura (Zhao et al., 2019)	-	-	3.43%

Shape Features Extraction

Based on the aforementioned related works, shape features extraction is mainly based-on geometry features including: area, slope, perimeter, centroid, irregularity index, equivalent diameter, convex area, and solidity...etc. Table 4.2 summarises the benchmarking of existing methods.

Table 4.2: Examples of Existing Shape Features Extraction Methods

Methods	**Acc**	**Sen**	**Spe**	**MAE**
Fourrier Descriptor (Liu and Shi, 2011)	80%	-	-	-
RF-FCM (Xiao et al., 2013)	81.9%	38.9%	99.7%	-
SMM-NSCT (Zewail and Hag-ElSafi, 2017)	78.9%	-	-	21.43%
GLCM- voxel-based morphometry (VBM) (Xiao et al., 2017)	91.4%	99%	83.33%	-
SVM-RFE (Madusanka et al., 2019)	86.61%	75%	77.78%	-
VBM (Janakasudha and Jayashree, 2020)	93.8%	-	-	-

Towards efficiently using shape features, selected approaches need to meet essential key-points including: (1) identifiability, (2) transition, rotation, and scale invariance, (3) affine invariance, (4) noise resistance, (5) occultation invariance, (6) statistically independent, and (7) importantly reliable. Shape features extraction approaches can be categorised as the following:

- Counter-based methods
- Region-based methods
- Space and transform domain-based methods Information preserving and non-information preserving based methods

As per the results presented in (Liu and Shi, 2011), Fourier descriptor overpassed statistical descriptors by achieving over 80% Acc. In fact, Fourier descriptors are highly insensitive to translation, rotation, scale changes as well as the starting processing point. It has shown high

performances in case of identified objects (human face, vehicles ...etc). However, in case of medical imaging the shape of different RoIs in the input sample changes through time progression, age and gender factors as well. Therefore, this study considers the region focus as the main shape feature extraction approach by calculating the coordinates of all points belonging to a particular RoIs. It is defined as the following (Equation 4.17, 4.18a.a and 4.18b.b):

$$f_{RF} = (\bar{x}, \bar{y}) \tag{4.17}$$

where:

$$\bar{x} = \frac{1}{A} \sum_{(x,y) \in RoI} x \tag{4.18a}$$

$$\bar{y} = \frac{1}{A} \sum_{(x,y) \in RoI} y \tag{4.18b}$$

where A is the region area: $A = \sum_{(x,y) \in RoI} 1$.

Colour Features Extraction

This feature is based on coloured medical images. Several features extraction methods have been used in the literature. Colour feature extraction approaches includes global descriptors defined when the whole image is considered, and local descriptors when separated portions of the image are considered. CHKM, and Zernike chromaticity derived from chromaticity approach are considered highly robust colour features extraction methods. However, it is not the case of colour histogram method. Four main specifications are essential to consider a method as efficient and accurate:

- Storage space
- Scalability
- Computational time required
- Rotation invariance

Based on the aforementioned criteria, CHKM has been applied for colour features extraction. CHKM considers 224 different colours possibilities. The main process conveyed by CHKM is to select a colour, denoted cp, from 224 possibilities that reassemble the best to a particular pixel colour and update the latter with cp. This step is applied on each pixel towards classifying all the pixels of an image into k clusters. The outcome of it is defined by the mean of all pixels in each cluster. The final output of CHKM feature is (Equation 4.19):

$$f_{CHKM} = \frac{N_K}{N} \tag{4.19}$$

Where N is the total number of pixels localised in the image, and NK is the total number of pixels belonging to cluster K. This method efficiently shortens image retrieval time and improves its performance. Moreover, CHKM demonstrates a less computational time factor, a high robustness to noise and displacement invariance. Algorithm 2 demonstrates the proposed shape and colour features extraction model.

Algorithm 2: Shape and Colour Features Extraction

Data: RoIs (Region of Interests)

Result: f_{RF}, f_{CHKM}

Step1: Calculation of Region focus shape features based on the region area A

$(x,y) \Leftarrow$ coordinates of the pixel $\in$ A

for $(x,y) \in RoIs$ **do**

    $A \Leftarrow \sum 1$

for $(x,y) \in RoIs$ **do**

    $\bar{x} \Leftarrow \frac{\sum x}{A}$

    $\bar{y} \Leftarrow \frac{\sum y}{A}$

$f_{RF} \Leftarrow (\bar{x}, \bar{y})$

Step2: Calculation of colour histogram of K-mean where K represents the number of clusters

$N \Leftarrow$ total number of pixels

$N_k \Leftarrow$ total number of pixels in cluster k

$c_p \in 2^{24}$ colours possibilities

$c_{pixel} \Leftarrow$ current pixel colour

for $c_{pixel} \in cluter\ k \in K$ **do**

    if $c_{pixel} \in 2^{24}$ **then**

        $c_{pixel} \Leftarrow$ best matching colour (c_p)

$f_{CHKM} \Leftarrow \frac{N_K}{N}$

4.4.4 Deep Hidden Features Extraction Method

The idea of this work is to extract, in addition to HF features, DHF features. The latter features are difficult to be interpreted by human brain or visually identified. Hence, it can meet the computer understanding and can be extracted by deep networks, particularly DL approaches. Recently, DL has been used in several applications, including detection, classification, predic-

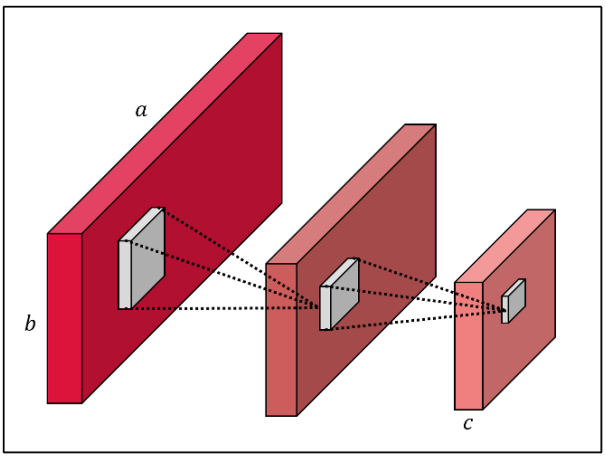


Figure 4.5: Image Dimensionality Reduction Through Convolutional Layer where (a) Feature Width, (b) Feature Height, (c) Number of Channels

tion...etc. Moreover, DL is being used for deep features extraction, particularly, CNN frameworks. The working principle of CNN is to extract features maps (FMs) of each input layer, for instance, the input of nth layer if the FMs extracted from the (n1)th layer. The shape of the input layer in CNN is defined as N*N*M, where N is the size of the FMs, and M represents the total number of channels considered. Figure 4.5 illustrates image size reduction using convolutional layers. DenCeption model, defined in the previous chapter, will be applied to extract DHF features. Resulted features will then be passed to the features weighting block of the proposed framework.

4.4.5 Features Weighting

Features Initialisation and Normalisation

Towards determining the approximate optimal degree of influence of each extracted feature, weighting presents a crucial step for relevant features selection. The weighting technique used in this work is to assign random initial weights. Let F be the matrix of extracted features and W is the associated weights vector defined as follows (Equation 4.20, 4.21):

$$F = [f_{ASM}, f_E, f_C, f_H, f_{Coa}, f_{Dir}, f_{RF}, f_{CHKM}, f_{DHF}] \tag{4.20}$$

$$W = [W_{ASM}, W_E, W_C, W_H, W_{Coa}, W_{Dir}, W_{RF}, W_{CHKM_f}, W_{DHF}] \tag{4.21}$$

After randomly assigning weights to each particular feature, a feature normalisation is applied which produces, subsequently, a normalised features weights having as values in the range of [0,1]. This conveys the following relationships (Equation 4.22):

$$F(x_i, w) = \sum_{j=1, w_j \in w}^{k} w_j x_{ij} \tag{4.22}$$

where w_j is the weight associated to x_{ij} feature $\in$ k , and k is the number of features.

Weights Regularisation

Several weighting techniques are introduced in the literature including, LR, RF classifier, Bayesian linear model...etc. The objective of considering each feature's importance is to adjust the allocated weights during the network's training process. The efficiency of each aforementioned technique is mainly linked to the size of the dataset being trained which most likely can cause overfitting, underfitting and vanishing problems, as mentioned earlier. The pre-definition of dataset classes is also an essential requirement for most of these techniques. To overcome this challenge, it is essential to consider using unsupervised learning approach. In this context, utilising SOM for weight regularisation offers distinct advantages over other ML models due to its unique characteristics. SOM excels in producing low-dimensional representations of high-dimensional data, making it ideal for weight regularisation tasks (Nazir et al., 2021).

Unlike traditional models, SOM utilises competitive learning, enabling it to efficiently capture complex patterns and relationships in the data without the need for labelled examples. The topological map created by SOM groups similar input data points together, facilitating effective weight generalisation by identifying common patterns and features within the data without explicit supervision (Khacef, Rodriguez, and Miramond, 2020). Additionally, SOM's competitive learning mechanism and neighbourhood relationships allow it to adapt to the underlying data distribution and capture intricate structures present in the input space, enhancing its capability for weight regularisation compared to other ML models. Initially, in the learning phase, SOM

associates W, as the random input weights vector with the artificial neurons, namely, units of the network. Then, each input feature vector $f \in F$ is presented to all units in the SOM. The unit with most similar weights to the input vector becomes the best matching unit, namely BMU. Based on the Euclidian distance, BMU is defined as follows (Equation 4.23):

$$BMU = arg_i min \| f - w_i \| \tag{4.23}$$

Once the BMU is calculated, the weight vector is updated as folows (Equation 4.24):

$$w_i(k+1) = w_i(k) + \delta(k)\Delta_i(BMU,k)(f - w_i(k)) \tag{4.24}$$

The SOM training iterations conclude when all features have been assigned updated weights. The feature vector with higher weight represents the feature with higher importance and vice-versa. The closer w_i is to zero, the more irrelevant the related feature is.

4.4.6 Features Fusion

Towards constructing a more robust features extraction outcome, capable of efficiently using multiple types of medical image and can lead to a high classification and prediction performances, the purpose of the proposed framework is to combine the HF features including those part of texture, shape, and colour, as well as fusing HF and DHF features as a following step. A set of experiments will be held to identify the optimal features combination that will feed into the disease classification stage.

High-Level Features Fusion

The first combination is presented by fusing texture and shape features considering their obtained weights from SOM. Therefore, there is no need to design a linear model with a fixed proportion and iteratively determining its value in order to update the fused features. Avoiding

that, the features fusion is presented as the following (Equation 4.25):

$$F_{texture-shape} = max(0, \sum_{i}^{k} w_i f_{texture} + \sum_{j}^{m} w_j f_{shape} + b_1) \tag{4.25}$$

where k is the number of texture features, m is the number of shape features, and b_1 is the bias. Same operation is applied for other considered combinations including: (i) shape-colour, (ii) texture-colour, and (iii) texture-shape-colour, defined as the following (Equation 4.26, 4.27, 4.28):

$$F_{shape-colour} = max(0, \sum_{i}^{m} w_i f_{shape} + \sum_{j}^{n} w_j f_{colour} + b_2) \tag{4.26}$$

$$F_{texture-colour} = max(0, \sum_{i}^{k} w_i f_{texture} + \sum_{j}^{n} w_j f_{colour} + b_3) \tag{4.27}$$

$$F_{texture-shape-colour} = max(0, \sum_{i}^{k} w_i f_{texture} + \sum_{i}^{m} w_i f_{shape} + \sum_{j}^{n} w_j f_{colour} + b_4) \tag{4.28}$$

where n is the number of colour features and (b_2, b_3, b_4) are the bias considered for shape-colour, texture-colour, and texture-shape-colour fusion operation, respectively. The updated weights are obtained by using feed-forward ANN. The resulted updated weights $W^{'}$ and features combination $F^{'}$ vectors are defined as follows (Equation 4.29, 4.30):

$$W^{'} = [W^{'}_{f_{texture-shape}}, W^{'}_{f_{shape-colour}}, W^{'}_{f_{texture-colour}}, W^{'}_{f_{texture-shape-colour}}] \tag{4.29}$$

$$F^{'} = [f_{texture-shape}, f_{shape-colour}, f_{texture-colour}, f_{texture-shape-colour}] \tag{4.30}$$

HF and DHF Features Fusion

At this stage the optimal HF features combination is considered. Let $Foptimal = f_{i,j}$ where i,j are the optimal selected HF features fusions $\in$ F. Following the fusion strategy applied for HF

features, $F_{optimal}$ and DHF combination is defined as the following (Equation 4.31):

$$F_{F_{optimal}-DHF} = max(0, \sum_{i}^{t} w_i^{'} F_{optimal} + \sum_{j}^{s} w_j f_{DHF} + b_5) \tag{4.31}$$

where t is the number of optimal features fusion, having as possible values k+m, m+n, k+n, or k+m+n, s is the number of DHF features, b_5 is the bias and w is the optimal features fusion weight vector. Updated weights are obtained using ANN.

4.5 Dataset

The selection of the BRATS MRI dataset (grey-scale, unlabelled) and the Retinal dataset (coloured, labelled) for training, testing, and validating the proposed feature extraction framework was a strategic decision driven by the need to ensure the model's robustness, adaptability, and generalisability across diverse medical imaging scenarios. Below is a thorough justification for the use of these two datasets.

4.5.1 BRATS MRI Dataset

Complexity and Realism in Medical Imaging

The BRATS MRI dataset, despite being unlabelled and composed of grey-scale images, provides a highly relevant and challenging environment for testing feature extraction models in the context of brain tumour analysis. The dataset includes multiple imaging modalities: T1, T1Gd, T2, and T2-FLAIR, acquired from different clinical settings, which introduces variability in imaging protocols and scanner characteristics. This variability closely mirrors real-world clinical scenarios where models must operate effectively across different imaging conditions without relying on pre-existing labels.

Benchmarking and Evaluation

BRATS is widely recognised as a benchmark in the medical imaging community, particularly for brain tumour segmentation and analysis. Utilising this dataset allows the proposed framework to be evaluated against established standards in the field, providing a clear indication of its performance in a critical area of medical diagnostics. The absence of labels in BRATS challenges the model to identify and extract meaningful features without the guidance of annotated data, testing its ability to generalise from raw, unlabelled inputs.

Enhancing Model Robustness

Incorporating BRATS MRI data in the training and validation phases is crucial for ensuring that the feature extraction framework can handle the inherent complexities of MRI scans, which are often characterised by subtle variations in intensity and texture. The model's ability to effectively process and analyse grey-scale images without labels demonstrates its robustness and adaptability, essential traits for deployment in diverse clinical environments where labelled data may not always be available.

Validation of Feature Extraction

The use of BRATS in the validation phase of the framework is particularly important because it allows the model to be tested on high-dimensional, unlabelled medical imaging data. This process helps to confirm that the extracted features are both relevant and useful for subsequent tasks, such as segmentation and classification, even in the absence of explicit labels. The validation against this challenging dataset reinforces the framework's capacity to generalise its feature extraction capabilities to other complex and unlabelled datasets.

4.5.2 Retinal Dataset

Detailed Labelling and Clinical Relevance

The Retinal dataset, composed of coloured and meticulously labelled Fundus images, serves as a crucial component in training and validating the proposed framework, particularly in the context of ophthalmological diagnostics. The detailed labelling of the dataset, which includes annotations for various stages of DR, provides a rich source of data for training the model to recognise and differentiate between normal and pathological conditions with high accuracy.

Complementary to BRATS

The Retinal dataset complements the BRATS MRI dataset by introducing a different type of medical imaging—coloured Fundus scansthat require the model to handle a completely different set of challenges, such as colour differentiation and higher image complexity. This diversity in data types ensures that the feature extraction framework is not overly specialised for a single imaging modality but is instead capable of adapting to a wide range of medical images.

Enhancing Generalisability

The use of a labelled dataset like the Retinal dataset allows for the training and fine-tuning of the model in a supervised learning context, which is essential for enhancing the model's accuracy and performance. The detailed labels provide ground truth data that the model can use to learn the correct associations between image features and clinical outcomes. This training process enhances the model's ability to generalise its learned features to other labelled datasets and real-world clinical settings.

Validation and Testing

Incorporating the Retinal dataset into the testing and validation phases allows for a rigorous evaluation of the feature extraction framework's effectiveness in a well-defined, labelled environment. This evaluation is crucial for assessing how well the model can extract relevant

features that correspond to clinically significant outcomes. The validation against labelled data also provides a benchmark for measuring the framework's performance in a controlled setting, ensuring that it meets the high standards required for clinical applications.

4.5.3 Combined Justification

Diverse Imaging Modalities

By using both the BRATS MRI and Retinal datasets, the research ensures that the proposed feature extraction framework is capable of processing and analysing a wide variety of medical images, ranging from grey-scale, unlabelled MRI scans to coloured, labelled Fundus images. This diversity is critical for developing a model that is both versatile and reliable across different medical fields.

Comprehensive Model Evaluation

The combination of these datasets allows for a comprehensive evaluation of the feature extraction framework across different stages of model development: training, testing, and validation. The BRATS dataset challenges the model to function without labels, testing its robustness, while the Retinal dataset allows for supervised learning and detailed validation, enhancing the model's overall accuracy and generalisability.

Addressing Multiple Clinical Needs

By incorporating datasets from different medical disciplines (neurology and ophthalmology), the research addresses a broader range of clinical needs, demonstrating the framework's potential utility in various healthcare contexts. This approach not only validates the framework across different datasets but also highlights its capability to contribute to multiple areas of medical diagnostics.

4.6 Conducted Experimentation and Research Evaluation Mechanism

To evaluate the effectiveness of the proposed features extraction method in comparison with existing works, an evaluation scheme of various measurement parameters is considered essential. For this purpose, two main experiments will be conducted to evaluate the capability of the proposed framework in handling different dataset cases including labelled and unlabelled data. The experiments conducted aimed to establish rigorous conditions that would thoroughly assess the efficacy of the feature extraction process. Additionally, the design of these experiments was inclusive of diverse dataset representations, covering both grey-scale and coloured medical images as inputs. The details of these experiments are structured as follows:

- Experiment 1 (Exp 1): involves grey-scale and unlabelled dataset, which will be conducted using BRAST dataset.
- Experiment 2 (Exp 2): covers coloured and labelled dataset, which will be carried out using the Retinal dataset.

The described experiments were conducted to: (1) assess if the proposed approach meets the established criteria for HF and DHF feature extraction and integration where an evaluation against its variants will be conducted, and (2) confirm its effectiveness in accurately performing feature extraction in comparison with benchmarking methods. To deepen the assessment of the proposed framework, the evaluation process broadens to encompass four critical criteria of evaluation to include responsiveness, adaptability, scalability, and reliability. This suggested evaluation mechanism is named RASR. These aspects will be explored following the execution of the two initial experiments and will be applied to both the variations of the proposed framework and the benchmarking methods. RASR's four assessment criteria are defined as follows:

- Responsiveness: refers to the duration needed to precisely identify pertinent features in response to dataset modifications. Essentially, it measures the efficiency of a feature ex-

traction model to process various types of medical images within a specified time frame, taking into account the parallel processing capabilities.

- Adaptability: defines the flexibility of a model to independently identifying essential features without reliance on external input, functioning efficiently in an unsupervised context.

- Scalability: refers to a model's capacity to effectively process medical images of any kind, regardless of their type, size, or complexity. It indicates the method's capability to handle datasets of varying dimensions, efficiently addressing both overfitting and underfitting concerns. Scalability is the attribute that allows the feature extraction process to accommodate the expanding volume of data seamlessly, thus maintaining consistent performance and responsiveness.

- Reliability: covers the model's ability to continue accurate processing despite the presence of faults. Should a bug arise at any stage, it will not impact the other components.

Each criterion of the RASR evaluation mechanism varies to include: (1) High, (2) Good, (3) Moderate, and (4) Passable, and (5) Low levels. Towards validating the accurate functionality of the proposed features extraction framework, several features extraction block variants will be involved in the testing mechanism to include: (1) HF only, (2) DHF only, and (3) HF-DHF fusion, denoted as Block1, Block2, and Block3 respectively. This will enable the validation of the importance of HF and DHF features fusion in enhancing the classification results. Table 4.3 summarises the different testing cases of Block1.

Table 4.3: Block 1 Variants - HF Fusions

Case	**Texture**	**Shape**	**Colour**
1	X	X	-
2	X	-	X
Continued on next page			

Table 4.3: Block 1 Variants - HF Fusions (Continued)

Case	Texture	Shape	Colour
3	-	X	X
4	X	X	X

To evaluate the classification outcomes of tested variants and benchmarking methods against the proposed method, a set of quantitative performance metrics is considered including: Sen, Spe, Acc, and MAE.

4.7 Results and Discussion

4.7.1 Proposed Method Versus its Variants

The testing of Block1, Block2, and Block3 was performed following Exp1 and Exp2 as mentioned above. For this purpose, each experiment served as a different testing approach using different medical imaging type to include MRI and Fundus as part of BRATS and Retinal datasets. Firstly, the testing of Block1 will only reflect the importance of HF features within the proposed framework where four different cases will be tested, optimised, and then classified using SVM. Secondly, Block2 will only focus on the contribution of DHF features in the final decision of the classifier model. Finally, the testing of Block3 will serve as the validation of the proposed framework where it merges the optimal combination of HF and DHF features ($F_{optimal}$) towards enhancing the final classification outcome.

BRATS Dataset: Grey-scale and Unlabelled Case

Table 4.4 presents the obtained performance metrics of the individual variants blocks to include Block1 and Block2.

Table 4.4: Individual Blocks Variants Testing using Performance Metrics: BRATS Dataset

Testing Block	**Processing Time (h:mm:ss)**	**Sen**	**Spe**	**Acc**	**MAE**
Block1 – HF Features Only					
Case 1 Features	8:15:00	73%	70%	72.4%	0.28
Case 2 Features	7:33:00	70%	68%	69.6%	0.31
Case 3 Features	6:44:00	68%	67%	67.8%	0.33
Case 4 Features	11:20:00	75%	71%	74.2%	0.26
Block2 – DHF Features Only	3:40:00	67.2%	53%	64.36%	0.37

In case of HF only features, the performance metrics indicate that the processing time varies significantly. In fact, Texture-Shape combination takes the longest time to reach more than 11 hours, but results in the highest Acc of 74.2% and lowest MAE of 0.26 among HF-only cases. This suggests that while the Texture-Shape-Colour feature set provides the most balanced performance, the trade-off in terms of computational efficiency is notable. On the other hand, the application of DHF only features resulted from DenCeption showed a significantly shorter processing time of less than 4 hours. Nevertheless, DHF-only features resulted the least Acc of 64.36% and the highest MAE of 0.37. This indicates that, despite the lower computational cost, relying solely on DHF may not be sufficient for accurate classification in the context of the proposed framework.

This also proves that the optimisation block does not perform well in case of lack of diverse features set. The individual HF features block generally provide a higher accuracy than DHF alone, which further indicates the need of diverse features and reinforces the avoidance of useless redundancy. However, the processing time of HF cases is also greater. This could suggest that HF features capture more informative characteristics pertinent to the classification task, although at a computational cost. The integration of DHF with HF features significantly increased Acc, particularly in the case of Texture-Shape features (97%) as present in Table 4.5.

Table 4.5: Integration Block Variant Testing using Performance Metrics: HF-DHF Fusion using BRATS Dataset

Testing Block3	**Processing Time**	**Sen**	**Spe**	**Acc**	**MAE**
Case 1 - DHF	10:30:00	98%	96%	97%	0.02
Case 2 - DHF	9:43:00	77%	76%	76.2%	0.24
Case 3 - DHF	8:35:00	73%	70%	72.4%	0.28
Case 4 - DHF	14:10:00	80%	78%	79.6%	0.21

This improvement can be attributed to the complementary nature of the different feature sets, where the depth and abstraction of DHF features enrich the discriminative power of HF textural and shape information. Therefore, not all features contribute equally to classification accuracy, and indiscriminate addition of features does not guarantee performance improvement. In fact, with the existence diverse and relevant features set, the optimal selection block plays its pivotal role. The latter can identify and utilise only those features that enhance classification, avoiding unnecessary computational cost and potential overfitting. These results also illustrate the importance of strategic feature selection. In fact, the highest accuracy is achieved not by the most complex model, but by the one that wisely combines relevant features.

In cases where the complexity does not translate into a corresponding accuracy enhancement, it suggests that some features may introduce redundancy or noise rather than discriminative information. The high processing time presented by most of the experiments impacted by two main factors. In fact, the unlabelled input images increased the processing time of the overall system as it affects the particular processing time of the overall system as it affects the particular processing time of DHF extraction. The DenCeption model took around four hours to proceed the deep DHF extraction of its own. Subsequently, this impacted remaining fusion experiments requiring parallel resources processing. In addition, the type of the processed MRI image (NifTii) is considered as a complicated input image, which has a drawback on its processing time. This, therefore, justifies the lower processing time of the experiments done on the

Retinal dataset, where the latter is composed of mainly .jpeg images which will be presented in the following sub-section.

Mapping the resulted performance metrics, it is evident that the specifications of the BRATS dataset have impacted the RASR outcomes of the proposed optimal features extraction method. In fact, its responsiveness to block testing scheme has not been successful where it comes to individual block testing, except Texture-Shape fusion. Moving forward into the integration testing of HF and DHF blocks, BRATS dataset showed a successful responsiveness, particularly in case of Texture-Shape-DHF, Texture-Colour-DHF, and Texture-Colour-Shape-DHF fusions. The lack of responsiveness presented by Shape-Colour fusion could be interpreted by the absence of Texture feature which represents a key feature in tumour disease detection and classification, which, as a result, impacted the responsiveness of the system processing.

The responsiveness gained by Texture-Colour-Shape-DHF fusion compared to Texture-Colour-Shape fusion is linked mainly to the increase of the Sen and Spe of the optimal features selection system. The increase of a deeper understanding of the input medical images, MRI in this case, impacted the performance of the system. As per this case, Texture-Colour-DHF also showed an evolutionary impact on the overall classification accuracy which helped into optimising the system processing. Despite that, the overall system is not considered as scalable given that it does not involve the shape features, which represents in this particular dataset a critical parameter that reflects the evolution of the tumour and helps in optimising its classification. Texture-Colour-DHF, on the other hand, is also classified as a non-scalable solution for the same reason. That said, Texture-Shape, Texture-Shape-Colour, Texture-Shape-DHF, and Texture-Shape-Colour-DHF have demonstrated a scalable solution by including key features positively impacting on the overall classification. Table 4.6 summarises the mapping of RASR evaluation mechanism on Block3 testing using BRATS dataset.

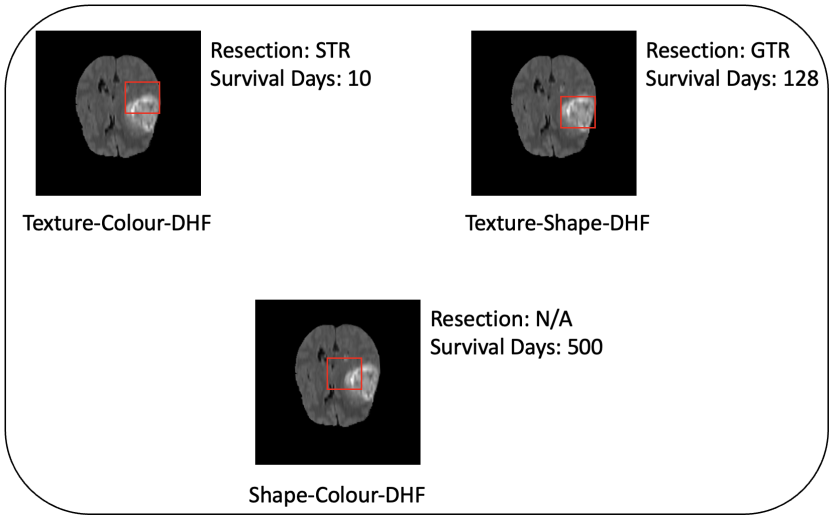


Figure 4.6: Critical Sample Testing of Block3 - BRATS Dataset

Table 4.6: Block3 Variants Evaluation using RASR for BRATS Dataset

Block3 Variant	**Responsiveness**	**Adaptability**	**Scalability**	**Reliability**
Case 1 - DHF	Moderate	High	Good	High
Case 2 - DHF	Good	Moderate	Moderate	Moderate
Case 3 - DHF	High	Low	Low	Low
Case 4 - DHF	Low	Passable	Good	Good

The mapping of Table 4.6 confirms the high importance of Texture-Shape fusion in the case of BRATS dataset. Towards validating the training stage of the proposed method against its variants, an example of testing MRI image is presented in Figure 4.6.

The figure shows the testing result of the tumour detection and classification alongside the indication of the resection status and the estimated survival days. The sample image belongs to a patient with GTR and 131 survival days. As per the results, Texture-Shape-DHF case has successfully identified the tumour with the correct specifications: GTR as resection status and approximate survival days of 128 which is strongly comparable with the original number. However, remaining blocks to include Texture-Colour-DHF and Shape-Colour-DHF have failed to identify the validation parameters resulting each (STR,10) and (N/A, 500) as values for resection status and approximate survival days respectively. Therefore, this validates the presented

results in Tables 4.4, 4.5 , and 4.6.

Justification of Extracted Features for BRATS MRI Dataset

- **Texture Features:** In MRI scans, particularly those used in brain tumour analysis, texture features are crucial because they help capture the intricate variations in the intensity patterns of the tissues, which are indicative of different pathological states. GLCM and Tamura methods were employed to extract features like texture contrast, energy, homogeneity, and entropy. These were selected because they provide valuable insights into the structural variations and are effective in distinguishing between normal and abnormal tissue in brain scans. Texture features played a significant role in the BRATS dataset, where the classification task heavily relied on identifying the irregular texture patterns associated with brain tumours. The results showed that texture features, when combined with shape features, achieved higher accuracy and reliability.

- **Shape Features:** The shape of a tumour or lesion in MRI scans is often a critical indicator of its type and stage. Hence, extracting shape-related features was essential for a comprehensive analysis. RF and other shape descriptors were used to quantify the area and perimeter of the regions of interest, such as tumours. These features were chosen because they help delineate the boundaries and physical dimensions of abnormal growths, which are key for tumour detection and classification. Shape features complemented texture features by providing geometric context, which is particularly important in cases where texture alone might not be sufficient to distinguish between similar tissue types.

- **Deep Hidden Features:** DHF, extracted through the DenCeption model, capture more abstract and complex patterns within the MRI scans that may not be apparent through texture and shape features alone. These features are crucial for enhancing the model's ability to identify subtle variations in the data. DHF were used to add depth to the feature set, improving the overall discriminative power of the model. The integration of DHF with texture and shape features resulted in a significant improvement in accuracy, as evidenced by the results showing that Texture-Shape-DHF fusions outperformed other

feature combinations.

Retinal Dataset: Colour and Labelled Case

Table 4.7 presents the outcomes of performance metrics tested on the individual variants blocks to include Block1 and Block2 using Retinal dataset.

Table 4.7: Individual Blocks Variants Testing using Performance Metrics: Retinal Dataset

Testing Block	**Processing Time (h:mm:ss)**	**Sen**	**Spe**	**Acc**	**MAE**
Block1 – HF Features Only					
Case 1 Features	5:30:00	71%	69%	70.6%	0.3
Case 2 Features	3:00:00	77%	73%	76.2%	0.24
Case 3 Features	2:50:00	69%	65%	68%	0.32
Case 4 Features	4:20:00	81%	80%	80.8%	0.19
Block2 – DHF Features Only	1:30:00	68%	60%	66.4%	0.34

The table shows that Texture-Shape features demonstrate moderate Sen (71%) and Spe (69%), leading to a fairly good Acc (70.6%) but at the expense of a longer processing time of more than 5 hours. Texture-Colour features, on the other hand, show improved accuracy of 76.2% and lower MAE of 0.24, suggesting a good balance of feature Sen, with a significantly reduced processing time. As per Table 4.7, Shape-Colour features underperform in all aspects, indicating that these features alone are less effective for the classification of DR using Fundus images. This was not the case for Texture-Shape-Colour features. In fact, the latte excel in all performance metrics to achieve 81% Sen, 80% Spe, 80.8%, Acc, and 0.19 MAE. However, this improvement was at the cost of a higher processing time, suggesting that while combining these features leads to better accuracy, it requires more computational resources. When comparing to

Block2 – DHF Features Only, which shows the least accuracy and highest mean average error, it is clear that HF, particularly Texture-Shape-Colour Features, are superior for this dataset. The fusion of the latter with DHF features have showed an outstanding level of performance across most of the evaluation metrics as presented in Table 4.8.

Table 4.8: Integration Block Variant Testing using Performance Metrics: HF-DHF Fusion using Retinal Dataset

Testing Block3	**Processing Time (h:mm:ss)**	**Sen**	**Spe**	**Acc**	**MAE**
Case 1 - DHF	3:30:00	84%	81%	83.4%	0.17
Case 2 - DHF	4:45:00	91%	88%	90.4%	0.03
Case 3 - DHF	2:55:00	78%	71%	76.6%	0.24
Case 4 - DHF	5:13:00	99%	98%	98.9%	0.01

In fact, the results of the features fusion block variant (Block3) proved that Texture-Shape Features with DHF has improved in all metrics compared to single HF and DHF blocks (Block1 and Block2 respectively), but with a processing time that suggests increased computational demand. Texture-Colour features with DHF show a better Acc (90.4%) and an a low MAE (0.03), indicating a highly reliable model, despite longer processing time of approximately 5 hours. However, Texture-Shape-Colour with DHF provide the highest Acc of 98.9%, Sen of 99%, Spe of 98% and exceptionally low MAE of 0.01. The reliability of this combination comes with a substantial cost of processing time but comparable to Texture-Colour-DHF fusion case. Shape-Colour features with DHF, on the other hand, slightly enhance performance metrics over single Block1 and Block2, but remain less effective than other DHF combinations. Therefore, it is evident that the addition of DHF to HF consistently improves model performance, but not always in a linear format. The Texture- Shape-Colour-DHF case stands out for achieving high accuracy with a reasonable processing time, suggesting an efficient balance of features usage.

Mapping the results of Retinal dataset through RASR evaluation mechanism (Table 4.9),

Texture-Shape-DHF fusion demonstrates good performance across all metrics, showing good responsiveness in adapting to dataset changes within a reasonable time. Its adaptability and scalability are also rated as good, meaning it can identify essential features independently and handle a variety of image types and sizes effectively.

Table 4.9: Block3 Variants Evaluation using RASR for Retinal Dataset

Block3 Variant	**Responsiveness**	**Adaptability**	**Scalability**	**Reliability**
Case 1 - DHF	Good	Good	Good	High
Case 2 - DHF	Moderate	High	Good	High
Case 3 - DHF	High	Moderate	Low	Passable
Case 4 - DHF	Moderate	High	High	High

The high reliability score indicates a robust model capable of maintaining accuracy even if faults arise. On the other hand, the Texture-Colour-DHF fusion, while having moderate responsiveness, excels in adaptability, indicating a strong ability well operate in unsupervised cases and to learn from the data without external inputs. Its good scalability and high reliability prove that while it may take slightly longer to process, it does so with high consistency and less vulnerability to errors. The results showed by Shape-Colour-DHF fusion are reflected on its high responsiveness, likely due to a less complex feature set allowing for faster processing times. However, it shows moderate adaptability, due to its lack of effectiveness in unsupervised contexts compared to other fusions. Its low scalability is caused by its difficulties in handling diverse datasets. This fusions impact on the overall framework reliability indicates that while generally dependable, it may not be as robust as other fusions against system faults. Finally, the Texture-Shape-Colour-DHF fusion, despite its moderate responsiveness, demonstrates high adaptability and scalability in handling complex feature sets and a wide range of image types and sizes effectively. Its high reliability also confirms its capability to maintain accurate processing through potential challenges.

Towards validating the obtained results in relation to the Retinal dataset, a critical sample of Fundus scan is shown in Figure 4.7 The latter shows the DR detection and stage estimation

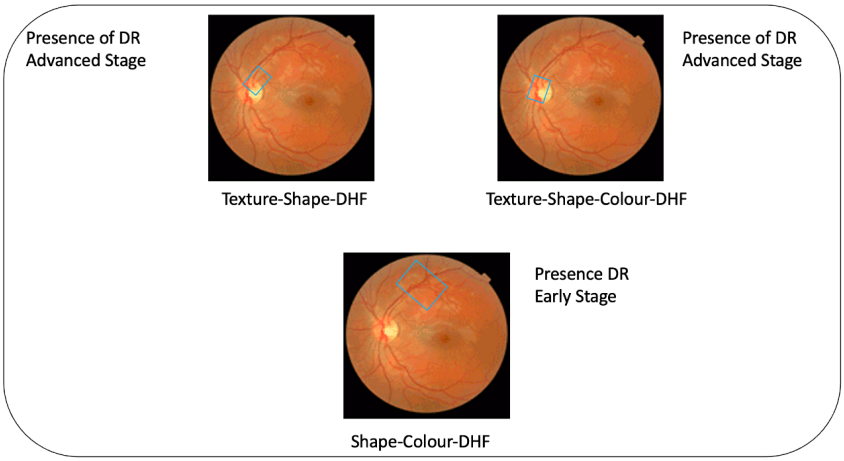


Figure 4.7: Critical Sample Testing of Block3 - Retinal Dataset

of the disease (Advanced stage in this particular sample).

As per the figure, Texture-Shape-Colour-DHF and Texture-Colour-DHF have successfully identified and detected the stage of the DR present in the scan, which was not the case for remaining experiments. This proves the importance of the extraction of key features that can support the system in the identification of critical samples and decreases the FN rates.

Justification of Extracted Features for Retinal Dataset

- **Texture Features:** In the context of retinal images, texture features are critical for identifying pathological changes such as retinal thickening, fluid accumulation, or hemorrhages, which are characteristic of diseases like DR. Similar to the MRI case, GLCM and Tamura were used to extract texture features. These were selected based on their ability to capture the fine structural details in retinal images, which are vital for accurate disease detection. Texture features played a pivotal role in differentiating between normal and abnormal retinal images, with the results showing that combinations involving texture features generally achieved higher Sen and Spe.

- **Shape Features:** These are important in retinal images to detect abnormalities in the geometry of the retina, such as irregular blood vessel patterns or distortions caused by macular edema. RF and other shape descriptors were used to extract features like area and perimeter. These were selected because they help in identifying distortions or en-

largements in the retinal structure that are indicative of DR. Shape features contributed to the robustness of the classification model, especially when combined with texture features, leading to improved accuracy and lower MAE.

- **Colour Features:** These features are particularly important in retinal images where colour variations can signify different stages of diseases like DR. For example, hemorrhages and exudates appear as distinct colour patterns in Fundus images. CHKM was used to extract colour features such as the mean colour value and standard deviation. These were selected to capture the full range of colour variations in the retinal images, which are crucial for detecting and classifying different stages of DR. Colour features were found to be highly effective in the Retinal dataset, with the combination of Texture-Shape-colour-DHF achieving the highest classification Acc of 98.9%. This indicates that colour, when combined with other features, provides a significant boost to the model's performance.

- **Deep Hidden Features:** In the case of retinal images, DHF extracted via the DenCeption model help in capturing complex patterns that are not easily identifiable through traditional feature extraction methods. These include deep variations in texture and colour that correlate with different disease stages. The inclusion of DHF was crucial in achieving high classification accuracy and reducing mean average error. The combination of Texture-Shape-colour-DHF was particularly effective, demonstrating the importance of integrating DL based features with traditional ones.

To compare Block3's performance on each experiment (including Exp1 and Exp2), ROC curve graph has been considered for both dataset cases (BRATS and Retinal respectively) as presented in Figure 4.8.a and 4.8.b.

For the BRATS dataset (Figure 4.8.a), the ROC curves show that all methods have a relatively high TPR (Sen), even as the FPR (1-Spe) increases. This indicates that the variants are generally effective at identifying true cases. The Texture-Shape-DHF combination appears to have a slightly higher curve, further proving that it offers the best balance between Sen and

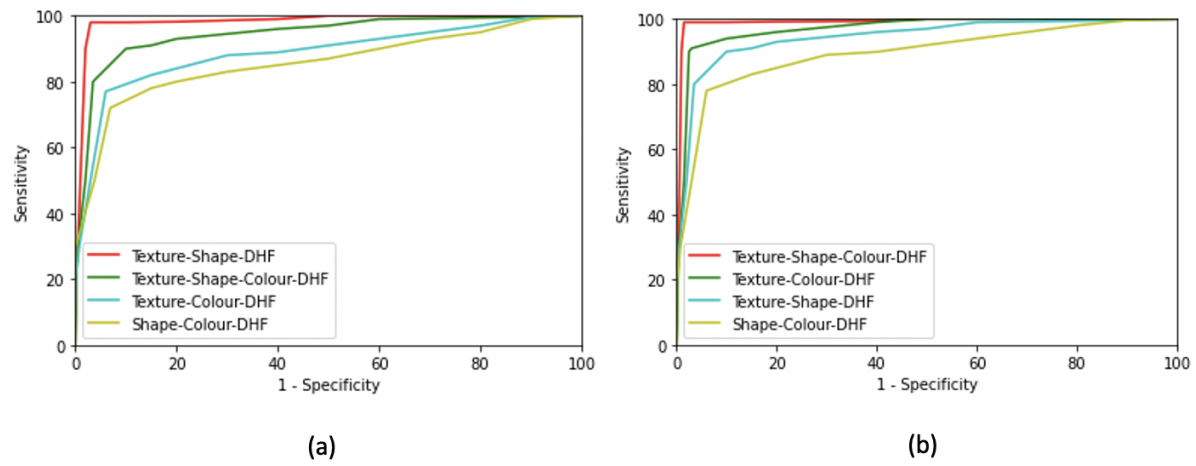


Figure 4.8: ROC Curve of Block3 Variants Testing: (a) BRATS Dataset, (b) Retinal Dataset

Spe among the tested fusions for this particular dataset. In contrast, for the Retinal dataset (Figure 4.8.b), the curves are much closer together, implying that the differentiation between the fusions is not as pronounced. Nevertheless, Texture-Shape-Colour-DHF leads slightly over Texture-Colour-DHF, indicating that the former combination of features is likely the most effective at classification for this particular dataset. These analyses further prove that that while combining multiple types of features with DHF may offer some advantage, the improvement is incremental rather than transformational. Moreover, the fact that the ROC curves of all methods are quite close together, especially in the Retinal dataset, could indicate a level of redundancy when adding more complexity to the feature set. This might imply that simpler combinations of features could be nearly as effective while being more computationally efficient. To further validate the best variants of each dataset, a comparison against benchmarking methods will be conducted in the following section.

4.7.2 Benchmarking Methods

The benchmarking methods selected for comparison with the proposed approach encompass a diverse range of techniques, from those that harness HF feature extraction to others that delve into DHF extraction, illustrating a comprehensive evaluation across the spectrum of feature analysis methodologies. In this context, authors in (Altaf et al., 2017) proposed a method

that integrates GLCM to calculate texture features such as entropy, energy, homogeneity, and correlation. Their method also used volumetric ratios of grey matter and white matter to cerebrospinal fluid, alongside clinical features to improve classification accuracy. Labelled MRI data has been used for multi-class AD and mild cognitive impairment diseases classification, namely AD and MCI respectively. The proposed framework combines both texture and clinical features, which has improved classification accuracy.

The application of multi-class classification offers an advantage to the proposal with an Acc of 79.8%, however, it is less than the Acc obtained through single class classification (94.8%). Additionally, the use of GLCM only for features extraction might pose computational challenges and could lead to overfitting due to its complexity and inherent high dimensionality. Although clinical features enhance the accuracy, they also increase the complexity of the model and the necessity for comprehensive clinical data, which may not always be available. Also, these clinical features might not solve the classification challenge between highly similar classes such as AD and MCI. Focusing only on HF features, a research proposed a method for glaucoma detection using texture features extraction (Kavya and Padmaja, 2017). In this context, GLCM has been used to calculate the occurrences of pixel pairs with specified values and relationships to determine the texture of an image. GLCM was applied alongside MRF technique to extract texture features, considering the changes in texture and intensity values in the Fundus images used. The medical images applied has the colour feature which was not considered for binary classification purpose. The latter was performed using SVM. The experiment was performed on labelled data containing annotations provided by ophthalmologists leveraged as ground truths for sensitivity and specificity calculations achieving an Acc of 86%. The method integrates multiple image segmentation techniques, which could potentially enhance the accuracy of the affected region extraction with the Fundus image. In addition, the use of both GLCM and MRF may provide a comprehensive understanding of the textural changes in the optic nerve head due to glaucoma, leading to more accurate detection.

The use of SVM for classification is a proven ML approach for binary classification tasks, which may yield good results in distinguishing between normal and glaucomatous images.

While the proposed framework presents high achievements in segmentation and classification accuracy, the use of texture-based features alone may not be sufficient to capture all the nuances of glaucoma progression. The reviewed work does not mention the use of colour information, which could be crucial for some types of glaucoma detection. The application of these methods may be computationally intensive and limit their scalability and reliability. A brain disease classification method using multi-feature fusion approach has been also proposed in (Qin and Wang, 2019). Similar to the work proposed in (Altaf et al., 2017), the solution addresses the classification of AD and MCI using structural MRI, as grey-scale scans used for the conducted experiments. The classification framework fuses three types of features: grey-matter volume from VBM, texture features from GLCM, and Gabor features. These features are intended to capture both 2D and 3D information from brain MRIs.

The process involves feature selection using an improved version of the SVM Recursive Feature Elimination (SVM-RFE) algorithm enhanced with a covariance method. The goal is to extract the most relevant features that can distinguish between normal controls (NC), AD, and MCI cases. The proposed method is tested on the public ADNI database. The proposed method is characterised by its multi-feature fusion which could improve the accuracy of AD and MCI classification compared to using single-feature methods achieving an Acc of 91.4% and 97% for AD and MCI, respectively. Improved covariance feature selection technique SVM-RFE has also played a pivotal role in selecting optimal subset of features, addressing the issue of overfitting that arises due to high dimensionality in MRI data. Conversely, the method suffers from several drawbacks to include: (1) complexity of the method where the fusion of multiple features and the improved feature selection process increase the computational complexity of the method, (2) lack of generalisability, in fact, while the method is tested on a specific dataset, its effectiveness on data from different sources or acquired with varying imaging parameters was not tested, and (3) lack of interpretability where the proposed feature selection approach was not well-defined. Correspondingly, authors in (Madusanka et al., 2019) proposed a multi-feature fusion technique for AD and MCI classification using MRI images.

The method combines texture and morphometric features derived from MRI, specifically

utilising Gabour filters, hippocampus morphometric analysis, and both 2D and 3D GLCM for feature extraction. The combination of 2D and 3D GLCM features makes the proposed method unique compared to the one proposed in (Xiao et al., 2017). In fact, the application of 3D GLCM captures spatial dependencies across MRI slices, helps providing more robust representation of brain structures than 2D analysis alone, though increasing its computational complexity. The classification was performed using an SVM only with a 10-fold cross-validation approach achieving an Acc of 86.61% and 78.95% for AD and MCI respectively. Nonetheless, the suggested approach did not perform as well as the comparable method introduced in (Qin and Wang, 2019). Additionally, the process of selecting the most informative features for classification is described as resource-intensive, which could impact the method's efficiency and ease of use in clinical settings.

Similar to (Qin and Wang, 2019), the proposed method lack of generalisability where it is crucial to validate the approach across diverse datasets to ensure its robustness and applicability. On a different note, Lin proposed A smart content-based image retrieval system based on colour and texture feature to enhance retrieval performance (Lin, Chen, and Chan, 2009). The proposed method has a great potential in its applicability on medical images data by considering three primary image features: Colour Co-occurrence Matrix (CCM), Difference Between Pixels of Scan Pattern (DBPSP), and CHKM. CCM calculates the probability of the same pixel colour occurrence between each pixel and its adjacent ones, DBPSP calculates differences between pixels according to motifs of scan patterns and converts it into probability occurrence, and CHKM classifies pixels into k-clusters based on colour similarity. The study employs Sequential Forward Selection (SFS) for feature selection to optimise feature sets for improved detection rates and computational efficiency. By integrating colour and texture features the proposed method has the potential to handle diverse input images, offering robust and adaptable performance achieving 92.2% Acc. Despite the showed performance, the proposed methods have several drawbacks to include: (1) high computational complexity, (2) Sen to noise variations with the input imaging requiring a flexible image pre-processing stage to avoid any lack of reliance on the texture and colour distributions, and (3) dependence on parameter tuning

which, in turn, affects the feature weights assignment and regularisation.

The benchmarking methods also involved works that exclusively utilise DHF extraction techniques, setting the stage for a detailed comparison and understanding of their contributions in the context of classification tasks enhancement. In this context, Nazir introduced an automated system for detecting DR and DMO from retinal images using a DL approach centered around a custom CenterNet model, incorporating DenseNet-100 for feature extraction (Nazir et al., 2021). This method first prepares a dataset with annotations to identify RoIs and then utilises the CenterNet model, enhanced with DenseNet-100, to localise and classify the disease lesions from annotated coloured Fundus images. The proposed framework is tested on challenging datasets, including APTOS-2019 and IDRiD, demonstrating high accuracy (97.93% and 98.10%, respectively) in disease detection and classification. The work presents a robust feature extraction approach incorporating DenseNet-100, which results the enhancement of the overall system's ability to recognise small lesions and deal with low-intensity and noisy images. The proposed method is also simple and efficient leveraging a one-stage detector (CenterNet), which offers a computationally efficient alternative to traditional two-stage detection methods while maintaining high performance. However, this impacts the scalability of the method to handle larger datasets.

The suggested framework presents lines of potential generalisation capability by using cross-dataset validation, though, both datasets represent the same type of medical imaging (Fundus). Despite the improvements in generalisation, there remains a potential of overfitting, especially when the method is trained on highly specific datasets, which could affect its performance on unseen data or under diverse conditions. Also, the non-consideration of other types of medical scans and grey-scale based images in addition to its heavy reliance on the quality of annotations for training, limit its adaptability. On the other hand, Dara proposed a CNN based feature extraction method for lung cancer classification dataset (TCGA-LUAD), using CT scans (Dara et al., 2018). The method aims to convert input data into a set of features to simplify subsequent learning and analysis processes. The proposed framework outlines an approach for automatic feature extraction using CNNs, alongside comparisons with MLP both

with and without manual feature extraction. The methodology employed relies on the convolutional and pooling layers of CNNs to process and extract features from medical images automatically.

The employment of CNNs for automatic feature extraction from CT scans proved the significant reduction of the pre-processing resources required and potentially uncover features that may not be evident through manual extraction methods. The proposed CNN based method has also showed a scalability by covering a relatively large dataset, which is beneficial given the increasing volume of medical imaging data. In addition to the resource-intensity required by the proposed method, it would have been beneficial if the validation experiments involved a different dataset with diverse image type and target disease which might have increased the reliability and adaptability of the proposed method. The paper also lacks any explanation regarding weights and parameters regularisation which raise a concern regarding overfitting risks. Table 4.10 summarises the characteristics of each of the reviewed works.

Table 4.10: Comparison of Benchmarking Feature Extraction Methods and Their Corresponding Imaging Data Type Coverage

Method	**Reference**	**Features**		**Imaging Data**		**Image Type**	
		HF	**DHF**	**Unlabelled**	**Labelled**	**Grey-scale**	**Coloured**
GLCM	(Altaf et al., 2017)	X			X	X	
GLCM-MRF	(Kavya and Padmaja, 2017)	X			X	X	

Continued on next page

Table 4.10: Comparison of Benchmarking Feature Extraction Methods and Their Corresponding Imaging Data Type Coverage (Continued)

Method	**Reference**	**Features**		**Imaging Data**		**Image Type**	
		HF	**DHF**	**Unlabelled**	**Labelled**	**Grey-scale**	**Coloured**
GLCM-VBM	(Xiao et al., 2017)	X			X	X	
SVM-RFE	(Madusanka et al., 2019)	X			X	X	
CHKM-CCM	(Lin, Chen, and Chan, 2009)	X			X	X	X
DenseNet-100	(Nazir et al., 2021)		X	X		X	X
CNN-MLP	(Dara et al., 2018)		X	X			X
Proposed Method		X	X	X	X	X	X

The conducted experiments will serve as the assessment framework in correspondence with the established evaluation criteria. This process affirms the effectiveness of the suggested approach in comparison to benchmarking methods. The responsive and adaptable nature of the proposed method has been enhanced by its scalability and dependability. All methods have been applied in accordance with Exp1 and Exp2 using BRATS and Retinal datasets, respectively. Tables 4.11 and 4.12 presents the outcome of the conducted experiments for bench-

marking methods against the proposed features extraction framework for BRATS and Retinal datasets, respectively.

For the BRATS dataset, Table 4.11 illustrates a comparative overview where the proposed method outperforms others across all performance metrics. Nazir's work applying DenseNet-100 and Lin's work with CHKM-CCM, demonstrate high accuracies of 90.5% and 89.1% respectively, where the proposed method's Acc achieves 97%. When considering Sen and Spe, which are critical in medical diagnostics to reduce FNs and FPs, the proposed method surpasses other methods with 98% and 96%, while the best competing methods, DenseNet-100 and CHKM-CCM, show a Sen of 92% and 87.2% and Spe of 89.7% and 89.7% respectively. Precision and F1-score prove that the proposed method at 96.3% and 97.14% exceeds all other methods. A no%distinction is also observed in the MAE where the proposed method achieves a minimal 0.09, significantly outperforming other methods, demonstrating that it not only identifies correct features more consistently but also with fewer errors, making it highly reliable for clinical applications.

In the context of the Retinal dataset, as shown in Table 4.12, similar patterns of performance have been identified. In fact, the proposed method achieves an outstanding Acc of 97%, while the best competing methods CHKM-CCM and GLCM-VBM reach 90.7% and 87.3% respectively. Sen and Spe are paramount in DR disease classification to ensure accurate patient diagnosis, and here too, the proposed method excels with 98% and 96% versus CHKM-CMM's 92% Sen and GLCM-VBM's 87% Spe. In precision and F1-score, the proposed method demonstrates a high reliability with scores of 96.3% and 97.14%, overcoming other methods. This distinction is pivotal with MAE of 0.09, reinforcing the proposed method's capacity to provide dependable and precise diagnoses over the others.

When critically comparing the individual feature sets (HF and DHF), it becomes clear that the integration of both does not always guarantee superior performance. For instance, in Table 4.11, the cases utilising HF features alone exhibit lower accuracies compared to the DHF features only. However, the integration of these feature sets, especially in cases where Texture, Shape, and Colour are combined with DHF, shows a significant increase in performance

metrics, indicating that an optimised combination of features can enhance accuracy and model reliability.

Table 4.11: Benchmarking Results of BRATS Dataset Versus Proposed Methods

Method	**Acc (%)**	**Sen (%)**	**Spe (%)**	**Precision (%)**	**F1-score (%)**	**MAE**
GLCM (Altaf et al., 2017)	68	70	65	67.3	68.6	0.33
GLCM-MRF (Kavya and Padmaja, 2017)	87.3	83.44	89	86	84.6	0.25
GLCM-VBM (Xiao et al., 2017)	71	75.3	68	72.3	73.4	0.3
SVM-RFE (Madusanka et al., 2019)	75.2	72	77	74.3	73.1	0.28
CHKM-CCM (Lin, Chen, and Chan, 2009)	89.1	87.2	89.7	88	87.5	0.2
DenseNet-100 (Nazir et al., 2021)	90.5	92	89.7	93	92.4	0.15

Continued on next page

Table 4.11: Benchmarking Results of BRATS Dataset Versus Proposed Methods (Continued)

Method	**Acc (%)**	**Sen (%)**	**Spe (%)**	**Precision (%)**	**F1-score (%)**	**MAE**
CNN-MLP (Dara et al., 2018)	88.7	90.1	89	90	90.04	0.26
Proposed Method	97	98	96	96.3	97.14	0.02

Table 4.12: Benchmarking Results of Retinal Dataset Versus Proposed Methods

Method	**Acc (%)**	**Sen (%)**	**Spe (%)**	**Precision (%)**	**F1-score (%)**	**MAE**
GLCM (Altaf et al., 2017)	73.8	75	72	74.7	74.8	0.4
GLCM-MRF (Kavya and Padmaja, 2017)	80.1	83	83.4	79	80.9	0.32
GLCM-VBM (Xiao et al., 2017)	87.3	86	87	85.8	85.8	0.21
SVM-RFE (Madusanka et al., 2019)	83	81.2	83.4	81	81.09	0.29

Continued on next page

Table 4.12: Benchmarking Results of Retinal Dataset Versus Proposed Methods (Continued)

Method	**Acc (%)**	**Sen (%)**	**Spe (%)**	**Precision (%)**	**F1-score (%)**	**MAE**
CHKM-CCM (Lin, Chen, and Chan, 2009)	90.7	92	89	91	91.4	0.11
DenseNet-100 (Nazir et al., 2021)	89.7	88.4	88	87.3	87.8	0.19
CNN-MLP (Dara et al., 2018)	83.7	85	83.1	81	82.9	0.24
Proposed Method	98.9	99	98	98.4	98.6	0.01

4.8 Conclusion

Disease classification requires specialist's expertise in locating inner areas of interest from medical images, particularly, grey-scale images (MRI) and coloured images (Fundus). That is, manual features extraction can be time consuming which might have side effects on the diagnosis and analysis process. To cope with this challenge, an automated features extraction and selection method is proposed. The framework is based on combining HF and DHF features towards achieving a high quality of medical analysis with optimised features set. The novel DL framework, DenCeption, has been applied for DHF extraction alongside HF features extraction techniques. The optimal combination of texture, shape, colour, and DHF has been used as an input to the classification model. The main aim of the proposed method is to create a

generalisable framework that can pick the best features combination based on the characteristics of the input dataset. Multiple experiments have been considered to test each possible features combination and reflect that on the proposed evaluation mechanism, RASR, to convey responsiveness, adaptability, scalability and reliability.

The conducted experiments have also been tested on benchmarking methods to validate the proposed approach. The proposed features extraction framework achieved outstanding results on both coloured/labelled and grey-scale/unlabelled based datasets, namely, BRATS and Retinal respectively. The novel method reached 97% for Texture-Shape-DHF combination and 98.9% for Texture-Shape-Colour-DHF combination, respectively. Despite the use of other features combinations, the high impact the aforementioned combinations provided helped in intensifying the responsiveness and reliability of the proposed framework by minimising the FPs and FNs that can occur. Considering the above results, the proposed framework can be scaled to be applied in real-time experiments. Hence, the potential application of its use in second and/or first clinical line. Moreover, a disease prediction model will be designed towards testing the proposed features extraction model on scenarios other than classification. The proposed prediction framework will be proposed in the following chapter.

In concluding this chapter, the complex interplay of HF and DHF has been explored, scrutinising their distinct contributions and combinations towards improving classification accuracy. The proposed feature extraction framework, marked by its responsiveness, adaptability, scalability, and reliability, has undergone thorough testing across two diverse datasets, each depicting a unique scenario. Moving forward, this framework will be integral to the predictive model in the next chapter. By integrating within this model, the framework aim to enhance the analysis influence of selected features on prediction outcomes. Furthermore, the chapter will covey a proposal of a new performance measurement tool, designed to assist in the nuanced selection of suitable metrics. This tool is aimed at accommodating problem specifics, application demands, and data accessibility, facilitating a comprehensive strategy for performance assessment and feature impact evaluation.

Chapter 5

Advancing Intelligent Medical Diagnostics with HyBoost: A Robust Predictive Framework for High-Dimensional Imaging Data

5.1 Introduction

In the previous chapter, the proposed adaptive feature extraction framework was introduced, initiating a substantial advancement in the domain of medical imaging analysis. The combination of the innovative DenCeption model, which, combined with a keen selection of pertinent features, is crucial in refining the learning task across diverse scenarios. The framework's skill in pinpointing the best features, customised for each particular situation, highlights its adaptability and efficiency. In the context of the current chapter, the emphasis shifts toward disease prediction, building on the foundation established by the cutting-edge approach to feature extraction. The core of this framework, with its ability to wisely select and apply the most influential features, is key in boosting the predictive accuracy of the proposed framework. By leveraging the strengths of DenCeption and carefully selected features, the proposed solu-

tion aims to expand the frontiers of disease prediction, providing a more detailed and effective solution to this complex challenge.

In the rapidly advancing field of automation in disease diagnosis, AI, particularly ML and DL, is playing an increasingly pivotal role. The nuanced utilisation of AI in analysing complex datasets for medical imaging, such as OCT, X-ray, and Fundus photography, is transforming the landscape of disease prediction and management. These technologies not only streamline the diagnostic process but also bring forth the potential for personalised patient care, underpinning the significance of integrating varied physiological and demographic features into the predictive models. The advent of ML and DL in medical imaging has prompted a substantial advancement in the quality of prediction. These advanced computational approaches have empowered clinicians with tools that offer an unparalleled depth of insight into the human anatomy. Particularly, in ophthalmology, where the detailed visualisation offered by OCT and Fundus photography is critical for early and accurate disease detection. X-ray imaging, too, has benefited from AI, with ML algorithms now adept at discerning patterns and anomalies that may escape the human eye. This type of scanning has undergone a renaissance with ML/DL interventions, introducing a new level of precision in detecting skeletal anomalies and thoracic complications. The integration of demographic data and physiological features, such as age, gender, and systemic health markers, into AI models has been instrumental in enhancing the accuracy of disease prediction. The diversity of data captured in medical images, is key to the advancement of ML/DL applications. These characteristics inform the models, fostering an environment where the confluence of data types enriches the predictive accuracy. It is this synergy that underscores the value of ML/DL in medical imaging, not only in refining predictions but also in tailoring patient care. The promise held within these advancements is one of a future where AI-driven diagnostics are not only adjuncts but integral components of patient care, presenting an insight into the field of healthcare marked by precision, efficiency, and proactive intervention. The combination of ML and DL with the analysis of medical images has led to significant changes in diagnosis methods. The arrival of these advanced technologies marks the beginning of a period characterised by the merging of high computing power and complex algorithms, resulting

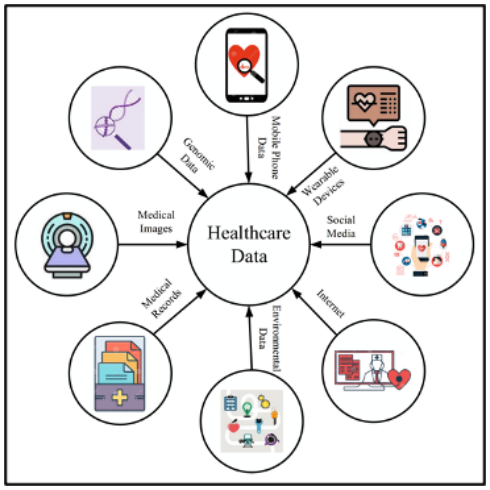


Figure 5.1: Power of ML and DL in Healthcare Data Handling (Rahmani et al., 2021).

in improved medical diagnosis processes and notable progress in how patients are cared for (Figure 5.1).

Medical imaging is a critical component of modern healthcare, providing clinicians with non-invasive means to visualise the inner workings of the body. As such, the quality and interpretability of these images are paramount. The evolution of ML/DL applications within this domain has been pivotal, offering unprecedented accuracy and efficiency in image analysis. These tools have the ability to learn from vast amounts of data, identifying patterns and anomalies with exceptional precision. Their application ranges broadly across various imaging modalities, including Fundus photography, OCT, and X-ray imaging—each serving a unique purpose in disease diagnosis and management. Fundus photography, which captures the back of the eye, is essential for diagnosing conditions such as DR and glaucoma (Das, Biswas, and Bandyopadhyay, 2022). DL models, trained on thousands of labelled images, can detect implicit changes in the retina, providing critical information that may not be immediately evident to the human eye (Mukherjee and Sengupta, 2023) (Figure 5.2).

This capability enhances early disease detection, which is crucial for conditions where early intervention can prevent or delay progression (Bala, Sharma, and Goel, 2023). OCT imaging, which offers a cross-sectional view of the eye, is indispensable for diagnosing retinal diseases (Sun, Yang, Tang, Ng, and Cheung, 2021). DL algorithms, through layer segmentation and feature analysis, have dramatically improved the speed and accuracy of OCT interpretation (Ting

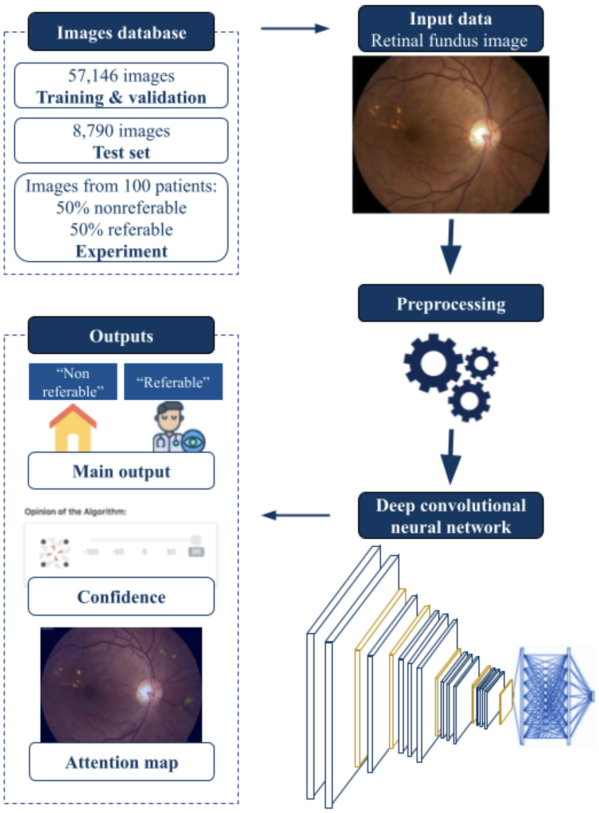


Figure 5.2: DR Detection Using Non Referable and Referable Fundus Imgaes: DL Vs Specialists (Noriega et al., 2021).

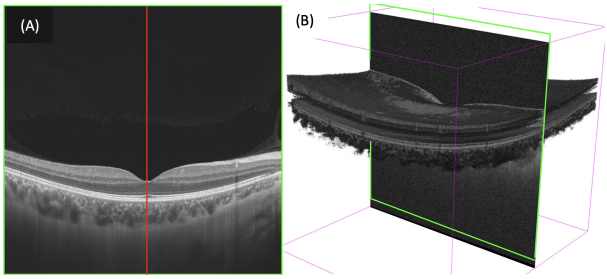


Figure 5.3: OCT Scan Dimensional Representation: (A) Axial Scan, (B) Cross-sectional Scan, and (C) OCT Volume (Khan, Sohail, Zahoora, and Qureshi, 2020).

et al., 2019). This rapid, automated analysis facilitates immediate clinical decision-making, which is particularly beneficial in high-volume, resource-constrained settings (Li, Ran, Cheung, and Prince, 2023a) (Figure 5.3).

X-ray imaging, one of the oldest forms of medical imaging, has also seen significant enhancements with the application of ML/DL. These technologies have shown great promise in detecting pathologies such as fractures, lung nodules, and signs of diseases like pneumonia

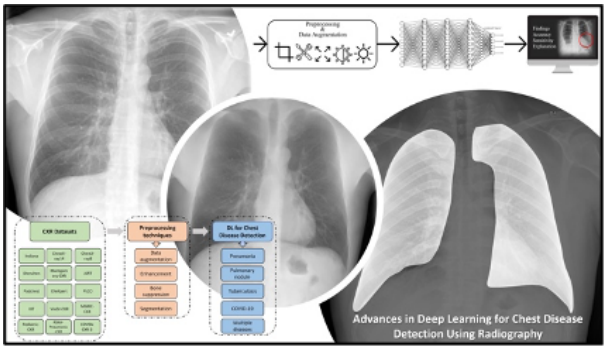


Figure 5.4: Automated Approach for X-ray Analysis (Ait Nasser and Akhloufi, 2023).

and tuberculosis (Sharma and Guleria, 2023c). The algorithms can prioritise cases, detect implicit or complex conditions, and even predict disease progression based on historical data (Ait Nasser and Akhloufi, 2023) (Figure 5.4).

The impact of ML/DL-based frameworks on disease prediction cannot be overstated. These frameworks analyse medical images with an efficiency and accuracy that surpass traditional methods, often identifying early-stage diseases before they manifest clinically. By enabling the early prediction of diseases, ML/DL technologies can dramatically influence patient outcomes, reduce the burden on healthcare systems, and streamline the workflow for healthcare professionals. Moreover, as these technologies continue to evolve, they constantly refine their predictive capabilities, learning from new data to adapt and improve. The potential for ML/DL to integrate with emerging imaging technologies and electronic health records presents an opportunity for a comprehensive approach to patient care, where predictive analytics can lead to personalised treatment plans and proactive health management (Kumar, Kumar, Deb, Ungureşan, and Muresan, 2023). The advent of ML and DL in medical image processing has marked a significant milestone in the field of diagnostics. With a focus on imaging modalities like Fundus, OCT, and X-ray, ML/DL frameworks have profoundly impacted disease prediction, offering a preview into a future where AI-powered diagnostics enhance every aspect of patient care. As researchers continue to leverage these technologies, they edge closer to a healthcare paradigm characterised by early detection, precision medicine, and improved prognostic outcomes.

5.2 Research Contributions

In this context, this research tackles challenges and advancements in medical image analysis, focusing on the complexity of high-dimensional medical imaging data. This work highlights the need for sophisticated analytical approaches to interpret these complex datasets effectively. The main contributions of this work are outlined as follows:

- **Development of a Scalable and Versatile Prediction Framework**: This work introduces a novel framework capable of efficiently adapting to various medical imaging data. The design focuses on scalability, ensuring that the framework can handle diverse and large datasets effectively. It is versatile enough to be applicable across different medical conditions, making it a significant contribution to the field of medical diagnostics.

- **Creation of an Innovative Hybrid Predictive Model**: A key contribution of this research is the introduction of a hybrid predictive model that synergises the strengths of multiple ML algorithms. This model is designed to optimise both the efficiency and accuracy of the predictive process, making it a substantial advancement over traditional single-algorithm models. Its hybrid nature allows for a more robust analysis of complex medical data, leading to more precise and reliable predictions.

- **Integration of Demographic and Physiological Features for Enhanced Performance**: Another major contribution is the incorporation of additional demographic and physiological features into the predictive model. This integration is pivotal as it allows for a more comprehensive analysis of patient data, taking into account a wider range of variables that can influence disease outcomes. By including these extra features, the model's ability to predict diseases is significantly enhanced, offering a more holistic approach to medical diagnostics. This aspect of the research not only improves the accuracy of disease prediction but also paves the way for more personalised and effective patient care.

The structure of this chapter is organised in the following manner:

- Section 3 delves into a discussion about works pertinent to the study.

- Section 4 describes the prediction framework being proposed and explains the algorithm behind the hybrid predictive model recommended.
- Section 5 presents the rationale behind the selection of the HyBoost model.
- Section 6 covers the datasets used in this chapter.
- Section 7 outlines the experimental procedures undertaken within the scope of this research.
- Section 8 illuminates features extraction and sample testing.
- Section 9 engages in a discussion on the results achieved and the process of benchmarking these findings. The section also shows the impact of demographic and physiological features.
- Finally, the chapter draws to a close with Section 10, which serves as the conclusion.

5.3 Related Works

In the discipline of healthcare, the integration of DL techniques has prompted a significant transformation, offering not only improved diagnostic accuracy but also improved efficiency and cost-effectiveness in the diagnosis and prediction of various medical conditions. Among the domains that have witnessed substantial progress, DL's impact on the field of ophthalmology and pulmonology is undeniably significant. The application of DL in these disciplines has revolutionised the way eye and lung related diseases are diagnosed, monitored, and treated. Within this broader landscape, the focus of this related works section centers on the profound implications of DL for DR, DMO, and lung related diseases. These diseases significantly affect a substantial portion of individuals with health conditions, making their early and accurate diagnosis crucial for effective treatment.

This section will delve into a selection of recent studies that exemplify the intersection of DL and diseases-related applications. These studies collectively illuminate both the potentials

and challenges that arise when AI meets the complexities of the aforementioned diseases. Consequently, they offer insights into the advancements, contributions, and critical gaps that exist within this rapidly evolving field hence their use as banchmarking state-of-the-art methods in the validation of the proposed framework. The reviewed papers highlight the multidimentional nature of DL applications in the field of DR, DMO and lung diseases, where, they address the pivotal issues of data diversity, model interpretability, and the necessity for more extensive datasets and external validation. Traversing through these studies, it becomes evident that DL's influence on these diseases diagnosis and prediction is both profound and promising. However, the literature also underscores the challenges that demand further attention, as discussed in the literature review chapter.

Ophthalmology has experienced a significant revolution, thanks to the adoption of advanced technologies, especially DL and ML. A broad range of eye diseases, impacting vision and the overall health of the eyes, has become a focal point in the current wave of AI-powered healthcare advancements. Through DL and ML methods, notable progress is being achieved in the early detection, continuous monitoring, and management of these eye conditions, leading to enhanced patient outcomes and better quality of life.

Predicting the response of patients with DMO to anti-VEGF treatment using pre-treatment OCT scans is a complex challenge addressed by authors in (Alryalat et al., 2022). They have developed a novel DL-based model for prediction purpose. The study employed a segmentation model based on the U-Net architecture, which was enhanced with squeeze excitation layers, inception modules, and multi-scale attention mechanisms, offering a nuanced approach to feature recalibration and spatial emphasis within the network. The model achieved high accuracy in segmenting DMO-related features in OCT images, with an Acc of 95.9%, AUC of 93.4%, Spe of 98.9%, Sen of 87.9%, precision of 80.7%, F1-score of 83.9%, and dice of 83.9%. For classifying patients as good or poor responders to anti-VEGF treatment, the study compared different DL models and found that including the predicted mask in the input layer improved classification Acc to 75%.

Real-world testing of the model showed 60% Acc in classifying response to treatment.

The model's accuracy was compared to ophthalmology trainees and specialists, with retina, specialists achieving the highest Acc at 86.3%. The study's main limitation is the relatively small sample size, which might limit the model's generalisability. A larger and more diverse dataset would provide stronger evidence of the model's effectiveness. While the model showed promise in classifying response to treatment, the Acc (60%) in real-world testing is relatively modest. Further refinement and validation are needed for practical clinical use. As mentioned above, DL models, especially complex ones like the proposed architecture can lack transparency and interpretability. Understanding how the model arrived at its predictions is essential for clinical acceptance. Something is worth mentioning that the model's performance in a specific population may not generalise well to other populations with different genetic and environmental factors. Validation or a more diverse dataset is crucial.

The research proposed in (Li et al., 2022a) advances upon the traditional Inception-V4 by proposing improvements that could potentially enhance pattern recognition capabilities, reflecting a trend towards more complex, deeper architectures. Instead of a binary classification of affected and non-affected eye scans, authors proposed a novel framework for DR and DMO classification using 8739 Fundus images (Li et al., 2022a). Their model was also tested on secondary data for testing purposes. Both of the used datasets have been independently reviewed and graded by ophthalmologists. Towards preventing overfitting problem, data augmentation techniques such as cropping, flipping, and rotation were applied to increase the heterogeneity of Fundus images. The authors employed an ensemble of five classification model instances based on improved Inception-V4 architecture. Each model learned different discriminative features, even when trained with the same data. The ensemble approach aimed to increase classification robustness. The model achieved high Sen (0.925), Spe (0.961) and AUC (0.992) for both non-referable DMO and referable DMO classification tasks on the primary test dataset. It performed slightly better than ophthalmologists in these tasks. As mentioned above, the models' generalisability was demonstrated by applying it to the Messidor-2 dataset, where it outperformed previously reported state-of-the-art methods in terms of AUC for DR and DMO detection. The impact of input image size on model performance was analysed, showing that

larger image sizes generally led to better performance. Sub-sampling experiments indicated that increasing the dataset size could further improve model accuracy (Li et al., 2022a). Comparing the ensemble model with five instances to a single model with larger input image size revealed improved performance with the ensemble approach and reduced time requirements. Despite the use of multiple datasets for training and testing/validation, which is a strength, the dataset's composition and quality could potentially introduce biases. The use of an ensemble approach is a notable feature, improving the model's robustness and generalisation. However, it would be interesting to explore the model's performance on other retinal conditions or expand its capabilities to identify a broader range of diseases.

DR, a widespread complication of diabetes, poses a significant risk of progressive visual deterioration and, in advanced stages, blindness. The importance of early detection and prompt treatment in controlling DR's advancement cannot be overstated. Historically, the diagnosis of DR has relied on ophthalmologists conducting manual inspections of Fundus images, a method that is not only time-intensive but also prone to variations between observers. However, the emergence of ML and DL technologies has ushered in a ground-breaking era for the prediction and analysis of DR through Fundus photographs. Automated algorithms are now capable of processing large collections of Fundus images, detecting subtle pathological changes that might be missed during manual reviews. These ML and DL approaches significantly improve diagnostic precision and offer a faster, more efficient approach to screening. The growing body of research on this subject underscores the revolutionary impact these technological advances are having on the screening and management of DR, as elaborated upon in the literature chapter.

A significant departure from Inception-based architectures is seen in (Wahab Sait, 2023), which adopts YoLo V7 for real-time object detection. It utilises a quantum marine predator algorithm for feature selection—merging evolutionary computation with quantum mechanics concepts—and employs the Adam optimiser to fine-tune MobileNetV3. Such a multi-aspected approach suggests a promising synergy between high-speed detection frameworks and sophisticated optimisation techniques. Addressing the mounting concerns surrounding DR, author introduced a agile deep-learning architecture tailored for DR severity grading, optimised for

limited computational prowess. By integrating pre-processing techniques with the Yolo-V7 feature extraction mechanism and the Quantum Marine Predator Algorithm (QMPA) for feature selection, the MobileNet-V3 model demonstrated robustness, producing accuracies more than 98% on two substantial datasets (Wahab Sait, 2023) . The model's streamlined design, optimised for swift computations, holds promise for its incorporation into mobile applications, potentially revolutionising remote healthcare solutions (Gupta, Thakur, and Gupta, 2023).

In evaluating these studies, it's evident that the nexus between computational techniques and ophthalmological diagnostics is strengthening. The focus is not just on achieving high accuracy but also on designing systems that are efficient and can be seamlessly integrated into real-world applications, particularly in remote healthcare. Refining these models, especially to excel in low-quality image scenarios, is the next frontier, promising to reshape the future of ocular diagnostics.

In contrast, the adoption of of EfficientNet B5 by a research proposed in (Paul and Talukder, 2023) underscores an interest in scalable architectures that meticulously balance model complexity and accuracy. EfficientNets have set new benchmarks in leveraging compound scaling; thus, their inclusion in comparative studies serves as an excellent touchstone for computational efficiency and performance. In this context, authors introduced a self-adaptive ensemble method for grading DR severity by stacking several dual attention mechanisms (Paul and Talukder, 2023). This dual attention model employs two distinct attention processes: one concentrates on lesion-specific areas, while the other learns correlations between spatial descriptors, effectively predicting DR severity levels. The study also introduces a self-adaptive meta-learner for stacking multiple dual attention models efficiently. When tested on the APTOS 2019 dataset, this approach surpassed many existing models, achieving an Acc of 97.78%.

The paper suggested in (Shimpi and Shanmugam, 2023) delves into the amplification of DR screening to avert blindness, underscoring the vitality of precocious detection via Fundus imaging. Notwithstanding the progress in CNN methodologies, overfitting persists as a problem. In this context, the study proposes an innovative multiclass AdaBoost strategy integrated with CNN-based categorisation to surmount overfitting and augment classification precision.

The exploration utilised the VGG-16 pretrained model for discerning features and employed the factor analysis technique for pre-processing DR snapshots. Concerning experimental outcomes and their analysis, the importance of accuracy, predominantly in skewed datasets, was accentuated and the accuracy calculation formula was interpreted. The research employed a conventional AdaBoost procedure encompassing 400 DTs, registering 87.45% training precision and 77.08% testing precision. In the AdaBoost model, the VGG 16 CNN was employed as a lone baseline estimator, with the peak testing accuracy attained using a 9-layer network, achieved at 91.05%. Furthermore, the AdaBoost CNN Classifier, leveraging a 9-layer VGG 16 CNN, showed an enhancement in Acc rates from 91.76% to 95.56%. Intriguingly, an increase in the estimator count inversely impacted accuracy. In the field of TL, the AdaBoost-VGG 16-CNN showcased its indispensability, elucidating that it expedited computation durations and heightened accuracy. When juxtaposed, the AdaBoost VGG-16 CNN surpassed both its single CNN estimator counterpart and the traditional AdaBoost equipped with a DT. Conclusively, the fusion of ensemble learning, specifically AdaBoost methodologies with CNN, manifests as an advancement in prediction and categorisation tasks. The integration of TL further amplifies accuracy while limiting computational demands. Such improvements are paramount for effective DR screening, influencing diabetes treatment and the prognosis for patients.

Reflecting on the literature, the increasing reliance on advanced ML and DL methods in detecting and classifying DR using retinal images. From using ensemble classifiers and attention mechanisms to incorporating hybrid optimisation algorithms, the literature demonstrates rapid advancements in automated DR diagnosis. Notably, the ResNet-based approaches appear to outperform other methods, suggesting its prominence in the field. However, while many of these methods have shown high accuracy rates, practical implementation in real-world scenarios, scalability, and cost-effectiveness are aspects that need further exploration. It's also essential to evaluate the model's adaptability to different datasets and real-world conditions. The literature opens doors to more streamlined and efficient prediction of DR, which can significantly impact patient care and management. However, continuous validation, especially with larger and more diverse datasets, is imperative to ensure the reliability and generalisability

of these models in clinical settings.

ML and DL have rapidly become transformative in the analysis of X-ray imagery. In recent years, a plethora of research has been devoted to harnessing these technologies for the diagnosis of various medical conditions, with their use in detecting pneumonia from X-ray images receiving notable focus. Traditional diagnostic practices, which rely heavily on the expertise of radiologists, are often prone to human error and subjective interpretations. In stark contrast, ML and DL algorithms present a more uniform and efficient method, capable of uncovering intricate patterns that might elude human observation. The expanding research in this area highlights the considerable promise of ML and DL techniques in improving both the precision and the expedience of pneumonia diagnostics.

Given the alarming mortality rates associated with pneumonia, especially in children, timely detection becomes paramount. X-rays, while instrumental, come with their set of challenges like potential misdiagnoses. A VGG-19 DL model was employed to tackle these issues, demonstrating promising results (Sharma and Guleria, 2023b). In fact, using a dataset of 5856 CXR, the model achieved an Acc of 93%, precision and recall of 0.931, F1-score of 0.931, and an AUC of 0.973. When benchmarked against other models, the proposed model excelled. Alongside pneumonia, X-rays can identify other conditions like emphysema, lung cancer, and tuberculosis. However, many DL models in this domain face computational and interpretability challenges. To address these concerns, a study in (Vetrithangam et al., 2023) proposed a modified ResNet152v2 DL model, emphasising high accuracy and efficient computation. Their refined model is designed for efficient pneumonia prediction from CXR, aiming for high accuracy, reduced complexity, and faster computation. When benchmarked, the model surpassed other methods, achieving 99.77% Acc, 99.86% Sen and precision, and 95.4% Spe.

5.3.1 Identified Challenges

DL's important role in disease diagnosis and prediction highlights both its potential and the challenges that accompany its integration into healthcare. While the application of advanced ML and DL techniques in analysing medical images has shown to enhance diagnostic accu-

racy and offer deeper insights into disease mechanisms, critical gaps in the current research landscape necessitate further examination. These technologies, despite their advancements, face issues related to model bias, overfitting, reproducibility, scalability, interpretability, and clinical relevance, alongside concerns about their sustainability and computational demands. Addressing these challenges is crucial for ensuring that the benefits of ML and DL are accessible across diverse medical settings and can contribute effectively to personalised treatment strategies. Identified challenges include:

- **Model Bias and Diversity:** High accuracy rates may conceal biases, particularly with non-diverse datasets, potentially skewing real-world performance.
- **Overfitting:** Models exceptionally tuned to specific datasets may not generalise well across the diverse spectrum of patient data.
- **Reproducibility and Scalability:** Many models are tested under controlled conditions, raising concerns about their performance in varied real-world settings.
- **Interpretability and Clinical Relevance:** The 'black-box' nature of some models may hinder their adoption by medical professionals who require understandable diagnostic pathways.
- **Sustainability and Computational Demands:** The computational intensity of some advanced DL models may not be feasible for every medical facility, especially in resource-constrained areas.

Exploring these gaps not only highlights areas for improvement but also underscores the necessity of developing robust, interpretable, and scalable DL and ML models that can operate effectively across the full spectrum of healthcare environments.

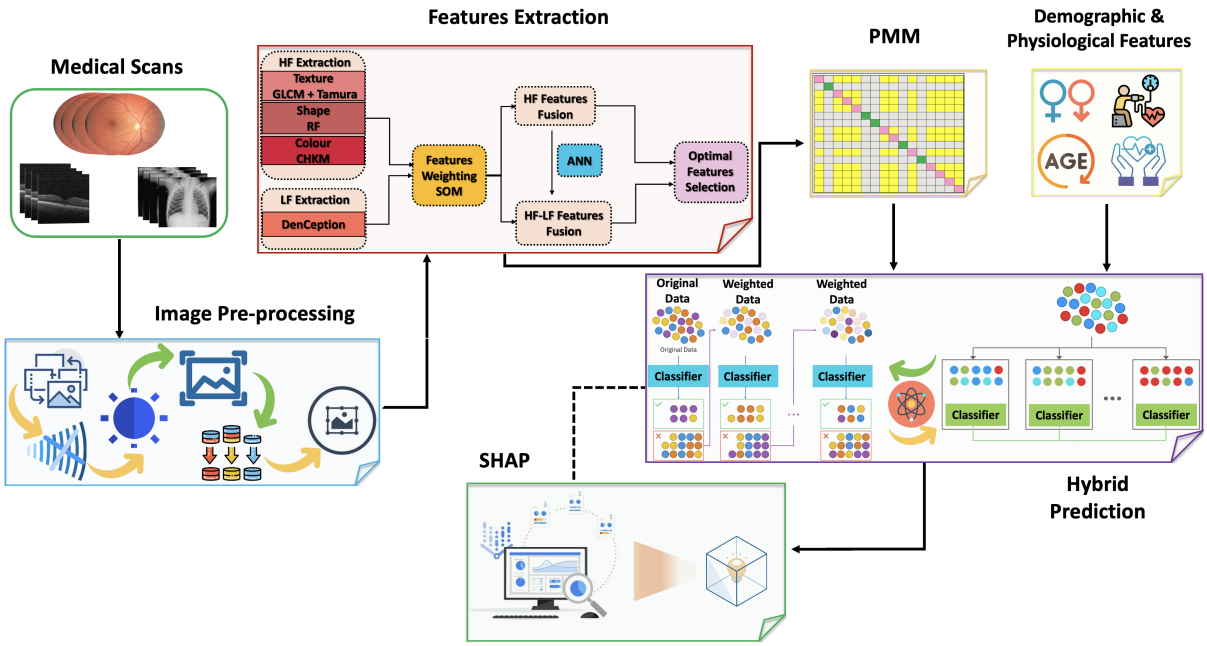


Figure 5.5: Proposed Prediction Framework.

5.4 Proposed Prediction Framework

In this section, a detailed explanation of the proposed prediction framework will be presented. The section will also convey the applied primary and secondary datasets as well as the various experiments performed to evaluate and validate the proposed framework. The importance of annotations, particularly labels, in providing additional information about features within a medical image is crucial in prediction tasks. This makes it easier for DL- and ML-based algorithms to understand and interpret medical images. The used datasets already fulfill this stage, thus, it has not been incorporated as a major block in the proposed prediction framework, as shown in Figure 5.5.

5.4.1 Block 1: Image Pre-Processing

Image pre-processing is an instrumental aspect in the pipeline of training DL models, particularly when working with medical datasets such as those involving DR, DMO, and pneumonia distribution. Multiple stages compose this block as illustrated in Figure 5.6.

Adjusting image dimensions through image resizing is a crucial step in pre-processing that en-

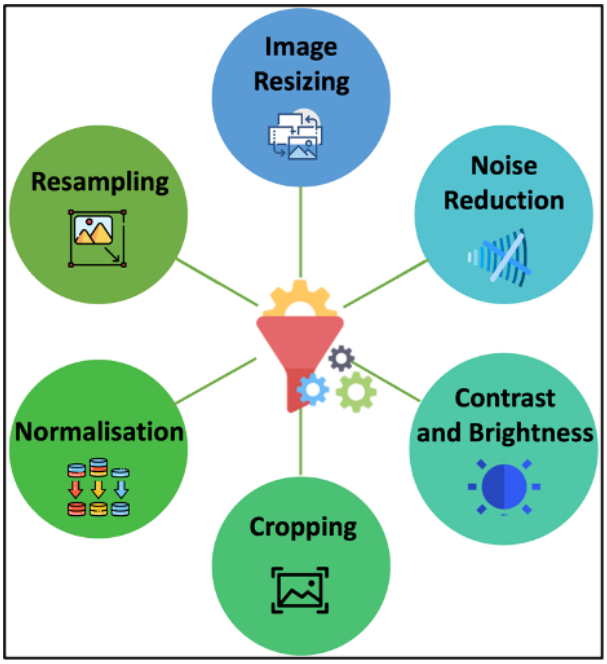


Figure 5.6: Image Pre-processing Steps.

sures consistent input to models and algorithms. Additionally, applying noise reduction techniques is important to remove noise and artifacts from the images. The process also involves enhancing and normalising the contrast of the image to improve the visibility of details and adjusting the brightness to make the medical image more appealing and suitable for analysis. Cropping is performed to remove non-essential parts of the medical image, thereby focusing on the RoI. Normalisation is another essential step where pixel values are scaled to a specific range to ensure consistency and compatibility with ML and DL algorithms. Lastly, resampling is used to change the image resolution by either upscaling or downscaling, which is necessary to adapt the image to different display and processing requirements.

5.4.2 Block 2: Features Extraction

The second major block is presented by features extraction (Loukil, Mirza, Sayers, and Awan, 2023). This block consists of extracting HF and DHF features from used medical images-based datasets. Various steps have been incorporated to finalise the list of features prepared for the prediction block. These steps are summarised as follows:

- Extracting HF features to include: texture contrast, texture energy, texture homogeneity, entropy, coarness, directionality, mean colour value, mean standard deviation value, shape area, and shape perimeter.
- Extract DHF features using DenCeption (Loukil and Salah, 2020).
- Automatic assignment of weights based on features importance in each used dataset.
- Features reduction and selection.

5.4.3 Block 3: Additional Features Incorporation

The third block consists of the incorporation of additional demographic and physiological features depending on the applied dataset. Below is a list of considered features:

- Age
- Sex
- SBP
- DBP
- Diabetic type and CRT for ophthalmology related datasets.

The inclusion of additional features in prediction tasks holds significant importance in medical and clinical data analysis. These auxiliary features provide valuable context and patient specific information that can greatly enhance the accuracy and clinical relevance of the predictive models. For instance, age is a fundamental factor influencing disease risk and progression, as health conditions often vary with age. Sex can also be a key determinant of disease prevalence and presentation. Additionally, blood pressure metrics are essential for predicting cardiovascular and hypertension-related outcomes. Disease type not only guides the choice of predictive models but also adds domain-specific knowledge to the analysis. Moreover, CRT measurements are

vital in ophthalmological related predictions. By incorporating these features, predictive models can be tailored to individual patients, leading to more precise risk assessments and treatment recommendations, ultimately improving patient care and healthcare decision-making.

5.4.4 Block 4: Prediction

The fourth major block covers the prediction/classification task. This block represents the core of the decision-making process, utilising a comprehensive set of features to provide precise and holistic insights. By combining HF, DHF, and supplementary patients' information, the framework achieves a multidimensional perspective on the patient's health and the associated risks. This integrated approach enhances the predictive power of the proposed framework, offering a more nuanced understanding of the underlying medical conditions and their potential outcomes. The HF features capture diverse visual patterns, while DHF features abstract complex information, and additional features contribute vital contextual details, allowing the prediction model to make informed decisions. With all these components working in concert, the prediction block becomes a powerful tool for risk assessment, disease diagnosis, and treatment recommendations offering a holistic and patient-centred approach to healthcare decision support, which is not part of this work's scope.

In the evolving landscape of automated prediction models, hybrid models that combine the strengths of multiple algorithms often emerge as powerful solutions to challenging prediction problems. In this work, one such promising combination is proposed by integrating XGBoost and AdaBoost, two renowned ensemble methods with distinctive advantages. XGBoost, an exemplar of gradient boosting techniques, is known for its ability to handle missing data, capture complex patterns, and deliver highly accurate predictions, especially with structured data. On the other hand, AdaBoost, standing for AdaBoost, operates by iteratively focusing on misclassified instances, thus ensuring that the model gives due attention to more challenging data points. By synthesising the capabilities of XGBoost and AdaBoost, the hybrid model aims to integrate XGBoost's advantages in delineating complex data relationships with AdaBoost's adaptive learning mechanism. Such a fusion not only augments the robustness of predictions

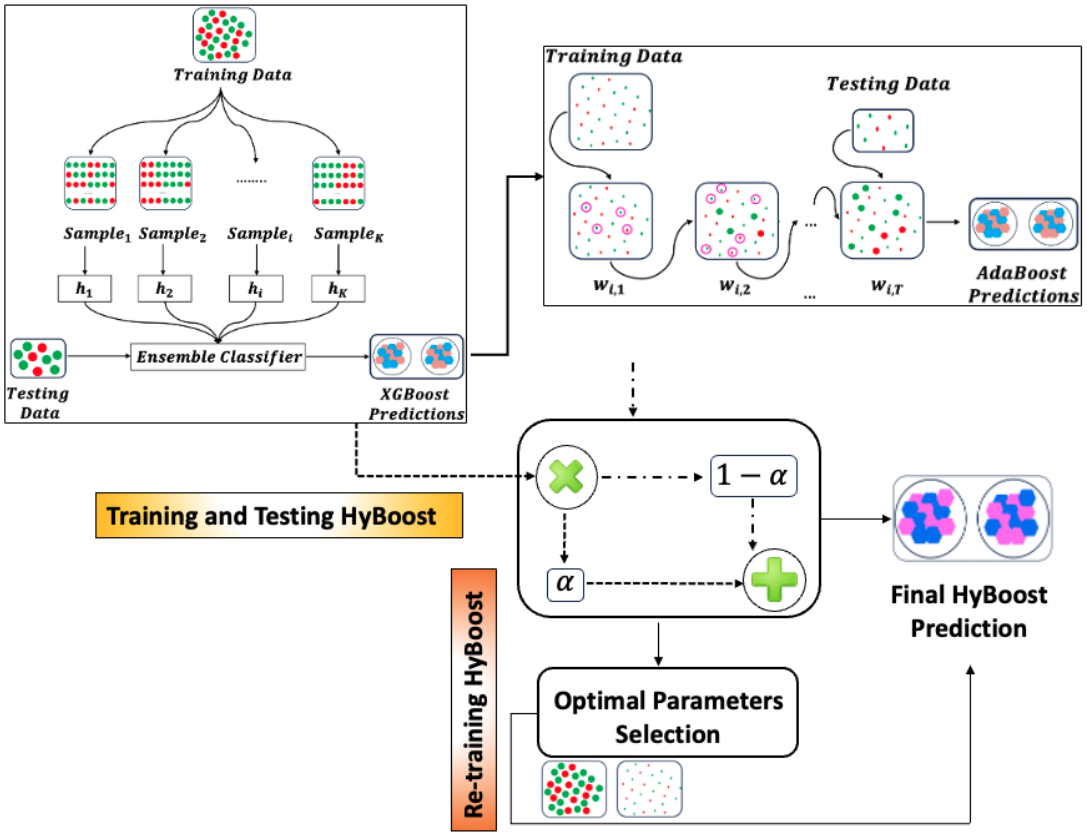


Figure 5.7: HyBoost Blocks: Training, Testing and Re-training Phases.

but also offers an innovative approach to capitalise on the complementary strengths of both algorithms. This combination is envisioned to outperform individual models, providing a sophisticated tool for the proposed predictive framework. Figures 5.7 and 5.8 shows the major steps constructing HyBoost. These steps are summarised in Table 5.1. The algorithm of the proposed hybrid model is illustrated in Algorithm 4.

During data preparation, the data is split into training and validation sets. This step is followed by feature engineering, normalisation, and the handling of any missing values. After this, AdaBoost is initialised by assigning equal weights to all instances in the training set. In every boosting iteration of AdaBoost, an XGBoost model is trained on the weighted dataset. This model then predicts the outcomes for the validation set. The weighted error rate of the XGBoost model is calculated, which in turn determines the influence that particular model has in the final prediction, based on its error rate. Weights are then updated — increased for instances that were incorrectly predicted and decreased for those correctly predicted. After

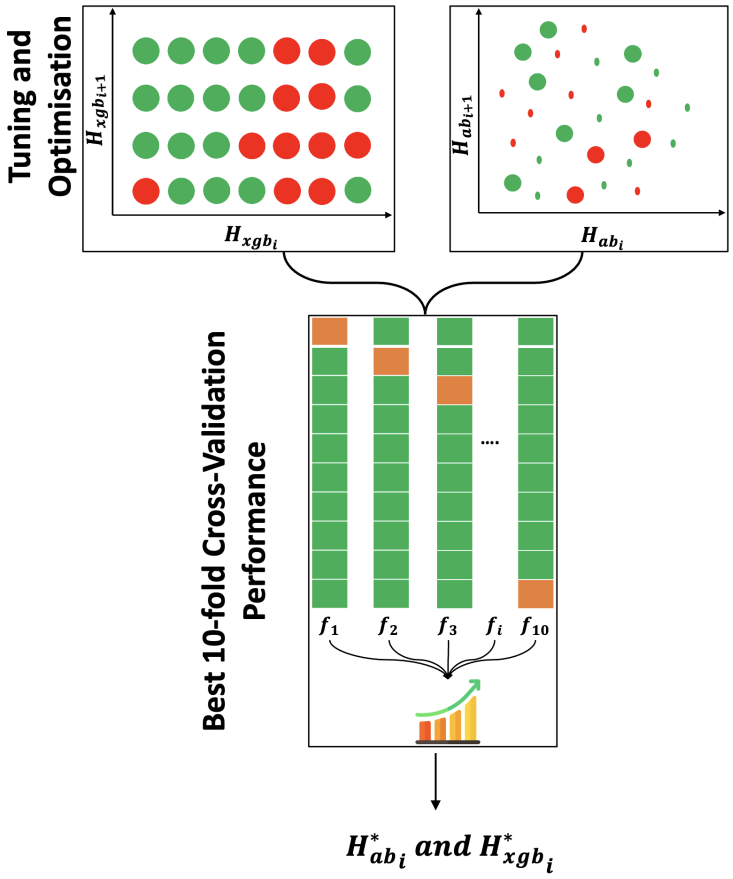


Figure 5.8: Optimal Parameters Selection Block of HyBoost Model.

adjusting, the weights are normalised. When making the final prediction for a new instance, each individual XGBoost model gives its prediction for the target. The conclusive prediction is a weighted vote, with weights being determined by the influence of each model. The outcomes of the prediction model is incorporated in the fifth block, namely feature rationale using Shapley Additive explanation (SHAP).

5.4.5 Block 5: Shapley Additive Explanation

The inclusion of SHAP within the proposed framework holds significant importance in post-prediction analysis. SHAP is an advanced interpretability technique that provides insights into the block box nature of DL/ML models, shedding light on the rationale behind a model's predictions. After the prediction block, it is crucial to understand why a specific decision was made, especially in the context of healthcare application. By employing SHAP values, the

framework explains the contribution of each feature, be it HF image details, DHF abstractions, or additional patient-specific information to the final prediction. This level of transparency is paramount for trust and accountability in the medical domain, where even slight misunderstanding or misinterpretation can have profound consequences. SHAP enables to grasp the driving factors behind each prediction, facilitating the assessment of model behaviour, the identification of potential biases, and the fine-tuning of the model for enhanced accuracy and fairness. This block will ensure that the decision-making process not only accurate but also comprehensible and justifiable, thereby, increasing its utility and trust worthiness in real world medical applications.

Table 5.1: HyBoost Major Blocks Description

Phase	**Step**	**Description**
HyBoost Training	XGBoost Training	• Initialisation of XGBoost with H_{xgb} and train on D_{train} to get the model M_{xgb} • Prediction on D_{train} using M_{xgb} to get y_{xgb} • Computation of residuals
	AdaBoost Training	Initialisation of AdaBoost with H_{ab} and train on D_{train} with the residuals to get the model M_{ab}.
Prediction		For each instance x_j in D_{test}: • Prediction using M_{xgb} to get $\hat{y}_{xgb_j}$ • Prediction using M_{ab} to get $\hat{y}_{ab_j}$ • Computation of final prediction Y_{final} using the weighted parameter $\alpha \in [0,1]$
Tuning and Optimisation		• Using 10-fold cross-validation to find optimal H_{xgb}, H_{ab}, and α.

Continued on next page

Table 5.1: HyBoost Major Blocks Description (Continued)

Phase	Step	Description
Optimal Parameters Selection		• Selection of H^*_{xgb}, H^*_{ab}, and α^* based on best 10-fold cross-validation performance.
Retraining		• Retraining M_{xgb} and M_{ab} on D_{train} using H^*_{xgb}, H^*_{ab}, and α^*.

5.4.6 Block 6: Adaptive Performance Evaluation

The final block consists of evaluating the performance of the proposed framework using an adaptive evaluation matrix (PMM) proposed in (Loukil, Mirza, and Sayers, 2023). PMM is a novel adaptive evaluation mechanism designed to enhance the performance evaluation of ML and DL models. This mechanism consists of three primary components: problem specification, task identification (including prediction and classification), and data characteristics (incorporating features, class distribution, and data balance specifications) as shown in Figure 5.9.

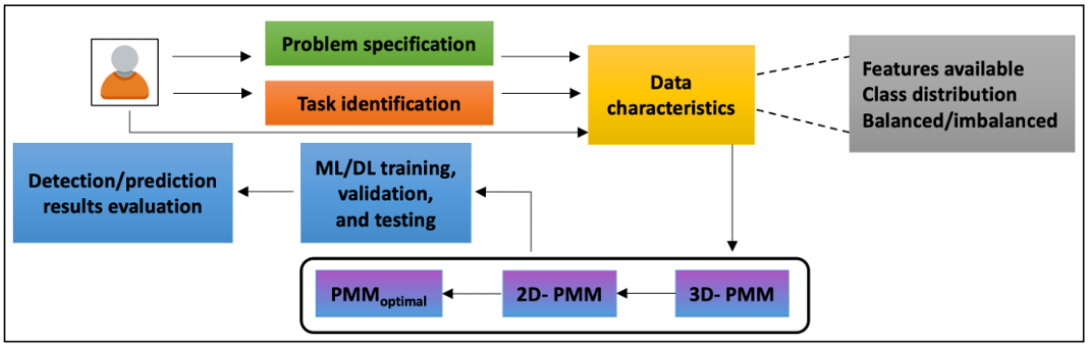


Figure 5.9: Proposed Performance Evaluation Framework

The approach relies on user input regarding the specifications of each component, making the method adaptable to various input scenarios. The mechanism then processes this input to form a three-dimensional PMM, where each dimension reflects the importance of the associated metrics. This importance is assesed using a correlation coefficient formula that assigns weights to each performance metric. The 3D-PMM is constructed using an algorithm that processes

each dimension in parallel. The primary purpose of the 3D-PMM tool is to identify evaluation parameters in a reliable and adaptive manner, resulting in an optimal set of metrics (PMM optimal vector). The suggested approach consists of three main components x_1, x_2, and x_3, respectively as follows:

- Problem specification,
- Task identification to include prediction, classification, etc...
- Data characteristics to include features, classes distribution, and balanced/imbalanced specification.

The consideration of these three components has the potential to efficiently enhance performance evaluation of ML/DL based models. The proposed evaluation mechanism consists of getting input from the user covering the specification of each component; this would vary depending on the input scenario which makes the proposed mechanism adaptive. As a follow-up step, provided features, related class distribution, as well as data balancing information are used as a base knowledge to the proposed 3D PMM matrix (3D-PMM). Each component has a specific contribution to the proposed 3D-PMM matrix final output, as shown in Figure 5.10. Each matrix dimension is characterised by *nxn* size, where *n* defines the total number of performance metrics. In fact, each dimension reflects the importance of involved metrics by assigning a correlation coefficient reflecting their weights, hence their importance, defined as below (Equation (5.1)):

$$corr_{coef}(y,z) = \frac{Cov(y,z)}{\sigma_y * \sigma_z} \tag{5.1}$$

$$= \frac{\sum[(y - mean(Y)) * (z - mean(Z))]}{\sqrt{[\sum(y - mean(Y))^2 * \sum(z - mean(Z))^2]}}$$

where y, z are the weights of the performance metric, Y, Z are the set of performance metrics for a particular dimension. The weights are defined as W_{Cii}, where $i \in 1..n$ of each dimension noted by, (x_1,x_1), (x_1,x_3), and (x_2,x_3), respectively. The 3D-PMM is implemented as shown

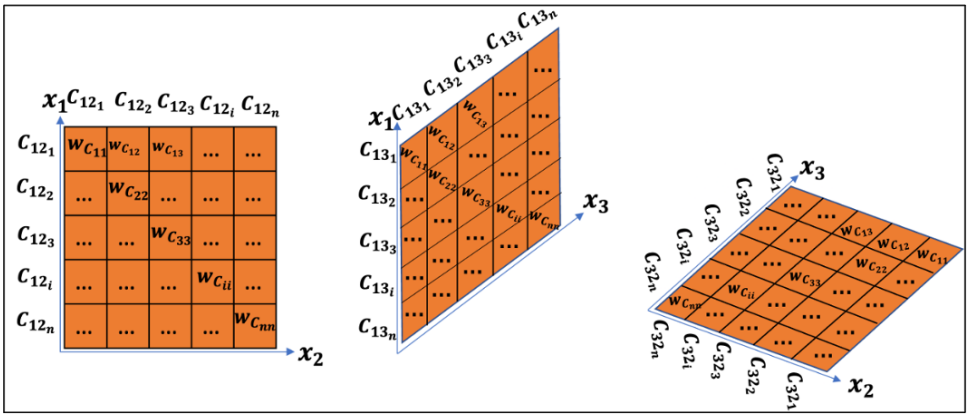


Figure 5.10: Three-Dimensional Representation of PMM Matrix

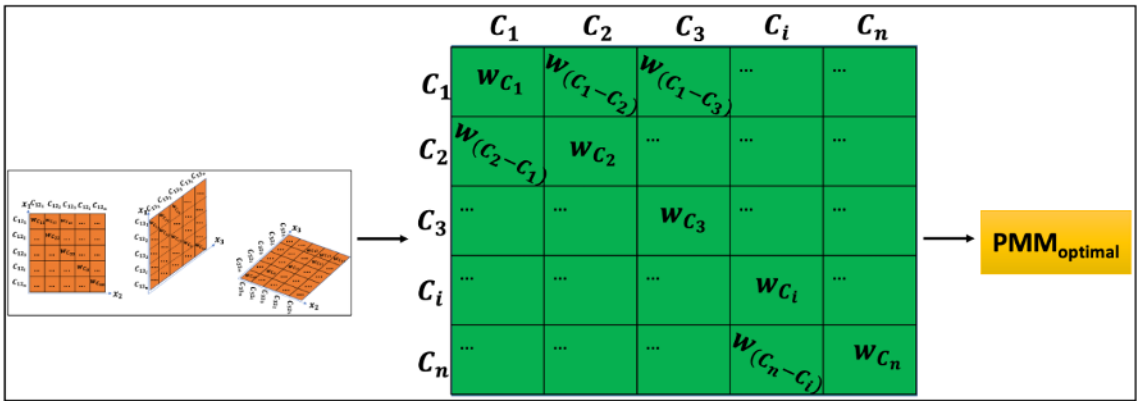


Figure 5.11: Conversion of Two-Dimensional PMM Matrix into Optimal Set of Performance Measurement Metrics Vector

in Algorithm 3, comprising three main parallel processes of each separate dimension.

3D-PMM tool intend to efficiently and adaptively identifying reliable and comprehensive evaluation parameters denoted by $PMM_{optimal}$ vector, as shown in Figure 5.11. The resulted optimal set of metrics is then used as part of the performance measurement of the given ML/DL based model.

5.5 Rationale Behind the Selection of the HyBoost Model

The selection and design of the HyBoost model were guided by a strategic and evidence-based approach, aimed at addressing specific challenges in medical image analysis and disease prediction. The decision to integrate XGBoost and AdaBoost into the HyBoost model was not arbitrary but based on a detailed assessment of their complementary strengths and their ability

Algorithm 3: Optimal Performance Evaluation Metrics

Input: $C_{(x_1,x_2)}, C_{(x_1,x_3)}, C_{(x_2,x_3)}$: set of evaluation metrics for each dimension
Output: $PMM_{optimal}$

$W \Leftarrow$ weight assignment function
$n \Leftarrow$ total number of metrics
$c \Leftarrow$ performance metric
$x_1 \Leftarrow$ problem specification
$x_2 \Leftarrow$ task identification
$x_3 \Leftarrow$ data characteristics
$C_{(x_1,x_2)} \Leftarrow [c_{121}, c_{122}, ..., c_{12n}]$: set of performance metrics for dimension (x_1,x_2)
$C_{(x_1,x_3)} \Leftarrow [c_{131}, c_{132}, ..., c_{13n}]$: set of performance metrics for dimension (x_1,x_3)
$C_{(x_2,x_3)} \Leftarrow [c_{231}, c_{232}, ..., c_{23n}]$: set of performance metrics for dimension (x_2,x_3)
$D_{12}, D_{13}, D_{23} \Leftarrow$ 2D matrix of (x_1,x_2), (x_1,x_3), and (x_2,x_3) dimensions, respectively
Step 1:
for $i \in \{1,\ldots,n\}$ **do**
    $W_{C_{12}}[i] \Leftarrow W(C_{12}[i])$
Step 2:
for $i \in \{1,\ldots,n\}$ **do**
    for $j \in \{1,\ldots,n\}$ **do**
        if $i == j$ **then**
            $W_{C_{12ii}} \Leftarrow corr_{coef}(C_{12i}, C_{12i})$
            $D_{12}[i,i] \Leftarrow W_{C_{12ii}}$
        else
            $W_{C_{12ij}} \Leftarrow corr_{coef}(C_{12i}, C_{12j})$
            $D_{12}[i,j] \Leftarrow W_{C_{12ij}}$
Repeat Step 1 and 2 for D_{13} and D_{23}
for $i \in \{1,\ldots,n\}$ **do**
    for $j \in \{1,\ldots,n\}$ **do**
        if $i == j$ **then**
            $PMM[i,j] \Leftarrow max(D_{12}[i,i], D_{13}[i,i], D_{23}[i,i])$
        else
            $PMM[i,j] \Leftarrow max(D_{12}[i,j], D_{13}[i,j], D_{23}[i,j])$
for $i \in \{1,\ldots,n\}$ **do**
    for $j \in \{1,\ldots,n\}$ **do**
        if $PMM[i,j] > 0$ **then**
            $PMM_{optimal}[i] \Leftarrow PMM[i,j]$
        else
            Continue
return $PMM_{optimal}$

to meet the complex requirements of medical diagnostics.

5.5.1 Addressing Identified Challenges in Disease Prediction

The development of the HyBoost model was motivated by the need to overcome the limitations of traditional predictive models in handling complex medical imaging data. In particular, issues such as model overfitting, bias, interpretability, and computational demands were identified as critical challenges in the application of DL and ML in healthcare. The integration of XGBoost and AdaBoost was carefully considered to address these challenges effectively.

5.5.2 Justification for XGBoost

XGBoost was chosen as a core component of the HyBoost model due to its proven ability to handle large, structured datasets with high dimensionality. XGBoost excels in capturing complex patterns and relationships within the data, which is essential for accurate disease prediction. Its regularisation techniques help to prevent overfitting, making it well-suited for medical datasets where the risk of overfitting is high due to the complexity and variability of the data. Moreover, XGBoost's efficiency in handling missing data and its scalability across different hardware environments make it a robust choice for medical applications.

5.5.3 Justification for AdaBoost

AdaBoost was selected to complement XGBoost in the HyBoost model due to its adaptive learning mechanism. AdaBoost focuses on improving the model's accuracy by iteratively reweighting misclassified instances, thereby enhancing the model's ability to learn from difficult or minority cases. This is particularly useful in medical imaging, where certain disease patterns may be less prevalent or harder to detect. The combination of AdaBoost with XGBoost enhances the model's overall robustness, ensuring that it can accurately predict outcomes even in challenging scenarios.

5.5.4 Integration of XGBoost and AdaBoost in HyBoost

The integration of XGBoost and AdaBoost into a hybrid model, HyBoost, was driven by the goal of combining their strengths to create a more powerful predictive tool. XGBoost's ability to model complex relationships and AdaBoost's focus on difficult cases result in a model that not only provides high accuracy but also generalises well across different datasets. This hybrid approach is particularly advantageous in medical diagnostics, where diverse data sources and patient characteristics can complicate predictions.

5.5.5 Practical Benefits of HyBoost

The HyBoost model offers several practical benefits that make it particularly useful for medical image analysis. By leveraging the complementary strengths of XGBoost and AdaBoost, HyBoost provides a more nuanced understanding of medical data, leading to more accurate and reliable disease predictions. This is further enhanced by the model's ability to incorporate additional patient-specific features, such as demographic and physiological data, which enrich the predictive process. The inclusion of SHAP values in the framework ensures that the model's decisions are transparent and interpretable, addressing a critical need for explainability in clinical settings.

5.5.6 Alignment with Research Goals

The selection of the HyBoost model aligns with the overarching goals of this research, which aims to advance the field of medical diagnostics through the development of robust, scalable, and interpretable predictive models. HyBoost's hybrid architecture is designed to meet the demands of modern healthcare environments, where accuracy, efficiency, and adaptability are paramount. The model's ability to handle diverse data and provide clear explanations for its predictions positions it as a valuable tool for clinicians, enhancing their ability to make informed decisions based on complex medical data.

5.6 Datasets

The datasets used in this chapter encompass a wide range of medical imagery and associated patient information, offering a rich foundation for various analyses, particularly in the domains of ophthalmology and pulmonology.

The selection of different datasets, including the Fundus dataset as the primary dataset, and the OCT and X-ray datasets as secondary datasets, was a deliberate choice aimed at thoroughly evaluating the adaptability and generalisability of the proposed models, particularly the HyBoost framework. Below is a detailed justification addressing the differences between the primary and secondary datasets and the rationale behind using these diverse datasets.

5.6.1 Fundus Dataset as the Primary Dataset

DR is a leading cause of blindness, and early detection is crucial. The Fundus dataset provides a rich set of labelled images, making it ideal for developing robust predictive models. The Fundus dataset was selected as the primary dataset because of its critical relevance to ophthalmology, particularly in the diagnosis of DR, a prevalent and visually debilitating condition. This dataset provides a balanced and comprehensive set of images that are ideal for training DL models, making it the cornerstone of the research in Chapter 5. The focus on Fundus images allowed for the development and fine-tuning of the HyBoost model in a specific, high-impact area of medical diagnostics. The dataset includes both normal and DR-affected images, along with demographic and physiological data, which provide a complex yet well-rounded foundation for training the model. As the primary dataset, Fundus images allow the research to focus on a specific medical condition, ensuring the model is trained on a dataset with clear clinical relevance and well-defined outcomes.

5.6.2 OCT and X-ray Datasets as Secondary Datasets: Selection Justification

The OCT and X-ray datasets were introduced as secondary datasets to evaluate the HyBoost model's performance across different medical imaging modalities and disease contexts. By testing the model on these additional datasets, the research aims to demonstrate the model's adaptability and ensure its robustness across diverse scenarios.

OCT Dataset

- **Complementary to Fundus Dataset:** OCT scans provide cross-sectional images of the retina, which complement the 2D Fundus images by offering additional depth information crucial for diagnosing retinal conditions such as DMO. This allows the model to learn and adapt to a different type of imaging data while staying within the domain of ophthalmology.
- **Testing Versatility:** By applying the model to OCT data, the research tests the model's ability to handle different image characteristics, such as variations in texture and structure that are distinct from those in Fundus images.

X-ray Dataset

- **Diverse Medical Field:** The X-ray dataset shifts the focus from ophthalmology to pulmonology, thereby expanding the scope of the model's application. This dataset includes chest X-rays used to diagnose conditions like pneumonia, which introduces a different type of imaging modality with unique challenges, such as detecting subtle variations in lung tissue.
- **Ensuring Generalisation:** Using the X-ray dataset helps in assessing the generalisation capabilities of the HyBoost model across entirely different types of medical imaging. The ability to accurately process and analyse X-ray images indicates the model's robustness and adaptability to various medical diagnostic tasks.

5.6.3 Rationale for Using Different Datasets

- **Addressing the Need for Adaptability:** The primary reason for using different datasets is to ensure that the models developed are not only effective within a single domain but are adaptable to various medical imaging contexts. This approach is crucial for creating versatile models that can be applied across different clinical scenarios, enhancing their practical utility in real-world settings.

- **Ensuring Comprehensive Model Evaluation:** By employing multiple datasets, the research provides a comprehensive evaluation of the HyBoost model. The primary dataset (Fundus) allows for the focused development of the model, while the secondary datasets (OCT and X-ray) offer additional layers of validation, ensuring that the model can maintain high performance across different types of medical images.

- **Mitigating Overfitting and Bias:** Utilising datasets from different medical fields reduces the risk of overfitting and helps to identify and mitigate any biases that might arise from training on a single type of data. This approach enhances the model's reliability and robustness, making it more likely to perform well in diverse clinical environments.

5.6.4 Visualisation of Additional Parameters Across Datasets: Focus on Fundus and OCT Datasets

Understanding the distribution of patient demographics and clinical parameters is crucial for interpreting the results of medical image analysis and the performance of predictive models. The following section present visual analyses of key parameters such as age distribution and CRT across different diabetic types, specifically in the context of Fundus images affected by DR and OCT scans affected by DMO. These visualisations provide insights into how these variables vary across different subtypes of diabetes and how they correlate with specific retinal conditions.

Figures 5.12.a and 5.12.b illustrate the age distribution of patients, categorised by diabetic

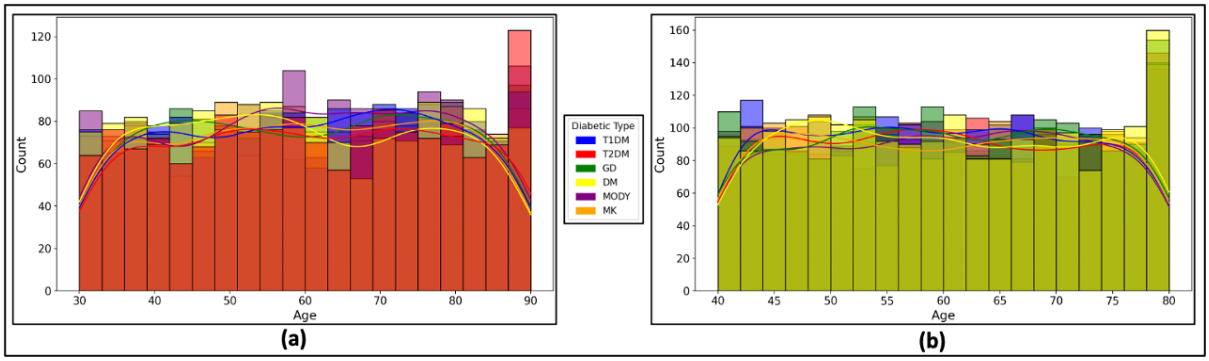


Figure 5.12: Age Distribution by Diabetic Type for - (a): Fundus Images Affected by DR and (b): OCT Scans Affected by DMO.

type, for Fundus images affected by DR and OCT scans affected by DMO, respectively. This analysis highlights the age-wise distribution of patients with ocular complications due to diabetes, showcasing significant trends across various diabetic subtypes. The data reveals that T2DM patients consistently form the largest group across different age ranges in both datasets, emphasising the widespread impact of T2DM on ocular health. Notably, while the Fundus dataset reflects a relatively stable number of DR-affected individuals aged 30 to 70, the OCT dataset shows an increase in DMO-affected individuals from 40 to 80 years. These trends underscore the critical role of T2DM in the progression of retinal diseases like DR and DMO across different age groups.

Further, Figures 5.13.a and 5.13.b depict the distribution of CRT values across different diabetic types for patients with DR, as observed in Fundus images, and for patients with DMO, as observed in OCT scans. The visualisations reveal a cohesive pattern in CRT values, predominantly around the 600 mark, across both datasets. Interestingly, the MODY diabetic subtype consistently exhibits lower median CRT values in both retinal conditions, suggesting a unique manifestation of retinal complications specific to this diabetic group. This detailed comparison not only highlights the similarities in CRT distribution across different diabetic types but also underscores the distinct retinal characteristics associated with specific subtypes, thereby providing a deeper understanding of the pathophysiological differences in diabetic retinal diseases.

Algorithm 4: Algorithm for HyBoost Hybrid Predictive Model

Data: $D_{train} = \{(x_i, y_i)\}_{i=1}^{N}$, $D_{test} = \{x_j\}_{j=1}^{M}$, M_{xgb} (XGBoost model), M_{ab} (AdaBoost model), H_{xgb}, H_{ab}, K, T, $\alpha \in [0,1]$

Result: Y^*_{final} (final best prediction)

Initialisation:

$\hat{y}_{xgb} \Leftarrow M_{xgb}$'s residuals (predictions)

$\hat{y}_{ab} \Leftarrow M_{ab}$'s predictions

$Y_{final} \Leftarrow$ final prediction prior to optimisation

Step 1: HyBoost Training Phase

Step 1.1: Train XGBoost

for $k \in \{1, \ldots, K\}$ **do**

$Ob_i \Leftarrow \sum_{i=1}^{N} l\left(y_i, \hat{y}_i^{(k-1)} + f_k(x_i) + \Omega(f_k)\right)$

$g_i \Leftarrow \frac{\partial}{\partial \hat{y}_i^{(k-1)}} l\left(y_i, \hat{y}_i^{(k-1)}\right)$

$h_i \Leftarrow \frac{\partial^2}{\partial (\hat{y}_i^{(k-1)})^2} l\left(y_i, \hat{y}_i^{(k-1)}\right)$

$\hat{y}_i^{(k)} \Leftarrow \hat{y}_i^{(k-1)} + \eta f_k(x_i)$

$M_{xgb}(x_i) \Leftarrow XGBoost(D_{train}, H_{xgb})$

residuals $\Leftarrow y_i - \hat{y}_{xgb}$

Step 1.2: Train AdaBoost

for $t \in \{1, \ldots, T\}$ **do**

$\alpha_t \Leftarrow \frac{1}{2} \ln\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right)$

$w_{i,t+1} \Leftarrow w_{i,t} * \exp(-\alpha_t y_i h_t(x_i))$

$M_{ab}(x_i) \Leftarrow AdaBoost(D_{train}, \text{residuals}, H_{ab})$

Step 2: HyBoost Prediction Phase

for $j \in \{1, \ldots, N\}$ **do**

$\hat{y}_{final_j} \Leftarrow \alpha * \hat{y}_{xgb_j} + (1-\alpha) * \hat{y}_{ab_j}$

Step 3: Tuning and Optimisation

Optimise $(H_{xgb}, H_{ab}, \alpha)$ with cross-validation

Step 4: Optimal Parameters Selection

$H^*_{xgb}, H^*_{ab}, \alpha^* \Leftarrow \arg\max_{H_{xgb}, H_{ab}, \alpha} \text{CrossValidationScore}(D_{train}, H_{xgb}, H_{ab}, \alpha)$

Step 5: Re-training Phase

for $j \in \{1, \ldots, M\}$ **do**

$Y^*_{final_j} \Leftarrow \alpha^* * M^*_{xgb}(x_j) + (1-\alpha) * M^*_{ab}(x_j)$

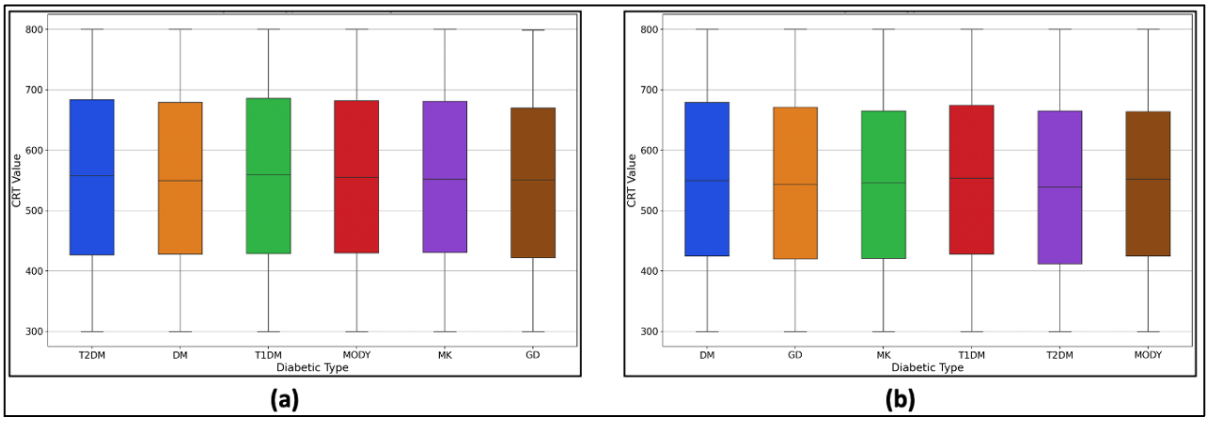


Figure 5.13: CRT Distribution by Diabetic Type for – (a): Fundus Images Affected by DR and (b): OCT Scans Affected by DMO.

5.7 Experimentation Scenarios

In the proposed disease prediction framework, a series of comprehensive experiments has been designed to ensure robustness and generalisability of the model. The aforementioned datasets, derived from Fundus and OCT imaging, will be at the forefront of the training phase. In contrast, the X-ray dataset, a modality distinct from the training datasets, will act as the validation set. This strategy ensures that the prediction model does not just fit the nuances of a particular imaging technique but extends its diagnostic proficiency across diverse medical imaging platforms, thus testing its generalisability. The first phase of experiments revolves around evaluating a variety of classifiers, namely RF, DT, XGBoost, and AdaBoost, as well as the proposed HyBoost model. To rigorously assess their performance, two distinct scenarios will be considered, as follows:

- **Baseline Scenario:** The classifiers will be trained on the raw imaging data, without any supplementation of demographic and physiological features. This provides an absolute performance metric, reflective of the fundamental capabilities of the classifiers when applied directly on imaging datasets.

- **Enhanced Scenario:** In this setup, the classifiers will be enhanced with additional demographic and physiological features, aiming to exploit the potential complementarities that these combined feature sets might offer.

Each of these scenarios will further devide into two experimental paths: (1) where the classifiers are deployed with their default hyperparameters and (2) where they undergo meticulous hyperparameter tuning. Such a methodical approach ensures that all possibilities have coevered to obtain the best possible model.

For benchmarking phase, SHAP values will be employed across top three well performed methods for post-classification evaluation. SHAP will delve deep into the results, interpreting and explaining the decision-making processes of the best method used in the benchmarking phase versus the proposed framework. This not only provides an interpretability layer over the predictive framework but also offers invaluable insights that might be pivotal for clinical validations.

Towards evaluating the results of each of the experiments, the robust PMM matrix will be applied. By carefully analysing each dataset's results, the aim to discern the most fitting performance metrics for each scenario, providing a detailed and refined understanding of the proposed models' strengths and areas of improvement. Taking into account disease prediction, classification, and the previously mentioned datasets for 3D-PMM, the resulting optimal vector comprises metrics to include: ROC curve, PR curve, AUC score, Precision, Recall, F1-score, accuracy, specificity, and LCE. These metrics will be utilised to evaluate all upcoming experiments.

Fine-tuning hyperparameters is a critical step in the development of the proposed prediction model. Hyperparameters, unlike model parameters that are learned during training, are pre-set configurations that can significantly influence the performance of a model. Properly tuned hyperparameters can make the difference between an average model and a highly accurate one. Without the right hyperparameters, even the most sophisticated algorithms might fail to provide satisfactory results or might take an inefficiently long time to train. On the other hand, with optimal hyperparameters, the same algorithms can achieve impressive performance in much less time. Therefore, comparing model performances with and without hyperparameter tuning is of paramount importance. Such a comparison provides insights into the potential enhancements brought about by tuning and highlights the necessity of investing time and resources into

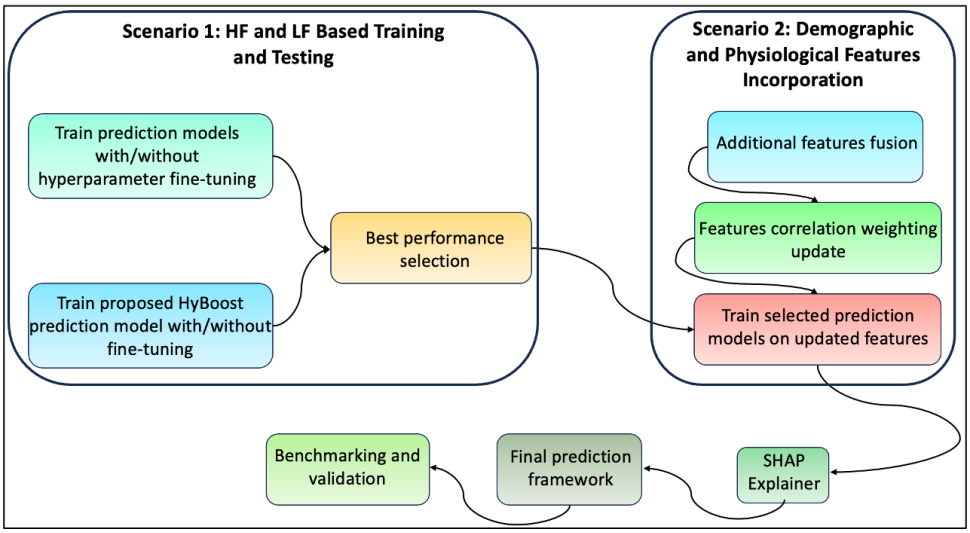


Figure 5.14: Experimentation Process.

this often-overlooked step. It demonstrates the direct impact of hyperparameter choices on a model's accuracy, efficiency, and overall effectiveness, emphasising the indispensable role they play in model optimisation. Table 5.2 provides a brief overview of the key hyperparameters for each tested model.

In the upcoming experiments, the aim is to demonstrate a compelling proposition: even when individual models like RF, DT, AdaBoost, and XGBoost are fine-tuned to their optimal performance, they might still fall short in comparison to the proposed hybrid model. By accurately adjusting hyperparameters and optimising each model, the purpose is to ensure an equitable basis field for comparison. The hypothesis suggests that the integrated fusion of AdaBoost and XGBoost in the proposed hybrid approach will consistently outperform these individual models, even at their best. This exploration is pivotal, shedding light on the potential benefits of integrating the strengths of multiple models into a cohesive hybrid system. Figure 5.14 illustrates the experiments done on each dataset.

Table 5.2: Hyperparameters Overview

Model	**Hyperparameters**	**Description**
RF	n estimators, max depth, min samples split, min samples leaf	Number of trees in the forest, Maximum depth of tree, Minimum samples required to split, Minimum samples at leaf node
DT	criterion, splitter, max depth, min samples split, min samples leaf	Function to measure the quality of a split, Strategy used, Maximum depth of tree, Minimum samples required to split, Minimum samples at leaf node
XGBoost	learning rate, n estimators, max depth, min child weight, gamma, subsample, colsample bytree	Step size shrinkage, Number of boosting rounds, Maximum depth, Minimum sum of instance weight, Minimum loss reduction, Fraction of samples, Fraction of features
AdaBoost	n estimators, learning rate, algorithm	Number of weak learners, Learning rate, Algorithm used
HyBoost	learning rate, n estimators, max depth, min child weight, gamma, subsample, colsample bytree, n estimators, learning rate, algorithm	Step size shrinkage, Number of boosting rounds, Maximum depth, Minimum sum of instance weight, Minimum loss reduction, Fraction of samples, Fraction of features, Number of weak learners, Learning rate, Algorithm used

5.8 Features Extraction and Sample Testing

5.8.1 Data Preparation: Image Pre-Processing Outcome

Figure 5.15 presents the original OCT, Fundus, and X-ray as well as the output of pre-processing step, respectively. As per the obtained results, image resizing ensures a uniform input shape for the neural, optimising memory usage and computational efficiency.

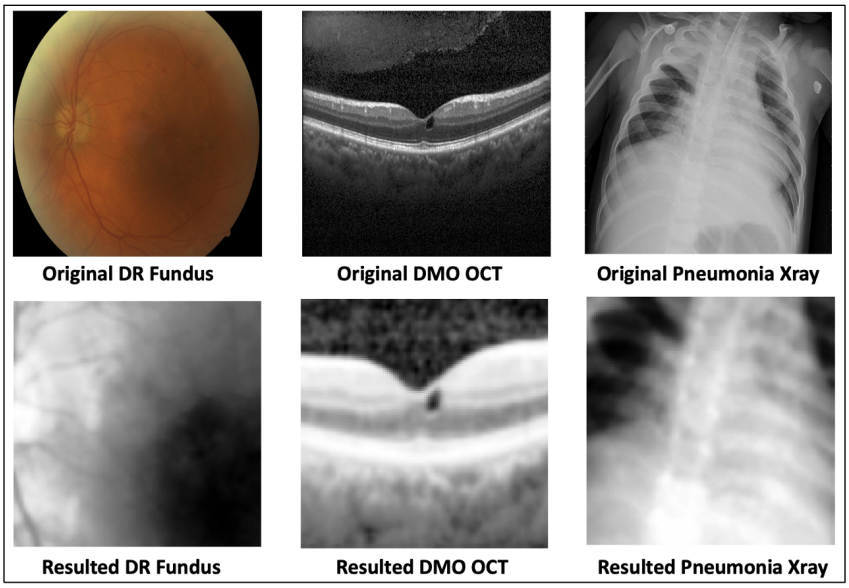


Figure 5.15: Image Pre-Processing Outcome.

Noise reduction, typically executed using methods like Gaussian blur, is pivotal for these datasets because it eliminates unwanted variations and artifacts that might not be representative of actual pathological changes. This step accentuates authentic features of the image, ensuring that the DL model is trained on real patterns rather than false noise. Next, contrast adjustment significantly improves the visibility of implicit features in retinal and lung images. Given that DR, DMO, and pneumonia can present with nuanced morphological alterations, enhancing the contrast through equalise hist method allows the model to discern and learn from these critical, yet slight, changes more effectively. Cropping, when applied prudently using extracted image contrast, centers the model's attention on the most relevant part of the image, such as the macula or the OD, ensuring that it's focusing on regions with the highest diagnostic value. Normalisation standardises pixel intensities across the dataset, ensuring consistent

and faster convergence during training, by mitigating issues related to scale disparities in the gradient updates. Lastly, resampling techniques are employed as a form of data augmentation, potentially increasing the robustness of the DL/ML models by enabling it to learn from diverse representations of the same feature. In summation, each of these pre-processing steps contributes to refining the input data for DL/ML models, ensuring not only efficient training but also enhancing their ability to generalise well on unseen medical images, which is vital for reliable and accurate disease prediction.

5.8.2 Features Extraction and Selection

High-level Features Extraction

As aforementioned, the process of extracting features starts mainly with HF extraction. By applying the process detailed in (Loukil et al., 2023), the resulted features are presented in the following. Towards understanding the relationships existing between these feature, correlation heatmap visualisation method was applied, particularly for disease affected data in each dataset. By identifying these correlations, it could provide better understanding of how these features interact and potentially influence the predictive power of the proposed prediction model. Features that are highly correlated might introduce multicollinearity, which can affect the stability and interpretability of the model. Therefore, it's essential to consider these relationships when training a ML/DL model, especially for tasks as critical as disease prediction. Three main heatmaps were produced to include: DR Fundus heatmap (Figure 5.16), DMO OCT heatmap (Figure 5.17), and pneumonia X-ray heatmap (Figure 5.18).

Observations

Figure 5.16 shows a strong negative correlation existing between Texture Contrast and Texture Homogeneity features, which indicates that as one feature increases, the other tends to decrease. This might suggest that areas with higher texture contrast tend to have less uniform textures. Coarseness, the mean values and standard deviations of RGB colours (BGR), on the

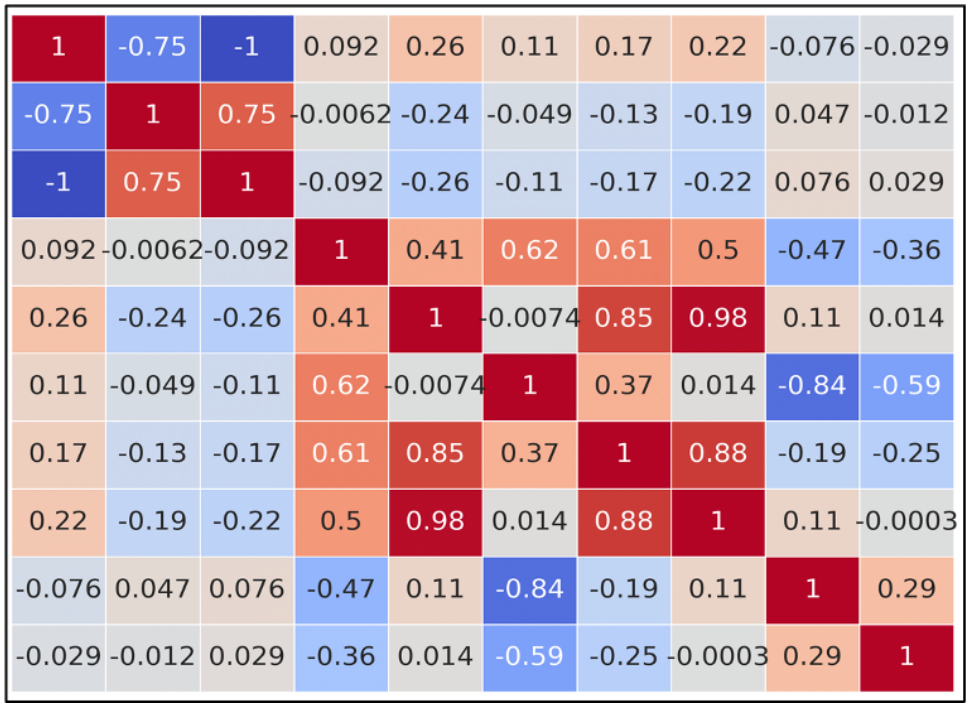


Figure 5.16: Correlation Heatmap for Fundus Dataset. The HF feature are from left to right (x-axis) and top to bottom (y-axis) as follows: Texture Contrast, Texture Energy, Texture Homogeneity, Entropy, Coarseness, Directionality, BGR, mean_values_std, Shape Area, Shape perimeter.

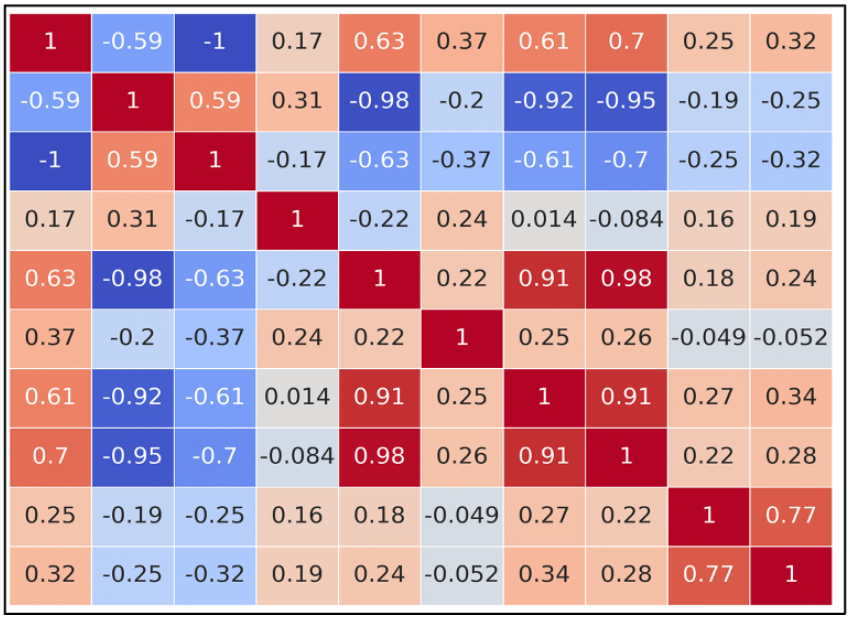


Figure 5.17: Correlation Heatmap for OCT Dataset. The HF feature are from left to right (x-axis) and top to bottom (y-axis) as follows: Texture Contrast, Texture Energy, Texture Homogeneity, Entropy, Coarseness, Directionality, BGR, mean_values_std, Shape Area, Shape perimeter.

other hand, are highly positively correlated. This may imply that as the coarseness of an image increases, there's also an increase in the average colour values and their variability. Directionality, particularly, showcases a strong negative correlation with the Shape Area, meaning that images with more defined directions or orientations might have smaller shape areas.

As per Figure 5.17, there is a significant negative correlation between Texture Energy and Coarseness. This suggests that smoother textures (lower coarseness) might be associated with higher energy levels in the image. The mean colour values are highly negatively correlated with Texture Energy and Texture Homogeneity. This could suggest that images with higher average colour values tend to have more varied uniform textures with lower energy levels. Shape Area and Shape Perimeter in this heatmap, on the other hand, showcase a strong positive correlation, indicating that as the area of the shape in an image increases, its perimeter tends to increase proportionally.

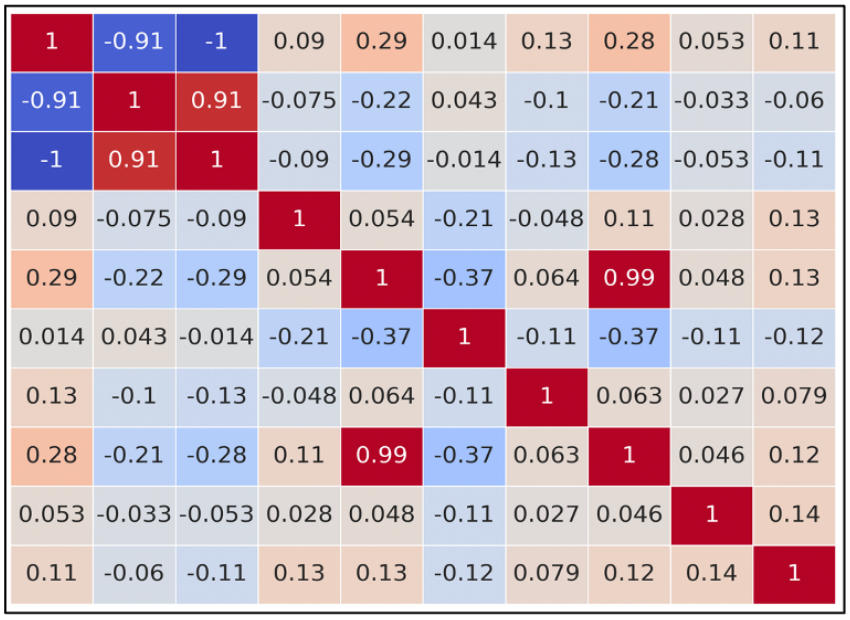


Figure 5.18: Correlation Heatmap for X-ray Pneumonia Dataset. The HF feature are from left to right (x-axis) and top to bottom (y-axis) as follows: Texture Contrast, Texture Energy, Texture Homogeneity, Entropy, Coarseness, Directionality, BGR, mean_values_std, Shape Area, Shape perimeter.

Figure 5.18 reveals that Texture Contrast and Texture Homogeneity again display a strong negative correlation, consistent with the first heatmap (Figure 5.16). Texture Energy and Coarseness are negatively correlated. This indicates that images with rougher textures might

have lower texture energy. The BGR values show a strong negative correlation with Texture Energy and a positive correlation with Coarseness. These correlations suggest a balance between colour uniformity and texture roughness in the dataset. A notable observation is the high positive correlation between BGR values and their standard deviations. This suggests that images with higher average colour values also tend to have a greater variability in colour. There is also a positive correlation between Shape Area and Shape Perimeter, similar to the second heatmap (Figure 5.17), indicating a proportional relationship between the two. Table 5.3 summarises the strong negative correlations presented by HF, their strong positive correlations, the presence or absence of weak correlations, as well as the implications for their feature selection resulted from each dataset.

Table 5.3: Datasets Heatmap Observations

Heatmap	Category	Observation
DR Fundus dataset	Strong Negative Correlations	• Texture Contrast with Texture Energy (-0.75) and Texture Homogeneity (-1.00) • Directionality with mean_values_colour (-0.84)
	Strong Positive Correlations	• Texture Energy with Texture Homogeneity (0.75) • Coarseness with *mean_values_Std* (0.98) • Entropy with Directionality (0.62), *mean_values_colour* (0.61), and *mean_values_Std* (0.50)

Continued on next page

Table 5.3: Datasets Heatmap Observations (Continued)

Heatmap	Category	Observation
	No or Weak Correlations	• Shape Area's correlation with Entropy (-0.47) and with Directionality (-0.84) • Shape Perimeter with Shape Area (0.29) • Coarseness and Directionality correlation (-0.01)
	Implications for Feature Selection	• Consideration of not using highly correlated features like Texture Contrast with Texture Homogeneity or Coarseness with *mean_values_Std* • Assessing the predictive power of weak correlations like Shape Area and Shape Perimeter
DMO OCT dataset	Strong Negative Correlations	• Texture Contrast with Texture Homogeneity (-1.00) • Texture Contrast with Texture Energy (-0.59) • Texture Energy with Coarseness (-0.98) and *mean_values_colour* (-0.92) • Texture Homogeneity with *mean_values_colour* (-0.61) and *mean_values_Std* (-0.70)

Continued on next page

Table 5.3: Datasets Heatmap Observations (Continued)

Heatmap	Category	Observation
	Strong Positive Correlations	• Texture Energy with Texture Homogeneity (0.59) • Coarseness with *mean_values_Std* (0.98) and Directionality (0.91) • *mean_values_colour* with *mean_values_Std* (0.91) • Shape Area with Shape Perimeter (0.77)
	No or Weak Correlations	• Weak correlations of Entropy with most other features • Negligible correlation between Directionality and *mean_values_colour* (0.01)
	Implications for Feature Selection	• Avoiding using highly correlated features to prevent multicollinearity. • Assessing the importance of weak correlations.
Pneumonia X-ray dataset	Strong Negative Correlations	• Texture Contrast with Texture Homogeneity (-1.00) and Texture Energy (-0.91)

Continued on next page

Table 5.3: Datasets Heatmap Observations (Continued)

Heatmap	Category	Observation
	Strong Positive Correlations	• Texture Homogeneity with Texture Energy (0.91) • Directionality with Coarseness (0.99)
	No or Weak Correlations	• Weak correlations of Entropy, *mean_values_colour*, and Shape Area with most other features • Weak correlations of Texture Contrast, Texture Energy, and Texture Homogeneity with features like *mean_values_Std*, *mean_values_colour*, and Shape Perimeter
	Implications for Feature Selection	• Exclusion of highly correlated features like Texture Contrast, Texture Homogeneity, and Texture Energy to reduce multicollinearity. • Exploration of the implications of Directionality and Coarseness before deciding on exclusions.

Analysis

Across all the heatmaps, there's a clear interplay between features like Texture Contrast, Texture Energy, Texture Homogeneity, and the colour metrics. These relationships provide vital information when constructing predictive models since they can heavily influence decisions

surrounding feature selection, engineering, and interpretation of models. Especially in the field of disease prediction, understanding these interrelations can augment the proposed prediction model's capability to detect implicit variations in images, which could be indicative of disease presence or progression. For the DR Fundus heatmap, one might contemplate not incorporating features that are highly correlated to avoid redundancy. Also, understanding the clinical implications of these features concerning DR Fundus images could enhance disease detection or progression insights. On the DMO OCT heatmap, the relationships between Texture Contrast, Texture Energy, and Texture Homogeneity could offer indicators about the disease's characteristics. Lastly, the pneumonia X-ray heatmap shows intriguing relationships, especially between texture metrics and the unique correlations observed with coarseness and directionality.

Deep Hidden Features Extraction

DenCeption played a critical role in extracting efficient DHF features. The processing resulted 6271 features with various importance across each dataset. Given the complexity of the generated features, unlike HF case, multiple correlation heatmaps are required for each dataset to get a better understanding of the different existing relationships. Due to the high number of features resulted, examples of sample testing's heatmaps are presented as follows for each dataset.

Figure 5.19.a represnts the heatmap of the testing sample from Fundus dataset. As shown, the diagonal represents the correlation of a feature with itself, and the value is always 1. Feature_3133 and Feature_3614 exhibit a strong negative correlation, represented by the deep blue square. This implies that when the value of one of these features increases, the other tends to decrease, and vice versa. This type of relationship might indicate that these features provide contrasting information about the dataset. Furthermore, Feature_2108, Feature_3746, and Feature_5637 appear to have a strong positive correlation with one another, as shown by the red regions in the heatmap. This suggests that these features might carry similar information. This confirm that when building the predictive model, it is crucial to consider not including all of these features where some of them can be excluded to avoid redundancy. Certain feature pairs

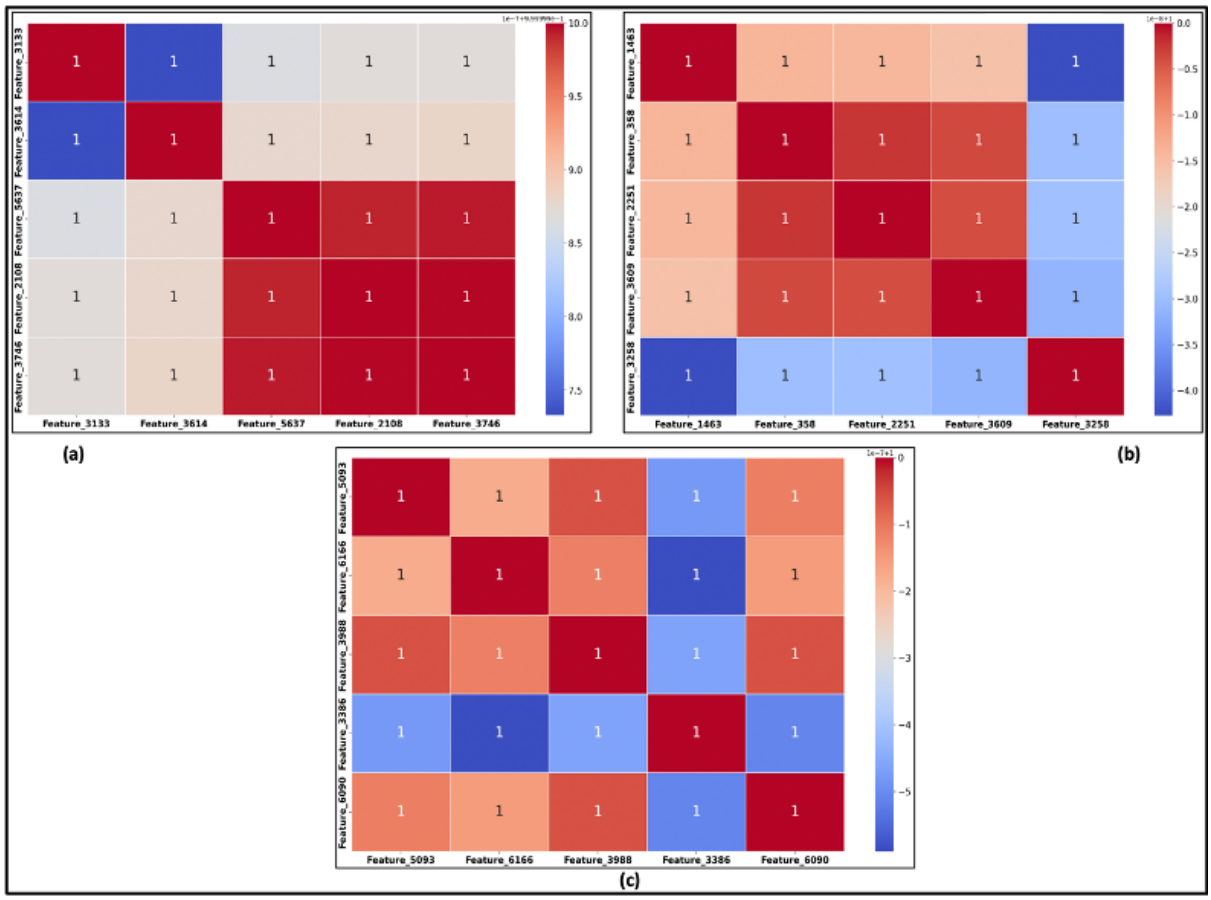


Figure 5.19: Correlation Heatmap Generated DHF Features for: (a) Fundus DR Dataset, (b) OCT DMO Dataset, (c) X-ray Pneumonia Dataset

like Feature_3614 with Feature_2108 and Feature_3614 with Feature_3746 have light-coloured regions, suggesting no or very weak correlation. This implies that these feature pairs might be independent of each other or that their relationship is not linear. On the other hand, some features present mixed correlation cases. In fact, Feature_5637 seems to have a strong positive correlation with Feature_3746 but only a weak to moderate correlation with Feature_3133. This kind of mixed relationship can offer insights into how certain features might interact with others differently.

On the other hand, Figure 5.19.b illustrates the testing sample for OCT dataset. As presented, Feature_1463 and Feature_358 exhibit a very strong positive correlation, as indicated by the bright red square. This implies that as one of these features increases in value, the other tends to increase as well, and vice versa. It can be concluded that these two features might

carry similar information and might be potentially redundant when used together in a model. On the other hand, some features do not present any kind of correlations. In fact, features such as Feature_2251 with Feature_1463 and Feature_2251 with Feature_358 show no or very weak correlation as indicated by the pale colour. This means changes in one feature don't necessarily predict changes in the other. The blue square between Feature_3609 and Feature_3258 indicates a strong negative correlation. This suggests that as one feature value increases, the other tends to decrease and vice versa. This is an interesting relationship and may signify that these features provide complementary information about the dataset. As per the heatmap, some features are with little to no variability. The near-white colour in some blocks indicates very low correlation, almost nearing zero. This can be seen between features like Feature_3609 with Feature_1463 and Feature_3609 with Feature_358. Such low correlations could be indicative of one or both features having little variability, or that there's no linear relationship between them.

Conversely, the correlation heatmap presented in Figure 5.19.c, presents the outcome of X-ray dataset. In fact, several pairs of features, such as Feature_5093 and Feature_6166, Feature_6166 and Feature_3988, and Feature_3988 and Feature_5093, have shown deep red regions indicating strong positive correlations. This means when one feature increases in value, the other tends to increase as well, suggesting that they may carry similar information. The feature pair Feature_3386 and Feature_5093 has a deep blue region which indicates a strong negative correlation. This suggests that when the value of one of these features increases, the other one tends to decrease. This type of relationship might be of interest as it shows the features are providing contrasting information. Some feature pairs, such as Feature_3386 with Feature_6166 and Feature_3386 with Feature_3988, exhibit light-coloured regions suggesting a weak to moderate correlation. This indicates that these feature pairs might have a less pronounced linear relationship. Feature_5093 appears to have a strong positive correlation with both Feature_6166 and Feature_3988, but Feature_3386 shows contrasting relationships with these features - a negative correlation with Feature_5093 and a weak correlation with the others.

As per the above analysis of each dataset, some features are strongly correlated (either pos-

itively or negatively), which suggests that they might be carrying redundant or complementary information. When building a model, the consideration of multicollinearity reduction is fundamental. The presence of features with no strong correlation to others might be valuable, as they could provide unique information to a predictive model. However, if a feature does not correlate with any other feature or the target variable, it might not add much predictive enhancement. This stage allowed for examining random subsets of features for similar patterns and better understand the DHF distribution in each particular case. The next step includes conducting feature importance and reduction technique SOM to further understand and optimise the dataset for model building.

Final DHF Features Set

Following the application of SOM technique on each dataset, feature selection has taken place based on the assigned weights to each feature. The neural network has been trained using a learning rate of 0.5 and sigma of 1.0. The training was done though 8000 epochs through different batches. The feature importance was performed based on the neurons' activations composing the neural network. The selected DHF was finalised with maximum aggregation of these features to ensure that: only relevant features are considered and redundant features are kept when needed. It is crucial to note that SOM's effectiveness isn't only about reducing dimensionality but ensuring the retained features maintain or enhance the prediction model's performance.

Towards critically evaluating the obtained results for each dataset it is essential to consider the following key components presented by SOM to include:

- Sensitivity to parameter settings
- Inherent characteristics of datasets
- Relevance and Domain Knowledge
- Reproducibility and Stability

Table 5.4 abridges the importance of each of the above components.

Table 5.4: SOM Key Components (Miljković, 2017)

Component	**Key Points**
Sensitivity to Parameter Settings in SOM	- Dependent on parameters like grid size, topology, learning rate, and neighbourhood function. - Need to run Sen analysis to evaluate robustness. - Observe changes in resulting feature sets.
Inherent Characteristics of Datasets	- Different imaging modalities have unique qualities and variances. - Affects results of feature reduction. - Understand dataset nature for better insight.
Relevance and Domain Knowledge	- Align selected features with domain knowledge. - Engage domain experts for evaluation of feature significance. - Check for exclusion of vital features or retention of redundant ones.
Reproducibility and Stability	- Ensure reproducibility given SOM's stochastic nature. - Be aware of potential variance in selected features when running multiple times.

After applying SOM to various datasets, the number of features retained varies. Specifically, for the X-ray dataset, the algorithm selected 1051 features, demonstrating its ability to condense the information while preserving the essential characteristics of the data. In the case of the OCT dataset, a larger set of 1500 features was selected, which may indicate the presence of more complex patterns or a higher degree of variability within the data. Lastly, for the Fundus dataset, SOM selected 1100 features, showcasing its adaptability and effectiveness across different types of datasets. By doing so, SOM helps in transforming the data into a more manageable and meaningful format, which is crucial for subsequent analysis and interpretation.

A vital piece in this evaluation is understanding the impact of feature reduction on the performance of the downstream task, prediction. If the model's performance remains consistent or even improves with fewer features, it speaks volumes about the SOM's efficiency and effectiveness in feature selection. Conversely, a noticeable performance process was followed to reassess the reduction process, as per the method in (Loukil et al., 2023). The reason behind the proposed process is that, while the significant feature reduction achieved through the SOM process is praiseworthy, the real test would lie in the application of these features in actual tasks. An evaluation of model performance metrics before and after the reduction is essential offering tangible insights into the feature selection process's true efficacy. In this work, the reassessment process focused mainly on comparing the reduced features versus the original set of features, where the former achieved promising results which confirms SOM's efficiency.

5.9 Prediction Results and Discussion

5.9.1 Baseline Scenario Results

This section sheds light on critical assessment of the predictive performance of the previously mentioned models, excluding any additional demographic and physiological features. This evaluation will serve as a foundation for filtering out the models with unsatisfactory performance, advancing only the outperforming models to the subsequent testing phase.

Fundus Dataset Prediction Results

The results obtained by training the prediction models including and excluding the hyperparameters fine-tuning is summarised in Tables 5.5 and 5.6. Without tuning, DT model displayed an Acc of 52.43%, signifying that approximately 52% of its predictions were correct. The precision stood at 52.80%, meaning that around 53% of its positive predictions were indeed positive, suggesting low FPs. The recall was at 61.30%, indicating that it correctly identified 61% of all actual positive cases, highlighting fewer FNs. The F1-Score, which harmoniously balances precision and recall, was recorded at 56.73%, where a score closer to 1 denotes better performance.

Upon hyperparameter tuning, there were noticeable shifts. The accuracy faced a marginal enhancement of 0.02%, settling at 52.45%. Conversely, precision decreased to achieve 51.88%, indicating an increase in FP predictions post-tuning. The recall increased to 63.60%, revealing a reduction in FNs and a better identification rate of actual positives. Lastly, the F1-Score, representing the equilibrium between precision and recall, rose to 57.14%, suggesting a slightly superior overall balance achieved through tuning.

Table 5.5: Fundus-DR Prediction Results without Fine-tuning

Model	**Acc (%)**	**Precision (%)**	**Recall (%)**	**F1-Score (%)**	**LCE**
XGBoost	61.08	62.58	58.45	60.44	0.45
RF	51.79	52.62	52.53	52.58	0.6
DT	52.43	52.80	61.30	56.73	0.55
AdaBoost	60.66	61.45	56.49	58.87	0.53
HyBoost	63.36	64.20	59.75	63.72	0.4

Table 5.6: Fundus-DR Prediction Results with Fine-tuning

Model	**Acc (%)**	**Precision (%)**	**Recall (%)**	**F1-Score (%)**	**LCE**
XGBoost	62.27	62.76	59.72	61.20	0.41
RF	54.41	54.33	53.43	53.88	0.5
DT	52.45	51.88	63.60	57.13	0.52
AdaBoost	63.41	63.92	60.91	62.40	0.39
HyBoost	64.96	65.28	63.25	62.39	0.36

Tuning DT model resulted in slight improvements in accuracy, recall, and F1-Score, at the cost of a minor drop in precision. Depending on the application, one might prioritise recall over precision or vice versa. For instance, in case of DR prediction, having a high recall might be

more crucial than having high precision. This is because missing a diagnosis (FN) could have severe implications, while a FP might lead to further tests but no immediate harm. However, generally, the trade-off between precision and recall should be determined based on the specific context and the consequences of making false predictions.

The results from RF model shows that the model's Acc is around 51.78%, which means it correctly predicts the outcome roughly 52% of the time. The precision indicates that when the model predicts a positive outcome, it is correct around 52.62%. Recall states that the model identifies 52.53% of all the actual positive outcomes. After being tuned, RF accuracy has risen to 54.41%, showing an improvement. Precision also rose to 54.33% with tuning, meaning the model has reduced the number of FPs. Though recall has seen a minor decrease, it's still in the same ballpark. The performance improvements after tuning, though modest, are significant. An increase in precision without a severe drop in recall often indicates a more reliable model. This is crucial, especially in applications where FPs can have grave consequences. However, it's worth noting that the overall performance metrics (like accuracy) are still just above 50%. In many real-world applications, such a performance might not be satisfactory. For instance, in DR detection and prediction, a high recall might be more crucial because missing out on TP cases can be life-threatening.

The performance of the AdaBoost model on the Fundus data shows no%improvements after the tuning process. An examination of the metrics reveals that accuracy increased from 60.66% in the untuned model to 63.41% in the tuned version. Likewise, precision faced an increase from 61.45% to 63.96%. The recall also demonstrated a positive shift, moving from 56.49% to 60.91%. Additionally, the F1-Score, improved from 58.87% to 62.40%. The corresponding changes in the confusion matrix values further substantiate the improved classification outcomes after tuning. These observations underscore the importance of model optimisation, emphasising that fine-tuning can significantly enhance performance, making the model more effective for specific datasets, particularly Fundus dataset.

The results from the XGBoost model without any tuning show a promising performance. In fact, the model achieved an Acc of 61.08%, which means that it correctly predicted the out-

comes for 61.08% of the samples in the test set. Similarly, the precision is found to be 62.58%. The recall, which indicates how many actual positives were identified, was at 58.45%. Lastly, the F1-Score, which is the harmonic mean of precision and recall and provides a single metric for model performance, is at 60.44%. On the other hand, when the XGBoost model was tuned, the performance improved across all metrics. The accuracy went up slightly to 62.27%, indicating a better overall prediction rate. The precision of the tuned model was 62.76%, a slight increase from the untuned model, showcasing that the model's positive identifications became more accurate post-tuning. The recall value also showed an increase to 59.72%, implying that the model became slightly better at identifying the actual positive samples. The F1-Score, a crucial metric for understanding a model's robustness, increased to 61.20%. Comparing the two sets of results, it's evident that tuning the XGBoost model provided an improvement, although slight, across all the metrics. The enhanced performance of the tuned model highlights the significance of optimising the model to achieve the best possible results. It's also worth noting that even minor improvements in performance can have a significant impact, especially when dealing with large datasets or critical applications.

In the case of the proposed HyBoost (untuned case), the accuracy is approximately 63.36%, meaning it correctly predicts the outcome 63.36% of the time which is still overpassing the performance of the other models solely (AdaBoost and XGBoost). The precision is approximately 64.20%. Recall, measuring the ratio of correctly predicted positive observations to the actual positives, is about 59.75%. The F1-Score is approximately 63.72%. In the case of tuned HyBoost, on the other hand, the accuracy indicates a correct prediction rate of 64.97%. The precision is slightly improved at about 65.28%. Moreover, the recall has slightly increased to approximately 63.25%. The F1-Score for the tuned model is 62.39%. Comparing both models, it's evident that tuning the HyBoost algorithm improved its accuracy and precision. In fact, there was a trade-off, with the recall noticeably increasing in the tuned model. This suggests that the tuned model is better at making correct predictions overall, and shows more Sen in identifying positive cases.

Before fine-tuning, the HyBoost model outperforms all others with an LCE of 0.4, suggest-

ing its hybrid approach is effective from the outset, while XGBoost also demonstrates strong predictive capability with a competitive LCE of 0.45. The RF model, however, exhibits the highest LCE at 0.6, indicating a need for refinement, and the DT model, with an LCE of 0.55, shows it has potential yet is outperformed by XGBoost and AdaBoost, which holds a moderate LCE of 0.53. After fine-tuning, improvements across the models are evident; XGBoost's LCE drops to 0.41, RF to 0.5, and DT's decreases slightly to 0.52, showcasing the benefits of model optimisation. Notably, AdaBoost's LCE diminishes to 0.39, indicating a significant enhancement, but it is HyBoost that exhibits the most substantial improvement to an LCE of 0.36, underscoring the impact of fine-tuning on its hybrid structure, which makes it a promising tool for Fundus-DR prediction.

Figure 5.20.a presents the AdaBoost model ROC curve without applied tuning. As illustrated, the AUC value reaches 0.65, signifying the model's moderate capacity to differentiate between the positive and negative classes. The curve's position above the diagonal underscores that the model performs better than mere random guessing. Conversely, Figure 5.20.b, where AdaBoost model is presented with tuning, showcases an improved AUC of 0.69. Its curve is more distanced from the diagonal, denoting enhanced performance in distinguishing the classes post-tuning. Ultimately, while tuning has improved the model's efficacy, as marked by the AUC ascent from 0.65 to 0.69, both metrics hint at potential avenues for further optimisation in the model's discriminative abilities.

The graph shown in Figure 5.20.c, with an AP of 0.65, depicts the AdaBoost model's performance without tuning. Here, precision values experience sharp fluctuations at higher recall rates, suggesting potential instability in certain threshold ranges. As transitioning to the second graph (Figure 5.20.d), where tuning has been applied, there is a noticeable enhancement in the PR curve, represented by an increased AP value of 0.70. The curve in the tuned model is smoother, indicating a more consistent performance across varying thresholds. This comparison underscores the significant impact of model tuning on improving the precision and recall trade-off, thereby enhancing the overall reliability and accuracy of the AdaBoost model in analysing Fundus data.

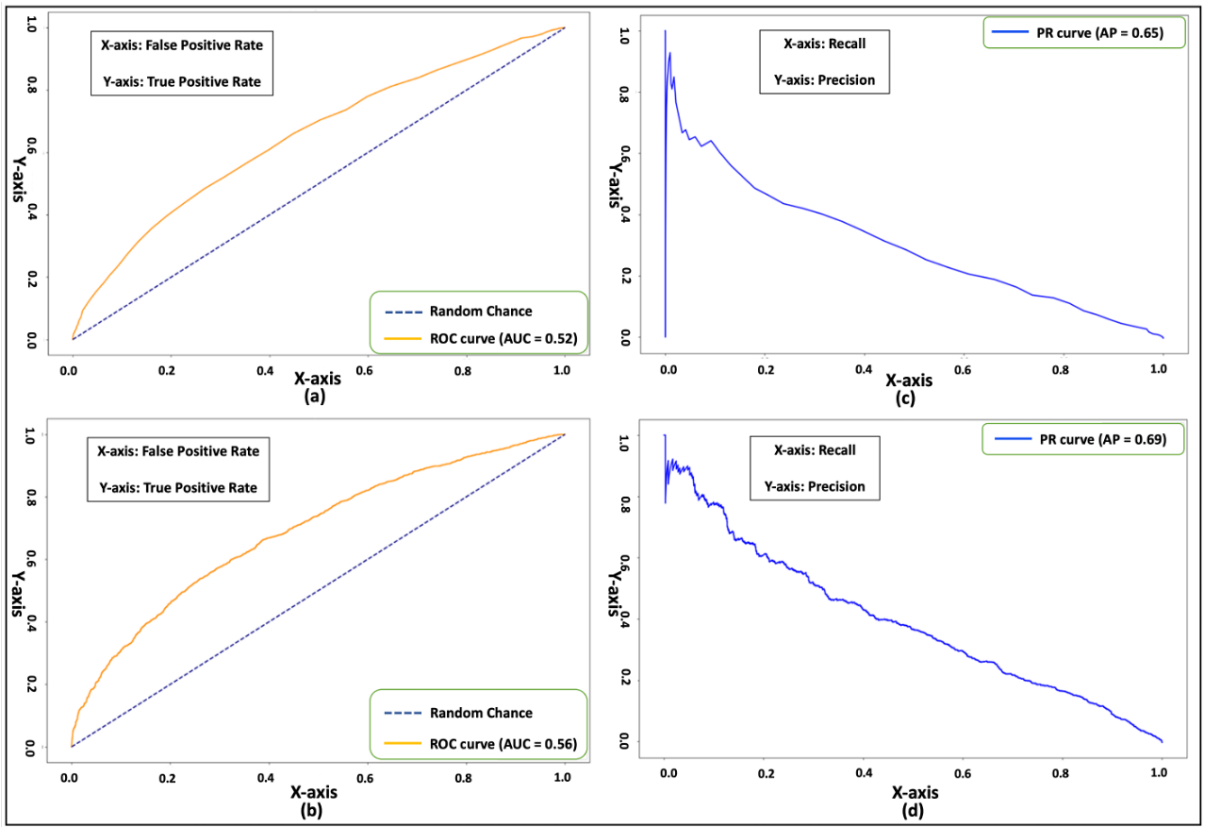


Figure 5.20: Fundus AdaBoost ROC (a) PR (c) Curves without Hyperparameters Fine-tuning and ROC (b) PR (d) Curves with Hyperparameters Fine-tuning.

Figure 5.21.a and 5.21.b show the XGBoost ROC curve without and with inclusion of hyperparameters fine-tuning.

The Fundus ROC curve of XGBoost model, without tuning, reveals an AUC of 0.65. This AUC value suggests moderate model performance, as an AUC of 0.5 equates to no discriminatory power (random guessing), while an AUC of 1.0 denotes a flawless model. This curve's position above the diagonal line of no discrimination indicates the model's capacity to differentiate between positive and negative classes. In contrast, when tuning is applied to the XGBoost Fundus ROC Curve, the AUC slightly improves to 0.67, again displaying the curve above the diagonal, which implies some degree of discriminatory power. To sum up, the XGBoost model's tuning led to a modest enhancement in discriminatory capability, with the AUC rising from 0.65 to 0.67. Nonetheless, both model variations showcase moderate efficacy, suggesting potential avenues for improvement.

Figure 5.21.c represents the PR curve for the XGBoost model without tuning. Observing the

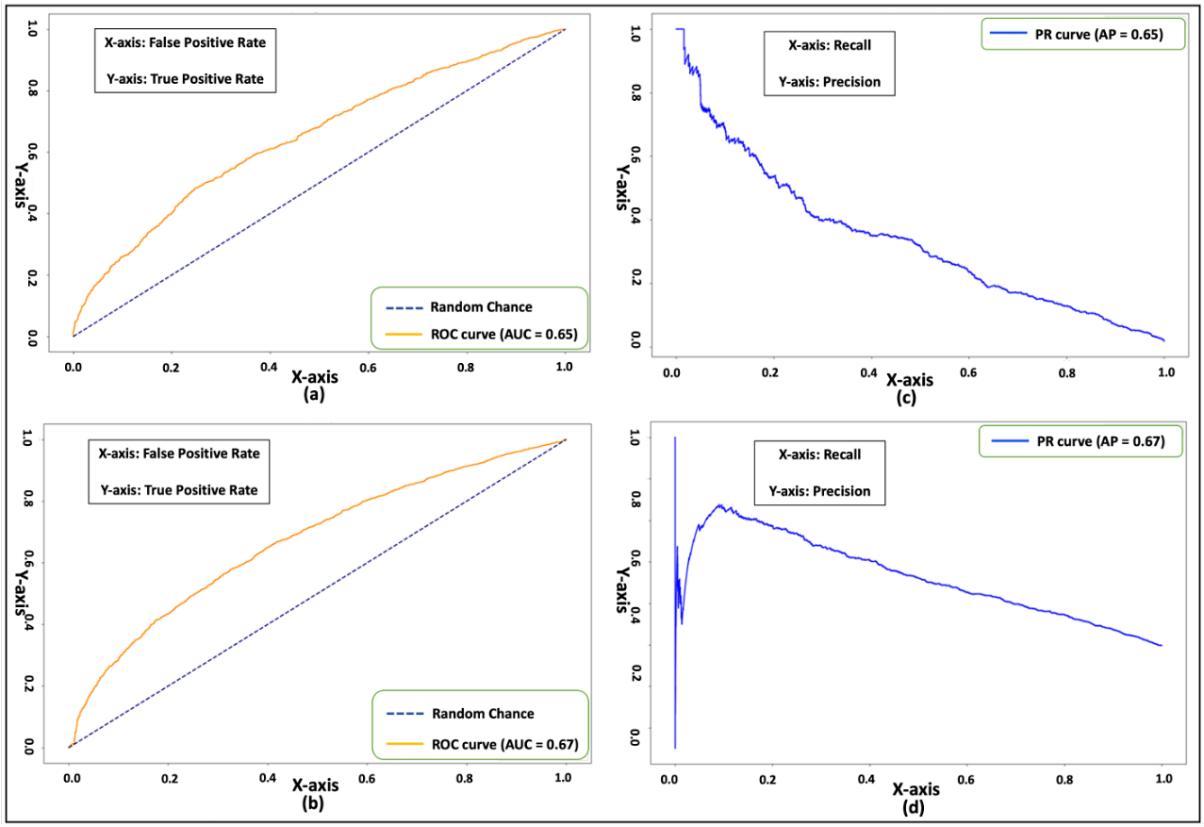


Figure 5.21: Fundus XGBoost ROC (a) PR (c) Curves without Hyperparameters Fine-tuning and ROC (b) PR (d) Curves with Hyperparameters Fine-tuning.

curve, it starts from a high precision level near 1.0 and gradually descends as recall increases. A smoother curve, particularly in the higher recall regions, usually implies a better-performing model. The AP score for this curve is 0.67. This score provides an aggregate measure of the model's performance across all classification thresholds and indicates that the model has a reasonably good balance between precision and recall. Figure 5.21.d, on the other hand, displays the PR curve post-tuning of the XGBoost model. The curve initiates with fluctuations in the precision, which may be due to overfitting or noise in the data. However, it stabilises after a recall of approximately 0.2. While the curve generally seems smoother than the first, the AP score is slightly lower at 0.66. Although the difference is minimal, it suggests that the tuning didn't substantially enhance the overall model performance in terms of the PR balance. When comparing the two curves, both the untuned and tuned XGBoost models have relatively similar performance in terms of precision and recall for the Fundus data. The slight decrease in the AP score after tuning may be due to overfitting or other model-specific factors. Nevertheless, both

curves display a good balance between precision and recall, with AP scores centric around 0.66 and 0.67.

As shown in Figure 5.22.a, the untuned version of the proposed HyBoost model exhibits an AUC of 0.68. This value suggests that the model possesses a moderate capability to discriminate between classes, but overpassing the performance of previous tested models. A perfect classifier would have an AUC of 1, while a completely random classifier would result an AUC of 0.5. Hence, with an AUC of 0.68, the non-tuned model is performing better than a random guess but has room for improvement. Contrastingly, Figure 5.22.b showcases the performance of the HyBoost model post-tuning, evident in its elevated AUC of 0.71. This enhancement, although marginal, is significant, implying that the process of tuning has optimised certain parameters of the model, thereby refining its classifying performance compared to AdaBoost and XGBoost. An AUC of 0.71 suggests that the tuned model holds a better discriminative ability than its non-tuned counterpart. In essence, while both models demonstrate commendable performance, the slight edge in AUC for the tuned version underscores the importance and potential benefits of fine-tuning the proposed hybrid model to better adapt to specific datasets, such as Fundus in this context.

The two PR curves, shown in Figure 5.22.c and 5.22.d, demonstrate the performance of the HyBoost model when applied to Fundus data under distinct configurations: before and after tuning. The initial figure represents the HyBoost model's performance without any tuning, resulting an AP score of 0.70. This suggests that the model demonstrates a commendable balance between precision and recall, but there is potential for further enhancement. In contrast, the subsequent figure shows the model's performance post-tuning, reflected in a slightly higher AP score of 0.71. This marginal increase in the AP score indicates that the tuning process has effectively optimised specific parameters of the model, improving its ability to maintain precision across varying recall thresholds. While the HyBoost model demonstrates robust performance in both scenarios, the enhanced AP score following tuning emphasises the significance of considering extra features that helps the model to optimise precision and recall, especially when dealing with specific datasets like Fundus.

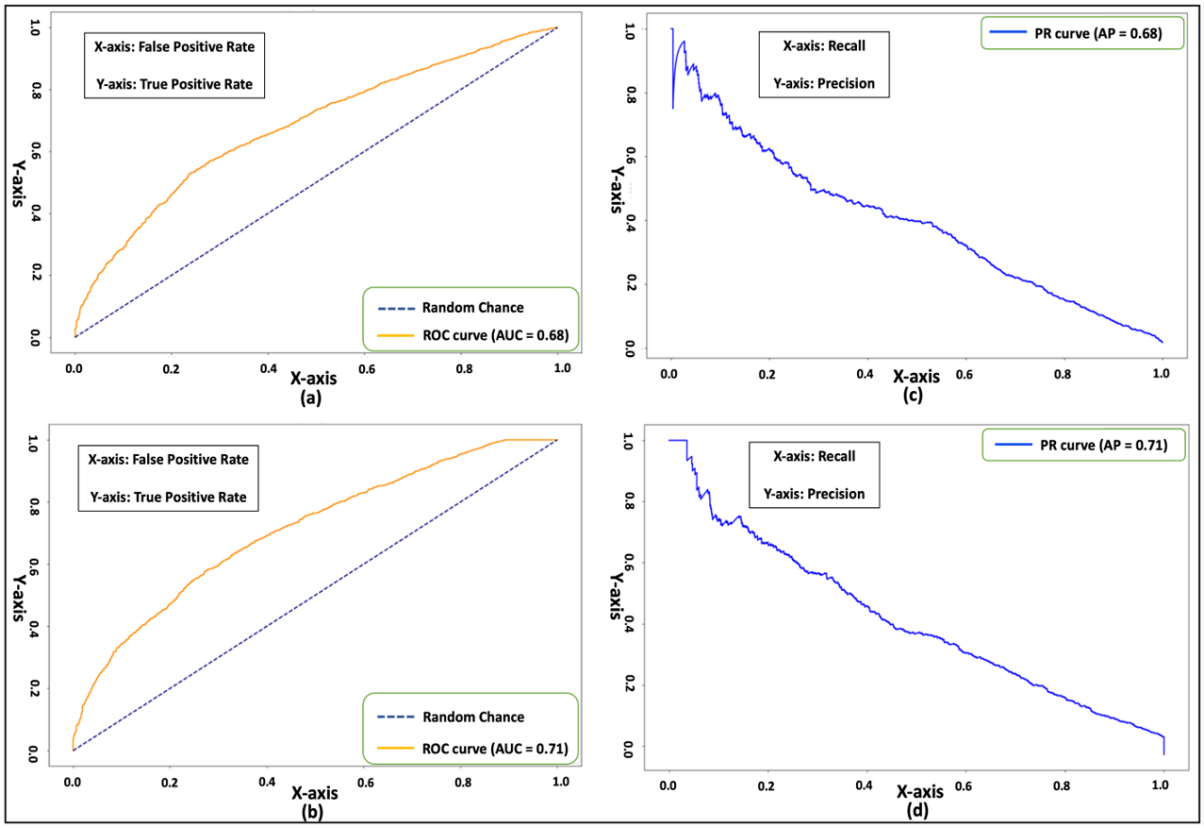


Figure 5.22: Fundus HyBoost ROC (a) PR (c) Curves without Hyperparameters Fine-tuning and ROC (b) PR (d) Curves with Hyperparameters Fine-tuning.

OCT Dataset Prediction Results

The results obtained by training the prediction models including and excluding the hyperparameters fine-tuning is summarised in Tables 5.7 and 5.8. To get deeper insights about the resulted values of top performed models training, the will be followed by a critical evaluation of the resulted ROC and PR curves

Table 5.7: OCT-DMO Prediction Results without Fine-tuning

Model	**Acc (%)**	**Precision (%)**	**Recall (%)**	**F1-Score (%)**	**LCE**
XGBoost	93.63	94.54	91.96	93.04	0.3
RF	90.79	90.55	89.55	90.05	0.35
DT	54.58	50.70	85.24	63.58	0.61
AdaBoost	93.01	94.62	90.00	92.25	0.32
HyBoost	96.62	95.09	93.6	96.27	0.21

Table 5.8: OCT-DMO Prediction Results with Fine-tuning

Model	**Acc (%)**	**Precision (%)**	**Recall (%)**	**F1-Score (%)**	**LCE**
XGBoost	94.09	94.6	92.34	93.9	0.26
RF	91.86	96.13	85.87	90.71	0.34
DT	61.41	55.22	87.59	67.74	0.6
AdaBoost	93.81	95.14	91.28	93.17	0.31
HyBoost	97.06	98.63	96.08	97.18	0.18

The tuned DT model has higher accuracy, precision, recall, and F1-score compared to the untuned DT model. After tuning, the DT model seems to perform better in all metrics. The improvement in the precision of the DT model after tuning is notable, which indicates a reduction in FPs. Both models have a high recall, suggesting that they are able to identify a large proportion of the actual positive instances. It seems that the tuning on the DT model has been effective, as it shows improvement in all the evaluation metrics.

Without tuning, the RF model's metrics resulted an Acc of 90.79%, signifying that it correctly predicted 90.79% of the samples. Its precision stood at 90.55%, meaning 90.55% of its positive predictions were accurate, while its recall at 89.55% indicated it correctly identified

this percentage of actual positive samples. The F1-Score, representing the harmonic mean of precision and recall, was 90.04%. Post-tuning with RF, metrics showed an improved accuracy of 91.86%, precision of 96.13% highlighting the model's improved reliability in positive predictions, and a slightly reduced recall of 85.87%, suggesting it missed some actual positives. However, the F1-Score slightly increased to 90.71%, indicating a better balance between precision and recall. In summary, the tuned model exhibited improved accuracy and precision, making its positive predictions more trustworthy. Despite the rise in precision, there was a slight decline in recall, pointing to the model's conservative stance in predicting positives. This trade-off, accentuated by the minor increment in F1-Score, means the model achieved a more balanced performance post-tuning.

Upon analysing the outcomes of the AdaBoost classifier, it's evident that the model's performance showed no%improvements post tuning. Initially, without any tuning, the classifier achieved an Acc of 93.02%, precision of 94.62%, recall of 89.99%, and an F1-score of 92.25%. However, after the tuning process, the model's accuracy rose to 93.81%, showcasing an increase of 0.79%. Similarly, the precision experienced an enhancement, moving up by 0.52% to reach 95.14%. The recall metric also showed a significant boost, rising by 1.29% to a value of 91.28%. Moreover, the F1-score, which harmonises precision and recall, observed a commendable ascent of 0.92%, culminating at 93.17%. In summary, the tuning procedure unequivocally optimised the performance of the AdaBoost classifier across all the metrics, underscoring the importance of model fine-tuning in achieving superior results.

OCT related results for the untuned XGBoost presented an accuracy rate of approximately 93.61%, suggesting that the model correctly predicted the outcomes in a significant majority of cases. The precision score of about 94.54% denotes that, of all the positive predictions made by the model, 94.54% were indeed correct. Meanwhile, the recall or Sen, which measures how many actual positives the model was able to capture, stands at approximately 91.96%. The F1-Score is about 93.04%, indicating a balanced performance between precision and recall. Moving on to the results after tuning XGBoost, there's a noticeable improvement where the accuracy rate has risen to approximately 96.09%. This shows an enhancement in the model's

performance in making correct predictions. The precision score has significantly increased to about 98.85%, meaning the model made very few FP predictions after tuning. The recall rate is around 93.44%, suggesting a slight improvement in detecting actual positive cases. Finally, the F1-Score for the tuned XGBoost stands at approximately 96.07%, showing an improved and balanced model performance between precision and recall. Tuning the XGBoost model has positively impacted its performance across all metrics. The results demonstrate the value of fine-tuning algorithms to enhance their predictive accuracy and reliability.

The untuned proposed HyBoost model achieved an accuracy of approximately 96.62%, showcasing its capability to correctly classify instances. Moreover, it exhibits a high precision of approximately 95.09%, signifying that the proportion of TP predictions among all positive predictions is quite high. In terms of recall, the model retrieves about 93.60% of the actual positive instances. The F1-Score, which is a harmonic mean of precision and recall, stands at approximately 96.27%, indicating a balanced performance between precision and recall. On the other hand, the HyBoost model post-tuning shows an enhanced performance. This advanced precision in classification post-tuning is evident with an accuracy soaring to approximately 97.06%. Its precision is also exemplary, standing at nearly 98.63%, indicating an even higher reliability in its positive predictions. The recall has seen an improvement, now capturing about 96.08% of actual positive instances. Complementing these metrics, the F1-Score reaches approximately 97.18%, revealing a very well-balanced model in terms of both precision and recall. Tuning the HyBoost model has evidently led to a noticeable enhancement in its performance across all metrics, making it a more reliable choice for OCT results classification, particularly in DMO prediction.

the HyBoost model stands out with the lowest LCE at 0.21, indicating a superior initial performance likely due to its advanced hybrid design. XGBoost follows with a commendable LCE of 0.3, while AdaBoost is not far behind, indicating an LCE of 0.32. The RF model's LCE is slightly higher at 0.35, suggesting some room for improvement. The DT model exhibits the highest LCE at 0.61, indicating that it might be less adept at handling the dataset without adjustments. Upon fine-tuning, all models exhibit improvements in LCE values. XGBoost shows

a marked decrease to 0.26, which underscores the effectiveness of fine-tuning in enhancing its predictive capabilities. RF also improves, although marginally, to an LCE of 0.34. The DT model sees a slight decrease in LCE to 0.6, indicating that while fine-tuning has had an effect, it remains the least effective model among those tested. AdaBoost's LCE marginally decreases to 0.31, suggesting that fine-tuning has a positive but limited impact on its performance.

In analysing the ROC curve shapes (Figure 5.23.a and 5.23.b), the untuned AdaBoost model's curve rapidly ascends towards the top-left corner, denoting a high Sen at a low FPR. Interestingly, the curve for the tuned AdaBoost model mirrors this behaviour, approaching the y-axis closely before turning horizontal. Both models exhibit an AUC value of 0.98, suggesting that their abilities to distinguish between positive and negative classes are almost impeccable and identical. This proximity to a perfect classifier score is commendable. The almost indistinguishable performance of the tuned and untuned AdaBoost classifiers on the OCT dataset is evident from their ROC curves and AUC values. This near-perfect alignment towards the top-left corner implies that both versions can achieve high Sen with minimal FPs, making them effective for the OCT dataset. The surprising similarity in performance between the two might suggest that the default AdaBoost classifier parameters were ideally suited for the OCT dataset, the tuning didn't substantially alter hyperparameters, or the dataset basically allows even basic models to excel. In conclusion, the consistency in performance implies that, if pressed for time or computational resources, the untuned model would be adequate. Yet, as the main aim is to maximise efficiency, delving into considering extra features has the potential to enhance the overall model's performance.

Figure 5.23.c displays an impressive AP of 0.98 with a curve that remains largely flat at the top, denoting consistent high precision across diverse recall levels, and signifying the untuned AdaBoost model's adeptness at differentiating between positive and negative classes. In contrast, Figure 5.23.d, while showcasing the same AP, reveals a slight decrease in precision as the recall approaches 1, possibly pointing to a few FP predictions at high recall for the tuned AdaBoost. Though both curves share an identical AP, indicating parallel overall efficacy, their minute shape differences, particularly at high recall points, suggest variances in prediction ten-

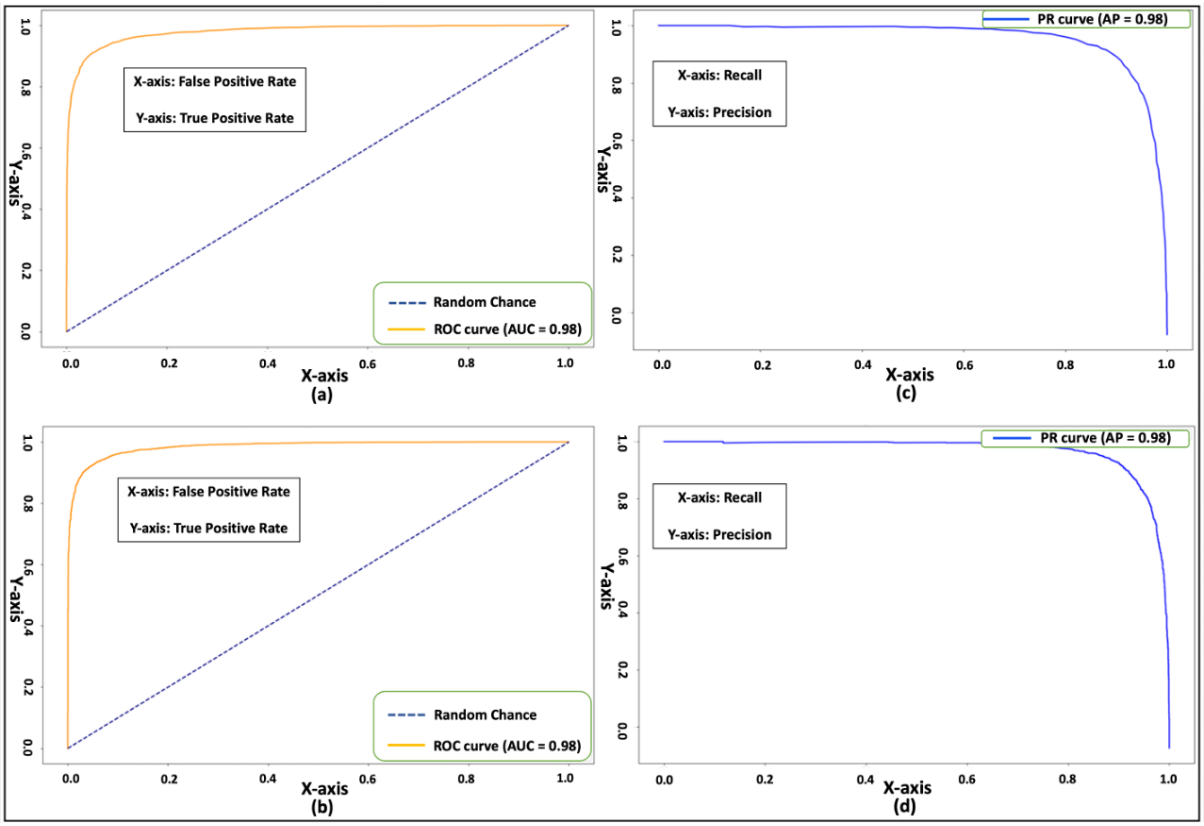


Figure 5.23: OCT AdaBoost ROC (a) PR (c) Curves without Hyperparameters Fine-tuning and ROC (b) PR (d) Curves with Hyperparameters Fine-tuning.

dencies. Based on the graph shown in Figure 5.23.d, the tuning seems to have preserved the aggregate performance but may have modified the model's reactions at specific recall intervals. It's imperative to weigh the trade-offs when making determinations grounded on these curves.

The untuned XGBoost ROC curve (Figure 5.24.a) exhibits impressive performance, with an AUC value of 0.98.

The AUC is a metric used to measure the overall performance of a classifier, with 1 indicating perfect classification and 0.5 denoting performance no better than random classification. An AUC of 0.98 suggests that the classifier is highly accurate in differentiating between the positive and negative classes. Figure 5.24.b, representing the XGBoost OCT ROC curve post-tuning, depicts an even more pronounced improvement. The AUC value here reaches 0.99, indicating almost perfect classification capabilities. XGBoost, an AdaBoost technique, appears to have enhanced the model's performance by fine-tuning it, as evidenced by the increase in AUC value from 0.98 to 0.99. While the OCT model without tuning already showcases strong

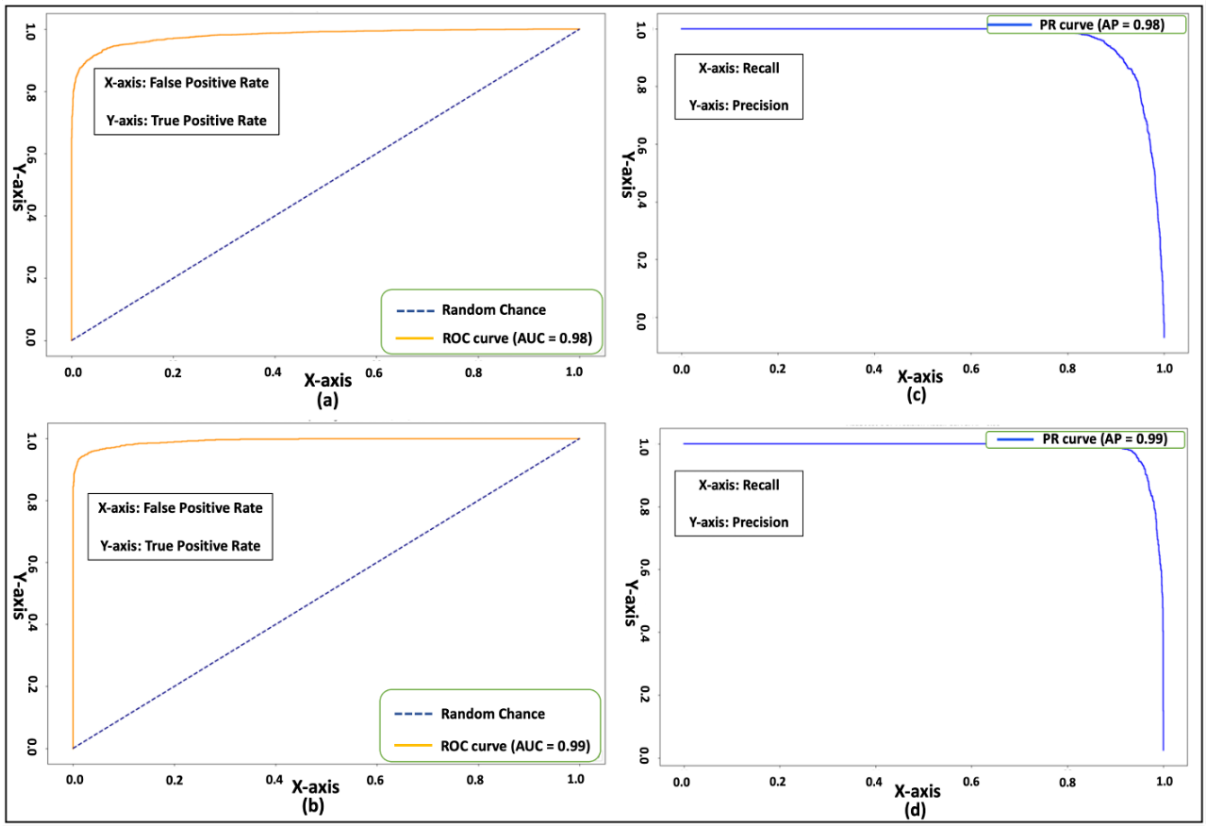


Figure 5.24: OCT XGBoost ROC (a) PR (c) Curves without Hyperparameters Fine-tuning and ROC (b) PR (d) Curves with Hyperparameters Fine-tuning.

classification performance with an AUC of 0.98, the application of the XGBoost algorithm further refines its accuracy, increasing the AUC to a significant 0.99.

For the untuned XGBoost model, the PR curve (Figure 5.24.c) starts off with a high precision but experiences a gradual drop as recall increases, finally culminating in a lower precision value as it approaches a recall of 1. The AP score for this curve is 0.98, which indicates a high level of model performance, especially considering that an AP score of 1 would be perfect. On the other hand, the PR curve for the XGBoost-tuned (Figure 5.24.d) model appears to maintain a consistently high precision for a larger span of recall values, eventually showing a slight decrease towards the end of the curve. This model has an even better AP score of 0.99, which signifies an improvement over the untuned model and is closer to optimal performance. XGBoost-tuned model displays superior performance in terms of precision and recall as compared to the untuned model, making it more reliable for predictions. The improvement in the AP score from 0.98 to 0.99 further corroborates this observation.

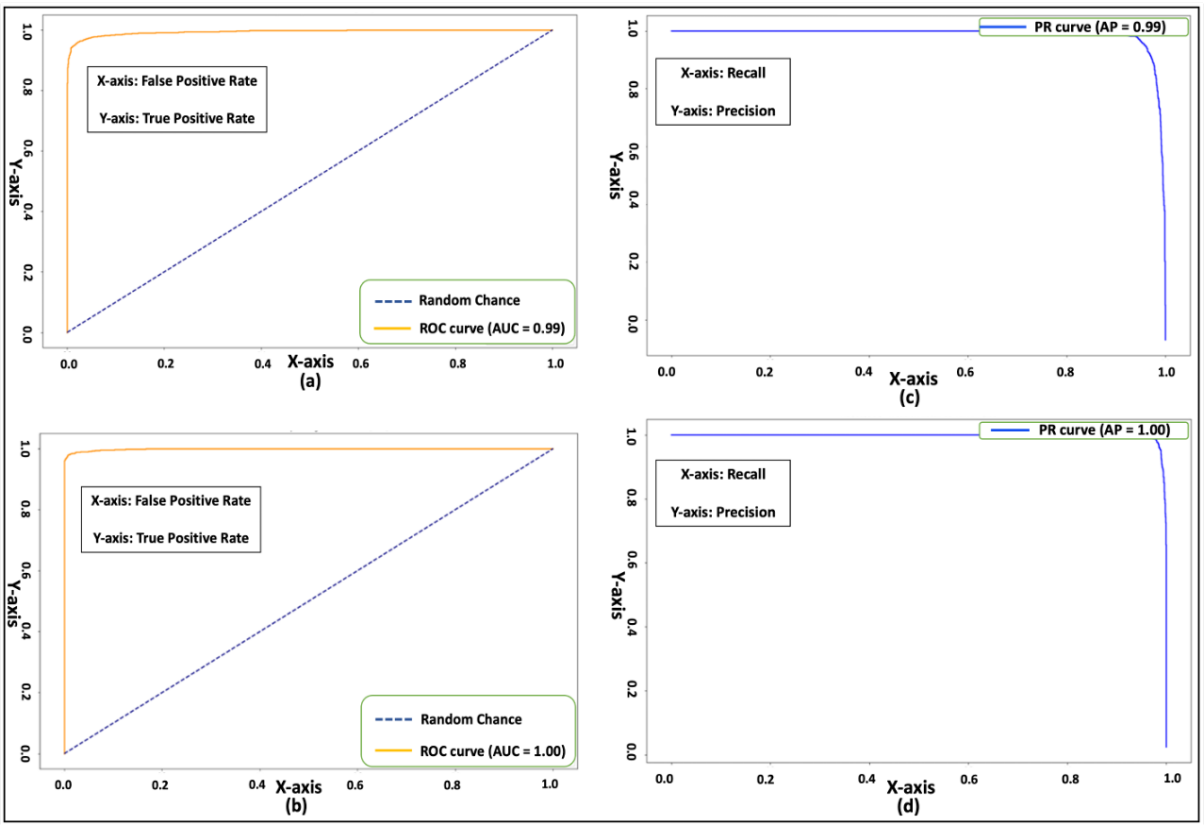


Figure 5.25: OCT HyBoost ROC (a) PR (c) Curves without Hyperparameters Fine-tuning and ROC (b) PR (d) Curves with Hyperparameters Fine-tuning.

The ROC curve, shown in Figure 5.25.a, corresponds to the HyBoost model without tuning. Upon examination, the ROC curve resulted an impressive AUC of 0.99. This indicates that the HyBoost model in its original state already possesses a high degree of classification capability, making very few mistakes when differentiating between the two classes. The second ROC curve (Figure 5.25.b) showcases the results after tuning the HyBoost model. Remarkably, this curve achieves an AUC of 1.00, signifying that the tuned HyBoost model provides flawless classification across all threshold levels. This perfect AUC indicates that with the tuning adjustments, the HyBoost classifier has been optimised to the extent that it makes zero classification errors, at least within the context of the data it was tested on. Both the untuned and tuned versions of the HyBoost model showcased exemplary performance in classifying the OCT data. The tuning process further enhanced the classifier's performance, achieving perfection as reflected in its ROC curve with an AUC of 1.00.

Confirming previous results, Figure 5.25.c showcases the PR curve of the untuned HyBoost

with an AP of 0.99. The curve is very close to the top-right corner of the graph, indicating a high level of precision throughout the range of recall values. However, there's a noticeable drop in precision as recall approaches 1.0, suggesting that while the model has an impressive overall performance, there are specific scenarios where it may produce FPs. On the other hand, Figure 5.25.d, presents a perfect AP of 1.00. The curve sits right at the top of the graph, meaning that the model maintains a precision of 1.0 across all recall values. This suggests that the tuned HyBoost model consistently produces accurate positive predictions, regardless of how many actual positive samples are being identified. The noticeable improvement from an AP of 0.99 in the first image to an AP of 1.00 in Figure 5.25.d implies that tuning the HyBoost parameters had a significant positive impact on the model's performance for the OCT dataset in general, and DMO prediction in particular. While both models demonstrate commendable performance, the tuned HyBoost model, as visualised in Figure 5.25.d, appears to provide perfect precision across all levels of recall, indicating a likely superior and more reliable classification performance for OCT results.

Validating a ML/DL model on an unseen dataset is of paramount importance, especially in the medical field. Unlike other domains, where errors might be acceptable or less critical, in medicine, even a slight misjudgement can have profound consequences on patient care, diagnosis, and treatment outcomes. Throughout the testing process, it became evident that different models reacted uniquely to each dataset, underscoring the complex nature of ML in healthcare. Each model, shaped by its underlying algorithms and architectures, showcased strengths and weaknesses when confronted with diverse data distributions present in the medical datasets. Some models that excelled with one dataset might have struggled with another, highlighting the criticality of diverse validation. This variability in performance across datasets reinforces the importance of the forthcoming X-ray dataset as the decisive stage. In fact, testing models on new, unseen data ensures that they are robust, generalisable, and can effectively handle real-world scenarios rather than just memorising patterns from their training data. Moreover, this validation acts as a rigorous filter to shortlist the most promising models. Only those that demonstrate superior performance on this validation will advance to the next phase of testing.

In this subsequent phase, the integration of demographic and physiological features will further refine the models, aiming to enhance their accuracy and utility. Ensuring rigorous validation in the preliminary stages guarantees that the foundation is strong before introducing these additional complexities.

X-ray Dataset Prediction Results

The results obtained by training the prediction models including and excluding the hyperparameters fine-tuning on X-ray dataset are summarised in Tables 5.9 and 5.10. The untuned DT classifier for X-ray data demonstrated an Acc of 72.21%, precision of 73.03%, recall of 96.11%, and an F1-Score of 82.99%. Post-tuning, there was a marked improvement across metrics, with accuracy increasing by approximately 7.72% to 79.93%, precision rising by around 7.23% to 80.26%, and the F1-Score augmenting by roughly 3.89% to 86.88%, despite a slight decrease in recall to 94.70%. These enhancements are further confirmed by the confusion matrices from both instances, which indicate an improvement in the counts of TPs and TNs, signalling superior classification performance. In summary, tuning has significantly enhanced the efficacy of the DT classifier in analysing X-ray data.

Table 5.9: X-ray without Fine-tuning

Model	**Acc (%)**	**Precision (%)**	**Recall (%)**	**F1-Score (%)**	**LCE**
XGBoost	95.61	96.89	96.89	96.89	0.3
RF	89.31	92.15	92.75	92.45	0.42
DT	72.21	73.03	96.11	82.99	0.5
AdaBoost	90.78	91.04	92.31	92.64	0.36
HyBoost	96.52	97.17	97.92	97.54	0.23

Table 5.10: X-ray with Fine-tuning

Model	**Acc (%)**	**Precision (%)**	**Recall (%)**	**F1-Score (%)**	**LCE**
XGBoost	94.76	96.75	95.74	96.24	0.32
RF	89.09	91.26	93.40	92.31	0.42
DT	79.94	80.27	94.70	86.89	0.47
AdaBoost	92.39	93.31	92.91	93.08	0.34
HyBoost	96.28	96.90	97.82	97.36	0.24

Without tuning, RF model's performance led to an accuracy of approximately 89.31%, precision of 92.15%, recall of 92.75%, and an F1-Score of 92.45%. However, with tuning (RF), these figures changed resulting in an Acc of 89.09%, precision of 91.26%, recall of 93.40%, and an F1-Score of 92.31%. Comparatively, there was a marginal decrease in Acc by 0.22% after tuning. Precision declined by nearly 0.89%, while recall showed a boost of roughly 0.65%. The F1-score, which provides a balance between precision and recall, experienced a minor decrease of 0.14% post-tuning. This suggests that while tuning enhanced recall, it made slight compromises on precision and accuracy. The choice between tuned and untuned models should be grounded in the analysis's specific objectives, in this case minimising FNs or positives. In addition, capturing a maximum number of TPs is pivotal, indicating a preference for higher recall, where the tuned model is more suitable. Conversely, for a more harmonised performance interplay between precision and recall, the untuned variant's marginally superior F1-score might be more appealing.

After hyperparameter tuning, the AdaBoost model showcased enhanced performance across all metrics. Specifically, accuracy increased from 90.78% to 92.39%, precision showed a modest rise from 91.04% to 93.31%, recall increased from 92.31% to 92.91%, and the F1-Score, elevated from 92.64% to 93.08%. This progression underlines the value and benefits of tuning in ML. To interpret, the elevated accuracy implies the model made more accurate predictions post-tuning. The improvement in precision suggests a higher proportion of correctly identified

positive cases, while the improvement in recall indicates the model's enhanced ability to detect authentic positives. Therefore, the tuning led the model to make better predictions, diminish errors, and attain a more harmonised proficiency in detecting TP instances.

The XGBoost model, without tuning, accurately identifies both positive and negative classes approximately 95.61% of the time, boasting a high precision and recall rate of 96.89%. However, it still faces the risk of misclassifying some instances, which in the context of medical X-ray results can be crucial. Tuning, typically employed to enhance model performance via hyperparameter adjustments, led to a slight decrease in precision in this case, with recall dropping more substantially. This may indicate the post-tuned model's cautious stance in predicting positives, evident from the rise in FNs pre-tuning. Given the gravity associated with X-ray classifications, any misclassification carries significant consequences. Thus, understanding the balance between precision and recall is vital: while high precision supports for the reliability of positive predictions, it might miss out on certain positive instances, and high recall, though capturing most positives, may include some false ones. Selecting between the tuned and untuned models should resonate with the diagnostic tool's objectives, pneumonia, as an example. If avoiding missed diagnoses is paramount, high recall should be prioritised. In addition, because ensuring the reliability of positive predictions is the goal, high precision should take precedence.

Untuned HyBoost model showcase significant performance with implicit distinctions: the untuned model slightly surpasses the HyBoost-tuned model with an Acc of 96.52% compared to 96.28%, a precision of 97.17% versus 96.90%, and an F1-Score of 97.54% against 97.36%. The recall values are nearly identical, with the non-tuned model marginally leading at 97.93% against the tuned model's 97.83%. These minor discrepancies suggest that, in this context, the untuned model may already be optimised, with its default parameters fitting the task well. Hence, the additional tuning didn't yield significant enhancements.

The HyBoost model once again exhibits exceptional performance with the lowest LCE of 0.23, suggesting its innate efficiency in handling X-ray image data. XGBoost also demonstrates strong performance with an LCE of 0.3. AdaBoost follows with an LCE of 0.36, and the RF

model records an LCE of 0.42, indicating that while effective, it could potentially benefit from further optimisation. The DT model presents the highest LCE at 0.5, which may reflect its tendency towards overfitting or lack of complexity required for the task. Upon fine-tuning, the results are somewhat unexpected; the XGBoost model's LCE slightly increases to 0.32. This could suggest overfitting during the fine-tuning process or that the initial parameters were already close to optimal for the dataset. RF maintains an LCE of 0.42, showing no improvement with fine-tuning, which might imply a limitation in the model's structure concerning the X-ray data. DT shows a slight improvement, reducing its LCE to 0.47, but still remains the highest amongst the models. AdaBoost's LCE decreases to 0.34, which is a positive indication of its responsiveness to fine-tuning.

Following a rigorous critical evaluation and performance assessment across multiple datasets during the training and validation stages of the initial scenario, it was determined that the DT and RF models underperformed in terms of their performance capabilities. As a result, they have been eliminated from consideration. This decision was reached upon contrasting their outcomes with the superior results demonstrated by AdaBoost, XGBoost, and the innovative hybrid model, HyBoost. Notably, the tuned AdaBoost, untuned XGBoost, and untuned HyBoost models have been cherry-picked for progression into the subsequent testing phase. This decision stemmed from the observation that both XGBoost and HyBoost, in their untuned states, showcased commendable results, with only marginal enhancements observed in their tuned iterations. However, in the case of AdaBoost, tuning made a monumental difference, notably enhancing prediction results for DR, DMO, and pneumonia.

The upcoming section will delve deeper into a comprehensive evaluation of these selected models. Herein, there will be an integration of both demographic and physiological features, amalgamating them with HF and DHF features to train and validate the selected predictive models. The crux of this analysis will revolve around each model's performance for every dataset, interpreted through their respective confusion matrices—a pivotal tool that offers a holistic view of the TPs, FPs, TNs, and FNs, thereby granting a comprehensive understanding of a model's performance nuances. Furthermore, the section will present key metrics such

as accuracy, Precision, Recall, Spe, and F1-score, providing a rounded perspective on each model's prowess. Concluding this section, a definitive decision will be made, spotlighting the most adept model, which will then be benchmarked within the final framework, juxtaposed against existing methodologies for a comprehensive comparison.

5.9.2 Enhanced Scenario Results: Validation on Selected Prediction Models

Prediction Performance of AdaBoost

The AdaBoost model applied to the Fundus dataset achieved an Acc of 78.5% in predicting DR (Table 5.11). This indicates that it was able to determine the presence or absence of DR correctly in about 78.5% of the examined instances, confirming the potential positive impact of the integration of demographic and physiological features where the absence of the latter features only achieved 63%. The model's Sen suggests that it correctly identifies 78.6% of all authentic positive cases against only 60.91% in the previous scenario. Conversely, its Spe indicates that 73.5% of TN cases were detected accurately. A crucial aspect of the model's performance is precision, which shows that, of all cases flagged as positive (Figure 5.26.a), around 71.1% truly were positive. The F1-score stands at 74.7%.

Table 5.11: AdaBoost Performance Metrics: Demographic and Physiological Features Case

	Acc (%)	**Precision (%)**	**Recall (%)**	**Spe (%)**	**F1-Score (%)**
Fundus	78.5	71.1	78.6	73.5	74.7
OCT	94.95	95.27	94.69	95.21	94.98
X-ray	93.9	94.9	93.2	94.6	94.03

This is a reasonable result, signifying a balanced handling of both FPs and FNs. This proves the balance addition caused by the incorporation of extra features without which the harmony between precision and recall was only around 62.40%. Although the AdaBoost model's per-

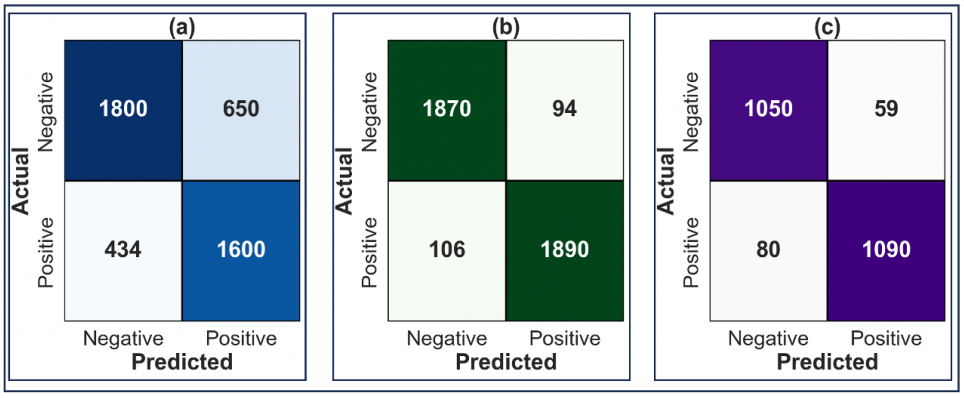


Figure 5.26: Confusion Matrix of AdaBoost for – (a): Fundus Dataset, (b): OCT Dataset and (c): X-ray Dataset.

formance on the Fundus dataset is commendable, there is always room for enhancement, especially in a medical setting by, for example, trying out different modelling methodologies leading to even better outcomes.

Delving into the OCT dataset, the AdaBoost model displayed some significant attributes (Figure 5.26.b). The model does not show a bias in its predictions as the numbers of TNs and TPs are fairly balanced, highlighting its well-adjusted nature. It's reassuring to note that the model registered a mere 94 FPs, indicating it does not rashly predict the disease's presence. This avoids unnecessary medical interventions that might otherwise be prescribed. However, a point of concern is the 106 FNs, which in medical contexts, could mean overlooking a disease. Such oversights might lead to severe repercussions, especially concerning DMO disease, which can culminate in irreversible vision damage if undetected. Both the precision and recall rates surpass 94% (Table 5.11), signifying both accurate and reliable predictions versus 93.31% and 92.9% in the previous scenario, respectively. Furthermore, the model presents an Acc slightly shy of 95%. Such a high figure is praiseworthy, but even minor percentages of misdiagnoses in medical scenarios can be consequential. Therefore, it's crucial to scrutinise FNs and positives. The performance showed by the AdaBoost on the OCT dataset overpassed the resulted metrics values of the precedent scenario, confirming the impact of considering age, gender, diabetic type, and blood pressure features when training the predictive model.

Switching focus to the X-ray validation dataset, the AdaBoost model showcased exemplary

performance in predicting pneumonia (Figure 5.26.c). With an overall accuracy achieved at 93.9% (Table 5.11), the model could correctly predict the disease's presence or absence in most cases. The model's Sen, standing at 93.2%, indicates that when pneumonia is genuinely present, it's detected most of the time, unlike the case of the absence of demographic and physiological features in the predictive model training. On the other hand, the model's Spe of 94.6% suggests a reliable prediction when the disease isn't present. Precision, another pivotal metric, sits at 94.9%, meaning that the majority of cases tagged as positive indeed are. The model also shines in its negative predictive value, correctly identifying 92.9% of authentic negative cases.

Despite these outstanding figures, there are areas of potential improvement. The FPR of 5.4% indicates occasional overdiagnosis, while the 6.8% FN rate suggests missed detection in certain instances. Despite these concerns, it is clear that there is a great improvement in minimising FP and FN rates compared to values of 6.5% and 7.1%, respectively where extra features are not considered. Both these figures, while relatively low, carry significant weight in medical contexts. Therefore, the model, while performing admirably, should be continuously refined to further pare down these errors.

Prediction Performance of XGBoost

The XGBoost model's performance on the Fundus images (Figure 5.27.a), primarily used for predicting DR, is quite satisfactory, achieving an accuracy slightly beyond 78% (Table 5.12). Precision and recall, both crucial metrics in clinical scenarios, show good values. Precision is an indication of how many of the positive predictions were actually correct, while recall (or Sen) indicates how many actual positive cases were identified correctly. When both these metrics show reasonable values, it means the model is striking a good balance between identifying true cases and avoiding false alarms, which was missing in previous tests done without considered demographic and physiological features in the training process. An F1-score, a harmonic mean of precision and recall, also points towards a solid balance between these two metrics. A higher F1-score indicates fewer FPs and negatives.

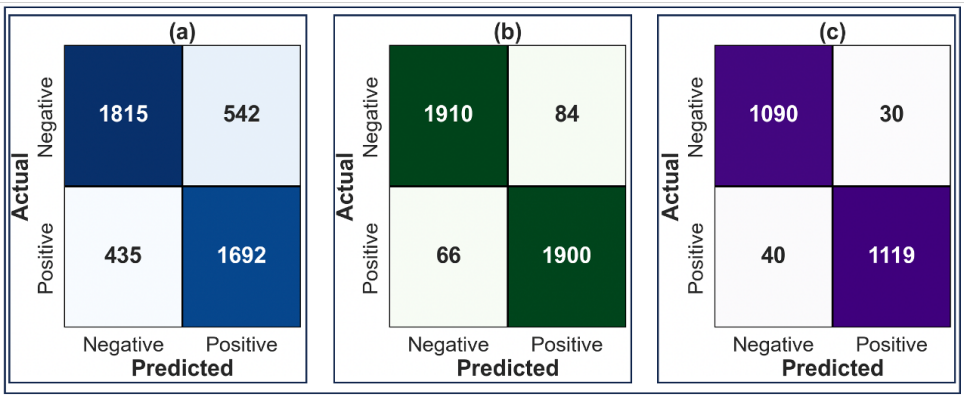


Figure 5.27: Confusion Matrix of XGBoost for – (a): Fundus Dataset, (b): OCT Dataset and (c): X-ray Dataset.

Table 5.12: XGBoost Performance Metrics: Demographic and Physiological Features Case

	Acc (%)	**Precision (%)**	**Recall (%)**	**Spe (%)**	**F1-Score (%)**
Fundus	78.16	75.72	79.54	77.01	77.59
OCT	96.61	95.77	96.65	95.78	96.2
X-ray	96.93	97.38	96.56	97.32	96.97

Despite these commendable metrics, there are areas of concern. The 542 FPs indicate instances where the model mistakenly predicted the presence of DR. The 435 FNs, on the other hand, are cases where DR went undetected. Both errors are concerning in a clinical setting, considering the significant implications of misdiagnosis. The choice between minimising FNs or FPs depends largely on the medical condition in question. In the case of DR, missing a diagnosis can have severe consequences, emphasising the importance of recall. This might necessitate adjustments to the model to enhance its recall, even if it comes at the cost of precision.

On the other side, the performance of the XGBoost model on the OCT dataset (Figure 5.27.b), used for detecting DMO disease, is impressive, with an accuracy close to 96.61% (Table 5.12), overpassing the accuracy achieved without the integration of extra features to reach only 93.63%. Sen or recall, which is approximately 96.65%, suggests the model's commendable ability to identify actual positive DMO cases. Such high recall is paramount in a clinical

setting where missing out on actual positive cases can have adverse patient outcomes. A Spe of 95.78% denotes the model's capability to correctly identify the negative cases or those without DMO. Precision, standing at around 95.77%, indicates that the vast majority of the model's positive predictions are indeed accurate. A FPR of 4.33% also means that out of 100 negative cases, the model mistakenly flags about four as positive. This confirms the impact of adding extra features into the training process where FPR was 5.46% also means that out of 100 negative cases the model wrongly flags more than 5 have DMO. The model's effectiveness isn't only restricted to accuracy. A well-balanced F1-score of 94.33% (versus 93.04%), suggests an equilibrium between precision and recall, essential for a reliable diagnostic tool. While the metrics suggest a robust model, real-world clinical scenarios demand a consideration of the tangible implications of FPs and negatives. The metrics should always be contextualised within the clinical utility and potential harm of misdiagnosis.

In the case of its application on the X-ray dataset, for model validation purpose, the XGBoost model shines here as well (Figure 5.27.c), showing an accuracy of approximately 96.93% (Table 5.12), versus only 95.61%. This high accuracy underscores the model's robust predictive capability for this task. The model's Sen/recall is around 96.56%, highlighting its efficacy in identifying true pneumonia cases. Such high recall is crucial to ensure patients with pneumonia are correctly diagnosed and receive appropriate treatment promptly. With a Spe of 97.32%, the model proves its mettle in identifying patients without pneumonia, ensuring that those who are disease-free aren't subjected to unnecessary treatments or interventions, unlike the previous scenario results where extra features were not taken into consideration imposing more FPs. In turn, this reduced the model's Spe. A precision of 97.38% is indicative of the model's robustness. This means that when pneumonia is predicted, it is correct in about 97.38 cases out of 100, proving an enhancement of about 0.49%. The balanced F1-score of 96.97% is again a testament to the model's balanced diagnostic capabilities, balancing both precision and recall effectively. The XGBoost model's robustness on the X-ray dataset makes it a promising tool for pneumonia diagnosis.

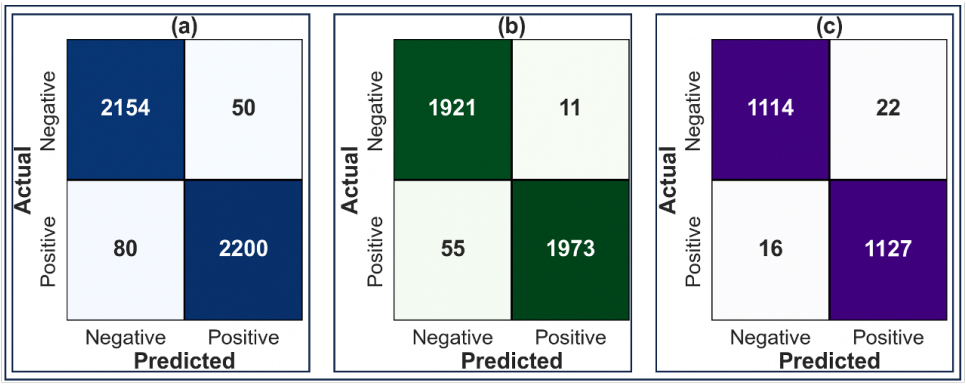


Figure 5.28: Confusion Matrix of HyBoost for (a): Fundus Dataset, (b): OCT Dataset and (c): X-ray Dataset.

Prediction Performance of HyBoost

On examining the results from the Fundus dataset, the hybrid predictive model (HyBoost) yields an impressive accuracy of approximately 96.66% (Table 5.13). Such a high accuracy underscores the model's adeptness in discerning between instances of DR and its absence. Moreover, the model's low FP count is commendable, as over diagnosis can often lead to unnecessary medical procedures and treatments, potentially burdening patients both financially and psychologically. The model's conservative approach in predicting positive cases ensures that unwarranted interventions are minimised. The results appear promising, where the HyBoost model addressed the FNs imperatively (Figure 5.28.a). In fact, the precision, recall, F1-score have risen by 33.6%, 38.85%, and 33.4%, respectively with a Spe score achieving 97.73%.

Table 5.13: HyBoost Performance Metrics: Demographic and Physiological Features Case

	Acc (%)	**Precision (%)**	**Recall (%)**	**Spe (%)**	**F1-Score (%)**
Fundus	96.66	97.78	96.49	97.73	97.12
OCT	98.33	99.45	97.29	99.43	98.35
X-ray	98.2	98.1	98.6	98.1	98.3

Transitioning to OCT dataset, the model stands out with a laudable Acc of 98.33% (Ta-

ble 5.13). Such a score indicates its proficiency in detecting DMO disease compared to only 96.28%. A closer inspection of the results illuminates the model's strengths, notably its high Sen of 97.29%. Sen holds particular gravitas in medical diagnostics, as the consequences of missing an authentic positive case can be severe. That said, a 2.71% chance of overlooking positive cases cannot be ignored (Figure 5.28.b), however, remaining better than the case of exclusion of extra features. On the Spe front, a score of 99.43% emphasises the model's finesse in reducing false alarms versus only 91.02%. The model's precision of 99.45% further instils confidence in its predictions, particularly when compared with only 91.04% resulted from previous scenario. With a nominal FPR, the model effectively limits misdiagnoses. In summation, the HyBoost model's performance on the OCT dataset is sterling. It appears poised to be a valuable ally in diagnosing DMO disease, but it's paramount to juxtapose its results with other diagnostic methodologies.

Delving into the X-ray validation dataset, the model continues to display prowess, boasting an Acc of 98.2% (Table 5.13) compared to only 96.52%. The precision and recall, standing at 98.1% and 98.6% respectively, further attest to the model's balanced performance. The near-equivalent values of precision and recall underscore the model's uniformity in predicting both positive and negative cases (Figure 5.28.c). The F1-score, harmonising both metrics, echoes this sentiment with a score of 98.3%. However, medical diagnostics is an arena where even minute discrepancies carry weight. The model's 16 FNs, while modest in comparison to the total sample size, underline a critical shortcoming. Overlooking authentic pneumonia cases can potentially deprive patients of essential care. The 22 FPs, on the other hand, may subject patients to superfluous treatments or tests, which can have psychological, physical, and financial ramifications.

Conclusively, while the HyBoost model exhibits commendable proficiency on the X-ray dataset in predicting pneumonia, the significance of even a handful of misdiagnoses cannot be understated. It's indispensable to continually monitor and validate the model's performance against fresh datasets and consider it in tandem with other diagnostic measures. In its entirety, the HyBoost predictor showcases the potential to be an instrumental tool in medical diagnos-

tics. Its results across the three datasets are promising, and with continuous refinements and validations, it can be a cornerstone in healthcare imaging analysis. However, it's essential to recognise that no model is flawless, and its integration should be done judiciously, keeping patient well-being at the forefront. Given its outperforming results compared to AdaBoost and XGBoost, the proposed hybrid predictive model will be considered as the final model for the suggested predictive framework to be then tested against state-of-the-art works (benchmarking).

5.9.3 Benchmarking: Comparative and Critical Discussion

Benchmarking and rigorous testing against established methods are critical for validating the efficacy and innovation of a proposed method within a research context. By engaging in comparative analysis with a range of notable approaches detailed in the literature review, researchers can demonstrate where their proposed method stands in the existing hierarchy of solutions, and identify specific areas of improvement or novelty. The discussed related works are going to be used in the following benchmarking process. The proposed method, which stands on the precipice of these diverse and powerful architectures, is meticulously designed to not only draw lessons from these preceding models but to also forge ahead with unique innovations. It is through this rigorous benchmarking process, placing the proposed framework in the crucible against the high-performing architectures mentioned, providing the ability to truly measure and articulate its contribution to the field—whether it be in enhancing accuracy, efficiency, generalisability, or computational cost-effectiveness. The ultimate aim is to provide a compelling case for the proposed solution by quantitatively and qualitatively demonstrating its superior and specialised performance on the same tasks used by the referenced works.

In this benchmarking study, the proposed method will be rigorously tested across three distinct datasets including Fundus, OCT and X-ray, to comprehensively evaluate its adaptability, scalability, and generalisability. This approach enables an assessment of the model's performance not only on datasets that align with the original modality types of the benchmarking methods but also on those that introduce new modality challenges. By including cases where

the dataset size differs significantly from the original training conditions of the benchmarking methods, the study aims to further validate the robustness and versatility of the proposed model. This thorough evaluation is intended to demonstrate the model's ability to maintain outstanding performance across various medical imaging modalities and varying dataset sizes, thereby establishing its superiority in the field of medical image analysis.

In the experiments conducted, a 10-fold cross-validation was meticulously applied, ensuring the robustness and credibility of the prediction results. Employing 10-fold cross-validation is paramount in the world of ML/DL, as it systematically divides the datasets into ten distinct subsets, using nine for training and one for testing in every iteration. This procedure is repeated ten times, ensuring that each subset is used for validation precisely once. Such a comprehensive approach minimises the risk of overfitting and enhances the generalisability of the model, making the results more reliable and less susceptible to data biases. In the forthcoming performance evaluation, the analysis will extend beyond conventional metrics to embrace a holistic view of model efficacy. accuracy, precision, recall, specificity and F1-score will serve as the foundational metrics resulted from PMM, capturing the essence of each model's predictive power. Complementing these, the processing time will be scrutinised, acknowledging that practical deployment of these algorithms demands not just precision but also efficiency.

Benchmarking Results Using Fundus Dataset

The Fundus dataset serves as a critical benchmark for evaluating the performance of various DL models in medical image analysis. The comparative analysis of different methods, including the proposed approach, reveals key insights into the strengths and weaknesses of each model, particularly in terms of accuracy, efficiency, and overall effectiveness as shown in Table 5.14.

Table 5.14: Benchmarking Results Using Fundus Dataset

Method	**Processing time (h:m:s)**	**Acc (%)**	**Precision (%)**	**Recall (%)**	**Spe (%)**	**F1-score (%)**
(Alryalat et al., 2022)	01:45:46	89.7	85.3	83.2	88.06	84.2
(Li et al., 2022a)	01:03:43	84.1	81.4	83.8	83	82.5
(Wahab Sait, 2023)	01:00:53	85.78	84.02	82	84.35	82.99
(Paul and Talukder, 2023)	01:10:00	80.1	78.13	79.9	79.98	79
(Sharma and Guleria, 2023b)	01:00:12	81.03	80.1	79.5	80.63	79.79
(Shimpi and Shanmugam, 2023)	00:56:42	69.73	63.2	64	65.06	63.5
(Vetrithangam et al., 2023)	00:56:42	69.73	63.2	64	65.06	63.5
Proposed Method	**00:52:25**	**96.66**	**97.78**	**96.49**	**97.73**	**97.12**

Alryalat et al.'s method demonstrates a respectable balance across all performance metrics, achieving an F1-score of 84.2% (Alryalat et al., 2022). The close alignment between precision (85.3%) and recall (83.2%) indicates a reasonable trade-off, suggesting that the model is well-calibrated for the task at hand. However, the major drawback of this approach is its significant processing time, which exceeds one and a half hours. This long duration may severely limit its practical applicability in real-world, time-sensitive environments, such as clinical settings where rapid diagnostics are essential. While the model's reliability is evident, there is a clear need for optimisation to enhance its efficiency without compromising accuracy.

The (Alryalat et al., 2022) method was originally developed and tested on a significantly smaller dataset consisting of OCT images. The original high accuracy (95.9%) reflects the model's optimisation for this specific modality. However, when applied to the Fundus dataset, which is larger and of a different modality, the accuracy drops to 89.7%, with a substantial processing time of over one and a half hours. The noticeable drop in accuracy and prolonged processing time indicates a lack of adaptability when transitioning from OCT to Fundus images. The model, while effective in its original domain, struggles to maintain the same level of performance when faced with a different imaging modality. The increased dataset size from 3,000 to 18,615 images exacerbates the model's inefficiencies, particularly in terms of processing time. This suggests that the method may not scale well with larger datasets. The significant decline in accuracy and F1-score highlights the model's limited generalisability. It was likely overfitted to the specific features of OCT images, which do not translate effectively to the different characteristics of Fundus images.

On the other hand, Li et al.'s method shows a moderate performance across all metrics (Li et al., 2022a), with a notable improvement in processing time compared to (Alryalat et al., 2022), reducing it to just over an hour. However, this gain in efficiency comes at the cost of lower accuracy (84.1%) and a reduced F1-score (82.5%). The method's precision (81.4%) and recall (83.8%) are balanced, but the trade-off in accuracy raises concerns about its applicability in medical image analysis, where precision is often more critical than speed. While the method offers better efficiency, the drop in overall performance may not be justifiable, particularly in scenarios where high accuracy is paramount.

The (Li et al., 2022a) method was originally designed for Fundus images, similar to the dataset used in this benchmarking study. Despite this alignment in modality, the accuracy drops significantly from the original 99.2% to 84.1% when tested on the larger Fundus dataset. Although the method is tested on the same modality (Fundus), the drop in accuracy suggests that it may not adapt well to datasets that are larger or more diverse than what it was originally trained on. The method was originally tested on a smaller dataset, and the significant performance decline when applied to the larger 18,615 image dataset indicates poor scalability. The

model may be overfitted to the specific characteristics of the original dataset. In addition, the reduced performance demonstrates limited generalisability, even within the same modality. This suggests that the method may be overly reliant on the specific data distribution of its original training set, struggling to maintain performance across different subsets of Fundus images.

Wahab Sait's method improves upon the balance of accuracy, precision, and recall compared to (Li et al., 2022a), with an accuracy of 85.78% and a precision of 84.02% (Wahab Sait, 2023). The processing time is further reduced to approximately one hour, enhancing its practicality for real-time applications. However, despite these improvements, the method still trails behind the proposed approach in all major performance metrics. The proposed method not only surpasses this model in terms of accuracy and F1-score but also achieves a shorter processing time, indicating that (Wahab Sait, 2023) could benefit from further optimisation, particularly in computational efficiency.

This method was originally designed for Fundus images, similar to the benchmark dataset, but with a smaller size. The accuracy decreases from 98% to 85.78% when tested on a larger Fundus dataset. The method shows a reasonable level of adaptability since it operates within the same modality. However, the drop in accuracy suggests that while the method is somewhat adaptable, it is not robust enough to handle the increased complexity or diversity of the larger dataset. The transition from 5,590 to 18,615 images leads to a decline in performance, indicating that the method struggles with scalability. The model may not be effectively managing the larger, more varied data. Moreover, the moderate decline in accuracy and F1-score indicates limited generalisability. The model may be overfitted to the specific subset of Fundus images used in the original study, leading to decreased performance when applied to a broader dataset.

Conversely, Paul and Talukder's method records the lowest performance among the evaluated models, with an Acc of 80.1% and an F1-score of 79% (Paul and Talukder, 2023). The significant drop in both precision (78.13%) and recall (79.9%) raises serious concerns about the reliability and robustness of this model, particularly in critical diagnostic settings where accuracy is very important. Additionally, the processing time of over one hour does not compensate for its lower performance, suggesting that the model requires substantial methodological revi-

sions. The combination of low accuracy and relatively long processing time makes this method less competitive compared to others in the study.

Similar to (Wahab Sait, 2023), this method was also originally trained on a smaller Fundus dataset. However, it shows a more significant drop in performance when applied to the larger Fundus dataset, with an accuracy decrease from 97.78% to 80.1%. The significant drop in accuracy and F1-score suggests a lack of adaptability, even within the same modality. The model fails to generalise effectively to the larger dataset. The method's poor performance on a larger dataset highlights its scalability issues. It likely struggles with the increased data diversity and volume, leading to a marked decline in performance. Additionally, the drastic reduction in performance underscores the method's limited generalisability. It appears to be highly specialised for the smaller, original dataset and does not perform well when applied to a broader or more complex set of Fundus images.

Contrarely, Sharma and Guleria's method is noteworthy for its short processing time of about one hour (Sharma and Guleria, 2023b). However, this efficiency comes with a trade-off in performance, as evidenced by its accuracy of 81.03% and an F1-score of 79.79%. While the model offers a commendable speed advantage, it does so at the expense of accuracy and reliability, making it less suitable for critical medical applications where precision is crucial. The trade-off between speed and accuracy is evident, and while the method may be suitable for less critical applications, it falls short when compared to the proposed method in both performance and overall effectiveness.

Originally developed for X-ray images, this method faces a significant challenge when applied to Fundus images, resulting in a notable drop in Acc from 93% to 81.03%. The method demonstrates limited adaptability, struggling to transition from X-ray to Fundus images. The performance decline indicates that the model is heavily tailored to the specific characteristics of X-ray images and does not transfer well to other modalities. While the original and benchmark datasets are relatively similar in size, the drop in performance suggests that the method may not scale effectively across different data modalities. The considerable decrease in accuracy and F1-score highlights the model's lack of generalisability. It is not versatile enough to generalise

across different types of medical images, performing poorly outside its original domain.

Shimpi and Shanmugam's method exhibits outstanding results with an accuracy of 93.7% and an F1-score of 80.51%, suggesting a robust model that performs well across various metrics (Shimpi and Shanmugam, 2023). However, it still lags behind the proposed method in overall performance and processing time. The model's longer processing time of approximately 1 hour and 28 minutes indicates potential inefficiencies in its computational approach, which could be a limiting factor in its practical application. While the model is strong in accuracy, the need for optimisation to improve efficiency and reduce processing time is apparent.

This method, originally designed for Fundus images with a dataset size close to the benchmark dataset, shows a smaller drop in performance, with accuracy decreasing from 95.56% to 93.7%. The method demonstrates relatively good adaptability within the same modality, maintaining a high level of performance despite the larger dataset. The performance stability suggests that the method scales reasonably well from 10,000 to 18,615 images, though there is a slight efficiency issue, as indicated by the longer processing time. The small decline in accuracy and F1-score indicates strong generalisability within the Fundus modality. The model can generalise effectively to a larger and potentially more diverse dataset, maintaining high performance.

Vetrithangam et al.'s method stands out for its exceptionally short processing time of less than one minute, the shortest among all the methods. However, this speed comes at a significant cost to performance, as the model achieves the lowest Acc (69.73%) and F1-score (63.5%) in the study. The trade-off between speed and performance is particularly pronounced here, suggesting that the model sacrifices too much in terms of accuracy to achieve its rapid processing time. In critical applications, such as medical diagnostics, this trade-off is unlikely to be acceptable, as it undermines the reliability and validity of the results.

Originally designed for a much smaller X-ray dataset, this method exhibits a drastic drop in performance when applied to the significantly larger and different Fundus dataset, with accuracy plummeting from 99.77% to 69.73%. The method shows a severe lack of adaptability, struggling to transition from X-ray to Fundus images. The significant decline in accuracy and

F1-score indicates that the model is not versatile and cannot effectively handle different imaging modalities. The substantial increase in dataset size from 1,485 to 18,615 images overwhelms the model, highlighting its poor scalability. The method appears optimised for small-scale datasets and cannot scale to handle larger, more complex data. The model's inability to generalise across different modalities and larger datasets is evident from the significant performance decline. It is highly specialised for the specific, small X-ray dataset it was originally trained on and does not perform well outside that narrow scope.

The proposed method sets a new benchmark for both performance and efficiency, achieving the highest Acc (96.66%) and F1-score (97.12%) among all the evaluated models. The method demonstrates an excellent balance between precision (97.78%) and recall (96.49%), indicating its robustness and reliability across various metrics. Furthermore, the processing time of just under one hour is the shortest among all the methods, highlighting the advanced optimisation techniques employed. The proposed method outperforms all other models, not only in terms of accuracy but also in computational efficiency, making it a significant advancement over the existing methods. This balance between high performance and speed marks the proposed method as a superior solution in the field of medical image analysis, capable of handling the demands of real-world applications effectively.

The proposed method consistently outperforms all the benchmarking methods across the Fundus dataset, achieving the highest Acc (96.66%) and F1-score (97.12%). This performance, coupled with the shortest processing time, highlights the method's superior adaptability, scalability, and generalisability. The proposed method demonstrates exceptional adaptability, maintaining top-tier performance across a different range of dataset sizes and modality types. Unlike other methods, it is not constrained by the specific characteristics of its training data, showcasing its versatility across varied imaging modalities. The method scales effectively from smaller to larger datasets without a significant loss in performance or efficiency. Its robust design ensures that it can handle large and diverse datasets, which is a critical requirement for real-world applications in medical imaging. The proposed method exhibits outstanding generalisability, maintaining high accuracy and F1-score across different datasets and modalities. Its ability

to generalise well across a wide range of Fundus images, regardless of size or complexity, positions it as a highly reliable tool for medical image analysis.

When comparing the benchmarking methods to each other and to the proposed method, it is clear that while some methods offer competitive accuracy (e.g., Shimpi and Shanmugam, 2023), they often do so at the expense of processing time, limiting their practical applicability. On the other hand, methods that prioritise efficiency, such as (Sharma and Guleria, 2023b) and (Vetrithangam et al., 2023), suffer from significant drops in accuracy, making them less reliable for critical applications.

The proposed method, in contrast, successfully balances both performance and efficiency, setting a new standard in the field. It addresses the limitations of the benchmarking methods by offering a model that is not only highly accurate but also efficient, making it well-suited for a wide range of medical imaging tasks where both speed and precision are essential. The comprehensive evaluation highlights the proposed method's superiority in terms of generalisability and adaptability, as it outperforms existing methods across all key metrics, making it a valuable contribution to the advancement of medical image analysis.

Benchmarking Results Using X-ray Dataset

The X-ray dataset serves as a vital benchmark for evaluating the adaptability, scalability, and generalisability of different DL models in medical image analysis. This section provides a detailed comparison of various benchmarking methods, critically evaluating their performance against each other and the proposed method as shown in Table 5.15.

Table 5.15: Benchmarking Results for X-ray Dataset

Method	**Processing time (h:m:s)**	**Acc (%)**	**Precision (%)**	**Recall (%)**	**Spe (%)**	**F1-score (%)**
(Alryalat et al., 2022)	01:53:10	88.00	85.2	86.9	87.3	86.00

Continued on next page

Table 5.15: Benchmarking Results for X-ray Dataset (Continued)

Method	**Processing time (h:m:s)**	**Acc (%)**	**Precision (%)**	**Recall (%)**	**Spe (%)**	**F1-score (%)**
(Li et al., 2022a)	01:15:00	82.30	80.1	81.42	82.0	80.75
(Wahab Sait, 2023)	01:03:32	86.30	84.8	85.07	85.5	84.93
(Paul and Talukder, 2023)	01:35:00	71.04	70.3	68.5	70.8	69.3
(Sharma and Guleria, 2023b)	00:50:12	87.00	85.8	86.03	87.6	85.91
(Shimpi and Shanmugam, 2023)	02:02:00	76.91	73.4	75.0	76.1	74.19
(Vetrithangam et al., 2023)	00:57:42	80.03	67.9	69.0	79.8	68.4
Proposed Method	00:23:40	**98.20**	**98.1**	**98.6**	**98.1**	**98.3**

The method from (Alryalat et al., 2022) demonstrates strong performance across all metrics with an F1-score of 86%, indicating a well-balanced model in terms of precision (85.2%) and recall (86.9%). However, the significant processing time of over one and a half hours limits its practical applicability, especially in scenarios where rapid diagnosis is critical. The method shows reasonable adaptability in transitioning from OCT to X-ray images, maintaining strong performance. However, the high processing time suggests that the model may not be optimised for efficiency when handling different modalities. While the accuracy remains high, the processing time indicates potential scalability issues, especially when applied to larger datasets or in time-sensitive environments. Additionally, the model generalises well from OCT to X-ray images, but the extended processing time highlights a need for optimisation to improve

efficiency while maintaining accuracy.

This work in (Li et al., 2022a) achieves a slightly lower Acc of 82.3% and an F1-score of 80.75% compared to (Alryalat et al., 2022). The processing time is shorter, which improves efficiency, but this comes at the cost of lower accuracy and F1-score, which may not be acceptable in scenarios where high precision is critical. The method shows limited adaptability, struggling to maintain high performance when transitioning from Fundus to X-ray images. The reduction in accuracy suggests that the method may not scale well when applied to datasets of a different modality or larger size. Moreover, the method generalises less effectively than (Alryalat et al., 2022), indicating that it may be too specialised for its original modality, making it less versatile across different types of medical images.

(Wahab Sait, 2023) strikes a good balance between accuracy, precision (84.8%), and recall (85.07%), surpassing (Li et al., 2022a) in overall performance. The processing time is also reduced, making it more practical for real-world use. However, it still trails behind the proposed method in all metrics. The method shows good adaptability, transitioning relatively well from Fundus to X-ray images, though not as effectively as the proposed method. While processing time is improved, the method still does not achieve the highest performance, indicating that further optimisation is needed for better scalability. Additionally, the model generalises well across different modalities, but its performance, while solid, does not match the proposed method, suggesting that its generalisability could be further enhanced.

The method in (Paul and Talukder, 2023) records the lowest Acc (71.04%) and F1-score (69.3%) among the benchmarks, raising significant concerns about its applicability in critical diagnostic settings. Despite having a relatively short processing time, the low performance suggests a need for substantial revisions to the model. The significant drop in performance indicates poor adaptability when transitioning from Fundus to X-ray images, suggesting that the model is not versatile enough to handle different modalities effectively. The method's low performance even on a smaller dataset suggests scalability issues, likely due to its over-reliance on the specific characteristics of its original dataset. The work generalises poorly to X-ray images, making it unsuitable for broader applications in medical image analysis where diverse

data types are encountered.

The proposed work in (Sharma and Guleria, 2023b) offers the shortest processing time among the benchmarks, which is advantageous. However, the trade-off is a reduction in Acc (87%) and F1-score (85.91%), making it less competitive with the proposed method. The method is relatively well-adapted to its original modality (X-ray), but the trade-off in accuracy for speed suggests that it may not be as versatile when higher accuracy is required. The short processing time indicates good scalability in terms of efficiency, but the reduction in accuracy suggests that the method may not scale well in terms of performance. Add to that, the model generalises well within the X-ray modality, but its performance still falls short of the proposed method, indicating room for improvement in balancing speed with accuracy.

(Shimpi and Shanmugam, 2023)'s work achieves reasonable Acc (76.91%) and F1-score (74.19%), indicating a robust model. However, the processing time is the longest among the benchmarks, which could limit its practical application. The method shows moderate adaptability but struggles with the transition from Fundus to X-ray images, as evidenced by the drop in performance and increased processing time. The method's long processing time indicates scalability issues, particularly when handling larger datasets or when efficiency is critical. Moreover, it generalises reasonably well, but the drop in performance and efficiency suggests that it may not be as robust across different modalities as the proposed method.

The last benchmarking method shows the lowest Acc (80.03%) and F1-score (68.4%) among the X-ray benchmarks, which, coupled with a moderate processing time, suggests that the model sacrifices too much in terms of performance for the sake of efficiency (Vetrithangam et al., 2023). The significant drop in accuracy and F1-score indicates poor adaptability, especially given that the original dataset was of the same modality (X-ray) but much smaller. The method struggles to scale from a smaller dataset to a larger one, leading to a noticeable decline in performance, which highlights its limitations in handling more extensive and complex datasets. Add to that, their method generalises poorly, particularly when moving to a larger dataset of the same modality, indicating that it is highly specialised and lacks the robustness needed for broader applications.

Contrarely, the proposed method achieves the highest Acc (98.2%) and F1-score (98.3%) among all benchmarks. It also has the shortest processing time, making it highly efficient and suitable for real-world applications. The proposed method shows exceptional adaptability, maintaining superior performance across different modalities, including X-ray images. Its ability to handle different types of medical images without significant loss in performance demonstrates its versatility. The method scales effectively from smaller to larger datasets, maintaining high accuracy and efficiency. This scalability is crucial for practical applications where datasets can vary in size. Moreover, it exhibits outstanding generalisability, performing consistently well across various datasets and modalities. Its robustness makes it a reliable tool for medical image analysis, capable of delivering accurate results in diverse clinical scenarios.

Benchmarking Results Using OCT Dataset

The robustness and versatility of various DL architectures are often gauged by their performance on standardised datasets, providing a common ground for comparison. In this context, the OCT dataset emerges as a significant benchmarking asset due to its complex, real-world medical imaging data. Each of the works under consideration has been meticulously tested against this OCT dataset. By evaluating these varied architectures on the OCT dataset, it provides the opportunity to draw comprehensive comparisons, assessing not only the models' accuracy and efficiency but also their ability to handle the intricacies of medical imaging data. The proposed method joins this line-up, having been subjected to the same rigorous testing regime on the OCT dataset, thereby ensuring that this research's findings and conclusions are grounded in a consistent and challenging real-world application. This comparative testing is not just a measure of performance but a testament to the advancements in the field and the potential of the proposed method to contribute meaningfully to medical image analysis.

For the top three performing methods, the inclusion of SHAP values will unveil the decision-making dynamics, clarifying which features most significantly sway the models' predictions. This dual approach, where performance metrics intersect with explainability analysis via SHAP, will provide a comprehensive understanding of the models' operational characteristics. The in-

sights gleaned will not only benchmark the models against the current state-of-the-art but will also illuminate the path forward for algorithmic refinement and application in real-world scenarios. Table 5.16 summarises the obtained results.

Table 5.16: Benchmarking Results Using OCT Dataset

Method	**Processing time (h:m:s)**	**Acc (%)**	**Precision (%)**	**Recall (%)**	**Spe (%)**	**F1-score (%)**
(Alryalat et al., 2022)	1:31:40	89.34	86.98	85.55	87.04	86.25
(Li et al., 2022a)	01:10:55	83.77	84.04	83.11	85.00	83.57
(Wahab Sait, 2023)	01:02:40	85.67	86.00	84.97	87.09	85.47
(Paul and Talukder, 2023)	01:21:33	77.65	79.50	76.00	78.46	79.50
(Sharma and Guleria, 2023b)	1:00:34	80.40	82.73	79.03	81.20	80.83
(Shimpi and Shanmugam, 2023)	1:30:10	93.53	94.65	90.18	94.34	92.35
(Vetrithangam et al., 2023)	00:57:60	69.66	71.60	68.00	72.40	69.75
Proposed Method	**00:50:09**	**98.33**	**99.45**	**97.29**	**99.43**	**98.35**

Starting with (Alryalat et al., 2022), their model exhibits a respectable balance across all metrics with an F1-score of 86.25%. The precision and recall are fairly close, indicating a reasonable trade-off between the two. However, a significant processing time of over one and a half hours may limit practical applicability in time-sensitive environments. Their model appears to be reliable, but there may be a need for optimisation to improve efficiency where time could

be a bottleneck for time-sensitive applications.

The work in (Li et al., 2022a) shows a reasonable performance, but their metrics are outperformed by other methods. They also exhibit a shorter processing time than (Alryalat et al., 2022), which indicates better efficiency, but this comes at the cost of lower accuracy and an F1-score. Although their processing time is reduced, the trade-off for accuracy may not be justified in medical image analysis where precision is paramount. The work in (Wahab Sait, 2023) demonstrates an improved balance of accuracy, precision, and recall over research in (Li et al., 2022a), though still trailing behind the proposed method. Their model also achieves a notable decrease in processing time. The method proposed in (Paul and Talukder, 2023) records the lowest accuracy and F1-score, which raises concerns about the reliability of their model, especially in critical diagnostic settings. Their processing time does not compensate for the lower performance, suggesting a need for a substantial methodological review. Authors in (Sharma and Guleria, 2023a) proposed a method that offers a promising processing time, the best among the related works, which is commendable. Nevertheless, the corresponding performance metrics, while fair, are not competitive with the proposed method. The approach in (Shimpi and Shanmugam, 2023) shows outstanding results compared to the related works with high accuracy and an excellent F1-score, suggesting a robust model. Yet, they are still eclipsed by the proposed method in overall performance and slightly in processing time, hinting at potential inefficiencies in their computational approach. The work in (Vetrithangam et al., 2023) has the shortest processing time, which is impressive. However, this comes at the expense of all performance metrics, with their method showing the lowest accuracy and F1-score. This indicates a significant trade-off between processing time and performance, which might not be acceptable in critical applications.

The proposed method demonstrates superior performance across all metrics with an impressive Acc of 98.33% and an F1-score of 98.35%, suggesting an excellent balance between precision and recall. Moreover, the processing time is the shortest among all the methods, which not only makes it highly efficient but also suggests the use of advanced optimisation techniques. The critical review of the related works centers around the need to balance pro-

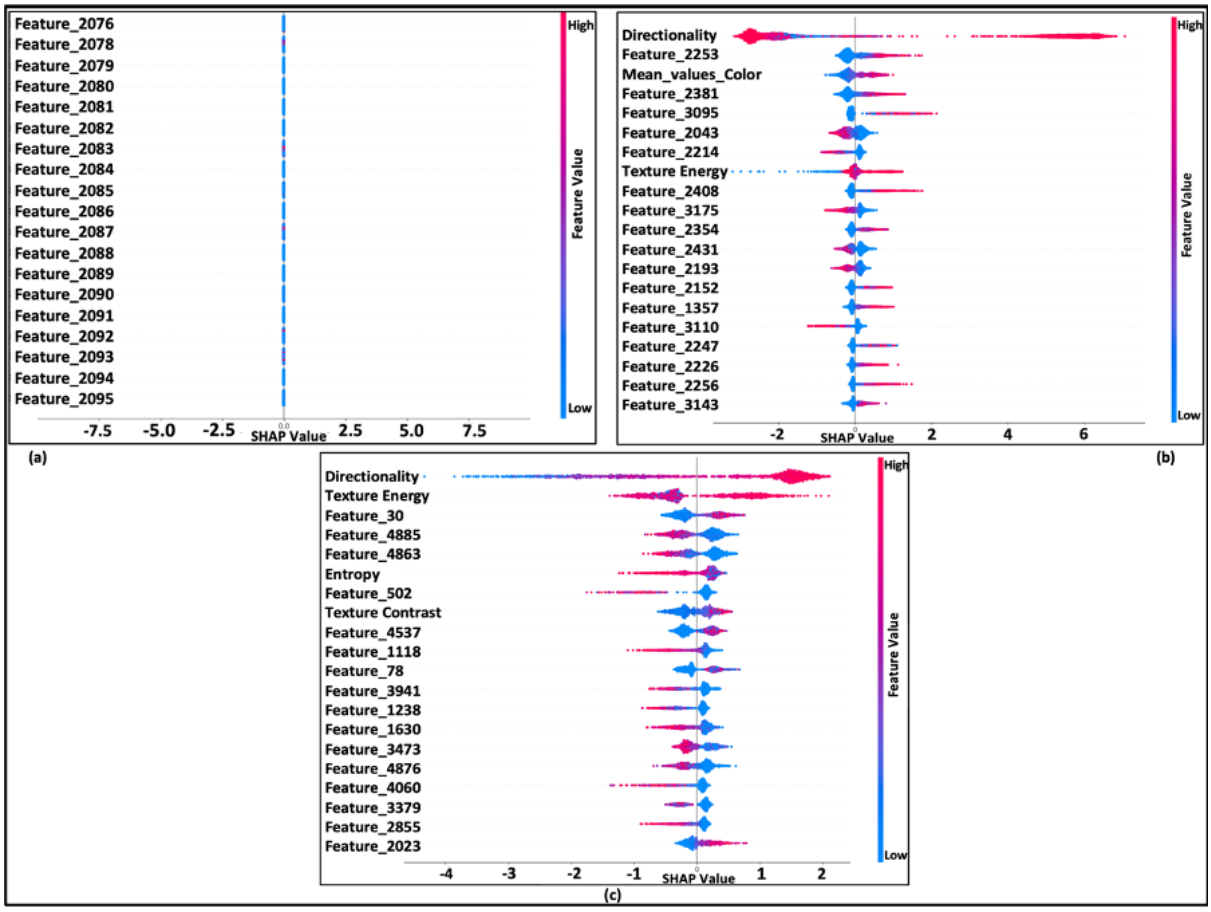


Figure 5.29: SHAP Values Explainer for (a): (Alryalat et al., 2022), (b): (Shimpi and Shanmugam, 2023), (c): Proposed Method.

cessing time with performance metrics—efficiency cannot come at the cost of effectiveness. Additionally, models must aim for a higher Spe without sacrificing Sen to be truly useful in varied operational scenarios. The proposed method sets a benchmark for both performance and efficiency, marking a significant advancement over the related works. It becomes a model example, showing that high accuracy and speed are achievable in concert.

To validate the obtained results analysis of SHAP values has taken place. Figure 5.29.a, c, and b show the SHAP values of related works (Alryalat et al., 2022), (Shimpi and Shanmugam, 2023), and the proposed method, respectively. Each figure shows the features importance of the first three folds for prediction.

As showing in Figure 5.28.a. in the first fold, Feature_2095 emerges as the primary driver with a predominantly positive influence on the model's output, as indicated by its largely pos-

itive SHAP values, suggesting that higher values of this feature push the model's prediction upward. Other features largely hover around a SHAP value of zero, implying minimal individual contribution to the model's predictions, with a uniform colour distribution indicating no dominant feature value affecting the impact. In the second fold, the scenario repeats with Feature_2095 maintaining its status as the most impactful feature, while other features show a negligible effect, although with a slight variability shown by the spread of SHAP values. The colour distribution remains consistent, showing no clear trend in feature values influencing the model. The third fold reiterates the pattern, with Feature_2095 again standing out for its positive impact, while the SHAP values for other features remain close to zero, reinforcing the minimal contribution narrative. The colour gradient holds steady, suggesting no direct correlation between the magnitude of feature values and their predictive impact.

Across all folds, the dominant influence of Feature_2095 positions it as a key predictor, while the lack of SHAP value variation for the other features points to their limited individual effect, raising the question of their overall importance or whether their potential effects are eclipsed by the predominant Feature_2095. The prediction model appears to have overlooked HF features, which might embody complex abstractions or data combinations potentially more predictive of outcomes, suggesting a missed opportunity for improving performance. This absence in significant SHAP value positions hints at a lack of comprehensive feature engineering that could have unveiled nuanced data patterns, possibly pointing to an overly simplistic model. Even with consistent SHAP value distributions across folds, the expected regularity in the significance of HF features is lacking, signifying underfitting. Moreover, the model's potential failure to capture interactions between features, particularly if HF features were designed to encapsulate such interactions, could limit its ability to exploit the data's full structure. The prominence of a single feature across folds raises concerns about bias and over-reliance on this feature, possibly compromising model robustness. Additionally, the model's generalisability could be questioned, as HF features are often key to adapting to new data variations. Lastly, the proximity of most features to zero in SHAP values suggests a potential redundancy or noise within the feature set, implying that the model might benefit from discarding these less infor-

mative features in favour of more impactful, high-level ones.

The analysis of the SHAP value plots reveals some critical insights into the model trained on OCT images by (Shimpi and Shanmugam, 2023). Directionality stands out as the most influential feature, exerting a strong positive influence on the model's predictions, as seen by the concentration of dots on the far right (Figure 5.29.b). This is followed by Feature_2253, which, although significant, displays more variability in its effect on the model's output, indicating that the model's response to this feature can vary. Mean_values_colour is also recognised as an important feature, but like Feature_2253, it shows a varied impact on the predictions. The spread of SHAP values across zero for each feature suggests that a feature's value could either increase or decrease the model output, with the specific effect dependent on the interplay with other features. Texture Energy, while having a less pronounced effect, shows a consistent and positive influence across all folds.

In terms of model understanding and diagnostics, it becomes evident that the model considers Directionality, Feature_2253, and Mean_values_colour as important features, as indicated by their prominence in the SHAP summary plot. The model demonstrates consistency in the impact of these features across different folds, which points to the stability of the model's behaviour and suggests that it is not overly sensitive to the specific data subset on which it is trained. Additionally, the varied distribution of SHAP values for certain features suggests complex interactions that could be influencing the model's predictions, highlighting the need for a nuanced understanding of how different features contribute to the model's decisions.

The reliance of the model on mean_values_colour for the OCT dataset raises concerns while the images are indeed grey-scale based. In such images, colour information, typically spread across RGB channels, should be non-existent, making colour a seemingly irrelevant feature for the model to focus on. This situation suggests that there might have been an oversight during the feature extraction and engineering phase; the colour feature should have been either removed or properly processed during the pre-processing stage. As a result, the model is likely to benefit from a feature selection approach that prioritises attributes more pertinent to grey-scale images, such as texture, edges, and contrast. This unexpected emphasis on colour also puts a

spotlight on the data pre-processing procedures, indicating that the colour feature warrants a closer review and potentially should be excluded from the dataset to ensure it does not introduce meaningless information. Additionally, this reliance on an ostensibly irrelevant feature could be a sign of overfitting, implying that the model might not generalise well to new, unseen data because it could be learning from noise instead of extracting significant patterns necessary for robust predictions.

The proposed model exhibits noteworthy consistency in the impact of the top features across all three folds, with 'Feature_30', 'Feature_4863', 'Feature_4885', and 'Feature_502' showing a high impact on the model output, suggesting that the method used for feature extraction and selection is stable and reliable (Figure 5.28.c). In the field of 'Directionality' and 'Texture Energy', the features, particularly 'Feature_30', have a strong positive SHAP value, implying they significantly contribute to the positive class predictions in the model, indicative of key characteristics in the OCT images relevant for the task at hand. The distribution of SHAP values indicates that the model is capturing a diverse array of effects from the features, with high-density regions in the figures denoting areas where features have a more uniform impact on the model output, and the sparser regions may represent more complex, non-linear relationships. The 'Texture Contrast' features show varied influence on model predictions, with some having positive and others negative SHAP values, reflecting a sophisticated model that accounts for different types of textures in OCT images, crucial for accurate diagnostics.

The clarity in the visualisation of SHAP values aids in understanding the decision-making process of the ML model, reinforcing the credibility and transparency of the analytical approach. Moreover, the method's ability to discern important features from an OCT dataset, which often contains complex and high-dimensional data, speaks to the effectiveness of the feature engineering and selection process utilised in the methodology. Overall, the results suggest that the model and methodology employed are capturing essential patterns and details in the data, vital for the accurate prediction of outcomes, and the consistency and depth of the SHAP analysis across all folds are particularly praiseworthy, demonstrating the reliability and potential clinical applicability of the approach in the analysis of OCT images.

5.10 Conclusion

The culmination of this research with the DenCeption feature extraction framework and the HyBoost hybrid predictive model has led to a significant advancement in the analytical capabilities for medical images interpretation, to include Fundus, OCT, and X-ray. The DenCeption framework has proved to be a robust method for feature extraction, ensuring consistent impact from the top features across different data folds, thereby affirming its stability and reliability. The extracted features have exhibited strong positive SHAP values, demonstrating their substantial contribution to the predictive prowess of the HyBoost model. The hybrid nature of HyBoost has allowed for the effective utilisation of various analytical strengths, capturing a wide spectrum of effects from the medical imaging datasets as evidenced by the distribution of prediction values. This diversity in feature influence is crucial for the nuanced understanding of those medical images, which are often characterised by complex textures and patterns. The positive and negative SHAP values for particular features, reflect the model's sophisticated handling of varying features, underlining its diagnostic precision.

Moreover, the clear visualisation of predictive values has not only facilitated a deeper understanding of the decision-making process inherent in the HyBoost model but has also strengthened its credibility and transparency. This interpretability is essential for clinical applications where practitioners' trust in automated systems is paramount. The effectiveness of DenCeption in high-dimensional data processing further underscores the utility of our feature engineering in refining the quality of inputs for superior outcomes. In conclusion, the DenCeption and HyBoost tandem have collectively shown promising results, capturing essential data patterns that are crucial for the accurate prediction of clinical outcomes based on different tested medical images types. The reliability and interpretability of the results, coupled with strong performance metrics, pave the way for their potential clinical applicability.

Chapter 6

Conclusion and Future Work

In the field of medical image processing, traditional techniques have traditionally been the foundation, although with limitations in automation, responsiveness, and reliability. These manual and semi-automated methodologies demand significant human intervention, often leading to inconsistencies in interpretation and analysis. As the medical field gravitates towards more automated solutions, ML and DL have emerged as promising avenues. However, these advanced techniques are not without their challenges. Interpretability, scalability, and adaptability remain critical hurdles, particularly given the complex nature of medical imaging data, which varies dramatically with each disease type. Furthermore, the lack of a standardised evaluation mechanism for ML and DL models in image processing exacerbates these challenges, coupled with the validation obstacles posed by the insufficiency of comprehensive datasets.

In response to these prevalent issues, the medical imaging community has witnessed a shift towards hybrid models that combine the strengths of various computational approaches to enhance classification and prediction capabilities. Yet, despite these advancements, the interpretability and scalability of such models remain in question. The alteration of algorithms and model architectures has not been paralleled by a comprehensive revision of the foundational framework supporting these models, leaving a significant gap in the pursuit of an effective solution.

6.1 Performance Overview of the Research Contributions

Addressing these gaps, the DenCeption model was designed as part of this thesis. DenCeption represents a significant advancement, transcending the limitations of existing ML and DL methodologies by offering a versatile and robust framework mainly tested and validated for medical image processing. The rigorous evaluation of DenCeption, through extensive training and testing phases, has uniquely demonstrated its superior performance across a multitude of metrics. Achieving an unprecedented Acc of 91.3%, DenCeption sets a new paradigm in the efficiency and effectiveness of hybrid models in tackling complex classification tasks. Its variants, including DenCeption-201, DenCeption-161, and DenCeption-121, further underscore the model's flexibility, each delivering commendable accuracies around 89%. In stark contrast, state-of-the-art models such as the ResNet-Inception lag significantly behind, with an Acc of only 73.4%.

One of the remarkable aspects of DenCeption is its nuanced Sen, especially evident in the DenCeption-201 variant, which boasts a 90% Sen rate. This adaptability in recognising positive instances, although slightly lower than DenCeption's peak Sen of 93%, is a testament to the model's refined predictive capabilities. Conversely, the DenCeption-HTB-NInC variant illustrates the criticality of integrating InC modules within the HTB structure, a modification that significantly enhances the model's ability to detect positive cases, underscoring the importance of architectural innovation in improving model performance.

Furthermore, DenCeption's precision rate of 94% and the highest F1-score of 93.4% highlight its unparalleled capacity in accurately identifying and classifying instances, reaffirming its status as a benchmark in medical image analysis. The precise fine-tuning of DenCeption's architecture, particularly the strategic incorporation of InA and InB modules within the HDB block, has been pivotal in optimising its feature extraction and representation capabilities.

The transition from training and testing to the validation phase marked a critical stage in this research, rigorously assessing DenCeption's generalisability, robustness, and practical applicability. This validation process, through a thorough examination of errors, computational

complexity, and scalability, was instrumental in establishing DenCeption's viability for real-world applications. This phase solidified DenCeption's role as a pivotal influence in the field of medical image analysis, positioned to address the complex challenges of disease classification with newfound precision and reliability.

The work conducted this thesis signifies a pivotal advancement in the domain of medical image processing, particularly addressing the critical aspect of features extraction. A thorough investigation into current methodologies unveiled a significant reliance on traditional techniques and a growing interest in automated ML and DL models. However, these approaches often fall short in terms of responsiveness, reliability, interpretability, scalability, and adaptability, especially when faced with the multifaceted nature of medical images. The essence of features, HF or DHF, plays a transformative role in the efficacy of classification and prediction models. Their quality, quantity, and nature not only influence a model's performance but also its ability to remain unbiased among diverse and sometimes imbalanced datasets.

Addressing these challenges, the research introduced a ground-breaking features extraction framework that inventively combines both HF and DHF. This framework, powered by the novel DenCeption model, signifies an advancement towards automating the feature extraction process, ensuring a high performance of medical analysis with an optimised set of features. The importance of this approach becomes even more pronounced in the context of manual features extraction – a task that demands considerable time and expertise, especially when dealing with complex grey-scale and coloured medical images like MRIs and Fundus images.

The DenCeption, GLCM, Tamura, RF, CHKM backed with MRF-EPM for segmentation framework emerged from a comprehensive analysis and rigorous testing across a spectrum of experiments, designed to evaluate the effectiveness of various features combinations. This thorough approach resulted an adaptive and scalable framework which is capable of selecting the most conducive features combination tailored to the specific requirements of the input dataset. Such an endeavour was not only theoretical; it was substantiated through experimentation evidence showcasing the framework's capability in handling both labelled and unlabelled datasets, achieving an impressive 97% Acc for the Texture-Shape-DHF combination and an

unprecedented 98.9% for the Texture-Shape-Colour-DHF combination.

These accomplishments underscore not only the technical relevancy of the proposed framework but also its practical implications in enhancing the responsiveness and reliability of medical image analysis. By minimising FPs and negatives, the framework promises a new era of precision in disease classification and diagnosis, potentially revolutionising the way medical imaging is approached.

This thesis, therefore, is presented as a testament to the potential of integrating HF and DHF in a consistent and optimised manner, leveraging the strengths of the DenCeption model to redefine the standards of features extraction in medical image processing. Delving into the contemporary landscape of medical research, it becomes apparent that DL's role in diagnosing and predicting diseases is both significant and expanding. Yet, despite the promising advancements, the reviewed literature highlights areas requiring further exploration and refinement. The application of advanced ML and DL methodologies in understanding complex medical imagery has indeed revolutionised diagnostic precision, enriching our comprehension of diseases and tailoring treatment pathways more effectively.

DL and ML methodologies are increasingly recognised for their superiority over traditional diagnostic methods, heralding a new epoch in medical science. Nonetheless, the highly achieved accuracy rates presented by numerous studies require a critical and deeper examination. accuracy metric, while important, occasionally obscures underlying model biases or overfitting, particularly with datasets lacking in diversity.

The discussion around model reproducibility and scalability underscores a critical concern presented by the presumption that these models will exhibit similar efficacy across heterogeneous medical contexts is optimistic. However, real-world healthcare settings challenged with disparities in technology, patient demographics, and image quality, would potentially decrease a model's performance outside controlled research environments.

Furthermore, an overemphasis on technical capabilities may dominate essential aspects of diagnostics like the clarity of a model's rationale and its pertinence to clinical settings. The adoption of models in clinical practice is contingent not just on their diagnostic accuracy but

also on their explainability, without which clinicians may hesitate to rely on them for critical health decisions. Additionally, the need for advanced model architectures must also balance with practical concerns of sustainability and computational requirements. In fcat, advanced DL models often require substantial computational resources which is inconvenient in case of under-resourced settings.

To address these critical challenges, this thesis introduces a novel predictive framework integrating the newly designed HyBoost hybrid model. This innovative model employs the strengths of AdaBoost and XGBoost within a comprehensive feature extraction framework mainly through the DenCeption model. This integration has resulted a substantial increase in medical image analysis, extending its capability to encompass diverse imaging types such as Fundus, OCT, and X-ray images. The DenCeption framework distinguishes itself as an adaptive feature extraction mechanism, ensuring the pivotal features' consistent influence across varied data scenarios, thereby certifying its robustness and dependability. The identified features, showcasing strong positive SHAP values, highlight their essential contribution to enhancing the predictive Acc of the HyBoost model.

The hybrid nature of HyBoost enables the combination of distinct analytical strengths. This feature diversity is pivotal in deconstructing the complex textures and patterns within medical images for prediction purposes. The prediction framework's precision highlights the efficient management of features usage. Furthermore, the transparent representation of predictive values not only deepens the understanding of HyBoost's decision-making but also enhances its scalability and adaptability. Additionally, the introduction of the PMM matrix as a novel tool for performance evaluation marks a strategic advance, customising the assessment of the models to the nuanced requirements of specific medical imaging data and diagnostic tasks in question.

6.2 Research Contributions Against Chosen Datasets

This section provides a comprehensive comparison of the performance of all contributions across the chosen datasets (summary provided in Table 6.1), offering a high-level overview of

the contributions while highlighting their strengths and areas for improvement. Each contribution addresses specific gaps in medical image processing, and their combined impact offers a holistic solution to the challenges faced in this field.

6.2.1 Contribution 1: Design of a Novel DL-Based Hybrid Model (DenCeption) Against MRI Dataset

The BRATS MRI dataset, known for its complexity in brain tumour segmentation, served as a robust testing ground for DenCeption model as part the first contribution. DenCeption demonstrated high performance, achieving superior accuracy (91.3%), sensitivity (93%), specificity (93.7%), precision (94%), F1-score (93.4%), and a low MAE (0.2), indicating its effectiveness in handling the variability inherent in brain tumour images.

- **Positive Influence:** The complexity and unlabelling of the BRATS dataset provided a rigorous environment for testing DenCeption, allowing it to showcase its advanced feature extraction capabilities and robustness. The inclusion of patient-specific data, such as age and resection status, further validated the model's relevance in real-world clinical settings.

- **Negative Influence:** Despite these positive outcomes, the dataset's focus on a single medical condition (brain tumours) may limit the generalisability of the results. The improvements observed, while significant, are incremental when compared to existing models like DenseNet-121. This suggests that while DenCeption offers advantages, particularly in complex medical image classification, further testing on diverse datasets is necessary to fully exploit its potential.

- **Critical Insight:** The development of DenCeption addresses key gaps in feature extraction and model robustness. However, its generalisability to other medical imaging modalities remains untested at Chapter 3 stage. Hence, Future research done in Chapter 4 and 5 involveed validating DenCeption across a broader range of datasets to confirm its applicability in diverse medical conditions.

6.2.2 Design of an Adaptive and Scalable Features Extraction Framework Against MRI and Retinal Datasets

The second contribution focuses on a hybrid feature extraction framework that combines HF and DHF features to improve medical image classification accuracy and reliability. This framework was validated using two datasets: MRI BRATS and Retinal, demonstrating notable performance improvements across all metrics. The results reveal that the framework outperforms traditional methods, achieving high accuracy (up to 98.9% in the Retinal dataset, Case 4) and low MAE (0.01), while dynamically adjusting to different imaging conditions and data complexities.

- **Positive Influence:** The diverse nature of these datasets underscores the versatility of the proposed framework. The MRI BRATS dataset tested its ability to handle complex 3D data, while the Retinal dataset evaluated its performance in detecting retinal diseases. The adaptability and scalability of the framework were clearly demonstrated, ensuring its reliability across different medical imaging scenarios.

- **Negative Influence:** Despite the promising results, the Retinal dataset's relatively small size (1,000 images) may affect the reliability of the findings. The framework's performance might be overestimated due to the limited variety of the Retinal dataset compared to larger datasets. Additionally, the increased processing time observed in certain cases (e.g., 14:10:00 in Case 4 - DHF for the BRATS dataset) may limit its real-time applicability, a crucial factor in clinical settings where timely diagnosis is essential.

- **Critical Insight:** The framework's adaptability across both MRI and Retinal datasets is promising, but the results should be interpreted with caution, particularly given the small size of the Retinal dataset. Expanding the testing to include larger and more diverse retinal datasets would provide a more comprehensive evaluation of the framework's effectiveness. Hence, a larger Fundus dataset was considered in Chapter 5. The trade-off between accuracy and computational efficiency remains a key consideration, especially

in clinical environments.

6.2.3 Design of a Novel Evaluation Mechanism for DL Models Against all Datasets

This contribution introduces a novel evaluation mechanism designed to enhance the reliability and validity of DL model assessments, particularly in the medical imaging domain. The mechanism incorporates correlation operations and random weight assignments to provide a systematic and objective approach to selecting evaluation metrics.

- **Positive Influence:** The mechanism's validation on diverse datasets (BRATS MRI, Retinal, Fundus, OCT, X-ray) highlights its potential in tailoring evaluation metrics to specific problem domains. This approach ensures that the most relevant metrics are identified, leading to more accurate assessments of model performance.
- **Negative Influence:** The generalisability of the evaluation mechanism depends on the diversity and quality of the datasets used. If the datasets are not representative of the wide range of medical imaging scenarios, the mechanism's effectiveness could be compromised.
- **Critical Insight:** The evaluation mechanism is a significant contribution to medical imaging, where accurate model assessment is crucial. However, further testing across more diverse datasets and application areas is necessary to fully validate its effectiveness. While the mechanism provides a structured approach to evaluating DL models, additional empirical evidence across various domains would strengthen the claim of its broad applicability.

6.2.4 Design of an Intelligent and Robust Predictive Framework Against Fundus, OCT and X-ray Datasets

This proposed predictive framework addresses the limitations of traditional predictive models by incorporating DenCeption, HyBoost, as well as vital patient demographic and physiological data, leading to significant performance improvements across various datasets, including Fundus, OCT, and X-ray.

- **Positive Influence:** The large size and diversity of the Fundus (18,615 images) and OCT (25,197 images) datasets provide a robust evaluation platform for the predictive framework. The inclusion of demographic and physiological data further enhances the model's predictive accuracy, making it highly relevant for real-world clinical applications. The results indicate that HyBoost outperforms existing models with accuracy rates reaching 98.33% for OCT and 98.2% for X-ray scans.

- **Negative Influence:** The X-ray dataset, while useful for extending the research into pulmonology, is relatively small (5,467 images), which might limit the reliability of the results. The framework's performance on X-ray data may not be as robust as on the larger Fundus and OCT datasets. Additionally, the increased complexity of the model may present challenges in terms of interpretability and computational efficiency.

- **Critical Insight:** The predictive framework shows significant promise across a range of datasets, particularly in ophthalmology and pulmonology. However, the varying sizes and complexities of the datasets mean that the results should be interpreted with some caution. Further testing on larger and more diverse X-ray datasets would provide a more complete picture of the framework's generalisability. The use of SHAP explainability analysis is a strong point, as it provides transparency in the model's decision-making process, which is crucial for practical adoption in clinical settings.

The diverse range of datasets used in this research, from the complex BRATS MRI dataset to the more varied Fundus and OCT datasets, demonstrates the adaptability and scalability of

the proposed models and frameworks. However, the varying sizes and complexities of these datasets introduce challenges which were considered when interpreting the results in all chapters of this research. By integrating these contributions, the thesis provides a robust framework for advancing medical image processing, with each contribution complementing the others to address the multifaceted challenges in this field. The performance metrics and detailed analysis presented offer significant advancements in medical diagnostics, while also recognising the areas where further validation and refinement are needed.

6.3 Limitations and Challenges

Throughout this thesis, several significant challenges were encountered, each posing a unique barrier to the research progress and the practical application of the findings. These challenges reflect broader issues within the field of DL and medical image processing, underscoring the complex compromise between technological capabilities, data availability, and expert validation in advancing healthcare innovations.

Firstly, the need for powerful hardware to conduct image training processes cannot be overstated. DL models, particularly those designed for medical image analysis, require substantial computational resources to process and learn from large datasets. The limitation in hardware capabilities directly impacted the scope and speed of this research. High-performance GPUs are essential for training DL models efficiently. However, the accessibility and cost associated with such advanced hardware often pose significant problems, potentially impacting research progress and limiting the complexity of models that can be explored.

Secondly, the availability of clinical medical data represents a critical challenge. Access to diverse and extensive clinical datasets is crucial for training robust models capable of generalising well to real-world scenarios. However, ethical considerations, privacy concerns, and logistical issues often restrict the availability of such data. Most clinical datasets are guarded due to patient confidentiality agreements, making it difficult for researchers to obtain varied and representative samples. This research relied on publicly available medical imaging datasets, which,

while invaluable, may not fully capture the diversity and complexity of real-world clinical data. The discrepancy between publicly available datasets and the various conditions encountered in clinical practice limits the validation and applicability of the proposed frameworks and models, potentially affecting their performance in real-world settings. This limitation also raises concerns about the generalisability of the models when deployed in diverse clinical environments, where patient demographics, imaging equipment, and protocols vary significantly.

Lastly, the absence of expert validation for the proposed framework poses a significant limitation. Expert insights, especially from medical professionals familiar with the nuances of disease diagnosis and medical imaging, are crucial for validating and refining computational models. Without validation from clinicians, the clinical relevance and trustworthiness of the models remain uncertain. This lack of validation could hinder the adoption of the models in real-world clinical settings, where trust in the technology is paramount. Moreover, the challenge of data availability indirectly impacted the possibility of expert validation, as clinicians typically require access to comprehensive and representative datasets to provide meaningful feedback. Despite the accurate results provided by the proposed solution, further accuracy and reliability can only be endorsed by medical professionals in regular clinical settings, where the models would need to demonstrate consistent performance across a range of real-world conditions.

6.4 Future Work

The future directions outlined in this thesis underscore a strategic roadmap towards refining and actualising the research into practical, impactful applications within the medical field. Each step is carefully designed to bridge the gap between theoretical models and their deployment in healthcare environments, focusing on enhancing the reliability and effectiveness of medical image analysis through DL.

The first initiative involves the continuation and expansion of the collaboration with the NHS Trust Gloucestershire research group. This partnership is pivotal for several reasons. Ac-

cess to a broader and more varied range of medical imaging data, specifically Fundus and OCT datasets, is crucial for the advancement of this research. The related diseases will then be the main focus of future works in order to further enhance the RASR related criteria of the proposed solution. Such datasets are invaluable for training more robust and accurate models by exposing them to a wider array of pathological conditions and imaging variances. The collaboration aims not only to secure these datasets but also to encourage an exchange of knowledge and expertise that can drive the refinement of the proposed DL frameworks. This collaboration is essential for addressing the limitation of expert validation, as it would provide the opportunity for medical professionals to rigorously test the models, offer feedback, and validate their effectiveness in real-world clinical scenarios. This would not only enhance the models' reliability and clinical relevance but also build trust among clinicians who are critical to the successful integration of these technologies into everyday medical practice.

The validation of the proposed solutions on these acquired datasets represents the next critical step. Validation is essential for assessing the models' performance and generalisability to real-world clinical scenarios. It involves a rigorous examination of the models' diagnostic accuracy, Sen, Spe, and other relevant metrics against a clinically sourced data. This process will address the current limitation of not being validated by any medical expert or in real clinical settings. By engaging clinicians in the validation process, the models can be refined to meet the practical needs of healthcare providers, ensuring that they are not only technically sound but also applicable and trustworthy in clinical environments. This iterative process of validation and refinement is fundamental to achieving a solution that is both scientifically robust and clinically relevant.

Finally, the deployment of the proposed framework in real-world clinical settings is the ultimate goal of this research. Transitioning from research prototypes to operational medical tools involves navigating a complex landscape of regulatory compliance, ethical considerations, and integration challenges. The deployment would require a comprehensive evaluation of the framework's compatibility with existing medical IT infrastructures, its adaptability to different clinical workflows, and its usability for healthcare professionals. Achieving successful deploy-

ment would also necessitate ongoing collaboration with clinicians and healthcare organisations to ensure that the technology meets the needs of both patients and providers. By addressing the current limitations through real-world testing and expert validation, this research aims to develop models that are not only accurate and efficient but also practical and reliable in diverse clinical settings.

Table 6.1: Summary Table of Contributions Performance Against Chosen Datasets

Contribution	**Dataset**	**Size**	**Modality**	**Acc (%)**	**Sen (%)**	**Spe (%)**	**Precision (%)**	**F1-Score (%)**	**MAE**	**Processing Time**
1: DenCeption	BRATS MRI	8,000	MRI	91.3	93	93.7	94	93.4	0.2	3:40:00
2: Feature Extraction	MRI BRATS & Retinal	8,000 (BRATS), 1,000 (Retinal)	MRI, Fundus	97 (BRATS Case 1-DHF) & 98.9 (Retinal Case 4-DHF)	98 (BRATS Case 1-DHF) & 99 (Retinal Case 4-DHF)	96 (BRATS Case 1-DHF) & 98 (Retinal Case 4-DHF)	-	-	0.02 (BRATS Case 1-DHF) & 0.01 (Retinal Case 4-DHF)	10:30:00 (BRATS Case 1-DHF) & 5:13:00 (Retinal Case 4-DHF)
3: Evaluation Mechanism	Multiple	N/A	MRI, Fundus, OCT, X-ray	N/A	N/A	N/A	N/A	N/A	N/A	N/A
4: Predictive Framework	Fundus, OCT, X-ray	18,615 (Fundus), 25,197 (OCT), 5,467 (X-ray)	Fundus, OCT, X-ray	96.66 (Fundus) & 98.33 (OCT) & 98.2 (X-ray)	-	97.73 (Fundus) & 99.43 (OCT) & 98.1 (X-ray)	97.78 (Fundus) & 99.45 (OCT) & 98.1 (X-ray)	97.12 (Fundus) & 98.35 (OCT) & 98.3 (X-ray)	-	-

Appendix A

Appendices: Algorithms

I declare that these algorithms are my own work.

A.1 Algorithm 1 - Texture Features Extraction

Algorithm 5: Texture Features Extraction

Data: RoIs (Region of Interests)

Result: $f_{ASM}, f_E, f_C, f_H, f_{Coa_{Z,BEST}}, f_{Dir}$

$n \Leftarrow$ number of levels
$M \Leftarrow n^2 > 0$
for $i \in \{0,\ldots,n\}$ **do**
  for $j \in \{0,\ldots,n\}$ **do**
    for $k \in \{0,\ldots,M\}$ **do**
      $GLCM_f(i,j) \Leftarrow \frac{2}{M}\sum occ(i,j)$
      $f_{ASM} \Leftarrow \sum\sum GLCM_f(i,j)^2$
      $f_E \Leftarrow -\sum\sum GLCM_f(i,j) * \log(GLCM_f(i,j))$
      $f_C \Leftarrow \sum\sum (i,j)^2 * GLCM_f(i,j)$
      $f_H \Leftarrow \sum\sum \frac{GLCM_f(i,j)}{1+|j-i|}$
$pix(i,j) \Leftarrow$ intensity value of the pixel at location (i,j)
$S_Z \Leftarrow 2^{2Z}$ where $Z \in [0:5]$
$N \Leftarrow$ normalisation factor
$\theta \Leftarrow$ quantisation angular position
$m \Leftarrow$ number of peaks
$\psi_k \Leftarrow$ angles window associated with the k^{th} peak.
$H_{Dir} \Leftarrow$ edge histogram
$M_w \Leftarrow$ measurement window
for $i,j \in \{0,\ldots,n\}$ **do**
  for $k = i - 2^{Z-1} - 1$ *to* $i + 2^{Z-1}$ **do**
    for $k = j - 2^{Z-1} - 1$ *to* $j + 2^{Z-1}$ **do**
      $Coa_Z \Leftarrow \sum \frac{pix(i,j)}{M_w}$
$A_{Z,V}(i,j) \Leftarrow \mid Coa_{Z,V}(i, j+2^{Z-1}) - Coa_{Z,V}(i, j-2^{Z-1}) \mid$
$A_{Z,H}(i,j) \Leftarrow \mid Coa_{Z,H}(i+2^{Z-1}, j) - Coa_{Z,H}(i-2^{Z-1}, j) \mid$
if $A_{Z,V}(i,j) > A_{H,V}(i,j)$ **then**
  $S_{Z,BEST}(i,j) \Leftarrow S_{Z,V}$
  $f_{Coa_{Z,V}} \Leftarrow \frac{Coa_{Z,V}}{S_{Z,BEST}}$
else
  $S_{Z,BEST}(i,j) \Leftarrow S_{Z,H}$
  $f_{Coa_{Z,H}} \Leftarrow \frac{Coa_{Z,H}}{S_{Z,BEST}}$
for $k = 1$ *to* m **do**
  for *each angle* $\theta \in \psi_k$ **do**
    $\rho \Leftarrow \sum(\theta - \theta_k)^2 * H_{Dir}(\theta)$
$f_{Dir} \Leftarrow 1 - N \cdot m \cdot \rho$

A.2 Algorithm 2 - Shape and Colour Features Extraction

Algorithm 6: Shape and Colour Features Extraction

Data: RoIs (Region of Interests)

Result: f_{RF}, f_{CHKM}

Step1: Calculation of Region focus shape features based on the region area A

$(x,y) \Leftarrow$ coordinates of the pixel $\in$ A

for $(x,y) \in RoIs$ **do**

    $A \Leftarrow \sum 1$

for $(x,y) \in RoIs$ **do**

    $\bar{x} \Leftarrow \frac{\sum x}{A}$

    $\bar{y} \Leftarrow \frac{\sum y}{A}$

$f_{RF} \Leftarrow (\bar{x}, \bar{y})$

Step2: Calculation of colour histogram of K-mean where K represents the number of clusters

$N \Leftarrow$ total number of pixels

$N_k \Leftarrow$ total number of pixels in cluster k

$c_p \in 2^{24}$ colours possibilities

$c_{pixel} \Leftarrow$ current pixel colour

for $c_{pixel} \in cluter\ k \in K$ **do**

    if $c_{pixel} \in 2^{24}$ **then**

        $c_{pixel} \Leftarrow$ best matching colour (c_p)

$f_{CHKM} \Leftarrow \frac{N_K}{N}$

A.3 Algorithm 3 - Optimal Performance Evaluation Metrics

Algorithm 7: Optimal Performance Evaluation Metrics

Input: $C_{(x_1,x_2)}, C_{(x_1,x_3)}, C_{(x_2,x_3)}$: set of evaluation metrics for each dimension
Output: $PMM_{optimal}$

$W \Leftarrow$ weight assignment function
$n \Leftarrow$ total number of metrics
$c \Leftarrow$ performance metric
$x_1 \Leftarrow$ problem specification
$x_2 \Leftarrow$ task identification
$x_3 \Leftarrow$ data characteristics
$C_{(x_1,x_2)} \Leftarrow [c_{121}, c_{122}, ..., c_{12n}]$: set of performance metrics for dimension (x_1,x_2)
$C_{(x_1,x_3)} \Leftarrow [c_{131}, c_{132}, ..., c_{13n}]$: set of performance metrics for dimension (x_1,x_3)
$C_{(x_2,x_3)} \Leftarrow [c_{231}, c_{232}, ..., c_{23n}]$: set of performance metrics for dimension (x_2,x_3)
$D_{12}, D_{13}, D_{23} \Leftarrow$ 2D matrix of (x_1,x_2), (x_1,x_3), and (x_2,x_3) dimensions, respectively
Step 1:
for $i \in \{1,\ldots,n\}$ **do**
    $W_{C_{12}}[i] \Leftarrow W(C_{12}[i])$
Step 2:
for $i \in \{1,\ldots,n\}$ **do**
    for $j \in \{1,\ldots,n\}$ **do**
        if $i == j$ **then**
            $W_{C_{12ii}} \Leftarrow corr_{coef}(C_{12i}, C_{12i})$
            $D_{12}[i,i] \Leftarrow W_{C_{12ii}}$
        else
            $W_{C_{12ij}} \Leftarrow corr_{coef}(C_{12i}, C_{12j})$
            $D_{12}[i,j] \Leftarrow W_{C_{12ij}}$
Repeat Step 1 and 2 for D_{13} and D_{23}
for $i \in \{1,\ldots,n\}$ **do**
    for $j \in \{1,\ldots,n\}$ **do**
        if $i == j$ **then**
            $PMM[i,j] \Leftarrow max(D_{12}[i,i], D_{13}[i,i], D_{23}[i,i])$
        else
            $PMM[i,j] \Leftarrow max(D_{12}[i,j], D_{13}[i,j], D_{23}[i,j])$
for $i \in \{1,\ldots,n\}$ **do**
    for $j \in \{1,\ldots,n\}$ **do**
        if $PMM[i,j] > 0$ **then**
            $PMM_{optimal}[i] \Leftarrow PMM[i,j]$
        else
            Continue
return $PMM_{optimal}$

A.4 Algorithm 4 - Algorithm for HyBoost Hybrid Predictive Model

Algorithm 8: Algorithm for HyBoost Hybrid Predictive Model

Data: $D_{train} = \{(x_i, y_i)\}_{i=1}^{N}$, $D_{test} = \{x_j\}_{j=1}^{M}$, M_{xgb} (XGBoost model), M_{ab} (AdaBoost model), H_{xgb}, H_{ab}, K, T, $\alpha \in [0,1]$

Result: Y^*_{final} (final best prediction)

Initialisation:

$\hat{y}_{xgb} \Leftarrow M_{xgb}$'s residuals (predictions)

$\hat{y}_{ab} \Leftarrow M_{ab}$'s predictions

$Y_{final} \Leftarrow$ final prediction prior to optimisation

Step 1: HyBoost Training Phase

Step 1.1: Train XGBoost

for $k \in \{1, \ldots, K\}$ **do**

$Ob_i \Leftarrow \sum_{i=1}^{N} l\left(y_i, \hat{y}_i^{(k-1)} + f_k(x_i) + \Omega(f_k)\right)$

$g_i \Leftarrow \frac{\partial}{\partial \hat{y}_i^{(k-1)}} l\left(y_i, \hat{y}_i^{(k-1)}\right)$

$h_i \Leftarrow \frac{\partial^2}{\partial (\hat{y}_i^{(k-1)})^2} l\left(y_i, \hat{y}_i^{(k-1)}\right)$

$\hat{y}_i^{(k)} \Leftarrow \hat{y}_i^{(k-1)} + \eta f_k(x_i)$

$M_{xgb}(x_i) \Leftarrow XGBoost(D_{train}, H_{xgb})$

residuals $\Leftarrow y_i - \hat{y}_{xgb}$

Step 1.2: Train AdaBoost

for $t \in \{1, \ldots, T\}$ **do**

$\alpha_t \Leftarrow \frac{1}{2} \ln\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right)$

$w_{i,t+1} \Leftarrow w_{i,t} * \exp(-\alpha_t y_i h_t(x_i))$

$M_{ab}(x_i) \Leftarrow AdaBoost(D_{train}, \text{residuals}, H_{ab})$

Step 2: HyBoost Prediction Phase

for $j \in \{1, \ldots, N\}$ **do**

$\hat{y}_{final_j} \Leftarrow \alpha * \hat{y}_{xgb_j} + (1-\alpha) * \hat{y}_{ab_j}$

Step 3: Tuning and Optimisation

Optimise $(H_{xgb}, H_{ab}, \alpha)$ with cross-validation

Step 4: Optimal Parameters Selection

$H^*_{xgb}, H^*_{ab}, \alpha^* \Leftarrow \arg\max_{H_{xgb}, H_{ab}, \alpha} \text{CrossValidationScore}(D_{train}, H_{xgb}, H_{ab}, \alpha)$

Step 5: Re-training Phase

for $j \in \{1, \ldots, M\}$ **do**

$Y^*_{final_j} \Leftarrow \alpha^* * M^*_{xgb}(x_j) + (1-\alpha) * M^*_{ab}(x_j)$

Appendix B

Appendices: Code

B.1 Pre-processing and Segmentation - Features Extraction Framework

```
import numpy as np
import cv2
from skimage.restoration import estimate_sigma
import bm3d
from SimpleITK import N4BiasFieldCorrection, GetArrayFromImage,
    GetImageFromArray
from sklearn.mixture import GaussianMixture

class PreprocessAndSegment:
    def __init__(self, image_path, ground_truth_path):
        self.image_path = image_path
        self.ground_truth_path = ground_truth_path

    def ground_truth_extraction(self):
        image = cv2.imread(self.image_path)
        ground_truth = cv2.imread(self.ground_truth_path,
            cv2.IMREAD_GRAYSCALE)
```

```
        _, binary_mask = cv2.threshold(ground_truth, 127, 255,
            cv2.THRESH_BINARY)
        binary_mask = binary_mask // 255
        return binary_mask

    def image_denoising(self, image):
        sigma_est = np.mean(estimate_sigma(image, multichannel=True))
        denoised_image = bm3d.bm3d(image, sigma_est)
        return denoised_image

    def bias_field_correction(self, image):
        sitk_image = GetImageFromArray(image)
        corrector = N4BiasFieldCorrection()
        corrected_image = corrector.Execute(sitk_image)
        corrected_image = GetArrayFromImage(corrected_image)
        return corrected_image

    def mrf_em_segmentation(self, image):
        gray_image = cv2.cvtColor(image, cv2.COLOR_BGR2GRAY)
        gmm = GaussianMixture(n_components=2, covariance_type='tied',
            max_iter=100, random_state=42)
        reshaped_image = gray_image.reshape((-1, 1))
        gmm.fit(reshaped_image)
        em_segmented =
            gmm.predict(reshaped_image).reshape(gray_image.shape)

        mask = np.zeros(gray_image.shape, np.uint8)
        mask[em_segmented == 1] = 1
        bgdModel = np.zeros((1, 65), np.float64)
        fgdModel = np.zeros((1, 65), np.float64)
        mask, bgdModel, fgdModel = cv2.grabCut(image, mask, None,
            bgdModel, fgdModel, 5, cv2.GC_INIT_WITH_MASK)

```

```
        mrf_segmented = np.where((mask == 2) | (mask == 0), 0,
            1).astype('uint8')
        return mrf_segmented

    def execute(self):
        ground_truth_mask = self.ground_truth_extraction()
        image = cv2.imread(self.image_path)
        denoised_image = self.image_denoising(image)
        corrected_image = self.bias_field_correction(denoised_image)
        segmented_image = self.mrf_em_segmentation(corrected_image)

        return segmented_image, ground_truth_mask
```

B.2 High Level Features Extraction: Texture, Shape, Colour Features

```
import numpy as np
import cv2
from sklearn.cluster import KMeans

class HighLevelFeaturesExtractor:
    def __init__(self, n_levels, n_clusters):
        self.n_levels = n_levels
        self.n_clusters = n_clusters

    def texture_features_extraction(self, roi):
        n = self.n_levels  # Number of levels
        M = n ** 2  # Number of gray levels in the image

        # Initialize matrices to store feature calculations
        GLCMf = np.zeros((n, n))
        fASM = fE = fC = fH = 0
```

```
        # Step 1: Compute GLCM and derive texture features
        for i in range(n):
            for j in range(n):
                for k in range(M):
                    occ = np.sum(roi == k)
                    GLCMf[i, j] = (2 / M) * occ  # GLCM calculation

                # Feature calculations
                fASM += GLCMf[i, j] ** 2
                fE += GLCMf[i, j] * np.log(GLCMf[i, j] + 1e-10)  # Add
                    small constant to avoid log(0)
                fC += (GLCMf[i, j] ** 2) * GLCMf[i, j]
                fH += GLCMf[i, j] / (1 + abs(i - j))

        # Initialize other required variables
        SZ = [2 ** z for z in range(6)]
        SZ_BEST = np.zeros_like(roi)
        fCoaZ_BEST = 0

        # Step 2: Compute coarseness (CoaZ)
        pix = roi  # Assuming roi is the pixel intensity map
        m, n = roi.shape

        # Initialize arrays for directional coarseness
        CoaZ_V = np.zeros_like(roi)
        CoaZ_H = np.zeros_like(roi)
        AZ_V = np.zeros_like(roi)
        AZ_H = np.zeros_like(roi)

        for i in range(m):
            for j in range(n):
                for k in range(1, len(SZ)):
```

```
                    CoaZ_V[i, j] = np.sum(pix[max(0,
                        i-SZ[k]//2):min(m, i+SZ[k]//2), j]) / SZ[k]
                    CoaZ_H[i, j] = np.sum(pix[i, max(0,
                        j-SZ[k]//2):min(n, j+SZ[k]//2)]) / SZ[k]

                    AZ_V[i, j] = np.abs(CoaZ_V[i, j] -
                        CoaZ_V[i+SZ[k]//2, j]) if i+SZ[k]//2 < m else 0
                    AZ_H[i, j] = np.abs(CoaZ_H[i, j] - CoaZ_H[i,
                        j+SZ[k]//2]) if j+SZ[k]//2 < n else 0

                    if AZ_V[i, j] > AZ_H[i, j]:
                        SZ_BEST[i, j] = CoaZ_V[i, j]
                        fCoaZ_V = CoaZ_V[i, j] / SZ_BEST[i, j]
                    else:
                        SZ_BEST[i, j] = CoaZ_H[i, j]
                        fCoaZ_H = CoaZ_H[i, j] / SZ_BEST[i, j]

        # Step 3: Compute directional features
        fDir = 0
        N = 1  # Normalization factor
        m = 1  # Number of peaks, assuming m = 1 for simplicity here
        theta = 0  # Assuming an initial angle theta = 0, can be
            adjusted based on the specific problem

        for k in range(1, m+1):
            for angle in range(0, 180, 45):  # Adjust angles as needed
                rho = np.sum((theta - angle) ** 2) * fDir
                fDir = 1 - N * m * rho

        return np.array([fASM, fE, fC, fH, fCoaZ_BEST, fDir])

    def shape_and_colour_features_extraction(self, rois):
        # Step 1: Calculate Region Focus Shape Features
        A = len(rois)  # Area of the region (number of pixels)
```

```
        x_sum = y_sum = 0

        for (x, y) in rois:
            x_sum += x
            y_sum += y

        x_bar = x_sum / A  # Centroid x
        y_bar = y_sum / A  # Centroid

        f_RF = np.array([x_bar, y_bar])

        # Step 2: Calculate Colour Histogram using K-means Clustering
        N = len(rois)  # Total number of pixels

        # Assuming each pixel in `rois` has an associated color in RGB
            format (3 channels)
        colors = np.array([roi[2] for roi in rois])  # Extracting
            color information

        # Applying K-means clustering on the color data
        kmeans = KMeans(n_clusters=self.n_clusters, random_state=42)
        kmeans.fit(colors)
        labels = kmeans.labels_

        # Calculate f_CHKM, the proportion of pixels in each cluster
        f_CHKM = np.zeros(self.n_clusters)

        for k in range(self.n_clusters):
            Nk = np.sum(labels == k)  # Number of pixels in cluster k
            f_CHKM[k] = Nk / N  # Proportion of pixels in this cluster

        return f_RF, f_CHKM

    def extract(self, segmented_image):
```

```
        texture_features =
            self.texture_features_extraction(segmented_image)
        shape_features, color_features =
            self.shape_and_colour_features_extraction(segmented_image)

        high_level_features = np.concatenate((texture_features,
            shape_features, color_features))
        return high_level_features
```

B.3 Deep Hidden Features Extraction: DenCeption Model

```
import tensorflow as tf
from tensorflow.keras.layers import Conv2D, MaxPooling2D,
   AveragePooling2D, GlobalAveragePooling2D, Dense, Input, Concatenate
from tensorflow.keras.models import Model

class LowLevelFeaturesExtractor:
    def __init__(self, input_shape, classes):
        self.model = self.build_denception(input_shape, classes)

    def build_denception(self, input_shape, classes):
        # Define the InA module with dynamic filters
        def InA_module(x, n, l, t, k):
            branch1x1 = Conv2D(n, (1, 1), padding='same')(x)

            branch3x3 = Conv2D(l, (1, 1), padding='same')(x)
            branch3x3 = Conv2D(l, (3, 3), padding='same')(branch3x3)

            branch3x3dbl = Conv2D(t, (1, 1), padding='same')(x)
            branch3x3dbl = Conv2D(t, (3, 3),
                padding='same')(branch3x3dbl)
            branch3x3dbl = Conv2D(t, (3, 3),
                padding='same')(branch3x3dbl)
```

```
20
21          branch_pool = AveragePooling2D((3, 3), strides=(1, 1),
                padding='same')(x)
22          branch_pool = Conv2D(k, (1, 1),
                padding='same')(branch_pool)
23
24          return Concatenate()([branch1x1, branch3x3, branch3x3dbl,
                branch_pool])
25
26      # Define the InB module with dynamic filters
27      def InB_module(x, n, l, t, k, j):
28          branch1x1 = Conv2D(n, (1, 1), padding='same')(x)
29
30          branch7x7 = Conv2D(l, (1, 1), padding='same')(x)
31          branch7x7 = Conv2D(l, (7, 1), padding='same')(branch7x7)
32          branch7x7 = Conv2D(l, (1, 7), padding='same')(branch7x7)
33
34          branch7x7dbl = Conv2D(t, (1, 1), padding='same')(x)
35          branch7x7dbl = Conv2D(k, (7, 1),
                padding='same')(branch7x7dbl)
36          branch7x7dbl = Conv2D(j, (1, 7),
                padding='same')(branch7x7dbl)
37
38          branch_pool = AveragePooling2D((3, 3), strides=(1, 1),
                padding='same')(x)
39          branch_pool = Conv2D(j, (1, 1),
                padding='same')(branch_pool)
40
41          return Concatenate()([branch1x1, branch7x7, branch7x7dbl,
                branch_pool])
42
43      # Define the dense block function with different filter numbers
44      def hybrid_dense_block(x, block_number):
45
```

```
        # Define the basic sequence of layers
        def add_basic_sequence(x, n, l, t, k):
            x = Conv2D(64, (3, 3), padding='same')(x)
            x = InA_module(x, n, l, t, k)
            x = Conv2D(64, (3, 3), padding='same')(x)
            x = Conv2D(64, (3, 3), padding='same')(x)
            x = InB_module(x, inb_n, inb_l, inb_t, inb_k, inb_j)
            x = Conv2D(64, (3, 3), padding='same')(x)
        return x

        if block_number == 1:
            n, l, t, k = 24, 48, 8, 24
            inb_n, inb_l, inb_t, inb_k, inb_j = 24, 48, 8, 24, 48
              x = add_basic_sequence(x)
        elif block_number == 2:
            n, l, t, k = 128, 128, 96, 64
            inb_n, inb_l, inb_t, inb_k, inb_j = 128, 128, 96, 64,
               96
              for _ in range(2):
                   x = add_basic_sequence(x)
        elif block_number == 3:
            n, l, t, k = 256, 256, 64, 128
            inb_n, inb_l, inb_t, inb_k, inb_j = 256, 256, 64, 128,
               192
              for _ in range(5):
    x = add_basic_sequence(x)
      x = Conv2D(64, (3, 3), padding='same')(x)
      x = Conv2D(64, (3, 3), padding='same')(x)
        elif block_number == 4:
            n, l, t, k = 256, 256, 256, 128
            inb_n, inb_l, inb_t, inb_k, inb_j = 256, 256, 256,
               128, 192
      for _ in range(5):
    x = add_basic_sequence(x)
```

```
        x = Conv2D(64, (3, 3), padding='same')(x)
        x = Conv2D(64, (3, 3), padding='same')(x)
    else:
        raise ValueError(f"Block number {block_number} is not valid.")


# Define the transition block (remains unchanged)
def hybrid_transition_block(x, block_number):
    # Define the filter numbers according to the RA table provided
    if block_number == 1:
        ra_n, ra_l, ra_m, ra_k = 64, 48, 64, 24
    elif block_number == 2:
        ra_n, ra_l, ra_m, ra_k = 224, 128, 256, 64
    elif block_number == 3:
        ra_n, ra_l, ra_m, ra_k = 320, 256, 512, 128
    else:
        raise ValueError(f"Block number {block_number} is not valid for RA.")

    # Define the filter numbers according to the RB table provided
    if block_number == 1:
        rb_n, rb_l, rb_m, rb_k = 128, 128, 128, 96
    elif block_number == 2:
        rb_n, rb_l, rb_m, rb_k = 256, 256, 256, 128
    elif block_number == 3:
        rb_n, rb_l, rb_m, rb_k = 320, 512, 512, 256
    else:
        raise ValueError(f"Block number {block_number} is not valid for RB.")

```

```
            # Define the filter numbers according to the InC table provided
            if block_number == 1:
                inc_n, inc_t, inc_l, inc_k = 256, 192, 8, 24
            elif block_number == 2:
                inc_n, inc_t, inc_l, inc_k = 512, 256, 32, 32
            elif block_number == 3:
                inc_n, inc_t, inc_l, inc_k = 512, 256, 128, 128
            else:
                raise ValueError(f"Block number {block_number} is not valid for InC.")

            # Composition of layers in the transition block
    x = RA_module(x, ra_n, ra_l, ra_m, ra_k)
    x = Conv2D(64, (3, 3), padding='same')(x)
    x = RB_module(x, rb_n, rb_l, rb_m, rb_k)
    x = InC_module(x, inc_n, inc_t, inc_l, inc_k)
    x = AveragePooling2D((2, 2), strides=(2, 2), padding='same')(x)
    return x

        # DenCeption Model definition
        inputs = Input(shape=input_shape)

        # Initial Convolution and Pooling
        x = Conv2D(64, (7, 7), strides=(2, 2), padding='same')(inputs)
        x = MaxPooling2D((3, 3), strides=(2, 2), padding='same')(x)

        # Stacking Hybrid Dense Blocks and Transition Blocks
        x = hybrid_dense_block(x, block_number=1)
        x = hybrid_transition_block(x, block_number=1)
        x = hybrid_dense_block(x, block_number=2)
        x = hybrid_transition_block(x, block_number=2)
        x = hybrid_dense_block(x, block_number=3)
        x = hybrid_transition_block(x, block_number=3)
```

```
        x = hybrid_dense_block(x, block_number=4)

        # Global Average Pooling and Output
        x = GlobalAveragePooling2D()(x)
        outputs = Dense(classes, activation='softmax')(x)

        denception_model = Model(inputs, outputs)
        return denception_model

    def extract(self, image):
        return self.model.predict(np.expand_dims(image, axis=0))
```

B.4 Features Weighting

```
from minisom import MiniSom

class FeatureWeighting:
    def __init__(self, som_shape=(10, 10)):
        self.som_shape = som_shape

    def weight_features(self, high_level_features, low_level_features):
        som = MiniSom(self.som_shape[0], self.som_shape[1],
            high_level_features.shape[1] + low_level_features.shape[1],
            sigma=0.5, learning_rate=0.5)
        som.train_random(np.hstack([high_level_features,
            low_level_features]), 100)
        weights = som.get_weights()
        return weights
```

B.5 Features Fusion

```
from sklearn.neural_network import MLPClassifier

class FeaturesFusion:
    def __init__(self):
        pass

    def fuse_and_update(self, features, weights):
        combinations = [
            (0, 1), (0, 2), (1, 2),  # texture-shape, texture-colour, shape-colour
            (0, 1, 2)  # texture-shape-colour
        ]

        best_score = -np.inf
        best_combination = None

        for comb in combinations:
            selected_features = features[:, comb]
            weighted_sum = np.dot(selected_features, weights[comb])
            ann = MLPClassifier(hidden_layer_sizes=(100,), random_state=42)
            ann.fit(selected_features, weighted_sum)
            score = ann.score(selected_features, weighted_sum)
            if score > best_score:
                best_score = score
                best_combination = comb

        return best_combination, ann

    def combine_features_and_update(self, high_level_features, low_level_features, best_combination, ann_weights):
        combined_features = np.hstack([high_level_features[:, best_combination], low_level_features])
```

```
        ann = MLPClassifier(hidden_layer_sizes=(100,), random_state=42)
        ann.fit(combined_features, ann_weights)
        return ann
```

B.6 Classification Block

```
from sklearn.neural_network import MLPClassifier

class ClassificationBlock:
    def __init__(self):
        pass

    def classify(self, final_weights):
        classifier = MLPClassifier(hidden_layer_sizes=(100,),
            random_state=42)
        classifier.fit(final_weights, np.ones(final_weights.shape[0]))
            # Assuming binary classification for simplicity
        return classifier
```

B.7 Proposed Features Extraction Framework: Full Pipeline Execution

```
# Full Pipeline Execution
image_path = 'path_to_image_dataset'
ground_truth_path = 'path_to_ground_truth_mask'

# Preprocess and Segment
preprocessor = PreprocessAndSegment(image_path, ground_truth_path)
segmented_image, ground_truth_mask = preprocessor.execute()

# High-Level Features Extraction
```

```
high_level_extractor = HighLevelFeaturesExtractor(n_levels=8,
    n_clusters=5)
high_level_features = high_level_extractor.extract(segmented_image)

# Low-Level Features Extraction
low_level_extractor =
    LowLevelFeaturesExtractor(input_shape=segmented_image.shape,
    classes=2)
low_level_features = low_level_extractor.extract(segmented_image)

# Feature Weighting using SOM
feature_weighting = FeatureWeighting(som_shape=(10, 10))
weights = feature_weighting.weight_features(high_level_features,
    low_level_features)

# High-Level Features Fusion and ANN-based Weight Update
features_fusion = FeaturesFusion()
best_combination, ann_weights =
    features_fusion.fuse_and_update(high_level_features, weights)

# Combine High-Level and Low-Level Features and Update Weights
final_ann =
    features_fusion.combine_features_and_update(high_level_features,
    low_level_features, best_combination, ann_weights)

# Classification
classifier_block = ClassificationBlock()
classifier = classifier_block.classify(final_ann.coefs_)
```

B.8 Image Pre-processing: Prediction Framework

```
import cv2
import numpy as np
```

```
import matplotlib.pyplot as plt
from skimage import exposure

class ImagePreprocessor:
    def __init__(self, image_path):
        self.image_path = image_path
        self.image = cv2.imread(image_path, cv2.IMREAD_GRAYSCALE)

    def display_image(self, title, img):
        plt.imshow(img, cmap='gray')
        plt.title(title)
        plt.axis('off')
        plt.show()

    def resize_image(self, width=256, height=256):
        self.image = cv2.resize(self.image, (width, height))
        self.display_image('Resized Image', self.image)
        return self.image

    def noise_reduction(self, kernel_size=(5, 5)):
        self.image = cv2.GaussianBlur(self.image, kernel_size, 0)
        self.display_image('Noise Reduced Image', self.image)
        return self.image

    def adjust_contrast(self):
        self.image = cv2.equalizeHist(self.image)
        self.display_image('Contrast Adjusted Image', self.image)
        return self.image

    def crop_image(self, x=64, y=64, w=128, h=128):
        self.image = self.image[y:y+h, x:x+w]
        self.display_image('Cropped Image', self.image)
        return self.image

```

```
    def normalize_image(self):
        self.image = cv2.normalize(self.image, None, 0, 255,
            cv2.NORM_MINMAX)
        self.display_image('Normalized Image', self.image)
        return self.image

    def resample_image(self, downscale_size=(64, 64),
        upscale_size=(128, 128)):
        downscaled_image = cv2.resize(self.image, downscale_size)
        self.image = cv2.resize(downscaled_image, upscale_size)
        self.display_image('Resampled Image', self.image)
        return self.image

    def preprocess(self):
        self.display_image('Original Image', self.image)
        self.resize_image()
        self.noise_reduction()
        self.adjust_contrast()
        self.crop_image()
        self.normalize_image()
        self.resample_image()
        return self.image
```

B.9 Proposed HyBoost Predictive Model

```
import numpy as np
from sklearn.neural_network import MLPClassifier
from xgboost import XGBRegressor
from sklearn.ensemble import AdaBoostRegressor
from sklearn.model_selection import cross_val_score, GridSearchCV,
    train_test_split

class HighLevelFeaturesExtractor:
```

```

class LowLevelFeaturesExtractor:

class FeaturesFusion:


class HyBoostModel:
    def __init__(self):
        pass

def hyboost_predictive_model(D_train, y_train, D_test, H_xgb, H_ab, K,
   T, alpha):
    # Step 1: HyBoost Training Phase
    # Step 1.1: Train XGBoost
    M_xgb = XGBRegressor(**H_xgb)
    M_xgb.fit(D_train, y_train)
    y_xgb_pred = M_xgb.predict(D_train)

    # Residuals from XGBoost
    residuals = y_train - y_xgb_pred

    # Step 1.2: Train AdaBoost on residuals
    M_ab = AdaBoostRegressor(base_estimator=None, n_estimators=T,
       random_state=42, **H_ab)
    M_ab.fit(D_train, residuals)
    y_ab_pred = M_ab.predict(D_train)

    # Step 2: HyBoost Prediction Phase
    # Prediction on training set
    y_final_train = alpha * y_xgb_pred + (1 - alpha) * y_ab_pred

    # Prediction on test set
    y_xgb_test_pred = M_xgb.predict(D_test)
    y_ab_test_pred = M_ab.predict(D_test)
```

```
    y_final_test = alpha * y_xgb_test_pred + (1 - alpha) *
        y_ab_test_pred

    # Step 3: Tuning and Optimization
    def objective_function(alpha):
        return -np.mean(cross_val_score(M_xgb, D_train, y_train, cv=5,
            scoring='neg_mean_squared_error')) + \
                -np.mean(cross_val_score(M_ab, D_train, residuals,
                    cv=5, scoring='neg_mean_squared_error'))

    # Grid Search for alpha optimization
    grid_params = {'alpha': np.linspace(0, 1, 10)}
    grid_search = GridSearchCV(estimator=object,
        param_grid=grid_params, scoring=objective_function, cv=5)
    grid_search.fit(D_train, y_train)
    best_alpha = grid_search.best_params_['alpha']

    # Re-train with best alpha
    M_xgb.fit(D_train, y_train)
    y_xgb_final_pred = M_xgb.predict(D_train)
    M_ab.fit(D_train, y_train - y_xgb_final_pred)
    y_ab_final_pred = M_ab.predict(D_train)

    y_final_optimized_train = best_alpha * y_xgb_final_pred + (1 -
        best_alpha) * y_ab_final_pred

    # Final prediction on the test set
    y_final_optimized_test = best_alpha * M_xgb.predict(D_test) + (1 -
        best_alpha) * M_ab.predict(D_test)

    return y_final_optimized_test

high_level_features = HighLevelFeaturesExtractor
low_level_features = LowLevelFeaturesExtractor
```

```
68
69 fusion_model = FeaturesFusion()
70 best_combination, ann =
       fusion_model.fuse_and_update(high_level_features, np.random.rand(3))
71 combined_features, ann_weights =
       fusion_model.combine_features_and_update(high_level_features,
       low_level_features, best_combination, ann.coefs_[0])
72
73 # Features and weights are now defined as:
74 features = combined_features  # Combined high and low-level features
75 weights = ann_weights  # Resulted ANN weights from
       combine_features_and_update
76
77 # Step 2: Split the combined features into training and testing
       datasets
78 D_train, D_test, y_train, y_test = train_test_split(features,
       np.random.rand(features.shape[0]), test_size=0.2, random_state=42)
79
80 # Step 3: HyBoost model training and prediction
81 H_xgb = {
82     'max_depth': 3,
83     'learning_rate': 0.1,
84     'n_estimators': 100,
85 }
86
87 H_ab = {
88     'n_estimators': 50,
89     'learning_rate': 0.1,
90 }
91
92 K = 100
93 T = 50
94 alpha = 0.5
95
```

```
# Execute the HyBoost model with the split data
y_final = hyboost_predictive_model(D_train, y_train, D_test, H_xgb,
    H_ab, K, T, alpha)
print("Final Predictions on Test Data:", y_final)
```

B.10 Performance Measurement Matrix

```
import numpy as np

class OptimalPerformanceMetricMatrix:
    def __init__(self):
        pass

    def weight_assignment_function(self, metric):
        """
        Assigns a weight to a given metric.
        Modify this function to implement the specific weight
            assignment logic as per your requirements.
        """
        return np.random.random()  # For simplicity, returning a
            random weight

    def correlation_coefficient(self, x, y):
        """
        Computes the correlation coefficient between two arrays x and
            y.
        """
        return np.corrcoef(x, y)[0, 1]

    def optimal_performance_evaluation(self, C12, C13, C23):
        """
        Evaluates the optimal performance metric matrix given the
            correlation matrices C12, C13, and C23.
```

```
        Returns the optimal performance metrics.
        """
        n = len(C12)

        # Initialize matrices and vectors
        WC12 = np.zeros(n)
        D12 = np.zeros((n, n))
        D13 = np.zeros((n, n))
        D23 = np.zeros((n, n))
        PMM = np.zeros((n, n))
        PMM_optimal = np.zeros(n)

        # Step 1: Calculate weighted performance metrics for C12
        for i in range(n):
            WC12[i] = self.weight_assignment_function(C12[i])

        # Step 2: Populate D12, D13, D23 matrices
        for i in range(n):
            for j in range(n):
                if i == j:
                    WC12ii = self.correlation_coefficient(C12[i],
                        C12[i])
                    D12[i, j] = WC12ii
                else:
                    WC12ij = self.correlation_coefficient(C12[i],
                        C12[j])
                    D12[i, j] = WC12ij

        # Repeat the same steps for D13 and D23
        for i in range(n):
            for j in range(n):
                if i == j:
                    WC13ii = self.correlation_coefficient(C13[i],
                        C13[i])
```

```
                    D13[i, j] = WC13ii
                else:
                    WC13ij = self.correlation_coefficient(C13[i],
                        C13[j])
                    D13[i, j] = WC13ij

        for i in range(n):
            for j in range(n):
                if i == j:
                    WC23ii = self.correlation_coefficient(C23[i],
                        C23[i])
                    D23[i, j] = WC23ii
                else:
                    WC23ij = self.correlation_coefficient(C23[i],
                        C23[j])
                    D23[i, j] = WC23ij

        # Step 3: Calculate PMM matrix
        for i in range(n):
            for j in range(n):
                if i == j:
                    PMM[i, j] = max(D12[i, i], D13[i, i], D23[i, i])
                else:
                    PMM[i, j] = max(D12[i, j], D13[i, j], D23[i, j])

        # Step 4: Determine the optimal performance metrics
        for i in range(n):
            for j in range(n):
                if PMM[i, j] > 0:
                    PMM_optimal[i] = PMM[i, j]

        return PMM_optimal
```

B.11 SHAP Analysis

```
import shap


    def shap_analysis(self, M_xgb, M_ab, D_train, D_test):
        # SHAP analysis for XGBoost
        explainer_xgb = shap.Explainer(M_xgb)
        shap_values_xgb = explainer_xgb(D_test)
        print("SHAP Analysis for XGBoost:")
        shap.summary_plot(shap_values_xgb, D_test)

        # SHAP analysis for AdaBoost
        explainer_ab = shap.Explainer(M_ab, D_train)
        shap_values_ab = explainer_ab(D_test)
        print("SHAP Analysis for AdaBoost:")
        shap.summary_plot(shap_values_ab, D_test)
```

Bibliography

Abdou, M. A. (2022). 'Literature review: Efficient deep neural networks techniques for medical image analysis'. *Neural Computing and Applications*, 34.(8), pp. 5791–5812.

Ahmed, F, Nuwagira, B, Torlak, F, and Coskunuzer, B (2023a). 'Topo-CXR: chest X-ray TB and pneumonia screening with topological machine learning.' *CVF International Conference on Computer Vision Workshops (ICCVW). IEEE, Paris*. Vol. 10.

Ahmed, M. S. et al. (2023b). 'Joint diagnosis of pneumonia, COVID-19, and tuberculosis from chest X-ray images: A deep learning approach'. *Diagnostics*, 13.(15), p. 2562.

Ait Nasser, A. and Akhloufi, M. A. (2023). 'A review of recent advances in deep learning models for chest disease detection using radiography'. *Diagnostics*, 13.(1), p. 159.

Akella, P. L. and Kumar, R. (2023). 'An advanced deep learning method to detect and classify diabetic retinopathy based on color fundus images'. *Graefe's Archive for Clinical and Experimental Ophthalmology*, pp. 1–17.

Al-Haidri, W., Matveev, I., Al-Antari, M. A., and Zubkov, M. (2023). 'A Deep Learning Framework for Cardiac MR Under-Sampled Image Reconstruction with a Hybrid Spatial and k-Space Loss Function'. *Diagnostics*, 13.(6), p. 1120.

Alaskar, H, Hussain, A, Almaslukh, B, Vaiyapuri, T, Sbai, Z, and Dubey, A. K. (2022). 'Deep learning approaches for automatic localization in medical images'. *Computational Intelligence and Neuroscience*, 2022.

Albawi, S., Mohammed, T. A., and Al-Zawi, S. (2017). 'Understanding of a convolutional neural network'. *2017 international conference on engineering and technology (ICET)*. Ieee, pp. 1–6.

Alotaibi, B. and Alotaibi, M. (2020). 'A hybrid deep ResNet and inception model for hyperspectral image classification'. *PFG–Journal of Photogrammetry, Remote Sensing and Geoinformation Science*, 88.(6), pp. 463–476.

Alryalat, S. A. et al. (2022). 'Deep learning prediction of response to anti-VEGF among diabetic macular edema patients: Treatment response analyzer system (TRAS)'. *Diagnostics*, 12.(2), p. 312.

Altaf, T., Anwar, S, Gul, N., Majeed, N, and Majid, M (2017). 'Multi-class Alzheimer disease classification using hybrid features'. *IEEE future technologies conference*.

Alzubaidi, L. et al. (2021). 'Review of deep learning: concepts, CNN architectures, challenges, applications, future directions'. *Journal of big Data*, 8, pp. 1–74.

Ammari, A., Mahmoudi, R., Hmida, B., Saouli, R., and Bedoui, M. H. (2023). 'Deep-active-learning approach towards accurate right ventricular segmentation using a two-level uncertainty estimation'. *Computerized Medical Imaging and Graphics*, 104, p. 102168.

Ananda, A., Ngan, K. H., Karabağ, C., Ter-Sarkisov, A., Alonso, E., and Reyes-Aldasoro, C. C. (2021). 'Classification and visualisation of normal and abnormal radiographs; a comparison between eleven convolutional neural network architectures'. *Sensors*, 21.(16), p. 5381.

Anwar, S. M., Majid, M., Qayyum, A., Awais, M., Alnowami, M., and Khan, M. K. (2018). 'Medical image analysis using convolutional neural networks: a review'. *Journal of medical systems*, 42, pp. 1–13.

Arel, I., Rose, D. C., and Karnowski, T. P. (2010). 'Deep machine learning-a new frontier in artificial intelligence research [research frontier]'. *IEEE computational intelligence magazine*, 5.(4), pp. 13–18.

Arena, P., Basile, A., Bucolo, M., and Fortuna, L. (2003). 'Image processing for medical diagnosis using cnn'. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 497.(1), pp. 174–178.

Arumugadevi, S and Seenivasagam, V (2016). 'Color image segmentation using feedforward neural networks with FCM'. *International Journal of Automation and Computing*, 13, pp. 491–500.

Atasever, S., Azginoglu, N., Terzi, D. S., and Terzi, R. (2023). 'A comprehensive survey of deep learning research on medical image analysis with focus on transfer learning'. *Clinical Imaging*, 94, pp. 18–41.

Athira, T. R. and Nair, J. J. (2023). 'Diabetic Retinopathy Grading From Color Fundus Images: An Autotuned Deep Learning Approach'. *Procedia Computer Science*, 218, pp. 1055–1066.

Azad, R. et al. (2022). 'Transdeeplab: Convolution-free transformer-based deeplab v3+ for medical image segmentation'. *International Workshop on PRedictive Intelligence In MEdicine*. Springer, pp. 91–102.

Bach Cuadra, M, Duay, V., and Thiran, J.-P. (2015). 'Atlas-based segmentation'. *Handbook of Biomedical Imaging: Methodologies and Clinical Research*, pp. 221–244.

Baid, U., Ghodasara, S., Mohan, S., Bilello, M., Calabrese, E., Colak, E., et al. (2021). 'The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification'. *arXiv preprint arXiv:2107.02314*,

Bala, R., Sharma, A., and Goel, N. (2023). 'Comparative Analysis of Diabetic Retinopathy Classification Approaches Using Machine Learning and Deep Learning Techniques'. *Archives of Computational Methods in Engineering*, pp. 1–37.

Bankman, I. (2008). *Handbook of medical image processing and analysis*. Elsevier.

Bao, H., Zhu, Y., and Li, Q. (2023). 'Hybrid-scale contextual fusion network for medical image segmentation'. *Computers in Biology and Medicine*, 152, p. 106439.

Basha, S. S., Dubey, S., Pulabaigari, V., and Mukherjee, S. (2020). 'Impact of fully connected layers on performance of convolutional neural networks for image classification'. *Neurocomputing*, 378, pp. 112–119.

Belhadi, A., Djenouri, Y., Diaz, V. G., Houssein, E. H., and Lin, J. C.-W. (2022). 'Hybrid intelligent framework for automated medical learning'. *Expert Systems*, 39.(6), e12737.

Bernardes, R., Serranho, P., and Lobo, C. (2011). 'Digital ocular fundus imaging: a review'. *Ophthalmologica*, 226.(4), pp. 161–181.

Bianco, S., Cadene, R., Celona, L., and Napoletano, P. (2018). 'Benchmark analysis of representative deep neural network architectures'. *IEEE access*, 6, pp. 64270–64277.

Billot, B. et al. (2023). 'SynthSeg: Segmentation of brain MRI scans of any contrast and resolution without retraining'. *Medical image analysis*, 86, p. 102789.

Bourouis, S., Alroobaea, R., Rubaiee, S., and Ahmed, A. (2020). 'Toward effective medical image analysis using hybrid approaches—review, challenges and applications'. *Information*, 11.(3), p. 155.

Bozkurt, F. (2022). 'A comparative study on classifying human activities using classical machine and deep learning methods'. *Arabian Journal for Science and Engineering*, 47.(2), pp. 1507–1521.

Bressem, K. K., Adams, L. C., Erxleben, C., Hamm, B., Niehues, S. M., and Vahldiek, J. L. (2020). 'Comparing different deep learning architectures for classification of chest radiographs'. *Scientific reports*, 10.(1), p. 13590.

Carnagie, J. O., Prabowo, A. R., Budiana, E. P., and Singgih, I. K. (2022). 'Essential oil plants image classification using xception model'. *Procedia Computer Science*, 204, pp. 395–402.

Celik, G. (2023). 'Detection of Covid-19 and other Pneumonia cases from CT and X-ray chest images using deep learning based on feature reuse residual block and depthwise dilated convolutions neural network'. *Applied Soft Computing*, 133, p. 109906.

Charbuty, B. and Abdulazeez, A. (2021). 'Classification based on decision tree algorithm for machine learning'. *Journal of Applied Science and Technology Trends*, 2.(01), pp. 20–28.

Chen, S.-C., Chiu, H.-W., Chen, C.-C., Woung, L.-C., and Lo, C.-M. (2018). 'A novel machine learning algorithm to automatically predict visual outcomes in intravitreal ranibizumab-treated patients with diabetic macular edema'. *Journal of clinical medicine*, 7.(12), p. 475.

Chen, S. and Guo, W. (2023). 'Auto-encoders in deep learning—a review with new perspectives'. *Mathematics*, 11.(8), p. 1777.

Chen, X., Wang, M., Ling, J., Wu, H., Wu, B., and Li, C. (2024). 'Ship imaging trajectory extraction via an aggregated you only look once (YOLO) model'. *Engineering Applications of Artificial Intelligence*, 130, p. 107742.

Chen, Y., Ashizawa, N., Yeo, C. K., Yanai, N., and Yean, S. (2021). 'Multi-scale self-organizing map assisted deep autoencoding Gaussian mixture model for unsupervised intrusion detection'. *Knowledge-Based Systems*, 224, p. 107086.

Cheng, B., Xiao, B., Wang, J., Shi, H., Huang, T. S., and Zhang, L. (2020). 'Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation'. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5386–5395.

Chieregato, M. et al. (2022). 'A hybrid machine learning/deep learning COVID-19 severity predictive model from CT images and clinical data'. *Scientific reports*, 12.(1), p. 4329.

Chollet, F. (2017). 'Xception: Deep learning with depthwise separable convolutions'. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1251–1258.

Chouhan, S. S., Kaul, A., and Singh, U. P. (2019). 'Image segmentation using fuzzy competitive learning based counter propagation network'. *Multimedia Tools and Applications*, 78.(24), pp. 35263–35287.

Chowdhary, C. L. and Acharjya, D. P. (2020). 'Segmentation and feature extraction in medical imaging: a systematic review'. *Procedia Computer Science*, 167, pp. 26–36.

Conghua, X., Song, Y., Zhu, Y., and Wang, L. (2005). 'A Grid-based Approach to Extracting Irregular Features of Medical Images'. *Computer Engineering and Applications*.

Conghua, X., Yuqing, S., and Jinyi, C. (2006). 'A new method of semantic feature extraction for medical images data'. *Wuhan University Journal of Natural Sciences*, 11.(5), pp. 1152–1156.

Cuevas-Rodriguez, E. O. et al. (2023). 'Comparative study of convolutional neural network architectures for gastrointestinal lesions classification'. *PeerJ*, 11, e14806.

Daghrir, J., Tlig, L., Bouchouicha, M., and Sayadi, M. (2020). 'Melanoma skin cancer detection using deep learning and classical machine learning techniques: A hybrid approach'. *2020 5th international conference on advanced technologies for signal and image processing (ATSIP)*. IEEE, pp. 1–5.

Dai, B., Bai, F., Sun, W., Huang, Y., and Wang, W. (2018). 'Using random forest algorithm for breast cancer diagnosis'. *2018 International Symposium on Computer, Consumer and Control (IS3C)*. IEEE, pp. 449–452.

Dara, S., Tumma, P., Eluri, N. R., and Kancharla, G. R. (2018). 'Feature extraction in medical images by using deep learning approach'. *International Journal of Pure and Applied Mathematics*, 120.(6), pp. 305–312.

Das, D., Biswas, S. K., and Bandyopadhyay, S. (2022). 'A critical review on diagnosis of diabetic retinopathy using machine learning and deep learning'. *Multimedia Tools and Applications*, 81.(18), pp. 25613–25655.

Deng, Y. et al. (2023). 'Automated CT pancreas segmentation for acute pancreatitis patients by combining a novel object detection approach and U-Net'. *Biomedical signal processing and control*, 81, p. 104430.

Desai, M. and Shah, M. (2021). ‘An anatomization on breast cancer detection and diagnosis employing multi-layer perceptron neural network (MLP) and Convolutional neural network (CNN)’. *Clinical eHealth*, 4, pp. 1–11.

Desikan, R. S. et al. (2006). ‘An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest’. *NeuroImage*, 31.(3), pp. 968–980.

Duong, C. N., Luu, K., Quach, K. G., and Bui, T. D. (2019). ‘Deep appearance models: A deep boltzmann machine approach for face modeling’. *International Journal of Computer Vision*, 127, pp. 437–455.

Duta, I. C., Liu, L., Zhu, F., and Shao, L. (2020). ‘Pyramidal convolution: Rethinking convolutional neural networks for visual recognition’. *arXiv preprint arXiv:2006.11538*,

Ecabert, O. et al. (2008). ‘Automatic model-based segmentation of the heart in CT images’. *IEEE transactions on medical imaging*, 27.(9), pp. 1189–1201.

El-Shafai, W., Abd El-Samie, F. E., Soliman, A. M., and Mostafa, M. G. (2024). ‘Traditional and deep-learning-based denoising methods for medical images’. *Multimedia Tools and Applications*, 83.(17), pp. 52061–52088.

Emara, T. H. M., Afify, H. M., Ismail, F. H., and Hassanien, A. E. (2019). ‘A modified inception-v4 for imbalanced skin cancer classification dataset’. *2019 14th International Conference on Computer Engineering and Systems (ICCES)*. IEEE, pp. 28–33.

Fotopoulos, D., Filos, D., Xinou, E., and Chouvarda, I. (2023). ‘Towards Lung Cancer Staging via Multipositional Radiomics and Machine Learning.’ *BIOSIGNALS*, pp. 317–324.

Ganesan, N, Venkatesh, K, Rama, M., and Palani, A. M. (2010). ‘Application of neural networks in diagnosing cancer disease using demographic data’. *International Journal of Computer Applications*, 1.(26), pp. 76–85.

Gao, Y., Ma, S., Liu, J., Liu, Y., and Zhang, X. (2021). 'Fusion of medical images based on salient features extraction by PSO optimized fuzzy logic in NSST domain'. *Biomedical Signal Processing and Control*, 69, p. 102852.

Garcea, F., Serra, A., Lamberti, F., and Morra, L. (2023). 'Data augmentation for medical imaging: A systematic literature review'. *Computers in Biology and Medicine*, 152, p. 106391.

Gargeya, R. and Leng, T. (2017). 'Automated identification of diabetic retinopathy using deep learning'. *Ophthalmology*, 124.(7), pp. 962–969.

Gavrishchaka, V., Yang, Z., Miao, R., Senyukova, O., et al. (2018). 'Advantages of hybrid deep learning frameworks in applications with limited data'. *International Journal of Machine Learning and Computing*, 8.(6), pp. 549–558.

Gichoya, J. W. et al. (2022). 'AI recognition of patient race in medical imaging: a modelling study'. *The Lancet Digital Health*, 4.(6), e406–e414.

Girdhar, N., Sinha, A., and Gupta, S. (2023). 'DenseNet-II: An improved deep convolutional neural network for melanoma cancer detection'. *Soft Computing*, 27.(18), pp. 13285–13304.

Gu, J. et al. (2018). 'Recent advances in convolutional neural networks'. *Pattern recognition*, 77, pp. 354–377.

Gulshan, V. et al. (2016). 'Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs'. *jama*, 316.(22), pp. 2402–2410.

Guo, X., Gichoya, J. W., Trivedi, H., Purkayastha, S., and Banerjee, I. (2023). 'MedShift: Automated Identification of Shift Data for Medical Image Dataset Curation'. *IEEE Journal of Biomedical and Health Informatics*,

Gupta, S., Thakur, S., and Gupta, A. (2023). 'Comparative study of different machine learning models for automatic diabetic retinopathy detection using fundus image'. *Multimedia Tools and Applications*, pp. 1–32.

Ha, V. K. et al. (2020). 'Optimized highway deep learning network for fast single image super-resolution reconstruction'. *Journal of Real-Time Image Processing*, 17, pp. 1961–1970.

Hazarika, R. A., Maji, A. K., Sur, S. N., Paul, B. S., and Kandar, D. (2021). 'A survey on classification algorithms of brain images in Alzheimer's disease based on feature extraction techniques'. *IEEE Access*, 9, pp. 58503–58536.

He, K., Zhang, X., Ren, S., and Sun, J. (2016a). 'Deep residual learning for image recognition'. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778.

He, K., Zhang, X., Ren, S., and Sun, J. (2016b). 'Identity mappings in deep residual networks'. *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV*. Vol. 14. Springer International Publishing, pp. 630–645.

Howard, A. G. et al. (2017). 'Mobilenets: Efficient convolutional neural networks for mobile vision applications'. *arXiv preprint arXiv:1704.04861*,

Howarth, P. and Rüger, S. (2004). 'Evaluation of texture features for content-based image retrieval'. *International conference on image and video retrieval*. Springer, pp. 326–334.

Hu, J., Shen, L., and Sun, G. (2018). 'Squeeze-and-excitation networks'. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7132–7141.

Hu, M., Zhang, J., Matkovic, L., Liu, T., and Yang, X. (2023). 'Reinforcement learning in medical image analysis: Concepts, applications, challenges, and future directions'. *Journal of Applied Clinical Medical Physics*, 24.(2), e13898.

Huang, G., Chen, D., Li, T., Wu, F., Van Der Maaten, L., and Weinberger, K. Q. (2017a). 'Multi-scale dense networks for resource efficient image classification'. *arXiv preprint arXiv:1703.09844*,

Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. (2017b). 'Densely connected convolutional networks'. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708.

Huda, W. and Abrahams, R. B. (2015). 'X-ray-based medical imaging and resolution'. *American Journal of Roentgenology*, 204.(4), W393–W397.

Huerga, C. et al. (2021). 'Role of correlated noise in textural features extraction'. *Physica Medica*, 91, pp. 87–98.

Iandola, F. N., Han, S., Moskewicz, M. W., Ashraf, K., Dally, W. J., and Keutzer, K. (2016). 'SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5 MB model size'. *arXiv preprint arXiv:1602.07360*,

Ikuta, M. and Zhang, J. (2022). 'A deep convolutional gated recurrent unit for CT image reconstruction'. *IEEE Transactions on Neural Networks and Learning Systems*, 34.(12), pp. 10612–10625.

Iqbal, S., N. Qureshi, A., Li, J., and Mahmood, T. (2023). 'On the analyses of medical images using traditional machine learning techniques and convolutional neural networks'. *Archives of Computational Methods in Engineering*, 30.(5), pp. 3173–3233.

Janakasudha, G and Jayashree, P (2020). 'Early detection of Alzheimer's disease using multi-feature fusion and an ensemble of classifiers'. *Advanced Computing and Intelligent Engineering*. Springer, pp. 113–123.

Janiesch, C., Zschech, P., and Heinrich, K. (2021). 'Machine learning and deep learning'. *Electronic Markets*, 31.(3), pp. 685–695.

Jena, B., Saxena, S., Nayak, G. K., Saba, L., Sharma, N., and Suri, J. S. (2021). 'Artificial intelligence-based hybrid deep learning models for image classification: The first narrative review'. *Computers in Biology and Medicine*, 137, p. 104803.

Jeyakumar, V. and Kanagaraj, B. (2019). 'A medical image retrieval system in PACS environment for clinical decision making'. *Intelligent Data Analysis for Biomedical Applications*. Elsevier, pp. 121–146.

Jeyaraj, P. R. and Nadar, E. R. S. (2019). 'Deep Boltzmann machine algorithm for accurate medical image analysis for classification of cancerous region'. *Cognitive Computation and Systems*, 1.(3), pp. 85–90.

Jiang, H. et al. (2023). 'A review of deep learning-based multiple-lesion recognition from medical images: classification, detection and segmentation'. *Computers in Biology and Medicine*, 157, p. 106726.

Jiang, J., Trundle, P, and Ren, J. (2010). 'Medical image analysis with artificial neural networks'. *Computerized Medical Imaging and Graphics*, 34.(8), pp. 617–631.

Kaur, P., Singh, G., and Kaur, P. (2018). 'A review of denoising medical images using machine learning approaches'. *Current medical imaging*, 14.(5), pp. 675–685.

Kavya, N and Padmaja, K. (2017). 'Glaucoma detection using texture features extraction'. *2017 51st Asilomar Conference on Signals, Systems, and Computers*. IEEE, pp. 1471–1475.

Keerthana, C., Tejasree, P., Rao, M. V. S., Kumar, R. S. P., and Yalla, P. (2023). 'Constructive Analysis on Prediction and Detection of Diabetic Retinopathy (DR) using Machine Learning Algorithms-A Generic Survey'. *2023 7th International Conference on Computing Methodologies and Communication (ICCMC)*. IEEE, pp. 240–245.

Khacef, L., Rodriguez, L., and Miramond, B. (2020). 'Improving self-organizing maps with unsupervised feature extraction'. *Neural Information Processing: 27th International Conference, ICONIP 2020, Bangkok, Thailand, November 23–27, 2020, Proceedings, Part II 27*. Springer, pp. 474–486.

Khan, A., Sohail, A., Zahoora, U., and Qureshi, A. S. (2020). 'A survey of the recent architectures of deep convolutional neural networks'. *Artificial intelligence review*, 53, pp. 5455–5516.

Khanna, M., Singh, L. K., Thawkar, S., and Goyal, M. (2023). 'Deep learning based computer-aided automatic prediction and grading system for diabetic retinopathy'. *Multimedia Tools and Applications*, 82.(25), pp. 39255–39302.

Kim, S., An, S., Chikontwe, P., and Park, S. H. (2021). 'Bidirectional rnn-based few shot learning for 3d medical image segmentation'. *Proceedings of the AAAI conference on artificial intelligence*. Vol. 35. 3, pp. 1808–1816.

Koné, I. and Boulmane, L. (2018). 'Hierarchical ResNeXt models for breast cancer histology image classification'. *Image Analysis and Recognition: 15th International Conference, ICIAR 2018, Póvoa de Varzim, Portugal, June 27–29, 2018, Proceedings*. Vol. 15. Springer International Publishing, pp. 230–237.

Koonce, B. (2021). 'SqueezeNet'. *Convolutional Neural Networks with Swift for TensorFlow: Image Recognition and Dataset Categorization*. Apress, pp. 73–85.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). 'Imagenet classification with deep convolutional neural networks'. *Advances in neural information processing systems*, 25.

Kumar, G. and Bhatia, P. K. (2014). 'A Detailed Review of Feature Extraction in Image Processing Systems'. *2014 Fourth International Conference on Advanced Computing Communication Technologies*, pp. 5–12. doi: 10.1109/acct.2014.74.

Kumar, K., Kumar, P., Deb, D., Unguresan, M.-L., and Muresan, V. (2023). 'Artificial intelligence and machine learning based intervention in medical infrastructure: a review and future trends'. 11.(2), p. 207.

Kumar, Y. and Gupta, S. (2023). 'Deep transfer learning approaches to predict glaucoma, cataract, choroidal neovascularization, diabetic macular edema, drusen and healthy eyes: an

experimental review'. *Archives of Computational Methods in Engineering*, 30.(1), pp. 521–541.

Kunhimon, S., Shaker, A., Naseer, M., Khan, S., and Khan, F. S. (2023). 'Learnable Weight Initialization for Volumetric Medical Image Segmentation'. *arXiv preprint arXiv:2306.09320*,

Kurita, T. (2019). 'Principal component analysis (PCA)'. *Computer vision: a reference guide*, pp. 1–4.

Kutan, F., KUTBAY, U., and ALGIN, O. (2023). 'Automated Cerebral Vessel Segmentation Using Deep Learning for Early Detection of Cerebrovascular Diseases'. *2023 5th International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA)*. IEEE, pp. 1–9.

Lai, Z. and Deng, H. (2018). 'Medical image classification based on deep features extracted by deep model and statistic feature fusion with multilayer perceptron'. *Computational intelligence and neuroscience*, 2018.

Lasker, A., Ghosh, M., Obaidullah, S. M., Chakraborty, C., and Roy, K. (2023). 'LWSNet-a novel deep-learning architecture to segregate Covid-19 and pneumonia from x-ray imagery'. *Multimedia Tools and Applications*, 82.(14), pp. 21801–21823.

Li, D., Ran, A. R., Cheung, C. Y., and Prince, J. L. (2023a). 'Deep learning in optical coherence tomography: Where are the gaps?' *Clinical & Experimental Ophthalmology*,

Li, F, Liu, Z, Chen, H, Jiang, M, Zhang, X, and Wu, Z (2019). 'Automatic detection of diabetic retinopathy in retinal fundus photographs based on deep learning algorithm. Translational Vision Science and Technology 8 (6)(2019)'. *Translational Vision Science Technology*, 8.(6), pp. 4–4.

Li, F. et al. (2022a). 'Deep learning-based automated detection for diabetic retinopathy and diabetic macular oedema in retinal fundus photographs'. *Eye*, 36.(7), pp. 1433–1441.

Li, Q., Cai, W., Wang, X., Zhou, Y., Feng, D. D., and Chen, M. (2014). 'Medical image classification with convolutional neural network'. *2014 13th international conference on control automation robotics & vision (ICARCV)*. IEEE, pp. 844–848.

Li, S., Liu, F., Jiao, L., Chen, P., and Li, L. (2022b). 'Self-supervised self-organizing clustering network: A novel unsupervised representation learning method'. *IEEE Transactions on Neural Networks and Learning Systems*, 35.(2), pp. 1857–1871.

Li, X. et al. (2023b). 'HAL-IA: A Hybrid Active Learning framework using Interactive Annotation for medical image segmentation'. *Medical Image Analysis*, 88, p. 102862.

Li, Y., Zhao, J., Lv, Z., and Li, J. (2021a). 'Medical image fusion method by deep learning'. *International Journal of Cognitive Computing in Engineering*, 2, pp. 21–29.

Li, Y. et al. (2023c). 'Self-supervised anomaly detection, staging and segmentation for retinal images'. *Medical Image Analysis*, 87, p. 102805.

Li, Z., Liu, F., Yang, W., Peng, S., and Zhou, J. (2021b). 'A survey of convolutional neural networks: analysis, applications, and prospects'. *IEEE transactions on neural networks and learning systems*, 33.(12), pp. 6999–7019.

Lin, C.-H., Chen, R.-T., and Chan, Y.-K. (2009). 'A smart content-based image retrieval system based on color and texture feature'. *Image and vision Computing*, 27.(6), pp. 658–665.

Linchundan (2019). *Fundus Image 1000 Dataset*. https://www.kaggle.com/datasets/linchundan/fundusimage1000. Kaggle dataset, Accessed: 2024.

Litjens, G. et al. (2017). 'A survey on deep learning in medical image analysis'. *Medical image analysis*, 42, pp. 60–88.

Liu, C. et al. (2019). 'Automatic segmentation of the prostate on CT images using deep neural networks (DNN)'. *International Journal of Radiation Oncology* Biology* Physics*, 104.(4), pp. 924–932.

Liu, J. and Shi, Y. (2011). 'Image feature extraction method based on shape characteristics and its application in medical image analysis'. *International Conference on Applied Informatics and Communication*. Springer, pp. 172–178.

Liu, L. Y.-F., Liu, Y., and Zhu, H. (2020). 'Masked convolutional neural network for supervised learning problems'. *Stat*, 9.(1), e290.

Liu, X., Pang, Y., Jin, R., Liu, Y., and Wang, Z. (2022). 'Dual-domain reconstruction network with V-Net and K-Net for fast MRI'. *Magnetic Resonance in Medicine*, 88.(6), pp. 2694–2708.

Loukil, Z., Mirza, Q. K. A., and Sayers, W. (2023). 'A Novel and Adaptive Evaluation Mechanism for Deep Learning Models in Medical Imaging and Disease Recognition'. *2023 10th International Conference on Future Internet of Things and Cloud (FiCloud)*. IEEE, pp. 270–277.

Loukil, Z., Mirza, Q. K. A., Sayers, W., and Awan, I. (2023). 'A deep learning based scalable and adaptive feature extraction framework for medical images'. *Information Systems Frontiers*, pp. 1–27.

Loukil, Z. and Salah, A.-M. (2020). 'Toward Hybrid Deep Convolutional Neural Network Architectures For Medical Image Processing'. *2020 International Conference on Innovation and Intelligence for Informatics, Computing and Technologies (3ICT)*. IEEE, pp. 1–6.

Ma, N., Zhang, X., Zheng, H.-T., and Sun, J. (2018). 'Shufflenet v2: Practical guidelines for efficient cnn architecture design'. *Proceedings of the European conference on computer vision (ECCV)*, pp. 116–131.

Madusanka, N., Choi, H.-K., So, J.-H., and Choi, B.-K. (2019). 'Alzheimer's Disease classification based on multi-feature fusion'. *Current Medical Imaging*, 15.(2), pp. 161–169.

Majumdar, S., Pramanik, P., and Sarkar, R. (2023). 'Gamma function based ensemble of CNN models for breast cancer detection in histopathology images'. *Expert Systems with Applications*, 213, p. 119022.

Mall, P. K. et al. (2023). 'A comprehensive review of deep neural networks for medical image processing: Recent developments and future opportunities'. *Healthcare Analytics*, p. 100216.

McNeely-White, D., Beveridge, J. R., and Draper, B. A. (2020). 'Inception and ResNet features are (almost) equivalent'. *Cognitive Systems Research*, 59, pp. 312–318.

Mei, S., Li, X., Liu, X., Cai, H., and Du, Q. (2021). 'Hyperspectral image classification using attention-based bidirectional long short-term memory network'. *IEEE Transactions on Geoscience and Remote Sensing*, 60, pp. 1–12.

Mendonca, A. M. and Campilho, A. (2006). 'Segmentation of retinal blood vessels by combining the detection of centerlines and morphological reconstruction'. *IEEE transactions on medical imaging*, 25.(9), pp. 1200–1213.

Messaoudi, H., Belaid, A., Salem, D. B., and Conze, P.-H. (2023). 'Cross-dimensional transfer learning in medical image segmentation with deep learning'. *Medical image analysis*, 88, p. 102868.

Miao, T., Liu, H., Yu, H., Wang, R., Chen, Z., and Wang, Y. (2022). 'An improved lightweight RetinaNet for ship detection in SAR images'. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 15, pp. 4667–4679.

Miljković, D. (2017). 'Brief review of self-organizing maps'. *2017 40th international convention on information and communication technology, electronics and microelectronics (MIPRO)*. IEEE, pp. 1061–1066.

Mingqiang, Y., Kidiyo, K., Joseph, R., et al. (2008). 'A survey of shape feature extraction techniques'. *Pattern recognition*, 15.(7), pp. 43–90.

Mishra, N. and Singh, A. (2022). 'A Deep Learning Approach for Detecting Diabetic Macular Edema through Analyzing Retinal Images'. *Mathematical Statistician and Engineering Applications*, 71.(3s), pp. 233–242.

Mishra, S. P. and Rahul, M. R. (2021). 'A comparative study and development of a novel deep learning architecture for accelerated identification of microstructure in materials science'. *Computational Materials Science*, 200, p. 110815.

Montazer, G. A., Giveki, D., Karami, M., and Rastegar, H. (2018). 'Radial basis function neural networks: A review'. *Comput. Rev. J*, 1.(1), pp. 52–74.

Mooney, P. T. (2018a). *Chest X-Ray Pneumonia*. https://www.kaggle.com/datasets/paultimothymooney/chest-xray-pneumonia. Accessed: April 2024.

Mooney, P. T. (2018b). *Kermany2018: OCT and Chest X-Ray Images for Classification*. https://www.kaggle.com/datasets/paultimothymooney/kermany2018. Accessed: April 2024.

Mukherjee, N. and Sengupta, S. (2023). 'Application of deep learning approaches for classification of diabetic retinopathy stages from fundus retinal images: a survey'. *Multimedia Tools and Applications*, pp. 1–61.

Mukhlif, A. A., Al-Khateeb, B., and Mohammed, M. A. (2023). 'Incorporating a novel dual transfer learning approach for medical images'. *Sensors*, 23.(2), p. 570.

Mushtaq, G. and Siddiqui, F. (2021). 'Detection of diabetic retinopathy using deep learning methodology'. *IOP Conference Series: Materials Science and Engineering*. Vol. 1070. 1. IOP Publishing, p. 012049.

Mutlag, W. K., Ali, S. K., Aydam, Z. M., and Taher, B. H. (2020). 'Feature extraction methods: a review'. *Journal of Physics: Conference Series*. Vol. 1591. 1. IOP Publishing, p. 012028.

Nakayama, Y., Lu, H., Li, Y., and Kamiya, T. (2020). 'WideSegNeXt: semantic image segmentation using wide residual network and NeXt dilated unit'. *IEEE Sensors Journal*, 21.(10), pp. 11427–11434.

Nazir, T. et al. (2021). 'Detection of diabetic eye disease from retinal images using a deep learning based CenterNet model'. *Sensors*, 21.(16), p. 5283.

Ngo, T. A., Lu, Z., and Carneiro, G. (2017). 'Combining deep learning and level set for the automated segmentation of the left ventricle of the heart from cardiac cine magnetic resonance'. *Medical image analysis*, 35, pp. 159–171.

Nixon, M. S. and Aguado, A. S. (2019). *Feature Extraction and Image Processing for Computer Vision*. Academic Press.

Noriega, A. et al. (2021). 'Screening diabetic retinopathy using an automated retinal image analysis system in independent and assistive use cases in Mexico: randomized controlled trial'. *JMIR formative research*, 5.(8), e25290.

O'shea, K. and Nash, R. (2015). 'An introduction to convolutional neural networks'. *arXiv preprint arXiv:1511.08458*,

Painuli, D., Bhardwaj, S., and köse, U. (2022). 'Recent advancement in cancer diagnosis using machine learning and deep learning techniques: A comprehensive review'. *Computers in Biology and Medicine*, 146, p. 105580.

Panayides, A. S. et al. (2020). 'AI in medical imaging informatics: current challenges and future directions'. *IEEE journal of biomedical and health informatics*, 24.(7), pp. 1837–1857.

Pang, S. and Yang, X. (2016). 'Deep convolutional extreme learning machine and its application in handwritten digit classification'. *Computational intelligence and neuroscience*, 2016.(1), p. 3049632.

Paul, L. and Talukder, K. H. (2023). 'Blindness Risk Prediction caused by Diabetic Retinopathy from Retinal Image'. *2023 International Conference on Electrical, Computer and Communication Engineering (ECCE)*. IEEE, pp. 1–6.

Pawar, S. P. and Talbar, S. N. (2022). 'Two-Stage Hybrid Approach of Deep Learning Networks for Interstitial Lung Disease Classification'. *BioMed Research International*, 2022.(1), p. 7340902.

Peng, H. and Long, F. (2001). ‘A Bayesian learning algorithm of discrete variables for automatically mining irregular features of pattern images’. *Proceedings of the Second International Conference on Multimedia Data Mining*, pp. 87–93.

Pleiss, G., Chen, D., Huang, G., Li, T., Maaten, L. van der, and Weinberger, K. Q. (2017). ‘Memory-efficient implementation of densenets’. *arXiv preprint arXiv:1707.06990*,

Pratella, D., Ait-El-Mkadem Saadi, S., Bannwarth, S., Paquis-Fluckinger, V., and Bottini, S. (2021). ‘A survey of autoencoder algorithms to pave the diagnosis of rare diseases’. *International journal of molecular sciences*, 22.(19), p. 10891.

Qaid, T. S., Mazaar, H., Al-Shamri, M. Y. H., Alqahtani, M. S., Raweh, A. A., and Alakwaa, W. (2021). ‘Hybrid deep-learning and machine-learning models for predicting COVID-19’. *Computational Intelligence and Neuroscience*, 2021.

Qin, X. and Wang, Z. (2019). ‘Nasnet: A neuron attention stage-by-stage net for single image deraining’. *arXiv preprint arXiv:1912.03151*,

Raaj, R. S. (2023). ‘Breast cancer detection and diagnosis using hybrid deep learning architecture’. *Biomedical Signal Processing and Control*, 82, p. 104558.

Rahmani, A. M. et al. (2021). ‘Machine learning (ML) in medicine: Review, applications, and challenges’. *Mathematics*, 9.(22), p. 2970.

Rajesh, G, Raajini, X. M., Sagayam, K. M., and Dang, H. (2020). ‘A statistical approach for high order epistasis interaction detection for prediction of diabetic macular edema’. *Informatics in Medicine Unlocked*, 20, p. 100362.

Ramraj, S., Nagamalai, D, Pandian, S, and Vimala, J (2016). ‘Experimenting XGBoost algorithm for prediction and classification of different datasets’. *International Journal of Control Theory and Applications*, 9.(40), pp. 651–662.

Rasti, R. et al. (2020). ‘Deep learning-based single-shot prediction of differential effects of anti-VEGF treatment in patients with diabetic macular edema’. *Biomedical optics express*, 11.(2), pp. 1139–1152.

Ravì, D. et al. (2016). ‘Deep learning for health informatics’. *IEEE journal of biomedical and health informatics*, 21.(1), pp. 4–21.

Razzak, M. I., Naz, S., and Zaib, A. (2018). ‘Deep learning for medical image processing: Overview, challenges and the future’. *Classification in BioApps*, pp. 323–350.

Rokh, B., Azarpeyvand, A., and Khanteymoori, A. (2023). ‘A comprehensive survey on model quantization for deep neural networks in image classification’. *ACM Transactions on Intelligent Systems and Technology*, 14.(6), pp. 1–50.

Ronneberger, O., Fischer, P., and Brox, T. (2015). ‘U-net: Convolutional networks for biomedical image segmentation’. *International Conference on Medical image computing and computer-assisted intervention*. Springer, Cham, pp. 234–241.

Rundo, L. et al. (2019). ‘HaraliCU: GPU-powered Haralick feature extraction on medical images exploiting the full dynamics of gray-scale levels’. *International Conference on Parallel Computing Technologies*. Springer, pp. 304–318.

Rundo, L. et al. (2021). ‘A CUDA-powered method for the feature extraction and unsupervised analysis of medical images’. *The Journal of Supercomputing*, 77.(8), pp. 8514–8531.

Samee, N. A. et al. (2022). ‘Classification framework for medical diagnosis of brain tumor with an effective hybrid transfer learning model’. *Diagnostics*, 12.(10), p. 2541.

Savchenko, A. V. (2019). ‘Probabilistic neural network with complex exponential activation functions in image recognition’. *IEEE transactions on neural networks and learning systems*, 31.(2), pp. 651–660.

Sevinç, E. (2022). ‘An empowered AdaBoost algorithm implementation: A COVID-19 dataset study’. *Computers & Industrial Engineering*, 165, p. 107912.

Shakeri, E., Crump, T., Weis, E., Mohammed, E., Souza, R., and Far, B. (2023). 'Explaining eye diseases detected by machine learning using shap: A case study of diabetic retinopathy and choroidal nevus'. *SN Computer Science*, 4.(5), p. 433.

Sharma, S. and Guleria, K. (2023a). 'A Deep Learning-based model for the Detection of Pneumonia from Chest X-Ray Images using VGG-16 and Neural Networks'. *Procedia Computer Science*, 218, pp. 357–366.

Sharma, S. and Guleria, K. (2023b). 'A deep learning model for early prediction of Pneumonia using VGG19 and neural networks'. *Mobile Radio Communications and 5G Networks: Proceedings of Third MRCN 2022*. Springer Nature Singapore, pp. 597–612.

Sharma, S. and Guleria, K. (2023c). 'A systematic literature review on deep learning approaches for Pneumonia detection using chest X-ray images'. *Multimedia Tools and Applications*, pp. 1–51.

Sharp, G. et al. (2014). 'Vision 20/20: perspectives on automated image segmentation for radiotherapy'. *Medical physics*, 41.(5), p. 050902.

Shi, H., Lu, L., Yin, M., Zhong, C., and Yang, F. (2023). 'Joint few-shot registration and segmentation self-training of 3D medical images'. *Biomedical Signal Processing and Control*, 80, p. 104294.

Shimpi, J. K. and Shanmugam, P. (2023). 'Multiclass Adaptive Boosting Approach for Diabetic Retinopathy Prediction Using Diabetic Retinal Images'. *Traitement du Signal*, 40.(3).

Shinde, P. P. and Shah, S. (2018). 'A review of machine learning and deep learning applications'. *2018 Fourth international conference on computing communication control and automation (ICCUBEA)*. IEEE, pp. 1–6.

Siddique, N., Paheding, S., Elkin, C. P., and Devabhaktuni, V. (2021). 'U-net and its variants for medical image segmentation: A review of theory and applications'. *IEEE access*, 9, pp. 82031–82057.

Simonyan, K. and Zisserman, A. (2014). 'Very deep convolutional networks for large-scale image recognition'. *arXiv preprint arXiv:1409.1556*,

Siradjuddin, I. A. and Muntasa, A. (2021). 'Faster region-based convolutional neural network for mask face detection'. *2021 5th International Conference on Informatics and Computational Sciences (ICICoS)*. IEEE, pp. 154–159.

Siuly, S. and Zhang, Y. (2016). 'Medical big data: neurological diseases diagnosis through medical data analysis'. *Data Science and Engineering*, 1, pp. 54–64.

Sohn, I. (2021). 'Deep belief network based intrusion detection techniques: A survey'. *Expert Systems with Applications*, 167, p. 114170.

Srikantamurthy, M. M., Rallabandi, V. S., Dudekula, D. B., Natarajan, S., and Park, J. (2023). 'Classification of benign and malignant subtypes of breast cancer histopathology imaging using hybrid CNN-LSTM based transfer learning'. *BMC Medical Imaging*, 23.(1), p. 19.

Su, Z. et al. (2021). 'Pixel difference networks for efficient edge detection'. *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 5117–5127.

Sun, L., Zhang, L., and Zhang, D. (2019). 'Multi-atlas based methods in brain MR image segmentation'. *Chinese Medical Sciences Journal*, 34.(2), pp. 110–119.

Sun, Y., Wang, X., and Tang, X. (2013). 'Hybrid deep learning for face verification'. *Proceedings of the IEEE international conference on computer vision*, pp. 1489–1496.

Sun, Z., Yang, D., Tang, Z., Ng, D. S., and Cheung, C. Y. (2021). 'Optical coherence tomography angiography in diabetic retinopathy: an updated review'. *Eye*, 35.(1), pp. 149–161.

Sunkari, S., Sangam, A., Suchetha, M, Raman, R., Rajalakshmi, R., Tamilselvi, S, et al. (2024). 'A refined ResNet18 architecture with Swish activation function for Diabetic Retinopathy classification'. *Biomedical Signal Processing and Control*, 88, p. 105630.

Surya, J., Kashyap, H., Nadig, R. R., and Raman, R. (2023). 'Developing a Risk Stratification Model Based on Machine Learning for Targeted Screening of Diabetic Retinopathy in the Indian Population'. *Cureus*, 15.(9).

Syed, S. R. and MA, S. D. (2023). 'A diagnosis model for detection and classification of diabetic retinopathy using deep learning'. *Network Modeling Analysis in Health Informatics and Bioinformatics*, 12.(1), p. 37.

Szegedy, C., Ioffe, S., Vanhoucke, V., and Alemi, A. (2017). 'Inception-v4, inception-resnet and the impact of residual connections on learning'. *Proceedings of the AAAI conference on artificial intelligence*. Vol. 31. 1.

Szegedy, C. et al. (2015). 'Going deeper with convolutions'. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–9.

Tajbakhsh, N. and Suzuki, K. (2018). 'A comparative study of modern machine learning approaches for focal lesion detection and classification in medical images: BoVW, CNN and MTANN'. *Artificial Intelligence in Decision Support Systems for Diagnosis in Medical Imaging*, pp. 31–58.

Tajbakhsh, N. et al. (2016). 'Convolutional neural networks for medical image analysis: Full training or fine tuning?' *IEEE transactions on medical imaging*, 35.(5), pp. 1299–1312.

Tan, M. and Le, Q. V. (2019). 'Efficientnet: Rethinking model scaling for convolutional neural networks'. *International conference on machine learning*. PMLR, pp. 6105–6114.

Tan, M. and Le, Q. V. (2021). 'Efficientnetv2: Smaller models and faster training'. *International conference on machine learning*. PMLR, pp. 10096–10106.

Tan, M., Pang, R., and Le, Q. V. (2020). 'Efficientdet: Scalable and efficient object detection'. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10781–10790.

Tang, F. et al. (2021). ‘A Multitask Deep-Learning system to classify diabetic macular edema for different optical coherence tomography devices: a multicenter analysis’. *Diabetes Care*, 44.(9), pp. 2078–2088.

Tanlikesmath (2019). *Diabetic Retinopathy Resized*. https://www.kaggle.com/datasets/tanlikesmath/diabetic-retinopathy-resized. Accessed: April 2024.

Thanki, R. (2023). ‘A deep neural network and machine learning approach for retinal fundus image classification’. *Healthcare Analytics*, 3, p. 100140.

Ting, D. S. W. et al. (2019). ‘Artificial intelligence and deep learning in ophthalmology’. *British Journal of Ophthalmology*, 103.(2), pp. 167–175.

Trombini, M., Solarna, D., Moser, G., and Dellepiane, S. (2023). ‘A goal-driven unsupervised image segmentation method combining graph-based processing and Markov random fields’. *Pattern Recognition*, 134, p. 109082.

Tsai, H.-Y., Zhang, H., Hung, C.-L., and Min, G. (2017). ‘GPU-accelerated features extraction from magnetic resonance images’. *IEEE Access*, 5, pp. 22634–22646.

Tsiknakis, N. et al. (2021). ‘Deep learning for diabetic retinopathy detection and classification based on fundus images: A review’. *Computers in biology and medicine*, 135, p. 104599.

Umamaheswari, C, Bhavani, R, and Sikamani, D. K. T. (2018). ‘Texture and Color Feature Extraction from Ceramic Tiles for Various Flaws Detection Classification’. *International Journal on Future Revolution in Computer Science & Communication Engineering*, 4.(1), pp. 169–179.

Usman, T. M., Saheed, Y. K., Ignace, D., and Nsang, A. (2023). ‘Diabetic retinopathy detection using principal component analysis multi-label feature extraction and classification’. *International Journal of Cognitive Computing in Engineering*, 4, pp. 78–88.

Varadarajan, A. V. et al. (2020). ‘Predicting optical coherence tomography-derived diabetic macular edema grades from fundus photographs using deep learning’. *Nature communications*, 11.(1), p. 130.

Vasudeva, R and Chandrashekara, S. (2023). ‘An Image Classification and Retrieval Hybrid Model for Larger Healthcare Datasets using Deep Learning’. *Indian Journal of Science and Technology*, 16.(35), pp. 2796–2806.

Verma, B., McLeod, P., and Klevansky, A. (2009). ‘A novel soft cluster neural network for the classification of suspicious areas in digital mammograms’. *Pattern Recognition*, 42.(9), pp. 1845–1852.

Vetrithangam, D, SATVE, P. P., KUMAR, J. R. R., Anitha, P, Vidhya, S, and SAINI, A. K. (2023). ‘prediction of pneumonia disease from x-ray images using a modified resnet152v2 deep learning model’. *Journal of Theoretical and Applied Information Technology*, 101.(17).

Wahab Sait, A. R. (2023). ‘A Lightweight Diabetic Retinopathy Detection Model Using a Deep-Learning Technique’. *Diagnostics*, 13.(19), p. 3120.

Wang, J. et al. (2020). ‘Deep high-resolution representation learning for visual recognition’. *IEEE transactions on pattern analysis and machine intelligence*, 43.(10), pp. 3349–3364.

Wang, J., Li, X., and Cheng, Y. (2023). ‘Towards an extended EfficientNet-based U-Net framework for joint optic disc and cup segmentation in the fundus image’. *Biomedical Signal Processing and Control*, 85, p. 104906.

Wang, S., Zhang, Y., Guo, C., Wang, C., Ji, H., and Liu, Z. (2021). ‘A crop image segmentation and extraction algorithm based on mask RCNN’. *Entropy*, 23.(9), p. 1160.

Wieser, W., Biedermann, B. R., Klein, T., Eigenwillig, C. M., and Huber, R. (2010). ‘Multi-megahertz OCT: High quality 3D imaging at 20 million A-scans and 4.5 GVoxels per second’. *Optics express*, 18.(14), pp. 14685–14704.

Xia, L. et al. (2022). '3D vessel-like structure segmentation in medical images by an edge-reinforced network'. *Medical Image Analysis*, 82, p. 102581.

Xiao, K., Liang, A. L., Guan, H. B., and Hassanien, A. E. (2013). 'Extraction and application of deformation-based feature in medical images'. *Neurocomputing*, 120, pp. 177–184.

Xiao, Z., Ding, Y., Lan, T., Zhang, C., Luo, C., and Qin, Z. (2017). 'Brain MR image classification for Alzheimer's disease diagnosis based on multifeature fusion'. *Computational and mathematical methods in medicine*, 2017.

Xie, H., Fu, C., Zheng, X., Zheng, Y., Sham, C.-W., and Wang, X. (2023a). 'Adversarial co-training for semantic segmentation over medical images'. *Computers in biology and medicine*, 157, p. 106736.

Xie, L. et al. (2023b). 'Deep label fusion: A generalizable hybrid multi-atlas and deep convolutional neural network for medical image segmentation'. *Medical image analysis*, 83, p. 102683.

Xie, Q., Luong, M.-T., Hovy, E., and Le, Q. V. (2020). 'Self-training with noisy student improves imagenet classification'. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10687–10698.

Xie, S., Girshick, R., Dollár, P., Tu, Z., and He, K. (2017). 'Aggregated residual transformations for deep neural networks'. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1492–1500.

Xie, X., Pan, X., Zhang, W., and An, J. (2022). 'A context hierarchical integrated network for medical image segmentation'. *Computers and Electrical Engineering*, 101, p. 108029.

Xu, F. et al. (2022). 'Prediction of the short-term therapeutic effect of anti-VEGF therapy for diabetic macular edema using a generative adversarial network with OCT images'. *Journal of Clinical Medicine*, 11.(10), p. 2878.

Xu, W., Fu, Y.-L., and Zhu, D. (2023). 'ResNet and its application to medical image processing: Research progress and challenges'. *Computer Methods and Programs in Biomedicine*, 240, p. 107660.

Yadav, S. S. and Jadhav, S. M. (2019). 'Deep convolutional neural network based medical image classification for disease diagnosis'. *Journal of Big data*, 6.(1), pp. 1–18.

Yala, A., Lehman, C., Schuster, T., Portnoi, T., and Barzilay, R. (2019). 'A deep learning mammography-based model for improved breast cancer risk prediction'. *Radiology*, 292.(1), pp. 60–66.

Yamashita, R., Nishio, M., Do, R. K. G., and Togashi, K. (2018). 'Convolutional neural networks: an overview and application in radiology'. *Insights into imaging*, 9, pp. 611–629.

Yang, C, Lan, H, Gao, F, and Gao, F (2020). 'Deep learning for photoacoustic imaging: A survey'. *arXiv preprint arXiv:2008.04221*,

Yang, Y. et al. (2021). 'A comparative analysis of eleven neural networks architectures for small datasets of lung images of COVID-19 patients toward improved clinical decisions'. *Computers in Biology and Medicine*, 139, p. 104887.

Yasashvini, R, Panjanathan, R, Graceline, J. S., and Jani Anbarasi, L (2022). 'Diabetic retinopathy classification using CNN and hybrid deep convolutional neural networks'. *Symmetry*, 14.(9), p. 1932.

Yi, R., Tang, L., Tian, Y., Liu, J., and Wu, Z. (2023). 'Identification and classification of Pneumonia disease using a deep learning-based intelligent computational framework'. *Neural Computing and Applications*, 35.(20), pp. 14473–14486.

You, A. et al. (2022). 'Application of generative adversarial networks (GAN) for ophthalmology image domains: a survey'. *Eye and Vision*, 9.(1), p. 6.

Yousaf, F., Iqbal, S., Fatima, N., Kousar, T., and Rahim, M. S. M. (2023). 'Multi-class disease detection using deep learning and human brain medical imaging'. *Biomedical Signal Processing and Control*, 85, p. 104875.

Zewail, R. and Hag-ElSafi, A. (2017). 'Appearance-based Salient Features Extraction in Medical Images Using Sparse Contourlet-based Representation'. *International Journal of Image, Graphics and Signal Processing*, 9.(9), p. 1.

Zhang, C., Lu, W., Wu, J., Ni, C., and Wang, H. (2024). 'SegNet Network Architecture for Deep Learning Image Segmentation and Its Integrated Applications and Prospects'. *Academic Journal of Science and Technology*, 9.(2), pp. 224–229.

Zhang, J. and Feng, Z. (2019). 'Inception DenseNet With Hybrid Activations For Image Classification'. *2019 6th International Conference on Systems and Informatics (ICSAI)*. IEEE, pp. 1295–1301.

Zhang, S., Li, Z., Zhou, H.-Y., Ma, J., and Yu, Y. (2023). 'Advancing 3D medical image analysis with variable dimension transform based supervised 3D pre-training'. *Neurocomputing*, 529, pp. 11–22.

Zhang, X., Zhou, X., Lin, M., and Sun, J. (2018). 'Shufflenet: An extremely efficient convolutional neural network for mobile devices'. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6848–6856.

Zhang, X. et al. (2021). 'Automated detection of severe diabetic retinopathy using deep learning method'. *Graefe's Archive for Clinical and Experimental Ophthalmology*, pp. 1–8.

Zhang, Y. et al. (2022). 'Prediction of visual acuity after anti-VEGF therapy in diabetic macular edema by machine learning'. *Journal of Diabetes Research*,

Zhao, P., Li, C., Rahaman, M. M., and Xu, H. (2022). 'A comparative study of deep learning classification methods on a small environmental microorganism image dataset (EMDS-6):

from convolutional neural networks to visual transformers'. *Frontiers in Microbiology*, 13, p. 792166.

Zhao, T., Hoffman, J., McNitt-Gray, M., and Ruan, D. (2019). 'Ultra-low-dose CT image denoising using modified BM3D scheme tailored to data statistics'. *Medical physics*, 46.(1), pp. 190–198.

Zhao, X., Wu, Y., Song, G., Li, Z., Zhang, Y., and Fan, Y. (2018). 'A deep learning model integrating FCNNs and CRFs for brain tumor segmentation'. *Medical image analysis*, 43, pp. 98–111.

Zhao, Y., Wang, X., Che, T., Bao, G., and Li, S. (2023). 'Multi-task deep learning for medical image computing and analysis: A review'. *Computers in Biology and Medicine*, 153, p. 106496.

Zheng, J., Liu, H., Feng, Y., Xu, J., and Zhao, L. (2023). 'CASF-Net: Cross-attention and cross-scale fusion network for medical image segmentation'. *Computer Methods and Programs in Biomedicine*, 229, p. 107307.

Zhong, G., Ding, W., Chen, L., Wang, Y., and Yu, Y.-F. (2023). 'Multi-scale attention generative adversarial network for medical image enhancement'. *IEEE Transactions on Emerging Topics in Computational Intelligence*,

Zhou, Q., Wang, Q., Bao, Y., Kong, L., Jin, X., and Ou, W. (2022). 'LAEDNet: a lightweight attention encoder–decoder network for ultrasound medical image segmentation'. *Computers and Electrical Engineering*, 99, p. 107777.

Zhu, F., Wang, S., Li, D., and Li, Q. (2023). 'Similarity attention-based CNN for robust 3D medical image registration'. *Biomedical Signal Processing and Control*, 81, p. 104403.