Article

EMM-LC Fusion: Enhanced Multimodal Fusion for Lung Cancer Classification

James Barrett and Thiago Viana *

Cyber and Technical Computing, University of Gloucestershire, Cheltenham GL50 2RH, UK
* Correspondence: tviana1@glos.ac.uk

Abstract: Lung cancer (LC) is the most common cause of cancer-related deaths in the UK due to delayed diagnosis. The existing literature establishes a variety of factors which contribute to this, including the misjudgement of anatomical structure by doctors and radiologists. This study set out to develop a solution which utilises multiple modalities in order to detect the presence of LC. A review of the existing literature established failings within methods to exploit rich intermediate feature representations, such that it can capture complex multimodal associations between heterogenous data sources. The methodological approach involved the development of a novel machine learning (ML) model to facilitate quantitative analysis. The proposed solution, named EMM-LC Fusion, extracts intermediate features from a pre-trained modified AlignedXception model and concatenates these with linearly inflated features of Clinical Data Elements (CDE). The implementation was evaluated and compared against existing literature using F1 score, average precision (AP), and area under curve (AUC) as metrics. The findings presented in this study show a statistically significant improvement ($p < 0.05$) upon the previous fusion method, with an increase in F-Score from 0.402 to 0.508. The significance of this establishes that the extraction of intermediate features produces a fertile environment for the detection of intermodal relationships for the task of LC classification. This research also provides an architecture to facilitate the future implementation of alternative biomarkers for lung cancer, one of the acknowledged limitations of this study.

Keywords: deep learning; lung cancer; machine learning; multimodal fusion

Citation: Barrett, J.; Viana, T. EMM-LC Fusion: Enhanced Multimodal Fusion for Lung Cancer Classification. *AI* **2022**, *3*, 659–682. https://doi.org/10.3390/ai3030038

Academic Editor: Hiroyuki Yoshida

Received: 29 June 2022
Accepted: 3 August 2022
Published: 9 August 2022

1. Introduction

It has been recognised that lung cancer (LC) is one of the most common causes of cancer related deaths in the UK, and that it also accounts for 12.4% of all diagnosed cancers worldwide [1,2]. The diagnosis of LC commonly follows the identification of malignant nodule(s) within a low-dose CT scan [3]. However, only approximately 16.2% of people survive for more than 5 years after diagnosis [1]. Therefore, detecting malignant tumours within the lung parenchyma or bronchi in a cancer's early stages can increase the probability of effective treatment [2,4,5]. The pathophysiology of lung cancer is yet to be fully understood; however, academia within healthcare hypothesise that a continued exposure to carcinogens, such as cigarette smoke, leads to genetic mutations and impacts protein synthesis, ultimately causing LC [2]. The symptoms of this disease can include a persistent cough, weight loss, dyspnea, and chest pain; a patient presenting with these symptoms would be referred for a computed tomography (CT) scan to provide a diagnosis [6].

As medical opinion is still highly regarded, the emergence of machine learning (ML) and new techniques must work in conjunction with radiologists and doctors in order to enhance productivity and precision [7]. The combination of modalities, such as images and textual information, lends itself to the medical field. This study presents a fusion architecture designed to exploit the intermediate and intermodal relationships of CT scans and clinical data elements (CDEs), which include symptoms, clinician observations, and cancer history, in order to provide a more accurate classification of LC.

Current Methods

The traditional pathway for LC diagnosis combines a multitude of screening techniques including chest X-Ray scans, CT scans, and clinical biomarkers e.g., family history and smoking habits. Ref. [8] described three causes for the misdiagnosis of LC, namely observer error, tumour characteristics, and technical considerations. Indeed, 90% of missed LC cases occur from chest X-ray scans [8]. This highlights a fundamental problem with the current diagnostic methods for detecting LC. However, there is minimal discussion pertaining to the future direction of LC diagnosis regarding emerging technologies, such as ML. More recently, this discussion was extended with regard to ML, but [5] expressed similar views upon the current limitations; marking cancerous cells is difficult due to the variance of intensity in CT scans and misjudgement of anatomical structure by doctors and radiologists. Following these studies, and providing an impetus for future research, [9] commented upon the use of emerging biomarkers, such as volatile organic compounds, sputum, metabolomics, and genetics, in conjunction with the application of ML.

2. Literature Review

2.1. Machine Learning

Machine learning has been defined as "the extraction of knowledge from data" [7]. Within healthcare, various algorithms already outperform doctors and radiologists [10]. However, [11] addressed a number of ethical considerations and, although this study establishes a place for ML in healthcare, it also acknowledges some of the limitations that are present. Nevertheless, the discussion of these challenges lacks depth regarding causes and/or solutions. More recently, [12] extended this discussion, articulating the significance of unconscious bias, overreliance, and interpretability. Ultimately, ML solutions must be developed with respect to these limitations, as acknowledged within the evaluation of the findings presented in this study.

2.1.1. Deep Learning

Deep learning (DL) algorithms learn feature representations with multiple levels of abstraction [13]. Addressing this technology from the perspective of healthcare, the application ranges from the diagnosis of Alzheimer's to the prognosis of COVID-19 [14]. Many solutions adopt pre-defined networks and, in some cases, pre-trained weights to encourage faster convergence. Ref. [15] observed that tuning pre-trained weights can be very effective, allowing the network to adapt to the classification problem. This highlights the potential for utilising a pre-trained network for the task of LC classification. However, it does not acknowledge the potential limitations of this method. Ref. [16] articulated that their method was more effective when trained from scratch as opposed to fine-tuning VGG16 weights, as the gap between natural and pathological images was too large. This has been considered during the development of this study and, thus, pre-trained weights were only frozen where the distribution between training datasets were equal.

Ref. [17] stated that pre-processing methods improve upon the accuracy of healthcare predictions. This implies that pre-processing methods will improve LC classification. However, their study only investigated the effect of this on Type II diabetes and, thus, the assumption cannot be made that this will apply to LC. Providing a more suitable analysis, [18] proposed a DL neural network which utilised pre-processing steps to reduce the noise and dimensionality of the data, which subsequently improved the results. The critical comparison of these studies underlines the general consensus that applying considered pre-processing steps serves for a better feature representation and, thus, a more accurate model.

Ultimately, [19] established a specific set of pre-processing steps for the task of malignant nodule detection in CT imaging. This method was recognised by many scholars due to the exceptional results which it produced. However, for some of these steps, minimal justification was provided, specifically a number of arbitrary values which determine the inclusive nature of the applied mask. As a result, these steps may lack repeatability for alternative LC datasets whereby the distribution differs from that of the data used by [19]. Despite this, [20] reinforces this implementation, albeit on the same dataset. These steps have been used within the data pipeline of the proposed solution presented in Section 3. However, the implementation has been discussed with respect to the lack of justification provided by [19].

Following the natural progression of data manipulation, [19] also implemented a number of data augmentation techniques specific to malignant nodule detection in LC. Included within these steps was rotational and horizontal flip augmentation, thus, preserving the distribution of the data and improving model generalisation. Evidence presented by [21] confirms that over 60% of prior medical studies implemented basic augmentation techniques, thus, validating the decision to include the techniques proposed by [19] within the developed solution.

When screening at-risk patients for LC, CT is considered as one of the key methods [22]. As a result, convolution neural networks (CNNs) have been extensively researched in regard to the feature extraction of medical images [23]. In 2018, [24] achieved a sensitivity and specificity of 0.87 and 0.991, respectively, using a deep convolution neural network to detect LC nodules in CT scan images from the KDSB17 dataset. Acknowledged as a limitation, their method downsized the input images to 128 × 128 from 512 × 512 due to hardware constraints, which could have led to the loss of important features [24]. However, [25] extended this approach by implementing transfer learning with AlexNet, a pretrained variation of a CNN, achieving an accuracy of 96%. Despite this, the authors did not address what data they used, which negatively affects the validity and repeatability of this study. However, aligning with their predominant use, this study will adopt a CNN architecture for the feature extraction of CT scan images.

2.1.2. Multimodal Learning

Modality has been defined as the way in which something happens or is experienced [26]. Multiple modalities are inherently present within the realm of medicine [27] and, thus, multimodal learning can be highly effective within DL and healthcare, improving the accuracy, sensitivity, and specificity of some classification problems [28,29]. This establishes that the utilisation of multimodality can result in robust and accurate predictions. As seen within the current diagnosis pathway, the method of diagnosis includes the type, size and location, and overall clinical status of the patient [30]. However, scholars have commented upon the difficulty of exploiting Supplementary Data as opposed to just complementary data within DL multimodal models [26]. In contrast, [31] identified that the utilisation of multiple modalities provides superior results regarding the effect of Supplementary Data; a characteristic which is sought to be applied to the task of LC classification.

The abundance of fusion techniques has accelerated the growth of multimodal ML. Such techniques include joint and co-ordinated representations, as in [26]. Ref. [32] provided a critical comparison between these two techniques and presented equal arguments for both approaches. However, this balanced discussion is not reflected in the literature, as a consequence of the lack of ability to interpret more than two modalities within coordinated representations. In a more recent study, [33] presented a multimodal architecture, projecting intermediate features into a joint space for classification. This architecture enriched the feature representations with unimodal models including a CNN and stacked denoising autoencoders. This outperformed other methods, but only for the task of Alzheimer's disease (AD) classification. Thus, by taking inspiration from this approach, the model presented in this paper adopts similar steps.

In respect to LC, several modalities have been identified as good indicators, including CT/PET/MRI scans [22], clinical/metabolomic biomarkers [34], and volatile organic compounds [35]. However, current screening trials, such as NLST, oversimplify LC risk prediction, reducing the cost efficacy due to the primary use of low-dose CT scans. It has been recognised that the pre-test probability can be improved if other clinical biomarkers, such as cancer history, history of other diseases, and asbestos exposure, are used [36]. This provides evidence on the suitability of multimodal learning for LC classification. However, in contrast to AD, there is a lack of multimodal datasets suitable for the inclusion of these biomarkers. Ref. [20] utilised the NLST and VLSP datasets and applied a co-learning approach, achieving an AUC (area under curve) of 0.91. However, it was articulated that their approach could be improved with additional CDEs, thus, providing an impetus for the implementation of a novel dataset.

The Lung Cancer Screening (LUCAS) dataset, published in 2020, provides 830 samples, including 76 CDEs and CT scans for each patient. This dataset was also presented alongside a benchmark ML architecture with an F1 and AUC score of 0.25 and 0.702, respectively [37]. More recently, the SAMA model improved this to an F1 score of 0.341 with a standard deviation of 0.058 [38]. The additional CDE data, despite the smaller sample size, provides a solution to the aforementioned limitation expressed by [20]. However, the data presented by [37] lacked clarity regarding the categorical nature of the CDEs. Despite the critical comparison against more established datasets, [37] offers an alternative dataset to provide an incentive for the development of multimodal ML for LC classification.

While the aforementioned approach proposed by [33] yielded good results for AD classification, the implementation for LC must be further validated. Ref. [39] identified that simple concatenation of the output features of unimodal models may lack the depth needed to exploit intermodal interactions. This implies that the implementation presented by [33] may fail to exploit the full potential of multiple modalities. In contrast, [16] extracts intermediate features from a pre-trained CNN model. It was identified that extracting these multi-level features from various layers within a CNN provided a richer feature representation and higher AUC than those purely extracted from the last fully connected layer. This richer fusion technique was more effective at exploiting the complex multimodal associations within heterogenous data, and increased the accuracy from 83.6 to 91.1 [16]. The method presented by [16] provides an approach which can be applied to the task of LC classification, as demonstrated in the subsequent sections of this study.

Although the argument presented by [16] highlights the potential benefits of extracting a richer feature representation from the CT scan, there is a requirement to learn a good feature representation in every modality before information fusion. This establishes a need to increase the dimensionality of the CDEs. Ref. [16] used a denoising autoencoder to achieve this, with an architecture in which the dimension of the encoded layer is greater than the dimension of the input layer. This further questions the validity of the approach used by [33], as the down-sampling from successive layers within a CNN is also a process of information loss [16]. Acknowledging the approaches of both [16] and [33], the solution presented in this study aimed to improve the feature representation of each modality, prior to fusion, for the task of LC classification.

2.2. Summary

Despite the recent successes of multimodal fusion as presented by various authors, the literature review highlights several shortcomings pertaining to the specific features which are utilised for the task of LC classification. It can be observed that current solutions for the task of LC classification only utilise the last layer(s) of a pre-trained feature extractor prior to multimodal fusion. Acknowledging works from other diseases, this approach lacks the depth required to identify intermodal relationships [39]. From an architectural perspective, the specific contribution that this paper aims to provide addresses this gap, concatenating multi-level feature representations from pre-trained networks for the task of multimodal fusion, taking inspiration from works presented by [16].

Further motivating this study, the review of relevant literature brought to light the scarcity of papers which utilise the multimodal dataset presented by [37], whereby a larger foundation of work would support the argument for adopting multimodal learning towards LC classification with the intention of improving existing solutions, such as [19].

3. Materials and Methods

3.1. Research Strategy

The research strategy and selection criteria applied to Section 2 was refined to present the most relevant papers for discussion. Google Scholar and several other repositories including ResearchGate, IEEE, Arxiv, ScienceDirect, and PubMed were used to search for academic literature around the topics of medicine and machine learning, utilising combinations of the following keywords: "Lung Cancer", "Diagnosis", "Machine Learning", "Classification", "Multimodal", and "Fusion". Subsequent literature was constrained by several variables, namely validity, relevance, and accessibility. The outstanding papers were then filtered using the inclusion and exclusion criterion described in Table 1 to formulate a critical discussion which facilitated the process of identifying a gap in the literature.

Table 1. Literature inclusion and exclusion criterion.

Include	Exclude
Deep learning papers published since 2012	Exclude all papers published before 2002
Peer-reviewed Studies	Any thesis lower than master's
Multimodal solutions for other problems	Papers written in a language other than English

3.2. Development

The development of this model explores an alternative method to improve the classification of LC diagnosis by combining techniques identified within the literature review. This section details the tools, datasets and model architecture which have been implemented, providing where needed the justification for each decision and reasons for why alternative methods were not adopted.

3.2.1. Tools/Frameworks

Table 2 describes the tools that have been used throughout the development of this project. Supporting future iterations and the reproducibility of this study, their respective versions have also been provided. Despite many researchers regarding the differences between PyTorch and TensorFlow as personal preference, PyTorch was used, as it facilitated faster and more intuitively pythonic development, an observation also made by [40].

Table 2. Development tools.

Tool		Version	Use
Language	Python	3.9	Develop ML pipeline
Libraries	Nilearn	0.9.0	NifTI image handling
	Pandas	1.4.0	Dataset handling and sanitization
	TorchIO	0.18.76	Data augmentation
	Pytorch	1.11.0	ML network layers
	CUDA	11.3	GPU accessibility
	Scikit-learn	1.0.2	Evaluative functions

3.2.2. Dataset

To facilitate parallel co-learning and to align with the proposed model architecture, heterogeneous data pairings were required. Table 3 lists five common datasets that have been used to develop ML solutions for LC diagnosis. A specific selection criteria was developed to aid this process including the type of data that was available, the number of samples, and its accessibility.

Table 3. Comparison of datasets.

Dataset	Accessibility	CT Scans	Clinical Data	Sample Size	Appearance in Literature *
NLST	1	1	1	26254	1
KDSB17	0	1	0	1397	1
UCI	1	1	0	32	1
LUCAS	1	1	1	830	0
MCL	1	0	1	61	1

* Mentioned or referenced by at least 10 other relevant studies as of January 2022.

In contrast to some of the more prevalent datasets identified within the literature review, the LUCAS dataset provided greater compatibility for the multimodal nature of this research problem. However, there were several limitations which impacted the adoption of this dataset, including the sample size and distribution, characteristics which could cause overfitting and lead to questions about the generalisability of the results. Despite these limitations, there are few studies which develop upon the benchmark network proposed by the authors and, thus, a stimulus to contribute towards the literature is provided. Notwithstanding this, the aforementioned limitations have been acknowledged within the analysis of the results.

Pre-Processing and Data Augmentation

Ref. [20] and previously [41] mirrored the pre-processing steps defined by [19] and, thus, established the validity of this approach and justified its application within this solution. Interestingly, [37] omitted these pre-processing steps within their benchmark model, potentially limiting the performance of their solution. Consequently, applying these aforementioned steps within the ML pipeline provides an improvement upon previous classification tasks on the LUCAS dataset.

Addressing the aforementioned challenge of overfitting, data augmentation should alleviate this and improve generalisation [42]. The implemented steps aimed to reflect the distribution of the original data, utilising random flip and rotation, also aligning with the implementation presented by [19]

3.2.3. Model

Established by [16], multi-layer feature extraction can provide richer feature representations of an input image. To facilitate this and to improve upon previous works, this solution implemented three models, utilising a modular training scheme to extract richer feature representations and to improve the fusion technique previously observed throughout literature. The three models are listed as follows:

1. A Unimodal 3D-CNN classification of LC in CT scans (Modified AlignedXception);
2. A DAE for dimensionality increase in CDEs;
3. A feature fusion network, combining stages 1 and 2.

The two backbone models, including the DAE and modified AlignedXception network, are pre-trained on the same train data, and the respective inputs are forwarded to the feature fusion network. Figure 1 details the overall architecture of this model.

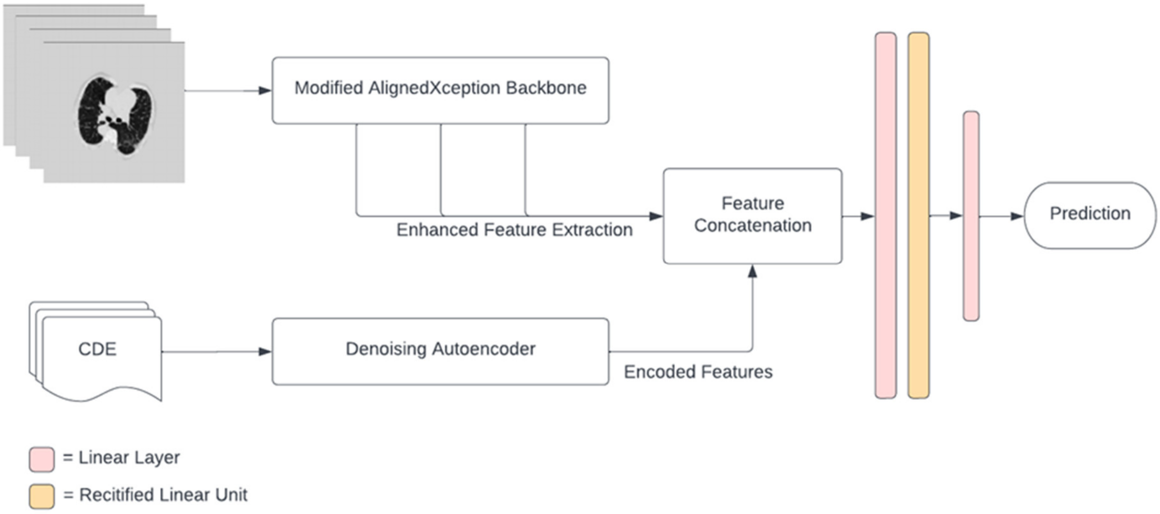

Figure 1. Model architecture schematic.

Denoising Auto Encoder Network

As previously discussed, the motivation for implementing the DAE was to increase the dimensionality and, thus, the significance of CDEs during fusion. This approach was adopted from [16], as it was observed to improve the performance of the fusion model. Replicating the architecture presented by [16] and encouraging the model to learn more generalised representations of the data, noise is added to the input during training using a dropout layer where $p = 0.2$ (Figure 2); this incentivises the model to be robust to missing data and prevents the model from learning an identity function [16]. However, during inference, all data points are fed into the DAE. Interestingly, there is a lack of work in the literature investigating the efficacy of inflating the dimensionality of data using DAEs to improve the fusion of multiple modalities; this observation was also made by [16].

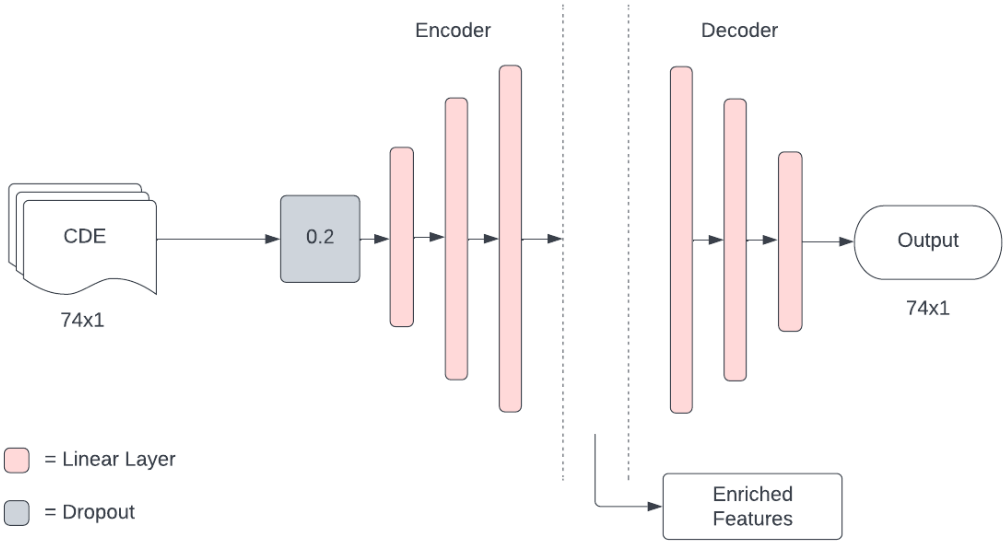

Figure 2. Denoising autoencoder schematic.

To stimulate a discussion upon the optimal architecture of the DAE for this solution, a number of models were developed varying the factor of inflation, where f is the number of features and n is the degree of inflation, such that inflation = f*n. Moreover, an additional model, omitting the DAE, is also implemented to evaluate its true effect on the performance of the model. The results of this are discussed in Section 5.

AlignedXception Model

Although this study implements the approach presented by [37] as a starting point, their original paper lacked depth and clarity regarding the justification for implementing a modified AlignedXception model. Thus, it was important to understand how this model compares to other architectures, such as Inception and VGG, and whether a more appropriate model could have been used.

Evaluating their novel architecture, [43] provided several pieces of evidence which motivated the keeping of this model. Interestingly, it was established that the AlignedXception model could outperform InceptionV3, VGG-16, and ResNet-152 while using a similar number of parameters. This indicates that each layer was able to learn a more effective feature representation and, thus, lends itself to the task of enhanced feature extraction. Additionally, the authors compared the schematic architecture to VGG16 and observed that in some respects they are similar, which aligns with previous studies for using VGG16 as a feature extractor, such as in [16,43]. Figure 3 provides the schematic of the developed AlignedXception architecture.

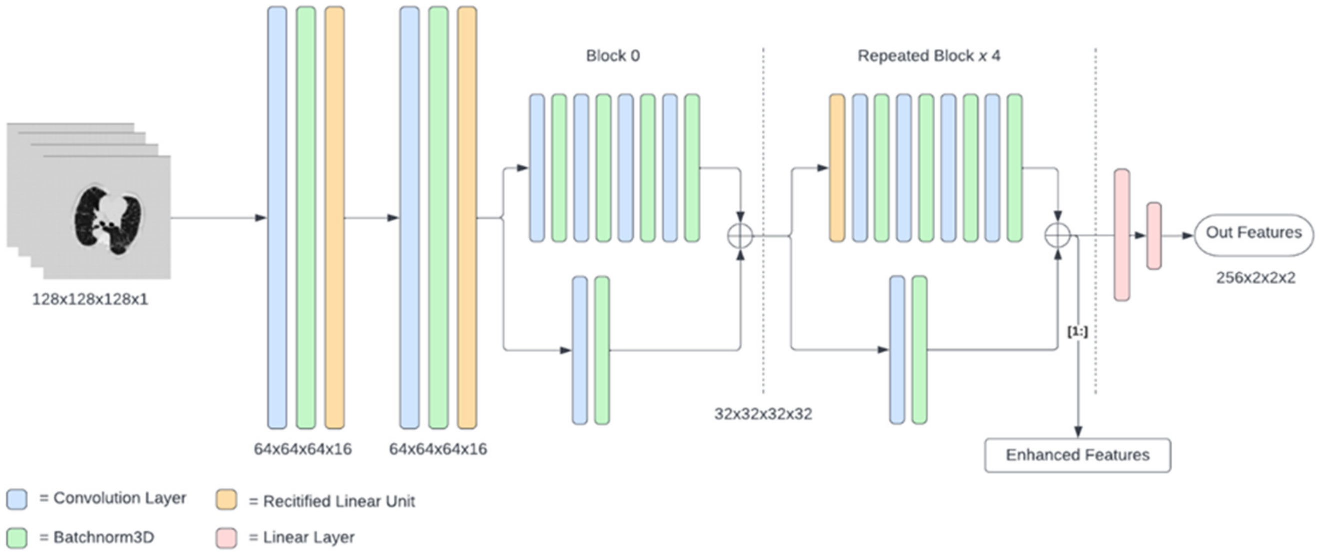

Figure 3. Schematic of the AlignedXception model.

3.3. Evaluation

To facilitate a comprehensive comparison to existing works, this study evaluates a range of model architectures. This included a baseline model replicating the implementation developed by [37] (control), an enhanced CT feature representation with simple linear inflation of CDE (EMM-LC), and an enhanced CT feature representation with DAE enriched CDE (EMM-LC DAE). It is important to acknowledge that the justification for re-implementing the approach presented by [37] is to identify how it performs compared to other solutions on exactly the same hardware in order to reinforce a valid comparison.

A lack of consistency within the evaluation metrics used between studies posed a challenge towards understanding how this novel approach compares. Therefore, to limit the subjective interpretability of these results, this study implements the evaluation metrics used by [37], which includes F1-Score and AUC in addition to average precision (AP) which was used by [38]. By implementing these metrics, subsequent comparative analysis can better substantiate claims that this method is superior, mitigating ambiguity within this study. To support this discussion and to expand on these results, an independent sample *t*-test is utilised to identify if any statistically significant improvements have been made.

In addition to the aforementioned evaluation metrics, sensitivity, specificity, and accuracy have also been implemented, as listed in Table 4. This was necessary in order to stimulate a critical discussion upon its implementation within current healthcare technologies, which commonly use these metrics, and whether it improves upon traditional diagnostic techniques. However, it must be acknowledged that accuracy will not independently form an argument for the adoption of this technique, as it lacks clarity and is significantly affected by the uneven distribution of positive and negative samples. Table 4 highlights this, whereby accuracy is the only metric to combine all values, obscuring specific performance insight.

Table 4. Evaluation metric equations.

Evaluation Metric	Formulae
Sensitivity	$\frac{TP}{(TP + FN)}$
Specificity	$\frac{TN}{(TN + FP)}$
Accuracy	$\frac{TP + TN}{(TP + TN + FP + FN)}$
F1 score	$\frac{TP}{TP + \frac{1}{2}(FP + FN))}$

TP, true positive; *FP*, false positive; *TN*, true negative; *FN*, false negative.

3.4. Summary

The focus of this methodology was to articulate the design implementation of the network architecture and to provide information to support the repeatability of this study. Crucially, it offers justification for the decisions made in regard to its development, and acknowledges ways in which limitations, regarding the evaluation, were mitigated.

4. Implementation

4.1. Pre-Processing

4.1.1. CT Scans

The pre-processing steps defined by [19] are considered by numerous scholars to be a standard technique for pre-processing CT scans for the task of LC detection. Despite the overwhelming adoption of these steps, some aspects lacked clarity, as discussed in Section 2. This ambiguity produced inaccurate results.

To address this limitation, several arbitrary values were altered within the pre-processing to improve the mask extraction of the image. Figure 4 provides evidence to highlight this improvement. Details of these changes have been omitted from the main body of work as it is not the primary focus of this study. By reducing the threshold of certain values, it must be assumed that the number of distracting features increased, a characteristic sought to be mitigated by [19]. The effect of these changes has been addressed in Section 6.

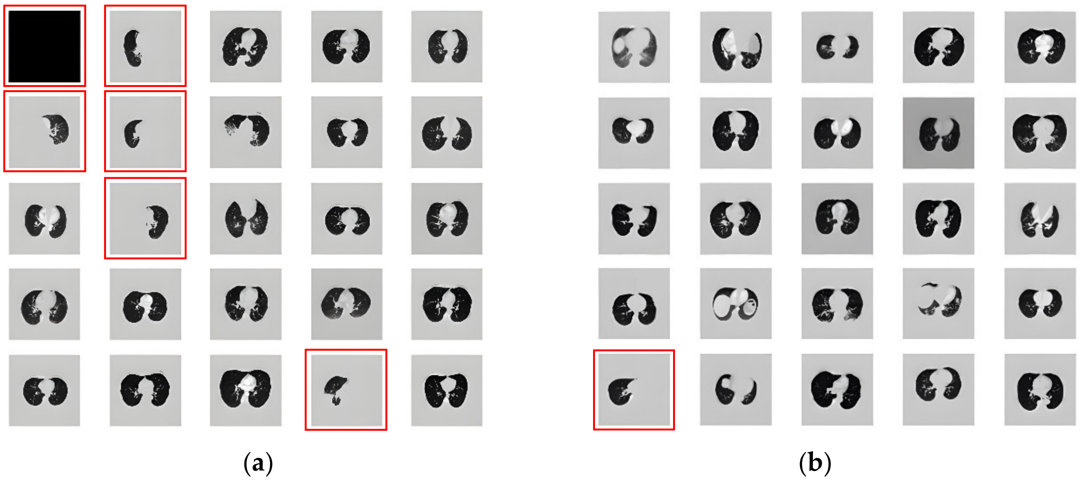

Figure 4. Random sample of pre-processed scans. (**a**) Previous method defined by [19] produces an error rate of 6/25. (**b**) Adapted solution, with new thresholding, produces an error rate of 1/25.

Irrespective of the improvements observed in Figure 4, a total of six scans failed, resulting in either one lung or no lungs being extracted from the scan. It was identified that the original image of these scans differed from the normal distribution, indicating that the scan was taken incorrectly or that there was a technical fault in the scanning procedure. This aligns with the observation made by [8] that errors can be attributed to image quality and or patient positioning/movement during a scan. To mitigate the impact that this had on the training procedure, any scans which failed the above process were removed from the training data in order to prevent the model learning poor features representations. Figure 5 illustrates the pre-processing of a single scan.

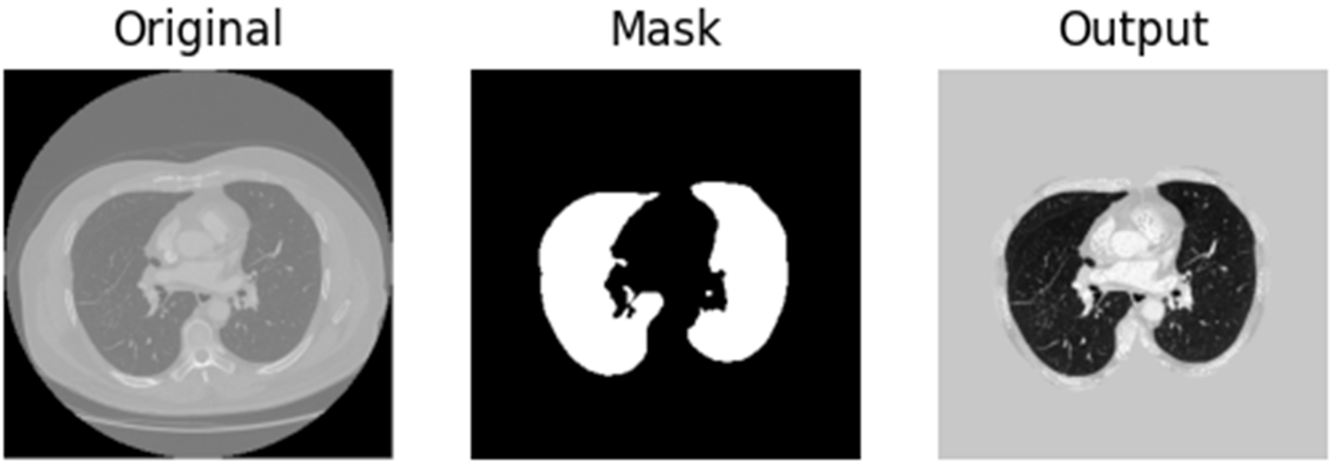

Figure 5. Data pre-processing.

4.1.2. Clinical Data Elements

The adoption of the multivariate dataset proposed by [37] presented a number of challenges which may have contributed to previously low performance metrics. Figure 6 depicts the correlation matrix between all attributes, several of which had no inputs, visually indicated by the missing correlation data which were removed from the dataset. As discussed in Section 2, some data was categorical but was not clarified within the paper and, thus, all values were clipped between 0 and 1.

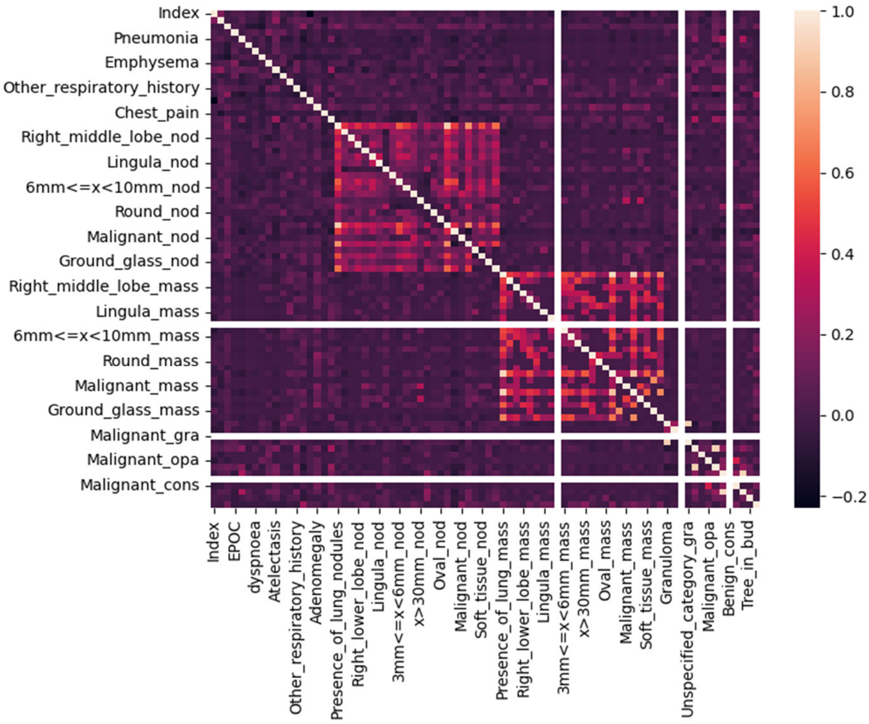

Figure 6. CDE Correlation Matrix.

4.2. Data Augmentation

As discussed in Section 3, the acquired dataset is small and, thus, vulnerable to overfitting. In order to mitigate this and to better generalise to the validation set, data augmentation was applied. Figure 7 illustrates the transformations applied using TorchIO for the 3D tensor manipulations. Specific augmentation techniques included horizontal flip, where $p = 0.5$, ±20° affine transformation, and resizing to 128^3, aligning with the basic augmentation implemented by [19].

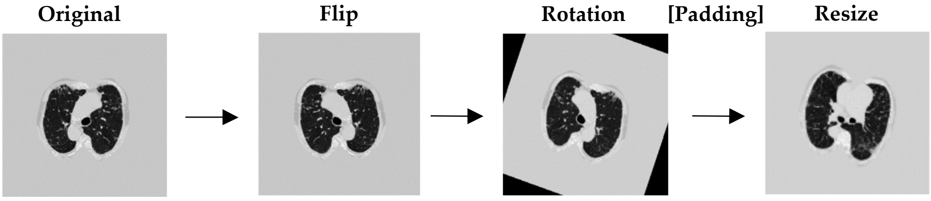

Figure 7. Data augmentation.

To conclude the augmentation steps and to align with the approach presented by [37], weighted sampling was implemented to balance the distribution of classes. However, the limitations of this method were not acknowledged by [37]. Ref. [42] stated that oversampling a minority class can lead to overfitting. During development, this was observed. However, this limitation was alleviated when combined with the other previously discussed augmentation techniques.

4.3. Training

The networks were trained on Google Colab Pro, utilising a high-end graphics processing unit (GPU) and storage; the hardware is detailed in Table 5. Several techniques were implemented in order to reduce training time, prevent overfitting, and enable repeatability, which involved using deterministic methods, larger batch sizes, simulated annealing, and reduced learning rate on plateau; these techniques were mostly employed by [37].

Table 5. Hardware.

Component	Type
GPU	Tesla P100-PCIE-16 GB
CPU	Intel(R) Xeon(R) CPU @ 2.30 GHz
RAM	32 GB
Storage	Google Drive 100 GB

Due to time constraints, it was not feasible to conduct statistical tests for every change made during development in order to justify design choices. Therefore, to validate any improvements made, deterministic methods were invoked. These values aimed to mitigate some indeterministic aspects of the learning phase which could potentially lead to varying results. This validates the improvements observed during testing. However, it must be acknowledged that these methods are not full proof, and some variability can still be sourced between GPU and central processing unit (CPU) extensions or different platforms [44]. For tests which determined the efficacy of the overall approach, such as for the simple multimodal (SMM) and enhanced multimodal 75k (EMM 75k) model, tests for statistical significance were calculated, utilising indeterministic methods.

4.4. Modified AlignedXception Model

4.4.1. Architectural Changes

The complexity of the architecture proposed by [37] was reduced to decrease training time and support smaller input sizes. From a hardware perspective, this freed up memory and consequently facilitated the use of batch normalisation, a technique that required sufficiently large batch size in order to reduce model error [45]. Table 6 details all the architectural changes made.

Table 6. AlignedXception architectural changes.

Change	Original	Proposed
Batch size	13	32
Image size	256^3	128^3
AlignedXception output features	512 × 4 × 4 × 4	256 × 2 × 2 × 2
Convolution input channels	32, 64, 128, 256, 256, 512	16, 32, 64, 128, 128, 256

Acknowledging these changes, a baseline model was re-implemented in order to provide grounds upon which an argument can be formed and, thus, to draw a direct comparison between the new fusion method and previous architectures in the same environment. This baseline model, referred to as SMM, mirrors the fusion method proposed by [37] and adopts the aforementioned architectural changes and data pre-processing/augmentation described in Sections 4.1 and 4.2. As a result, any significant difference against the new method can be directly attributed to the change in fusion technique.

4.4.2. Feature Extraction

The extended part of this architecture which distinguishes it from previous multimodal techniques on this dataset is the explicit extraction of intermediate features from pre-trained networks. The intermediate features are returned in the forward pass of the pre-trained network, consisting of 65,536, 8192, and 2048 features, respectively. Following this extraction, they are concatenated with the features obtained from the CDEs.

4.5. Denoising Autoencoder Model

Providing clarity to this implementation, all aspects described by [16] were included within this solution, including the specific values for dropout layers and L1 regularisation as to prevent the model from learning an identity function. Equation 1 provides clarification for the mean squared error (MSE) loss function with L1 regularisation.

$$L(w) = \frac{1}{N}\sum_{i=1}^{N}(f(x_i; w) - y_i)^2\lambda||w_1||$$

Equation 1. MSE Loss with L1 Regularisation.

Despite this clarity, there was ambiguity in regard to the training parameters, such as the learning rate or accepted performance metrics. However, finding better hyperparameters was outside the scope of this paper, and so it was approximated. The training configuration for this model has been provided in Table 7. Each model was trained for 200 epochs.

Table 7. DAE hyperparameters.

Hyperparameter	Value
Learning rate	0.001
Batch size	32
Lambda L1 regularization	0.001
Epoch	200

Aligning with the training methodology presented by [16] the CNN and DAE model were pre-trained for the use of intermediate feature extraction by freezing their weights. Despite [15] articulating the efficacy of tuning weights, the same training data was used for both models and, as a result, the weights were frozen, as both training datasets had equal distribution. To facilitate this, the back propagation of the gradient required to update the weights was not passed through the pre-trained models. In addition, BN and dropout were disabled by setting the pre-trained models to evaluation mode.

4.6. Summary

Facilitating the validity and repeatability of this study, all aspects of the implementation have been discussed, including the data augmentation, training environments, and hyperparameters. In addition, further clarity has been provided within Appendix A in regard to the implementation of the pre-processing steps.

5. Results

The purpose of this study was to identify and develop a method to better exploit intermediate features for the task of multimodal fusion in LC classification by focusing efforts on improving upon the implementation proposed by [37]. In order to contribute to the implementation of this approach and to validate the findings of this study, additional evaluation metrics were implemented, consisting of F1 score, AUC, and AP, as well as sensitivity and specificity. This provided a clearer understanding into the performance of the proposed solution from both a ML and medical perspective.

5.1. The Efficacy of Pre-Processing and Augmentation

Implementing a simple feature concatenation method in conjunction with the improved data pre-processing and augmentation methods yielded an average F1 score of 0.402, as shown in Table 8. In comparison to the original benchmark model proposed by [37] which achieved 0.25, the mean F1 score of the SMM is an improvement of 60.8%. Improvements are also noted in AUC and AP.

Table 8. Descriptive statistics of SMM.

Model	**F1**		**AUC**		**AP**	
	Mean	**Std**	**Mean**	**Std**	**Mean**	**Std**
LUCAS	0.25	--	0.702	--	--	--
SAMA	0.341	0.058	--	--	0.251	0.061
SMM	**0.402**	**0.04**	**0.843**	**0.036**	**0.419**	**0.074**

The evaluation of the SMM model was taken from five independently trained models to provide validity to these results. The standard deviation has also been provided within Table 8 to support subsequent t-tests to observe any statistically significant improvements that differing fusion methods provide. This model acts as a baseline to compare subsequent developments.

5.2. The effect of utilising intermediate features

As highlighted in Section 2, some scholars expressed the importance of improving the fusion method in order to better exploit intermodal relationships [39]. To explore this concept, intermediate features were extracted from a pre-trained modified AlignedXception model and concatenated with the simple linear features of the CDEs. In total, two additional tests were conducted, varying the number of extracted features. Five repeat tests were conducted for the EMM-LC 75k model to validate the results. The mean and standard deviation have been provided in Table 9.

Table 9. CT scan feature extraction, along with the SMM and EMM-LC 75K averaged results from five independent tests, presenting mean and std.

Model Name	**CT Features**	**F1**		**AUC**		**AP**	
		Mean	**Std**	**Mean**	**Std**	**Mean**	**Std**
SMM-LC	2048	0.402	0.040	0.843	0.036	0.419	0.074
EMM-LC 10k	10,240	0.400		0.876		0.876	
EMM-LC 75k	**75776**	**0.508**	**0.031**	**0.847**	**0.011**	**0.53**	**0.069**

The mean F1 score of 0.508 with a standard deviation of 0.031 significantly improves upon the original SMM fusion method by 26.37%. Figure 8 displays these results, highlighting that the intermediate pre-trained features improve the performance of the multimodal fusion network. In order to identify the statistical significance of this result, a paired t-test with equal variances was applied. This yielded a p-value of 0.0007, confirming the statistical significance to 99.93% confidence.

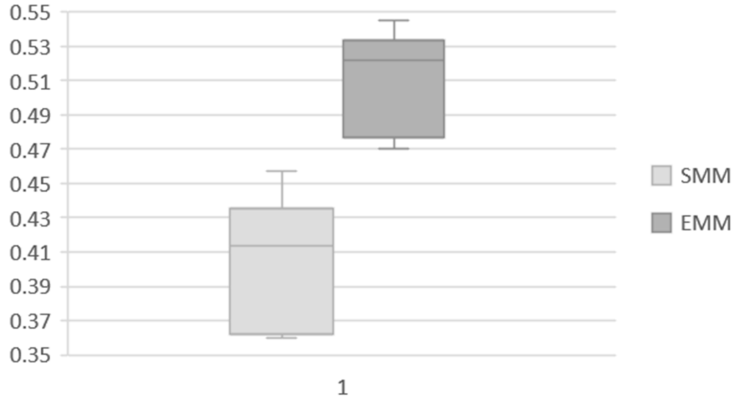

Figure 8. Box plot comparison of SMM and EMM fusion models.

In order to facilitate a critical discussion of the results obtained from both an ML and medical perspective, the specificity and sensitivity were calculated, observing 0.9615 and 0.5385, respectively, for the best EMM-LC 75k model.

5.3. Enriching CDEs with a DAE

Three tests were performed to provide a comprehensive analysis upon the effect of enriching the CDEs with a DAE, taking inspiration from the implementation provided by [16]. Although this study already investigated the efficacy of this approach at differing levels of inflation, it was a pre-requisite to re-implement these tests, as the complexity of data was increased to accommodate the added dimensionality of CT scans. Table 10 details the tests conducted.

Table 10. Description of tests to validate the implementation of DAE. Tests were implemented in conjunction with the same architecture as the EMM-LC 75K model.

CDE Description	CDE Features	Total Number of Features Concatenated
DAE inflation (×10)	740	76,516
DAE inflation (×20)	1480	77,256
DAE inflation (×30)	2220	77,996

Contradicting earlier studies, the implementation of the DAE to enrich the CDEs negatively affected the performance of the model, as highlighted in Figure 9. In addition, there was no significant improvement between the degrees of inflation. In an attempt to improve these results, regularisation techniques were reduced with the intention of preserving data whilst encouraging a rich feature representation. However, this did not have any noticeable impact on the results.

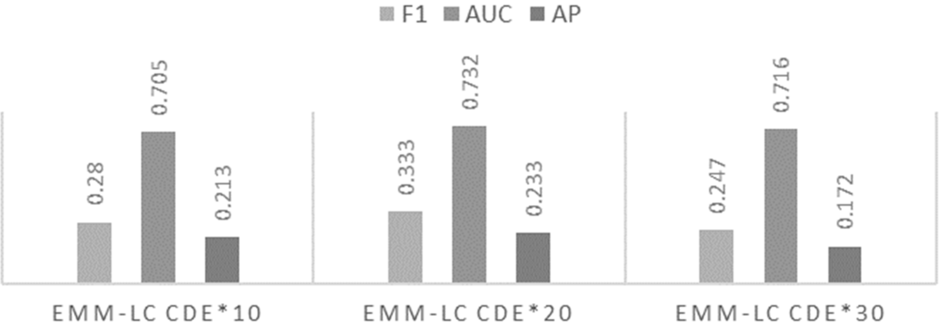

Figure 9. Varying the degree of inflation with DAE.

5.4. Summary

The results presented within this section prove with statistical significance that by enhancing the feature representation of the CT scan, the performance of the model can be improved. Figure 10 highlights a number of studies implementing solutions towards the same dataset. Notwithstanding the visible improvement, this was achieved with reduced model complexity and half the original image size. However, contradicting the implementation proposed by [16], using a DAE to enrich the CDEs had a negative impact on the performance of the model.

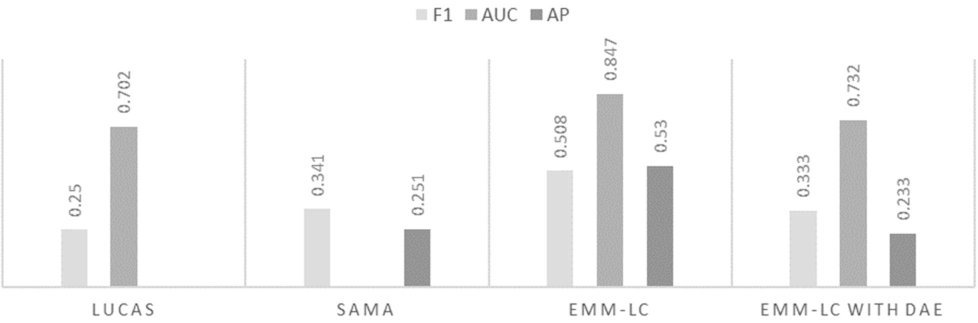

Figure 10. Comparison of EMM-LC Fusion against previous benchmark models.

6. Discussion

The objective of this study was to identify current ML solutions for LC diagnosis and to improve these by developing an architecture which utilises intermediate pre-trained features within the fusion of multiple modalities. The tests conducted aimed to provide empirical evidence to support the claim that the proposed fusion method is superior to existing techniques.

6.1. Main Findings

It was found that by utilising the pre-processing steps defined by [19], see Appendix A for detail, and BN, in addition to reducing the model complexity, the model could outperform previous benchmark models on the same validation set by 60.8% (Section 5.1). As part of the main focus of this study, it was also established that by increasing the number of intermediate features extracted from the pre-trained AlignedXception network, the model was able to better distinguish between cancerous and non-cancerous patients. The statistically significant improvement of 26.37% (Section 5.2) validated the new fusion architecture presented in this study. Interestingly, however, and contrary to the pre-existing literature, utilising a DAE to enrich the features of the CDEs negatively affected the predictive capabilities of the model.

6.2. Simple Multimodal Fusion

In order to draw a fair comparison between methods, a simple multimodal fusion architecture was re-implemented under the same conditions as the EMM-LC model. To improve upon this implementation, despite it not being the main focus of this study, the pre-processing, architectural changes, and training parameters increased the F1 score from 0.25 to 0.402 (Section 5.1). Although these results cannot be compared like for like, it is extremely indicative that these techniques improve upon previous literature. These results supported the initial hypothesis that justified the reasoning for utilising these techniques. Clarifying the cause for these improvements, [19] articulated that the pre-processing techniques removed distracting features; however, there is little evidence to establish the actual numerical improvement despite the prevalent adoption from other scholars. Batch size has a greater impact on the efficacy of batch normalisation (BN) compared to instance normalisation (IN). Ref. [45] clarified that IN has limited success in visual recognition tasks. Originally, instance normalisation was implemented, but the results observed when implementing BN align with the findings presented by [45].

Figure 11 shows the statistical improvements that this implementation provides regarding the observed F1 score, which were 0.25 and 0.402 for the LUCAS and SMM-LC model, respectively. It must be acknowledged that this method also exceeds in other quantitative measures, such as inference time, due to the decreased computational complexity from reducing the input image; however, the evaluation of this is outside the scope of this paper.

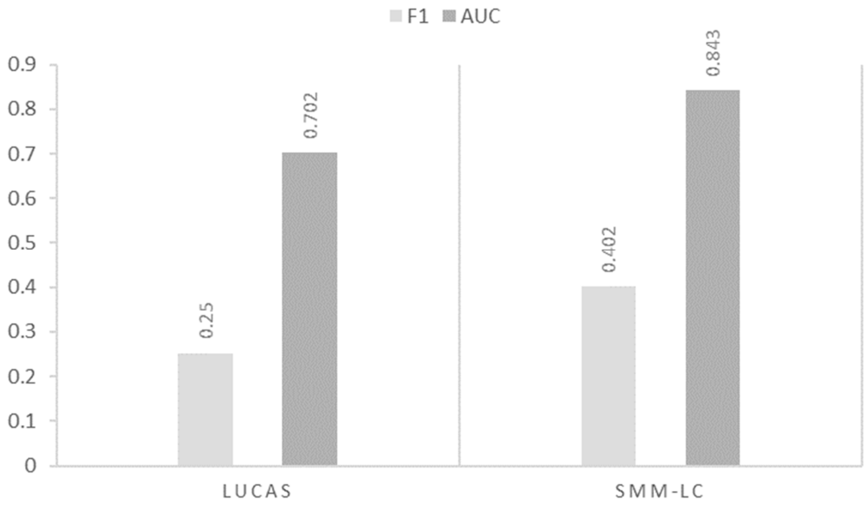

Figure 11. The SMM comparison against the LUCAS benchmark.

6.3. EMM-LC Fusion

The findings presented within this study provide a clear indication that by utilising an enhanced feature representation, the performance of the model can be improved significantly and, thus, satisfies one of the main objectives of this study. However, it is important to acknowledge how these findings contribute to the existing literature. The inspiration for this approach originated from the implementation presented by [16], a 'Richer fusion network', in which a multilevel feature representation provides a fertile environment for full multimodal fusion. Other scholars, particularly within the realm of semantic image segmentation, have also adopted similar approaches on account of the ability for CNNs to learn good feature representations from unstructured data [16]. However, for the task of LC classification, there was insufficient evidence to confirm that this approach has been used, thus, providing a valid contribution to the existing literature.

Notwithstanding the distinction in applications, the findings with regard to the multilevel feature representation aligns with [16] and improves upon the existing benchmark model developed by [37], as seen in Figure 12. Visually, this graph indicates that the architecture may further benefit from additional features. However, the existing literature suggests that by utilising all feature maps, no significant improvement is observed, and computational cost is greatly increased [16]. Therefore, this validates the implementation presented within this study.

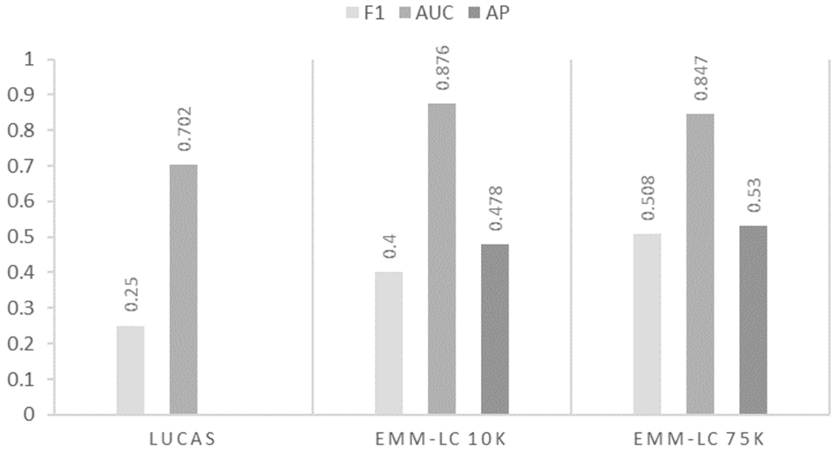

Figure 12. Varying degrees of CT feature enhancement.

The significance of the proposed architecture must be considered from a medical perspective in order to provide a realistic interpretation of the results. Presented in Section 5.2, the specificity and sensitivity metrics provide an in-depth understanding of the ability to detect and classify true positives and true negatives. The values show the efficacy of detecting negative samples with a specificity of 0.9615. However, a sensitivity of 0.5385 highlights the challenge of detecting positive samples.

6.4. EMM-LC DAE Fusion

Despite the significant improvements that this new fusion method provides, the argument presented by [16] suggests that the large imbalance of features would limit the performance potential of the model by overwhelming the low-dimensional clinical data by the high-dimensional CT data. However, implementing the DAE to enrich the CDEs did not yield the expected improvements.

The previous literature provides a strong indication that the decline in performance is a reflection on the choice of CDEs, as opposed to an architectural flaw. Ref. [46] articulated that the proliferation of data that is unrelated to the question of interest hinders the ability to detect real relationships and patterns. This signifies that the amplification of dimensionality inadvertently increases the number of redundant features. In relation to this solution, the implementation of the DAE increases the significance of variables which have little relevance to the task of LC classification, thus, worsening the performance of the model. The significance of these findings challenge the novel implementation proposed by [16] but do not negate it; the limitation stems from the data, not the approach. However, the existing literature, or lack thereof, regarding this method does also contribute to the argument that it is not an appropriate feature inflation method.

6.5. Summary

This section interprets and synthesises the findings presented in this study with respect to the existing literature. Moreover, this discussion establishes the contribution that these results provide and, crucially, identifies the cause for contradictory findings with respect to the relevant literature to support these claims. It also introduces some of the limitations in regard to the findings presented.

7. Limitations

The most prevalent limitation within this study stems from the quality of the data, a constraint that caused the conflicting results reported in previous literature. Figure 13 draws attention to this by illustrating the correlation between the 74 CDEs and LC. It has already been highlighted that features which do not positively contribute to the task in question prevent relationships and patterns from being learned [46]. Therefore, the implementation of the DAE would increase the significance of redundant features, thus, impairing the distinguishing capabilities of the multimodal model.

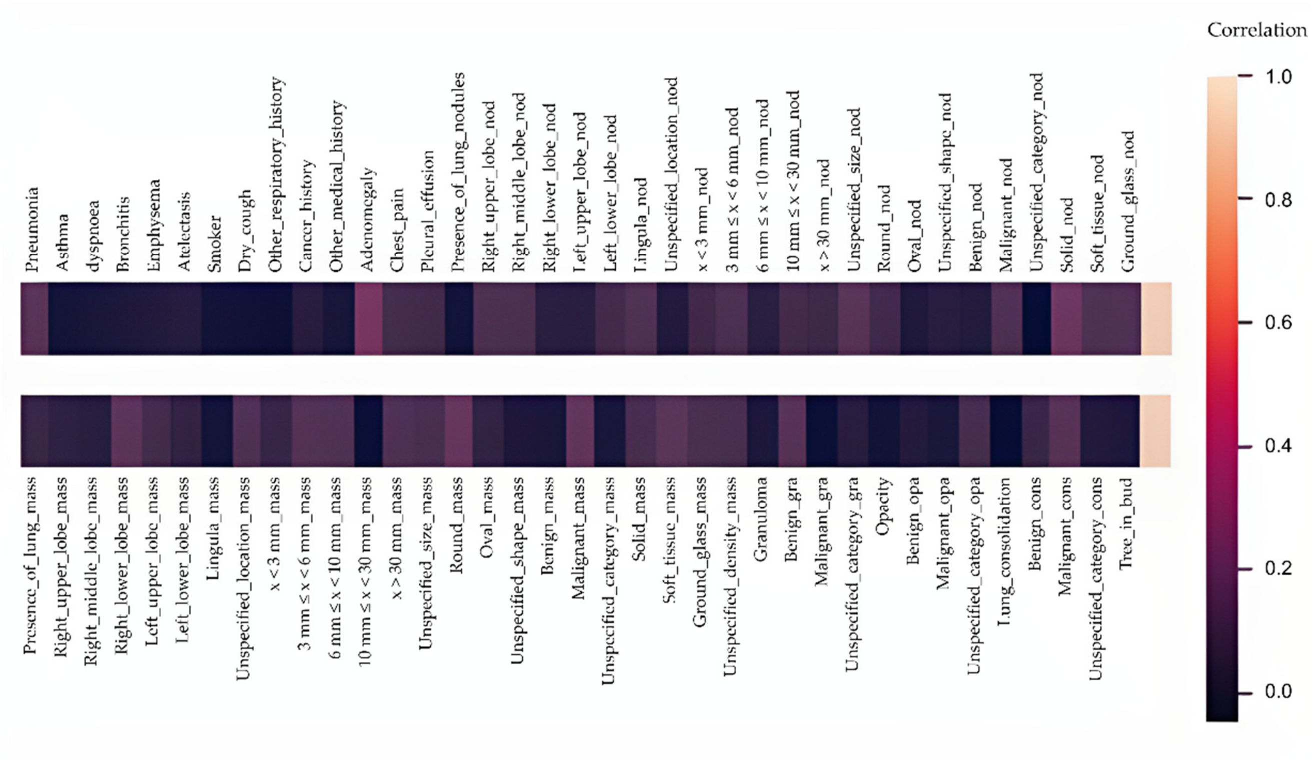

Figure 13. Correlation matrix of CDEs against cancer.

This indicates that the utilisation of alternative biomarkers may better anticipate the diagnosis of malignant tumours and ultimately limit the number of diagnostic procedures on benign nodules [9]. It has been well established that the use of biomarkers improves the performance of multimodal models, but this study clearly identifies that the way in which these are incorporated has a significant impact on the identification of intermodal relationships.

Continuing the discussion pertaining to the data used, the number of samples and the respective validation methods also presented challenges concerning the generalisability of the results. In total, the test data consisted of 170 samples, 13 of which were cancerous. The limited number of samples implies a statistical uncertainty with respect to the average test error [47]. Therefore, future work should evaluate this model using k-fold cross-validation, a method that was omitted due to time constraints. This approach would add credibility and generalisability to these results by taking the average test error across k-trials [47].

The literature suggests that the current method of image resampling, namely reducing the image size to 128^3, may have a tangible impact on the detective capabilities of the model, specifically for samples where the tumour is relatively small. Ref. [48] articulated that the method of resampling, utilising nearest neighbour interpolation, results in a serious loss of quality and thus, obscures small nodules. Although other sampling methods, such as linear interpolation may mitigate this limitation, ultimately, using the original dimensions would preserve all of the potentially significant features. It is recommended that future works investigate this in order to understand the potential impact that this has on the corresponding model.

With respect to the application of this approach in healthcare, this method lacks interpretability, the transparency of reasoning. Ref. [49] articulated that localised knowledge of a malignant nodule is required to fully harness the potential of novel biomarkers. However, acknowledging this limitation, and aligning with existing perceptions of ML in healthcare, new techniques should work in parallel with doctors and radiologists to enhance productivity and precision [7]. The significance of this recognises that despite a specificity and sensitivity of 0.9615 and 0.5385, respectively, the benefits of this approach become evident when used in conjunction with medical professionals. Work should continue to develop this and to improve the detection of LC, as it has been established that early treatment offers encouraging prognosis with a survival rate of 5 years at 71% [50].

8. Future Work

In order to develop this approach and improve its suitability within a clinical setting, several points must be further investigated with respect to the interpretability of the model, and the application of suitable biomarkers. A recommended pathway for the continuation of this proposed solution entails the implementation of class activation maps [51], preservation of scan resolution, and the addition of alternative biomarkers.

For the purpose of interpretability, the future direction of this architecture should facilitate the application of class activation maps in order to visualise significant relationships and patterns within CT scans. This would align with [46] with respect to the requirement for localised knowledge in order to fully harness the potential of novel biomarkers. It also raises an important consideration towards the efficacy of using whole scans prior to feature fusion. As previously discussed, data unrelated to the question of interest hinders the ability to detect real relationships and patterns [43]. Therefore, the identification and extraction of regions of interest, prior to feature fusion, proposes a natural progression for this architecture, in addition to the preservation of scan resolution to retain all significant features. Several key takeaways must be acknowledged from this paper and adopted into existing solutions, as it has been proved that multimodal learning, using enhanced feature representations, improves the accuracy of lung cancer classification.

9. Conclusions

This study set out to critically examine existing research within the field of multimodal deep learning and identify ways in which current methods for LC classification could be improved. The findings clearly indicated that there was a lack of consideration for the potential that intermediate, multimodal features may provide towards identifying LC in CT scans. Subsequent to the literature review, this paper aimed to contribute towards existing works by proposing an enhanced multimodal fusion method, named EMM-LC Fusion, which utilised intermediate, pre-trained features to improve the existing benchmark model applied to the LUCAS dataset.

The experiments confirmed that the extraction of multi-level features, for the purpose of multimodal fusion, improved the ability of the model to identify intermodal relationships, distinguishing true positive and true negative samples. The improvement in mean F1 score of 26.37% when using the EMM-LC fusion model is supported by the statistical significance of the results ($p < 0.05$). These findings complement earlier studies which have adopted this approach for other diseases, such as breast cancer [16], and AD [33]. This adds to the growing body of work which promotes the use of multiple modalities for disease diagnosis.

Although this study successfully demonstrated that enhancing feature representations of CT scans before fusion improves performance, the results obtained from applying a DAE to the CDE conflicted with findings presented by earlier studies [16]. This questions the rationality of utilising a DAE to enrich CDEs, as it was observed to exacerbate the performance of the model when compared to simple linear inflation. Upon the basis that these conflicting results do not invalidate this approach, a natural progression of this work is to incorporate and analyse alternative biomarkers for LC and to evaluate whether this provides any significant improvements. Ultimately, this study lays the groundwork for future research in the use of multimodal biomarkers for the task of LC diagnosis.

Supplementary Materials: All code is available at: https://github.com/jb4rr/EMM-LC-Fusion. Variations of this code are accessible under different branches.

Author Contributions: Conceptualization, JB. and TV.; methodology, JB. and TV.; software, JB.; validation, JB. and TV.; formal analysis, JB. and TV.; investigation, J.B.; writing—original draft preparation, J.B.; writing—review and editing, T.V. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest

Appendix A

Table A1. LUCAS Clinical data elements.

Pneumonia	Asthma	dyspnoea
Bronchitis	Emphysema	Atelectasis
Smoker	Dry_cough	Other_respiratory_history
Cancer_history	Other_medical_history	Adenomegaly
Chest_pain	Pleural_effusion	Presence_of_lung_nodules
Right_upper_lobe_nod	Right_middle_lobe_nod	Right_lower_lobe_nod
Left_upper_lobe_nod	Left_lower_lobe_nod	Lingula_nod
Unspecified_location_nod	x < 3 mm_nod	3 mm ≤ x < 6 mm_nod
6 mm ≤ x < 10 mm_nod	10 mm ≤ x < 30 mm_nod	x > 30 mm_nod
Unspecified_size_nod	Round_nod	Oval_nod
Unspecified_shape_nod	Benign_nod	Malignant_nod
Unspecified_category_nod	Solid_nod	Soft_tissue_nod
Ground_glass_nod	Unspecified_density_nod	Presence_of_lung_mass
Right_upper_lobe_mass	Right_middle_lobe_mass	Right_lower_lobe_mass
Left_upper_lobe_mass	Left_lower_lobe_mass	Lingula_mass
Unspecified_location_mass	x < 3 mm_mass	3 mm ≤ x < 6 mm_mass
6 mm ≤ x < 10 mm_mass	10 mm ≤ x < 30 mm_mass	x > 30 mm_mass
Unspecified_size_mass	Round_mass	Oval_mass
Unspecified_shape_mass	Benign_mass	Malignant_mass
Unspecified_category_mass	Solid_mass	Soft_tissue_mass
Ground_glass_mass	Unspecified_density_mass	Granuloma
Benign_gra	Malignant_gra	Unspecified_category_gra
Opacity	Benign_opa	Malignant_opa
Unspecified_category_opa	Lung_consolidation	Benign_cons
Malignant_cons	Unspecified_category_cons	Tree_in_bud

Appendix A.1. Pre-processing Steps

Although the analysis of the implemented pre-processing steps is outside the scope of this paper, it is important to acknowledge the changes made so that the repeatability of this study is not negatively affected. Table A2 details the changes made within the pre-processing steps defined by [19]. The following discussion presents the complete set of techniques that were employed to pre-process the CT scans.

Table A2. Pre-processing threshold value changes.

Threshold Name	Previous Value	Proposed Value
Vol_limit	[0.68, 7.5]	[0.0, 7.5]
Intensity_th	−600	−500
bg_patch_size	10	1
Eccen_th (binarize_per_slice)	0.99	0.999

area_th (binarize_per_slice)	30	5
Area_th (all_slice_analysis)	6e3	3e3

To reduce the computational complexity required to process the image, the scan is halved in size using a nilearn affine transform. Following this, a binary mask is extracted using the following steps:

1. Remove top slices;
2. Apply Gaussian filter (stdv = 1 px);
3. Binarize filter with a threshold of −500;
4. Remove all 2D components which are smaller than 5 mm^2 or have an eccentricity greater than 0.999;
5. Remove 3D volumes of more than 7.5 litres;
6. Remove components with average minimum distance of 62 mm from the centre;
7. Compute convex hull of image.

The values within the scan are clipped between −1200 and 600, and values outside of the mask are padded with the value of 170, the luminance of common tissue [19].

Appendix A.2. Tests

Table A3. Lowered regularisation for DAE.

Run Number	F1	AUC	AP
N10	0.296	0.694	0.231
N20	0.333	0.66	0.151
N30	0.324	0.668	0.186

References

1. Lung Cancer Statistics. Available online: https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/lung-cancer (accessed on 31 October 2021).
2. Siddiqui, F.; Vaqar, S.; Siddiqui, A.H. Lung Cancer. In *StatPearls*; StatPearls Publishing: Treasure Island, FL, USA, 2022.
3. Cainap, C.; Pop, L.A.; Balacescu, O.; Cainap, S.S. Early Diagnosis and Screening in Lung Cancer. *Am. J. Cancer Res.* **2020**, *10*, 1993–2009.
4. Bach, P.; Kelley, M.; Tate, R.; Mccrory, D. Screening for Lung Cancer—A Review of the Current Literature. *Chest* **2003**, *123*, 72S-82S. https://doi.org/10.1378/chest.123.1_suppl.72S.
5. Makaju, S.; Prasad, P.W.C.; Alsadoon, A.; Singh, A.K.; Elchouemi, A. Lung Cancer Detection Using CT Scan Images. *Procedia Comput. Sci.* **2018**, *125*, 107–114. https://doi.org/10.1016/j.procs.2017.12.016.
6. Beckles, M.A.; Rudd, R.M.; Spiro, S.G.; Colice, G.L. Initial Evaluation of the Patient with Lung Cancer: Symptoms, Signs, Laboratory Tests, and Paraneoplastic Syndromes. *Chest* **2003**, *123*, 97S–104S.
7. Akhil, J.; Samreen, S.; Aluvalu, R. The Future of Health Care: Machine Learning. *Int. J. Eng. Technol.* **018**, *7*, 23–25. https://doi.org/10.14419/ijet.v7i4.6.20226.
8. Del Ciello, A.; Franchi, P.; Contegiacomo, A.; Cicchetti, G.; Bonomo, L.; Larici, A.R. Missed Lung Cancer: When, Where, and Why? *Diagn. Interv. Radiol.* **2017**, *23*, 118–126. https://doi.org/10.5152/dir.2016.16187.
9. Seijo, L.M.; Peled, N.; Ajona, D.; Boeri, M.; Field, J.K.; Sozzi, G.; Pio, R.; Zulueta, J.J.; Spira, A.; Massion, P.P.; et al. Biomarkers in Lung Cancer Screening: Achievements, Promises, and Challenges. *J. Thorac. Oncol.* **2018**, *14*, 343–357. https://doi.org/10.1016/j.jtho.2018.11.023.
10. Davenport, T.; Kalakota, R. The Potential for Artificial Intelligence in Healthcare. *Future Healthc. J.* **2019**, *6*, 94–98. https://doi.org/10.7861/futurehosp.6-2-94.
11. Char, D.S.; Shah, N.H.; Magnus, D. Implementing Machine Learning in Health Care—Addressing Ethical Challenges. *N. Engl. J. Med.* **2018**, *378*, 981–983. https://doi.org/10.1056/NEJMp1714229.
12. Rajkomar, A.; Dean, J.; Kohane, I. Machine Learning in Medicine. *Engl. J. Med.* **2019**, *380*, 1347–1358. https://doi.org/10.1056/NEJMra1814259.
13. LeCun, Y.; Bengio, Y.; Hinton, G. Deep Learning. *Nature* **2015**, *521*, 436–444. https://doi.org/10.1038/nature14539.
14. Meng, L.; Dong, D.; Li, L.; Niu, M.; Bai, Y.; Wang, M.; Qiu, X.; Zha, Y.; Tian, J. A Deep Learning Prognosis Model Help Alert for COVID-19 Patients at High-Risk of Death: A Multi-Center Study. *IEEE J. Biomed. Health Inform.* **2020**, *24*, 3576–3584. https://doi.org/10.1109/JBHI.2020.3034296.
15. Taormina, V.; Cascio, D.; Abbene, L.; Raso, G. Performance of Fine-Tuning Convolutional Neural Networks for HEp-2 Image Classification. *Appl. Sci.* **2020**, *10*, 6940. https://doi.org/10.3390/app10196940.

16. Yan, R.; Zhang, F.; Rao, X.; Lv, Z.; Li, J.; Zhang, L.; Liang, S.; Li, Y.; Ren, F.; Zheng, C.; et al. Richer Fusion Network for Breast Cancer Classification Based on Multimodal Data. *BMC Med. Inform. Decis. Mak.* **2021**, *21*, 134. https://doi.org/10.1186/s12911-020-01340-6.
17. Misra, P.; Yadav, A. Impact of Preprocessing Methods on Healthcare Predictions. In Proceedings of the 2nd International Conference on Advanced Computing and Software Engineering, New York, NY, USA, 1 January 2019.
18. Lakshmanaprabu, S.K.; Mohanty, S.N.; Shankar, K.; Arunkumar, N.; Ramirez, G. Optimal Deep Learning Model for Classification of Lung Cancer on CT Images. *Future Gener. Comput. Syst.* **2019**, *92*, 374–382. https://doi.org/10.1016/j.future.2018.10.009.
19. Liao, F.; Liang, M.; Li, Z.; Hu, X.; Song, S. Evaluate the Malignancy of Pulmonary Nodules Using the 3D Deep Leaky Noisy-or Network. *IEEE Trans. Neural Netw. Learn. Syst.* **2019**, *30*, 3484–3495. https://doi.org/10.1109/TNNLS.2019.2892409.
20. Gao, R.; Tang, Y.; Khan, M.S.; Xu, K.; Paulson, A.B.; Sullivan, S.; Huo, Y.; Deppen, S.; Massion, P.P.; Sandler, K.L.; et al. Cancer Risk Estimation Combining Lung Screening CT with Clinical Data Elements. *Radiol. Artif. Intell.* **2021**, *3*, e210032. https://doi.org/10.1148/ryai.2021210032.
21. Chlap, P.; Min, H.; Vandenberg, N.; Dowling, J.; Holloway, L.; Haworth, A. A Review of Medical Image Data Augmentation Techniques for Deep Learning Applications. *J. Med. Imaging Radiat. Oncol.* **2021**, *65*, 545–563. https://doi.org/10.1111/1754-9485.13261.
22. Neroladaki, A.; Botsikas, D.; Boudabbous, S.; Becker, C.D.; Montet, X. Computed Tomography of the Chest with Model-Based Iterative Reconstruction Using a Radiation Exposure Similar to Chest X-Ray Examination: Preliminary Observations. *Eur. Radiol.* **2012**, *23*, 360–366. https://doi.org/10.1007/s00330-012-2627-7.
23. Gao, R.; Tang, Y.; Xu, K.; Kammer, M.N.; Antic, S.L.; Deppen, S.; Sandler, K.L.; Massion, P.P.; Huo, Y.; Landman, B.A. *Deep Multi-Path Network Integrating Incomplete Biomarker and Chest CT Data for Evaluating Lung Cancer Risk*; SPIE: Bellingham, DC, USA, 2021; Volume 11596, pp. 387–393.
24. Serj, M.F.; Lavi, B.; Hoff, G.; Valls, D.P. A Deep Convolutional Neural Network for Lung Cancer Diagnostic. *arXiv* **2018**, arXiv:1804.08170.
25. Agarwal, A.; Patni, K.; Rajeswari, D. Lung Cancer Detection and Classification Based on Alexnet CNN. In Proceedings of the 2021 6th International Conference on Communication and Electronics Systems (ICCES), Coimbatre, India, 8–10 July 2021; pp. 1390–1397.
26. Baltrusaitis, T.; Ahuja, C.; Morency, L.-P. Multimodal Machine Learning: A Survey and Taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *41*, 423–443. https://doi.org/10.1109/TPAMI.2018.2798607.
27. Sleeman, W.C., IV; Kapoor, R.; Ghosh, P. Multimodal Classification: Current Landscape, Taxonomy and Future Directions. *arXiv* **2021**, arXiv:2109.09020.
28. Zhang, D.; Wang, Y.; Zhou, L.; Yuan, H.; Shen, D. Multimodal Classification of Alzheimer's Disease and Mild Cognitive Impairment. *NeuroImage* **2011**, *55*, 856–867. https://doi.org/10.1016/j.neuroimage.2011.01.008.
29. James, A.P.; Dasarathy, B.V. Medical Image Fusion: A Survey of the State of the Art. *Inf. Fusion* **2014**, *19*, 4–19. https://doi.org/10.1016/j.inffus.2013.12.002.
30. Rivera, M.P.; Mehta, A.C. Initial Diagnosis of Lung Cancer* ACCP Evidence-Based Clinical Practice Guidelines (2nd Edition). *Chest* **2007**, *132*, 131S–148S.
31. Wu, Y.; Ma, J.; Huang, X.; Ling, S.H.; Su, S.W. DeepMMSA: A Novel Multimodal Deep Learning Method for Non-Small Cell Lung Cancer Survival Analysis. In Proceedings of the 2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Melbourne, Australia, 12 June 2021; pp. 1468–1472.
32. Guo, W.; Wang, J.; Wang, S. Deep Multimodal Representation Learning: A Survey. *IEEE Access* **2019**, *7*, 63373–63394. https://doi.org/10.1109/ACCESS.2019.2916887.
33. Venugopalan, J.; Tong, L.; Hassanzadeh, H.R.; Wang, M.D. Multimodal Deep Learning Models for Early Detection of Alzheimer's Disease Stage. *Sci. Rep.* **2021**, *11*, 3254–3267. https://doi.org/10.1038/s41598-020-74399-w.
34. Xie, Y.; Meng, W.-Y.; Li, R.-Z.; Wang, Y.-W.; Qian, X.; Chan, C.; Yu, Z.-F.; Fan, X.-X.; Pan, H.-D.; Xie, C.; et al. Early Lung Cancer Diagnostic Biomarker Discovery by Machine Learning Methods. *Transl. Oncol.* **2021**, *14*, 100907. https://doi.org/10.1016/j.tranon.2020.100907.
35. Pesesse, R.; Stefanuto, P.-H.; Schleich, F.; Louis, R.; Focant, J.-F. Multimodal Chemometric Approach for the Analysis of Human Exhaled Breath in Lung Cancer Patients by TD-GC × GC-TOFMS. *J. Chromatogr. B* **2019**, *1114–1115*, 146–153. https://doi.org/10.1016/j.jchromb.2019.01.029.
36. Burzic, A.; O'Dowd, E.L.; Baldwin, D.R. The Future of Lung Cancer Screening: Current Challenges and Research Priorities. *Cancer Manag. Res.* **2022**, *14*, 637–645. https://doi.org/10.2147/CMAR.S293877.
37. Daza, L.; Castillo, A.; Escobar, M.; Valencia, S.; Pinzón, B.; Arbelaez, P. *LUCAS: LUng CAncer Screening with Multimodal Biomarkers*; Spinger: Berlin/Heidelberg, Germany, 2020; pp. 115–124, ISBN 978-3-030-60945-0.
38. Roa, M.; Daza, L.; Escobar, M.; Castillo, A.; Arbelaez, P. SAMA: Spatially-Aware Multimodal Network with Attention For Early Lung Cancer Diagnosis. In *Multimodal Learning for Clinical Decision Support*; Syeda-Mahmood, T., Li, X., Madabhushi, A., Greenspan, H., Li, Q., Leahy, R., Dong, B., Wang, H., Eds.; Lecture Notes in Computer Science; Springer International Publishing: Berlin/Heidelberg, Germany, 2021; Volume 13050, pp. 48–58, ISBN 978-3-030-89846-5.
39. Sahu, G.; Vechtomova, O. Adaptive Fusion Techniques for Multimodal Data. *arXiv* **2021**, arXiv:1911.03821.

40. Chirodea, M.C.; Novac, O.C.; Novac, C.M.; Bizon, N.; Oproescu, M.; Gordan, C.E. Comparison of Tensorflow and PyTorch in Convolutional Neural Network—Based Applications. *13th International Conference on Electronics, Computers and Artificial Intelligence (ECAI), 2021, pp. 1-6,*
41. Wang, J.; Gao, R.; Huo, Y.; Bao, S.; Xiong, Y.; Antic, S.L.; Osterman, T.J.; Massion, P.P.; Landman, B.A. Lung Cancer Detection Using Co-Learning from Chest CT Images and Clinical Demographics. *Proc. SPIE Int Soc. Opt. Eng.* **2019**, *10949*, 365–371. https://doi.org/10.1117/12.2512965.
42. Shorten, C.; Khoshgoftaar, T.M. A Survey on Image Data Augmentation for Deep Learning. *J. Big Data* **2019**, *6*, 60. https://doi.org/10.1186/s40537-019-0197-0.
43. Chollet, F. Xception: Deep Learning with Depthwise Separable Convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, San Juan, PR, USA, 21–26 July 2017; pp. 1251–1258.
44. Reproducibility—PyTorch 1.11.0 Documentation Available online: https://pytorch.org/docs/stable/notes/randomness.html (accessed on 19 April 2022).
45. Wu, Y.; He, K. Group Normalization. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, German, 8–14 September 2018; pp. 3–19.
46. Altman, N.; krzywinski, M. The Curse(s) of Dimensionality. *Nat. Methods* **2018**, *15*, 397–400. https://doi.org/10.1038/s41592-018-0013-3.
47. Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; Massachusetts Institute of Technology: Cambridge, MA, USA, 2016; ISBN 978-0-262-03561-3.
48. Thevenaz, P.; Blu, T.; Unser, M. Image Interpolation and Resampling. In *Proceedings of the Handbook of Medical Imaging, Processing and Analysis*; Academic Press: Cambridge, MA, USA, 2000; pp. 393–420.
49. Blandin Knight, S.; Crosbie, P.A.; Balata, H.; Chudziak, J.; Hussell, T.; Dive, C. Progress and Prospects of Early Detection in Lung Cancer. *Open Biol.* **2017**, *7*, 170070. https://doi.org/10.1098/rsob.170070.
50. Shah, R.; Sabanathan, S.; Richardson, J.; Mearns, A.J.; Goulden, C. Results of Surgical Treatment of Stage I and II Lung Cancer. *J. Cardiovasc. Surg.* **1996**, *37*, 169–172.
51. Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; Torralba, A. Learning Deep Features for Discriminative Localization. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 2921–2929.