A Novel and Adaptive Evaluation Mechanism for Deep Learning Models in Medical Imaging and Disease Recognition

1st Zainab Loukil
Cyber and Technical Computing
University of Gloucestershire
Gloucestershire, UK
zloukil2@glos.ac.uk

2nd Qublai Khan Ali Mirza
Cyber and Technical Computing
University of Gloucestershire
Gloucestershire, UK
qalimirza@glos.ac.uk

3rd Will Sayers
Games Technologies
University of Gloucestershire
Gloucestershire, UK
wsayers@glos.ac.uk

***Abstract*—Recent advancements in deep learning (DL) and machine learning (ML) algorithms have led to their extensive use in various fields, including healthcare, finance, image processing, etc. However, selecting the most appropriate evaluation metrics for these algorithms remains a challenge. In this paper, we propose a novel evaluation metrics mechanism that takes into account the problem domain and the specific application area. The proposed 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 Performance Measurement Matrix (PMM) vector that reflects the most optimal measurement metrics. Our 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. We demonstrate the effectiveness of our mechanism using a case study on medical image processing applications.**

***Index Terms*—Evaluation metrics, Performance evaluation, Model evaluation mechanism, Deep Learing, Measurement matrix.**

I. Introduction

Artificial intelligence (AI) has rapidly advanced over the past decade, with Machine Learning (ML) and Deep Learning (DL) being two of its most prominent sub-fields. In healthcare domain, the impact of AI has been significant. In fact, ML and DL algorithms have been developed to aid in medical image analysis, drug discovery, disease diagnosis, and personalized treatment planning, among other applications [1]–[3].

For instance, DL has shown promising results in the detection and classification of various diseases from medical images, such as mammograms for breast cancer screening [4] and CT scans for lung cancer detection [5]. Additionally, ML/DL models have been developed to predict disease progression and treatment outcomes based on patient data, aiding in personalized treatment planning [6].

Despite the potential benefits of AI in healthcare, there are also concerns about its ethical implications and potential biases in algorithmic decision-making. Moreover, the adoption of AI in healthcare requires careful consideration of regulatory and legal frameworks, as well as the development of appropriate data privacy and security measures [4]. This potential must be balanced with careful consideration of ethical and regulatory implications to ensure equitable and effective use of these technologies. The impact of evaluation metrics on final decision-making resulted from ML/DL models, can be quite significant, as the choice of these parameters can influence the selection of the best-performing model and the assessment of its performance in real-world applications [6].

While existing performance evaluation mechanisms hold promise for advancement, there are still several challenges in evaluating the performance of ML and DL algorithms. One of the main challenges is the lack of a universal evaluation framework that can handle different types of data and algorithms. Moreover, there is a need for more comprehensive and interpretable evaluation metrics that can capture the complexities of the algorithm and the data [7]. Fig. 1 illustrates the six key characteristics an evaluation mechanism needs to incorporate.

In this paper, a novel adaptive and reliable performance measurement mechanisms is proposed, namely, performance measurement matrix (PMM). The main contribution of the suggested method is to evaluate the impact of PMM on ML/DL models testing, validation, and decision-making.

The structure of the paper is as follows:

- The first section covers a brief background on the use of evaluation methods in existing state of the art.
- The second section highlights the limitations of existing ML/DL evaluation mechanisms
- The third section details the proposed performance evaluation mechanism using PMM as well as performed experimentation. Results and discussion are also presented at the end of the section.
- The paper ends with a conclusion and further future directions.

II. Background

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

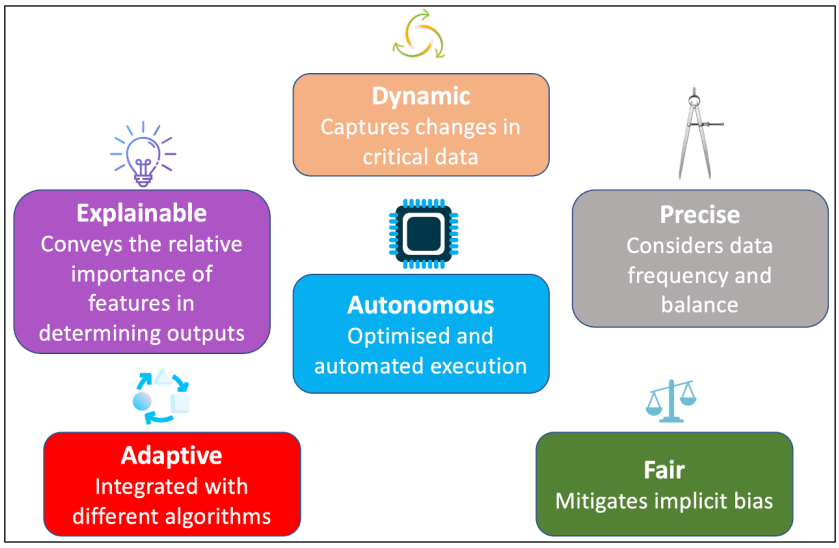

Fig. 1. Key Characteristics of an Evaluation Mechanism for DL/ML algorithms

- **Accuracy (Acc):** widely used in ML and DL applications measuring the proportion of correctly classified examples over the total number of examples.
- **Sensitivity (Sen):** Used in binary classification to measure the proportion of true positive examples that are correctly classified by the model.
- **Specificity (Spe):** Used in binary classification to measure the proportion of true negative examples that are correctly classified by the model.
- **Precision-Recall (PR):** Used to measure the trade-off between precision and recall at different classification thresholds.
- **F-measure (F-m):** Used in information retrieval for evaluating the performance of binary classification systems by combining the precision and recall metrics into a single score.
- **Accuracy Under Curve (AUC):** Used to measure the performance of a classifier by calculating the area under the receiver operating characteristic (ROC) curve.
- **Receiver Operating Characteristic (ROC) curve:** Provides a visualization of the trade-off between true positive and false positive rates (FPR).
- **Total Operating Characteristic (TOC) curve:** Provides a visualization of the trade-off between true positive and false negative rates (FNR).
- **Prevalence (P):** Used to measure the proportion of positive examples in the dataset.
- **Positive Predictive Value (PPV):** Used to measure the proportion of true positive examples among all the positive predictions made by the model.
- **Negative Predictive Value (NPV):** Used to measure the proportion of true negative examples among all the negative predictions made by the model.

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 I provides a summary comparison of aforementioned parameters.

In addition to the above, loss metrics are commonly used evaluation metrics for machine learning and deep learning 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 [20], [21]:

- **Mean Absolute Error (MAE):** used mainly for regression problems. It measures the average of the absolute differences between the predicted and actual values.
- **Dice Coefficient for multiple classes (DSC):** a statistical similarity, used in image segmentation tasks to evaluate the overlap between the predicted and ground-truth masks.
- **Categorical Cross-Entropy (LCE):** is used for multi-class classification problems. It measures the difference between the predicted probabilities and the actual class labels.
- **Dice Coefficient (DC):** is used in segmentation problems. It measures the overlap between the predicted and actual segmentation masks.
- **Soft Dice Loss (SDL):** is used as a loss function in training deep learning models for segmentation tasks. It measures the overlap between the predicted segmentation mask and the ground truth mask.

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

In recent years, several advanced evaluation mechanisms have been proposed. One notable example is adversarial evaluation, which involves generating adversarial examples to test the algorithm's robustness and generalizability [8], [9].

TABLE I
COMPARATIVE EVALUATION

Measurement parameter	Pros	Cons
Acc	Easy to understand and calculate	Can be misleading with imbalanced datasets
Sen	Focuses on true positive rate	Doesn't account for true negatives
Spe	Focuses on true negative rate	Doesn't account for true positives
PR	Good for imbalanced datasets	Doesn't account for true negatives
F-m	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/TOC curves	Useful for imbalanced datasets, where it may be more important to optimize 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.
P	Useful for understanding disease prevalence	Doesn't account for classification accuracy
PPV	Provides the likelihood of positive predictions being correct	Affected by class distribution
NPV	provides the likelihood of negative predictions being correct	Affected by class distribution

TABLE II
COMPARATIVE EVALUATION OF LOSS METRICS

DSC	Allowing the model optimization	Sensitive to small errors in predictions leading to over-fitting problem.
DC	Appropriate for segmentation problems	Can be biased towards predicting larger objects and does not consider false positives and false negatives.
SDL	Mitigate the vanishing gradient problem.	Sensitive to class imbalance and can be computationally expensive.
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
MAE	Less sensitive to outliers	Does not punish larger errors more severely than smaller errors

Another approach involves utilizing diversity metrics such as variation ratio, entropy, and inverse participation ratio to assess the diversity of the generated output [10]. Additionally, there have been efforts to develop interpretability and explainability metrics to evaluate the transparency and interpretability of machine learning (ML) and deep learning (DL) algorithms.

Numerous studies in the literature have employed various evaluation metrics to assess the performance of DL models. For instance, in an image classification study by Zhou et al. (2016), the top-1 and top-5 accuracy metrics were used to evaluate the DL model's performance [11]. Sokolova and Lapalme (2009) utilized metrics such as accuracy, precision, recall, and F1-score for classification tasks [12]. In the domain of medical diagnosis, the area under the receiver operating characteristic curve (ROC) is commonly employed to assess the performance of DL models [13]. In natural language processing, metrics like BLEU, ROUGE, and METEOR are frequently used [14]. Furthermore, hyperparameters such as learning rate, batch size, and regularization strength can significantly impact the performance of DL models, and evaluation metrics can be utilized to fine-tune these hyperparameters [15]. These studies underscore the importance of selecting appropriate evaluation metrics for specific tasks and emphasize the impact of these metrics on model selection, hyperparameter tuning, and real-world deployment.

The literature confirms that existing evaluation mechanisms suffers from limited Performance Assessment, where inadequate evaluation metrics may fail to provide a comprehensive and accurate assessment of the model's performance [16]. This can result in an incomplete understanding of its capabilities and limitations. Moreover, the inaccuracy of choosing the adequate evaluation metric can mislead conclusions. In fact, improperly chosen evaluation metrics can lead to misleading outcomes about the effectiveness or superiority of a particular model [17]. This can result in misguided decisions regarding model selection or deployment.

Javad et al proposed that inefficient resource allocation could have a potential impact on the evaluation mechanism of ML/DL models. For instance, if evaluation metrics do not align with the specific requirements and objectives of the task or application, resources such as time, computational power, and human effort may be wasted on evaluating irrelevant aspects or using inefficient evaluation techniques [18].

Yoonyoung et al. have proposed a bias reduction AI-based model. Their proposed evaluation mechanism lacks in comparability. In fact, inadequate evaluation metrics may hinder the comparability of different models or approaches [19]. This makes it challenging to draw meaningful comparisons and identify the most suitable model for a given task.

Based on the aforementioned overview, the selection of inadequate evaluation metrics can lead to four main drawbacks, as indicated by previous studies [12], [22], [23]:

- **Overfitting:** Evaluation metrics such as accuracy, precision, recall, and F1-score can be influenced by overfitting. Overfitting occurs when the algorithm performs well on the training data but poorly on the test data.
- **Imbalanced Data:** Evaluation metrics such as accuracy, precision, and recall can be highly dependent on class

distribution and can lead to misleading results if the dataset is imbalanced.

- **Lack of Robustness:** Some evaluation metrics can be highly sensitive to noise in the data and may not be robust to small changes in the input.
- **Lack of Interpretability:** Some evaluation metrics, such as AUC, do not provide any insight into the internal workings of the algorithm and can be difficult to interpret.

In the following section, a presentation of the proposed performance evaluation mechanism will be detailed.

III. Performance Measurement Matrix (PMM): Proposed Evaluation Mechanism

A. Mechanism

Towards overcoming aforementioned gaps in existng ML/DL performance evaluation methods, this work proposes a novel adaptive evaluation mechanism, as shown in Fig. 2. 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 three dimensional PMM matrix (3D-PMM).

Each component has a specific contribution to the proposed 3D-PMM matrix final output, as shown in Fig. 3. Each matrix dimension is characterised by nxn size, where n defines the total number of performance metrics. In fact, each dimension refelcts the importance of involved metrics by assigning a correlation coefficient reflecting their weights, hence their importance, defined as below (Eq(1)):

$$corr_{coef}(y,z) = \frac{Cov(y,z)}{\sigma_y * \sigma_z} \tag{1}$$

$$= \frac{\sum[(y - mean(Y)) * (z - mean(Z))]}{\sqrt{[\sum(y - mean(Y))^2 * \sum(z - mean(Z))^2]}}$$

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 in Algorithm 1, 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 Fig. 4. The resulted optimal set of metrics is then used as part of the performance measurement of the given ML/DL based model.

Algorithm 1: 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**
            $W_{C_{12_{ii}}} \Leftarrow corr_{coef}(C_{12_i}, C_{12_i})$
            $D_{12}[i,i] \Leftarrow W_{C_{12_{ii}}}$
        else
            $W_{C_{12_{ij}}} \Leftarrow corr_{coef}(C_{12_i}, C_{12_j})$
            $D_{12}[i,j] \Leftarrow W_{C_{12_{ij}}}$
Repeat Step 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 $PMM_{optimal}$

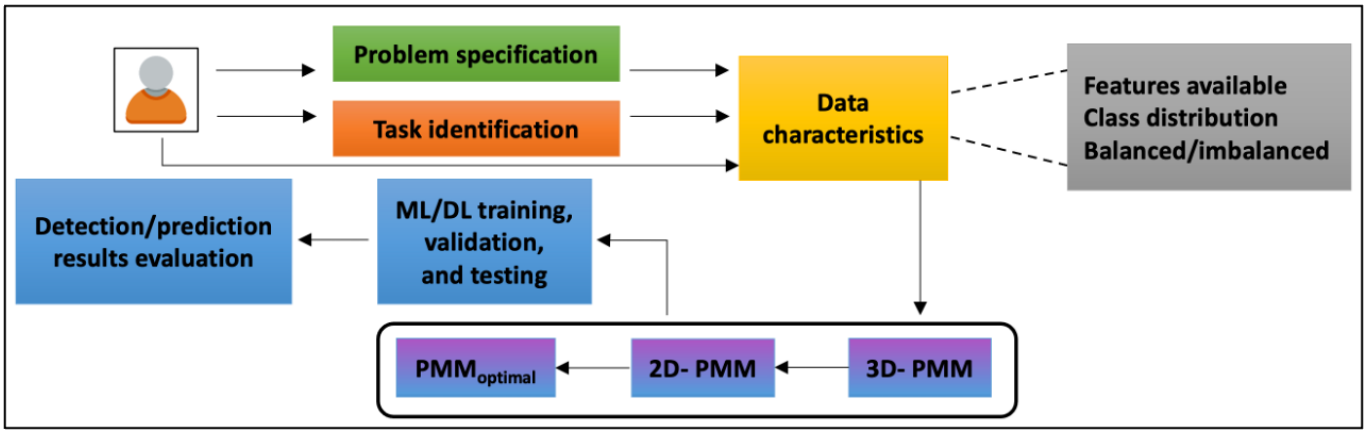

Fig. 2. Proposed Evaluation Mechanism

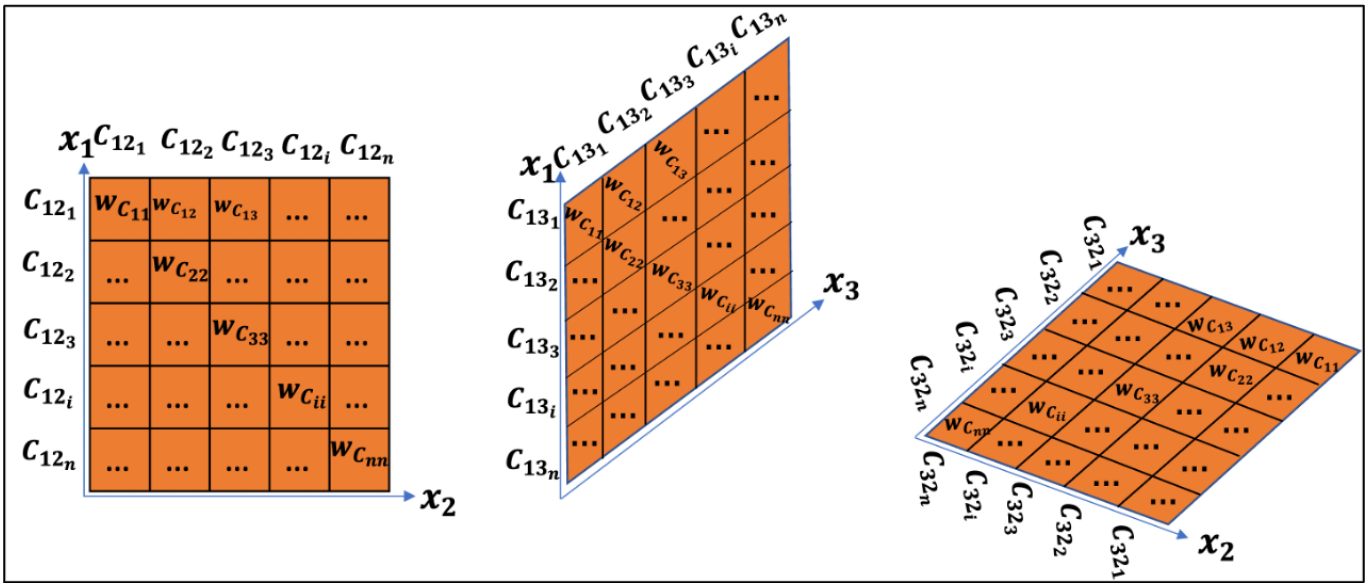

Fig. 3. Three-Dimensional representation of PMM matrix

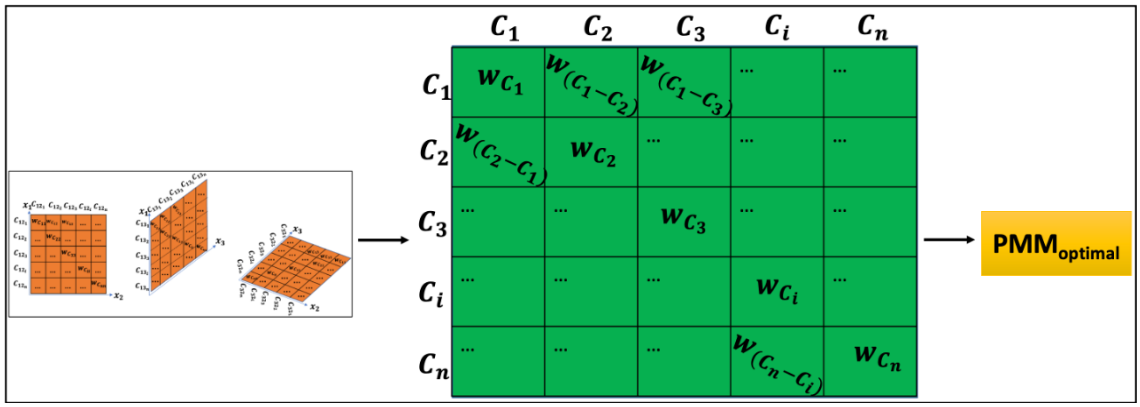

Fig. 4. Conversion of Two-Dimensional PMM matrix into optimal set of performance measurement metrics vector

B. The Testbed

To process and evaluate the proposed framework, GHNHSFS DR 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. Table III summarises the hardware specifications as well as used Python environment and MATLAB toolboxes.

C. Dataset and Experiments

The evaluation of the proposed performance assessment mechanism comprises three primary experiments, each with its own distinct set of component specifications. Table IV provides a summary of the conducted experiments.

The experiments primarily focus on detecting O6-methylguanine-DNA methyltransferase (MGMT) through Magnetic Resonance Imaging (MRI) scans in Experiment 2, as well as detecting and predicting Diabetic Macular Edema (DMO) through fundus scans in Experiments 1 and 3. The testing and validation of PMM tool involved two main datasets. One dataset includes BraTS data obtained from the RSNA-ASNR-MICCAI Brain Tumor Segmentation challenge 2021, consisting of 2,000 cases (8,000 MRI scans) available as NIfTI files (.nii.gz). These MRI scans encompass four different modalities: native (T1), post-contrast T1-weighted (T1Gd), T2-weighted (T2), and T2 Fluid Attenuated Inversion Recovery (T2-FLAIR) volumes, acquired from various institutions using different clinical protocols and scanners. To optimize computational time and processing resources, the NIfTI files were converted into .png format, taking into account the acquisition information of slice start and end. The retinal dataset includes both labeled and unlabeled versions, comprising over 10,000 fundus color scans of patients with DMO. Conversely, the BRATS dataset is a simple, unlabeled dataset composed of grayscale scans containing patients with MGMT.

Convolutional Neural Network (CNN), specifically the Densely Connected Network (DenseNet) model, was employed in all experiments for the training, testing, and validation stages. Subsequently, the ensuing discussion revolves around the optimal PMM 2D matrices, as well as the $PMM_{optimal}$ vector obtained for each respective scenario.

D. Results and Discussion

The variation in user input within the proposed evaluation mechanism influences the optimal parameters that are obtained. Specifically, the set of weights derived from each experiment differs due to factors such as the characteristics of the data, the specific task performed by the model, and the problem domain. Furthermore, these dimensions also affect the resulting correlations that establish connections between different metrics and assess the relationship among these parameters.

This justification is further supported by the obtained PMM 2D matrix for each scenario analysed. In Experiment 1, the optimal PMM vector excludes MAE, PR, F-m, DC, and LCE. However, the results differ slightly in the case of Experiment 2. Specifically, Acc, PPV, and NPV exhibit negative or null weights, leading to their exclusion from the final set. Experiment 3 similarly eliminates AUC, F-m, and Acc in addition to NPV and P metrics. These results demonstrate the impact of data characteristics, including class inclusion and balancing criteria, on the effective selection of DL performance measurement parameters. The chosen parameters establish a direct connection between their mathematical representation and the problem domain, as well as the specific task at hand. Figure **??** illustrates the 2-D PMM matrix for each of the aforementioned experiments.

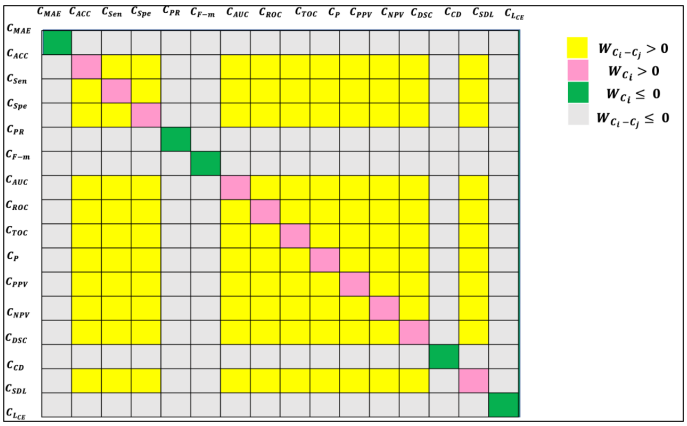

Fig. 5. Experiment 1 2D-PMM Results

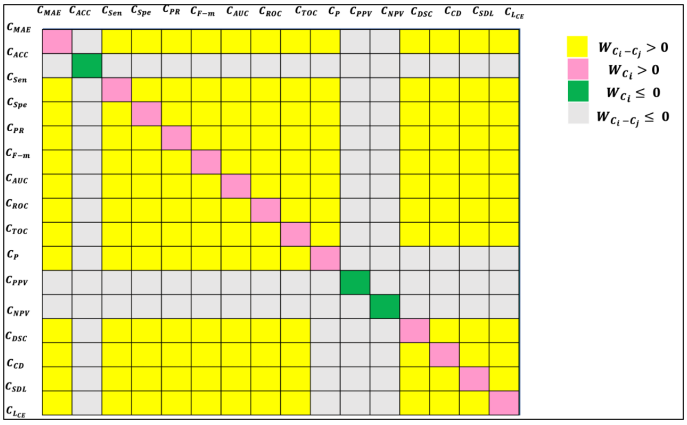

Fig. 6. Experiment 2 2D-PMM Results

The proposed evaluation mechanism successfully addresses the challenge of lacking standardization in evaluation metrics used for different ML/DL scenarios, particularly in healthcare applications. Overcoming this challenge has the potential to enhance developed applications. Without standardized metrics, comparing the performance of different algorithms across diverse datasets and applications becomes difficult. By considering correlations between evaluation metrics, the proposed

TABLE III
HARDWARE AND SOFTWARE SPECIFICATIONS

	Specifications
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
Software	- Python environment and MATLAB 2019b - 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

TABLE IV
EXPERIMENTATION SCENARIOS

Experiment	Problem Specification	Task Indentification	Task Characteristics
1	Disease detection	Classification	Coloured fundus images Labelled data Classes are balanced
2	Disease detection	segmentation	MRI gray-scale images Unlabelled data Classes are imbalanced
3	Disease prediction	Segmentation and classification	Coloured fundus images Unlabelled data Classes are imbalanced

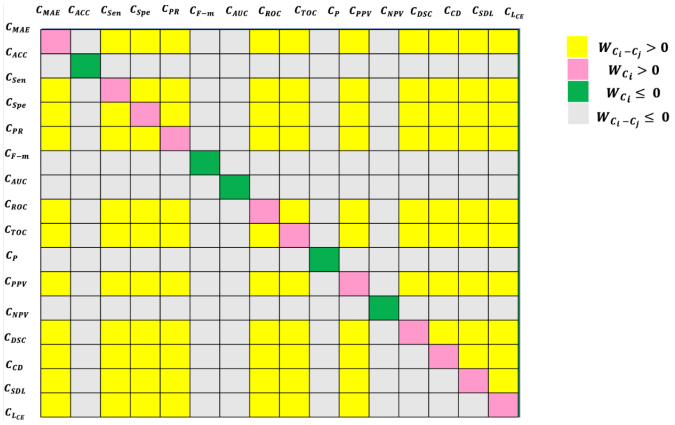

Fig. 7. Experiment 3 2D-PMM Results

mechanism overcomes the limited scope of existing evaluation mechanisms. Many existing metrics only focus on specific aspects of model performance, such as accuracy or precision, which may provide a narrow view of overall effectiveness and limit real-world applicability.

Experiment 1 and 3, which utilize the labeled Retinal dataset, face the challenge of bias inherent in the dataset. Consequently, the resulting DenseNet model may also exhibit bias, leading to inaccurate predictions and negative impacts on decision-making, especially in healthcare. Hence, according to the proposed evaluation mechanism, it is crucial to incorporate reliable and comprehensive metrics that align with the specific dataset requirements.

The results obtained from the proposed PMM tool demonstrate its ability to provide clear and precise interpretation and explainability of the usefulness of the DL algorithm. The tool effectively addresses critical problems related to disease detection and prediction in all three tested and validated scenarios. Additionally, it safeguards the integrity of the DL model, mitigating the risk of inaccurate predictions.

IV. CONCLUSION

In healthcare applications, the impact of evaluation metrics on both classic and automated methods is of utmost importance due to the potential life-threatening consequences of incorrect decisions. Selecting appropriate evaluation metrics becomes crucial to ensure the reliability and safety of algorithms for patient use [24].

During the testing phase, the choice of evaluation metrics significantly influences the accuracy and reliability of the results. The use of suitable metrics, such as sensitivity and specificity, plays a vital role in ensuring that the algorithm accurately identifies true positives and true negatives. This becomes particularly critical in healthcare applications, where the implications of missed or false diagnoses can have severe consequences [25].

This paper proposes a novel adaptive and reliable evaluation mechanism for ML/DL algorithms. The mechanism has demonstrated promising results in providing an optimized set of evaluation metrics that are suitable for both the training and validation stages. The PMM tool can be seamlessly integrated into various applications and possesses the capability to adapt based on the specific dimensions of the problem. Moreover, the proposed evaluation mechanism addresses the inherent challenges in healthcare applications, where the stakes are higher, and the consequences of incorrect decisions can be life-threatening. By carefully selecting evaluation metrics, the

mechanism ensures that the algorithm is not only accurate but also reliable and safe for patient use. It enables healthcare professionals to have confidence in the algorithm's performance and trust its output.

The testing phase plays a critical role in assessing the algorithm's performance, and the choice of evaluation metrics directly impacts the accuracy and dependability of the results. By utilizing appropriate metrics such as sensitivity and specificity, the mechanism ensures that the algorithm effectively identifies true positive and true negative cases. This level of accuracy is paramount in healthcare, as misdiagnoses or false diagnoses can lead to detrimental outcomes for patients. The proposed evaluation mechanism empowers healthcare practitioners with the tools to make informed decisions and improve patient care.

The significance of the proposed evaluation mechanism extends beyond individual experiments and scenarios. It represents a step forward in standardizing evaluation metrics for ML/DL algorithms in healthcare. This standardization facilitates comparisons between different algorithms, datasets, and applications, enabling researchers and practitioners to assess and benchmark performance effectively. It promotes transparency, reproducibility, and advancements in the field, ultimately leading to improved healthcare outcomes.

The PMM tool, which forms the cornerstone of the proposed evaluation mechanism, offers adaptability and versatility. It can be seamlessly integrated into various healthcare applications, accommodating the specific needs and requirements of different domains. Its adaptability ensures that the evaluation metrics align with the given dimensions of the problem, resulting in a comprehensive and tailored evaluation process.

In conclusion, the proposed adaptive and reliable evaluation mechanism addresses the critical challenges faced in evaluating ML/DL algorithms, particularly in healthcare applications. By selecting the most appropriate evaluation metrics and utilizing the PMM tool, this mechanism enhances the assessment of algorithm performance, promotes patient safety, and contributes to the advancement of the field. Its impact extends beyond individual experiments, offering a standardized approach to evaluation that improves comparability, reproducibility, and overall effectiveness in healthcare applications.

REFERENCES

[1] J. Olczak, et al., "Presenting artificial intelligence, deep learning, and machine learning studies to clinicians and healthcare stakeholders: an introductory reference with a guideline and a Clinical AI Research (CAIR) checklist proposal," Acta orthopaedica, vol. 92, no. 5, pp. 513-525, 2021.

[2] R. S. K. Vijayan, et al., "Enhancing preclinical drug discovery with artificial intelligence," Drug discovery today, vol. 27, no. 4, pp. 967-984, 2022.

[3] V. Chang, et al., "An artificial intelligence model for heart disease detection using machine learning algorithms," Healthcare Analytics, vol. 2, p. 100016, 2022.

[4] D. Ardila, A. P. Kiraly, S. Bharadwaj, B. Choi, J. J. Reicher, L. Peng, et al., "End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography," Nature Medicine, vol. 25, no. 6, pp. 954-961, Jun. 2019.

[5] C. Krittanawong, H. Zhang, Z. Wang, M. Aydar, and T. Kitai, "Artificial intelligence in precision cardiovascular medicine," Journal of the American College of Cardiology, vol. 75, no. 23, pp. 2996-3013, Jun. 2020.

[6] S. Wang, M. Zhou, Z. Liu, Z. Liu, Z. Hu, and Q. Dai, "A multi-view deep convolutional neural networks for lung nodule segmentation," in International Conference on Information Processing in Medical Imaging, Cham, Switzerland, Jun. 2016, pp. 63-75.

[7] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. Cambridge, MA, USA: MIT Press, 2016.

[8] S. Raschka and V. Mirjalili, Python Machine Learning, 3rd ed. Birmingham, UK: Packt Publishing, 2019.

[9] A. P. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms," Pattern Recognition, vol. 30, no. 7, pp. 1145-1159, Jul. 1997.

[10] L. A. Hendricks, B. Burns, K. Saenko, T. Darrell, and M. Rohrbach, "Women also snowboard: Overcoming bias in captioning models," in Proceedings of the European Conference on Computer Vision, Munich, Germany, Sep. 2018, pp. 793-809.

[11] Z. Obermeyer and E. J. Emanuel, "Predicting the future—big data, machine learning, and clinical medicine," New England Journal of Medicine, vol. 375, no. 13, pp. 1216-1219, Sep. 2016.

[12] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," Information Processing and Management, vol. 45, no. 4, pp. 427-437, Jul. 2009.

[13] T. Janssen and A. Smeulders, "Adversarial evaluation of deep neural networks," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, Jun. 2019, pp. 8622-8631.

[14] B. Kim, H. J. Kim, and G. Kim, "Learning to generate diverse and creative resumes through adversarial training," in Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, Jan. 2019, vol. 33, pp. 6839-6846.

[15] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning requires rethinking generalization," in International Conference on Learning Representations, New Orleans, LA, USA, May 2019.

[16] A. Saxena, et al., "Metrics for offline evaluation of prognostic performance," Int. J. Prognostics Health Manag., vol. 1, no. 1, pp. 4-23, 2010.

[17] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient over F1 score and accuracy in binary classification evaluation," BMC genomics, vol. 21, pp. 1-13, 2020.

[18] J. H. Joloudari, et al., "Resource allocation optimization using artificial intelligence methods in various computing paradigms: A Review," arXiv preprint arXiv:2203.12315, 2022.

[19] Y. Park, et al., "Comparison of methods to reduce bias from clinical prediction models of postpartum depression," JAMA Netw Open, vol. 4, no. 4, pp. e213909-e213909, 2021.

[20] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014.

[21] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in International Conference on Medical image computing and computer-assisted intervention, Springer, Cham, 2015, pp. 234-241.

[22] J. Gao, L. Ma, and W. Zhou, "A comprehensive review on performance evaluation metrics for regression, classification, and ranking problems in machine learning," Expert Systems with Applications, vol. 152, p. 113469, 2020.

[23] D. M. Powers, "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation," Journal of Machine Learning Technologies, vol. 2, no. 1, pp. 37-63, 2011.

[24] A. Al-Turaiki and N. Alajlan, "Evaluation metrics for machine learning algorithms in healthcare applications: A review," Journal of Healthcare Engineering, vol. 2021, article ID 6682347, 2021, doi: 10.1155/2021/6682347.

[25] A. Haque and M. H. Uddin, "Evaluation metrics for machine learning algorithms in healthcare applications: A review," Artificial Intelligence in Medicine, vol. 118, p. 102064, 2021, doi: 10.1016/j.artmed.2021.102064.