Interpretation of Artificial Intelligence Models in Healthcare

Interpretation of Artificial Intelligence Models in Healthcare: A Pictorial Guide for Clinicians

Ali Abbasian Ardakani[1]*, Omid Airom[2], Hamid Khorshidi[3], Nathalie J Bureau[4], Massimo Salvi[5], Filippo Molinari[5], U Rajendra Acharya[6]

[1] *Department of Radiology Technology, School of Allied Medical Sciences, Shahid Beheshti University of Medical Sciences, Tehran, Iran*

[2] *Department of Mathematics, University of Padova, Padova, Italy*

[3] *Department of Information Engineering, University of Padova, Padova, Italy*

[4] *Department of Radiology, Centre hospitalier de l'Université de Montréal, Montreal, Quebec, Canada*

[5] *Biolab, PolitoBIOMedLab, Department of Electronics and Telecommunications, Politecnico di Torino, Turin, Italy*

[6] *School of Mathematics, Physics and Computing, University of Southern Queensland, Springfield, Queensland, Australia*

∗ Corresponding author:

Ali Abbasian Ardakani

Address: Department of Radiology Technology, School of Allied Medical Sciences, Shahid Beheshti University of Medical Sciences, P.O.Box: 1971653313, Tehran, Iran.

E-mail: Ardakani@sbmu.ac.ir, A.ardekani@live.com. ORCID: 0000-0001-7536-0973

Interpretation of Artificial Intelligence Models in Healthcare: A Pictorial Guide for Clinicians

Highlights

- This paper provides a comprehensive guide for clinicians to interpret the outputs of AI models.
- Twenty-two evaluation metrics are reviewed for the evaluation and interpretation of AI models.
- Two ML models and two DL models are developed based on a breast cancer database to represent real-world situations.
- The metrics are used to evaluate and compare the performance of models.

Abstract

With the explosion of digital health records available in the healthcare industry, artificial intelligence (AI) models can play a more effective role in managing patients. Machine learning (ML) and deep learning (DL) techniques are two groups of methods used to develop predictive models that serve to improve the clinical processes in the healthcare industry. These models are also implemented in medical imaging machines to empower them with an intelligent decision system to aid physicians in their decisions and increase the efficiency of their routine clinical practices. The physicians who are going to work with these machines need to have an insight into what happens in the background of the implemented models and how they work. More importantly, they need to be able to interpret their predictions, assess their performance, and compare them to find the one with the best performance and fewer errors. This review aims to provide an accessible overview of key evaluation metrics for physicians without AI expertise. In this review, we developed four real-world diagnostic AI models (two ML and two DL models) for breast cancer diagnosis using ultrasound images. Then, twenty-two of the most commonly used evaluation metrics were reviewed uncomplicatedly for physicians. Finally, all metrics were calculated and used practically to interpret and evaluate the outputs of the models. Accessible explanations and practical applications empower physicians to effectively interpret, evaluate, and optimize AI models to ensure safety and efficacy when integrated into clinical practice.

Keywords: Explainable Artificial Intelligence**,** Deep Learning Models, Machine Learning Models, Clinical Translation, Healthcare, Radiomics.

1. Introduction

Artificial intelligence (AI) is defined as the science and technique used to develop intelligent systems, specifically computer programs, to mimic human behaviour (Jiang et al., 2017; McCarthy, 2007). In other words, AI gives machines the ability to undertake actions that rely on human intelligence and enables computers to act like humans when solving complex decision-making tasks (Allen, 2020; Janiesch et al., 2021). AI has been applied in game-playing, speech recognition, natural language processing, computer vision, and classification (McCarthy, 2007). It has been used to develop medical decision support systems for disease diagnosis, treatment design, drug interactions and discovery, image processing, and other tools that can assist physicians in their tasks (Manne and Kantheti, 2021; Yu et al., 2018). With the rapid increase in AI tools being developed and deployed in healthcare, particularly in medical imaging, proper evaluation of these models is critical prior to clinical use. Rigorous assessment ensures patient safety, optimized utilization of healthcare resources, and informed clinical decision-making. However, many healthcare professionals currently lack sufficient training in this area.

A lot of attention has been paid to the importance and the capability of AI in healthcare over the past 5 years (Secinaro et al., 2021). AI tools are being rapidly developed for tasks ranging from detection and diagnosis to image improvement and workflow enhancement. The recent improvements in AI have brought up this question and concern: will AI-empowered machines replace physicians in the future? The European Society of Radiology (ESR) addressed this concern in a white paper (Neri et al., 2019). The answer is simple: NO. AI tools will not be a replacement for physicians (especially radiologists), and they will only serve as an aiding tool to help them achieve more accurate results (Manne and Kantheti, 2021). Radiology is one of the fields in which a lot of effort has been made to develop AI tools according to the data available, including pattern recognition, spatial modelling, algorithmic scoring, and long-term surveillance (Harvey and Gowda, 2020). Although AI tools can make a great impact in the healthcare industry, there are some obstacles that must be solved to take full advantage of them. One of the most important problems is the lack of a standard evaluation of these tools in the regulations. To solve this problem, the Food and Drug Administration (FDA) started to update the regulations by creating evaluation guides for AI tools and designing a

completely new regulatory paradigm (Graham, 2016). In addition, the rapid creeping of AI in healthcare made the World Health Organization (WHO) publish a report on ethical concerns, opportunities, and challenges of AI (World Health, 2021). In the report, they recommended and adjusted policies and ethical principles in using AI in healthcare. Thereby, with the fast-growing medical AI tools and their consistency in healthcare, physicians need to keep up with them and share their knowledge of AI-based tool development (Miller, 2019). Collaborative efforts between clinicians and medical physicists in AI model development can yield significant benefits. By actively involving clinicians in the development process, AI models can be tailored to address specific clinical problems and incorporate domain-specific knowledge. This interdisciplinary collaboration fosters a better understanding of AI models, promotes trust, and enhances the clinical relevance and utility of the developed models.

Many of the AI systems are like black boxes. This means that there is no clear understanding of how they work, and the outputs should be analyzed and interpreted. Physicians need to understand the core of AI models and be able to interpret the outputs to avoid any errors (Paranjape et al., 2019). Interpretability is a crucial aspect of healthcare AI models as it allows clinicians to trust and validate the decisions made by these models. Understanding how AI models arrive at their predictions is essential for transparency, accountability, and patient safety. The interpretability of AI models not only enhances transparency but also plays a crucial role in building trust and acceptance among healthcare professionals. Moreover, it is essential for physicians to actively engage and share their knowledge with AI researchers to collaboratively develop practical AI tools that effectively address real clinical problems. The output of AI-developed models used in healthcare may seem non-transparent for health professionals (Lötsch et al., 2021). As a result, physicians require proper training in the field of AI to improve their work, reduce costs and errors, increase accuracy, and provide transparency in its use and outcomes (Paranjape et al., 2019). To overcome this limitation, in our previous paper, we shed light on radiomics-based black-box models and provided a pictorial guide for physicians to interpret radiomics features and use them in routine clinical practices (Abbasian Ardakani et al., 2022).

In this second educational paper, we aim to provide a comprehensive guide for clinicians to interpret and evaluate the outputs of diagnostic AI models in healthcare, enabling them to make informed decisions based on the model's predictions. We present and explain twenty-two of the most commonly used evaluation metrics to assess and interpret the outputs of AI tools. We aim to provide readers with enough background knowledge enabling them to

feel comfortable when working with diagnostic medical AI systems. These metrics are the confusion matrix-based features, the receiver operating characteristic (ROC) Curve, Precision-Recall Curve, Cumulative Gains Curve, Lift Curve, Decision Curve, Calibration Curve, and Grad-CAM (Fig. 1). In the following, each of these metrics will be explained, and interpreted using diagnostic real-world AI models. In this review paper, first, we evaluate and interpret the outputs of two machine-learning (ML) models. In the next step, two deep-learning (DL) based models are trained, and tuned and the outputs are analyzed and compared to get a comprehensive view of both DL and ML models.

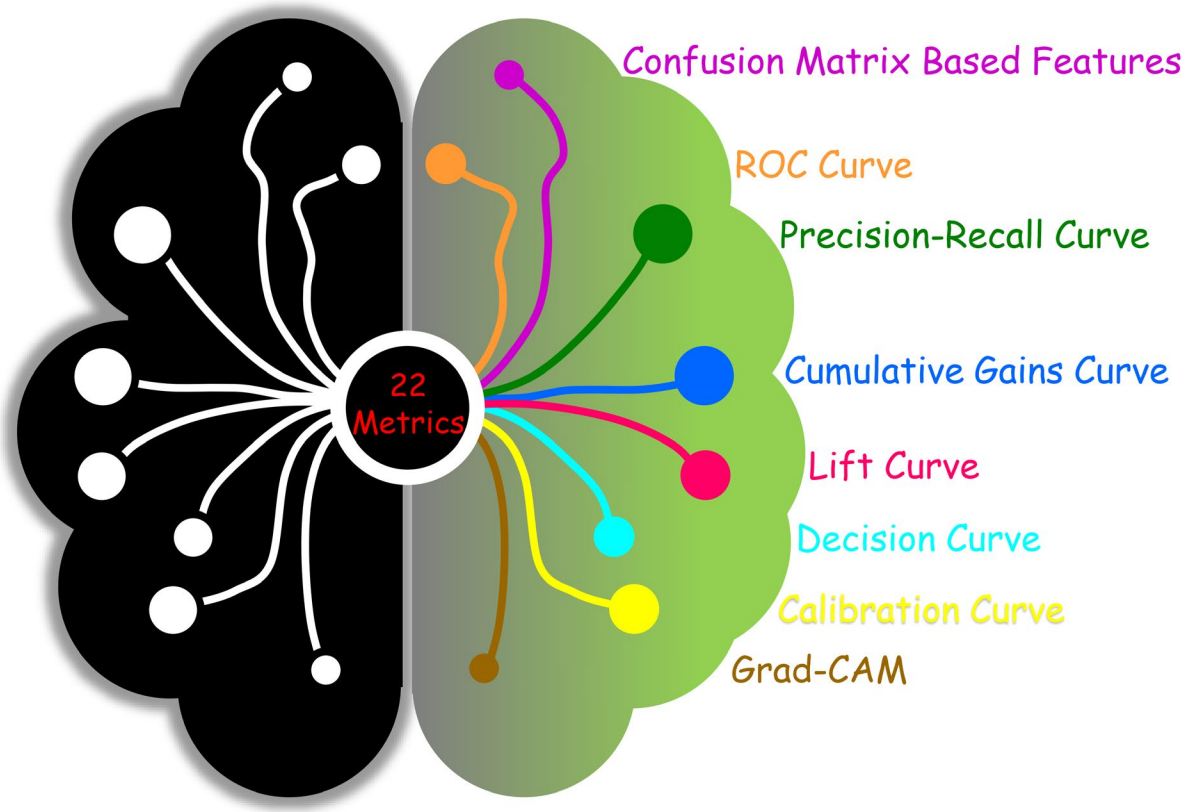


Fig. 1. Overview diagram of the aim and approach of this review to evaluate diagnostic (classification) medical AI systems.

2. Machine Leaning Models

In this section, two real models (model 1: support vector machine (SVM), model 2: K-nearest neighbors (KNN)) were developed based on radiomics features of 109 benign and 123 malignant breast lesions on ultrasound images (Abbasian Ardakani et al., 2023) to mimic real clinical diagnostic task. K-fold cross-validation was employed to improve the generalization

and robustness of the model. Additional information about the dataset, the models, the radiomics features as well as the equation of metrics are provided in the supplementary material.

2.1. Confusion Matrix Based Features, Group 1

Confusion matrix is a crucial tool for evaluating the performance of diagnostic models. It consists of four values: true positive (TP), false positive (FP), true negative (TN), and false negative (FN). TP and TN indicate positive and negative cases correctly diagnosed, respectively. Similarly, FP and FN are negative and positive cases wrongly diagnosed, respectively (Fig. 2A). In healthcare, positive cases are commonly assigned to malignancy or the presence of a disease and negative cases represent benignity or the absence of a disease. The following metrics are used to determine the performance of diagnostic models (Japkowicz and Shah, 2011): sensitivity, specificity, accuracy, precision, F1-Score, balanced accuracy, and geometric mean (G-mean). The confusion matrix is a simple yet effective approach for highlighting a model's strengths and weaknesses and guiding the direction of modelling improvements. The confusion matrix has pros and cons. The main advantages of this matrix and the extracted metrics are their simplicity and the information they provide about the distribution of correctly and incorrectly diagnosed cases. On the other hand, the confusion matrix does not present information about errors, calibration degree, and the benefit of a model according to the predicted values (probabilities). Therefore, to obtain a complete evaluation of a diagnostic model, it is recommended to use additional techniques alongside the confusion matrix (Bekkar et al., 2013; Japkowicz and Shah, 2011; Kubat et al., 1998; Saito and Rehmsmeier, 2015).

Interpretation

The confusion matrices for both models are shown in Fig. 2B. At a glance, it can be conceived that model 1 correctly diagnosed both positive and negative cases better. However, the main question is "How Much Better?" We need to calculate quantitative metrics extracted from confusion matrices to answer this question.

Fig. 2B shows the confusion matrices and calculated related metrics of models. Sensitivity (recall) defines the percentage of positive cases that are correctly diagnosed by a model (TP) to all positive cases (TP+FN). Thereby, as the number of TP cases of model 1 is greater than that of model 2, the sensitivity of model 1 is higher than that of model 2. On the other hand, specificity defines the percentage of negative cases that are correctly diagnosed by

the model (TN) to all negative cases (TN+FP). Thereby, as model 1 correctly diagnosed a higher number of negative cases (TN), the specificity of model 1 is higher than that of model 2. The accuracy quantifies the ratio of correctly predicted cases (TP+TN) to the total number of cases. Thereby, as model 1 predicted more cases correctly, the accuracy of model 1 is better than that of model 2. Precision computes the ratio of correctly identified positive cases (TP) to the total number of positive cases predicted by a model (TP+FP). Thereby, as both TP and FP cases of model 1 are greater and lower than model 2, respectively, model 1 indicates better precision.

In classification problems, the number of cases in the classes is not always equal. When the distribution of cases among the classes becomes too uneven, the issue of class imbalance arises. In other words, in binary classification problems with imbalanced datasets, the ratio of the majority class to the minority class will be greater than 1. Technically, any dataset with such a ratio greater than one can be considered an imbalanced dataset, but in most studies, class ratios of 5 to 1 or more have been considered as imbalance problems (Maldonado and López, 2014). In most real-world cases, the minority class is the class of interest, and any error in the classification of cases belonging to this class may lead to higher costs, such as in medical cases and the diagnosis of patients.

The F1-Score considers both sensitivity and precision by harmonically averaging them. When TP and TN cases are more sensitive to your aim or you have a balanced database, the accuracy would be a better metric, while when FP and FN are crucial or the database is imbalanced, the F1-Score reflects better information. Model 1 has a higher F1-Score than model 2, implying better model performance and a superior balance between precision and sensitivity. The balanced accuracy assesses the model's capability to correctly classify cases in all classes through the arithmetic averaging of sensitivity and specificity. Model 1 has a higher balanced accuracy value, indicating more accurate identification of both positive and negative cases. The G-mean (and also balanced accuracy) considers both sensitivity and specificity equally and measures the model's ability to correctly categorize cases (especially for imbalanced databases). It is calculated by taking the square root of the product of sensitivity and specificity. Model 1 has a higher G-mean value, which indicates a more accurate identification of both positive and negative cases.

The mentioned indices except accuracy are useful only for binary classification problems. Sensitivity and specificity are less affected by class imbalance but they do not

consider the rate of TN and TP, respectively. However, the accuracy provides an intuitive interpretation but may not be suitable for imbalanced datasets. Precision is better suited for balanced datasets with significant costs for FP. Compared to other metrics in this section, the F1-Score, balanced accuracy and G-mean help evaluate the output of AI models for imbalanced databases (Bekkar et al., 2013; Brodersen et al., 2010; Bush et al., 2008; Grandini et al., 2020; Japkowicz and Shah, 2011; Kubat et al., 1998; Powers, 2020; Saito and Rehmsmeier, 2015).

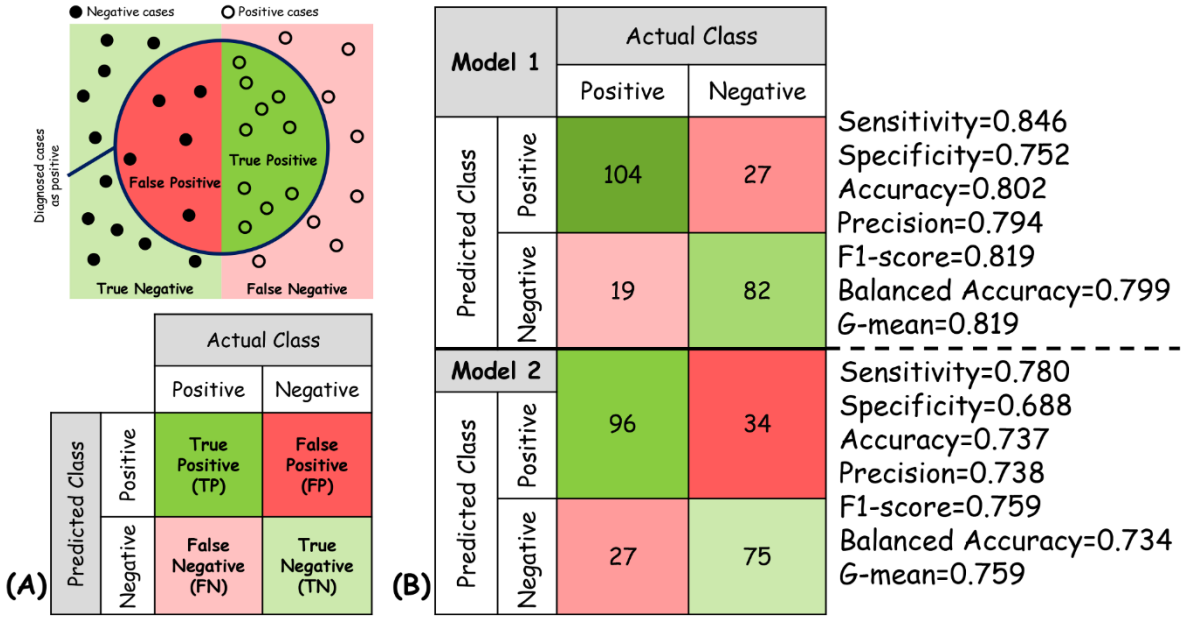


Fig. 2. A confusion matrix and the related values according to the model's prediction (A). Confusion matrices of the models 1 and 2 and their related evaluation metrics (B).

2.2. ROC Curve

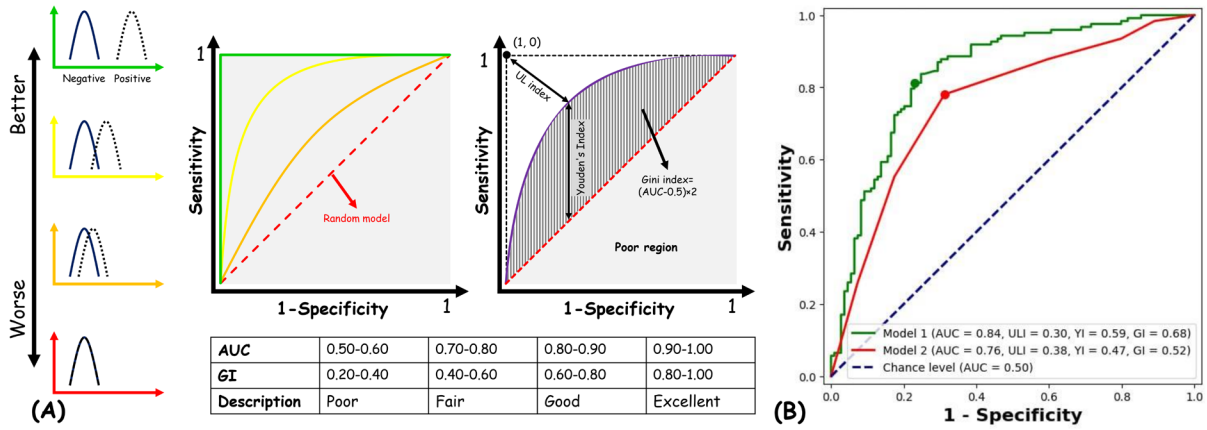
The receiver operating characteristic (ROC) curve analysis is commonly used for the evaluation of binary classification problems. The ROC curve illustrates the trade-off between the TP rate (TPR or sensitivity) and FP rate (FPR or 1-specificity) at various classification thresholds. Four different indices can be obtained from a ROC curve (Bradley, 1997; Engler et al., 2004; Fawcett, 2006; Hanley and McNeil, 1982; Koyama et al., 2016; Youden, 1950): 1) area under the ROC curve (AUC), 2) Upper-left index (ULI), 3) Youden's index (YI), and 4) Gini index (GI) (Fig 3A).

Interpretation

The AUC provides an overall summary of the model's performance across all possible thresholds. It is useful for both balanced and imbalanced datasets, as it balances the performance across all classes. However, it has limitations in disclosing specific performance at a threshold and is not suitable for applications when the costs of FP and FN are different (one of the FP or FN is more important than the other). Based on AUC values, model 1 outperformed model 2, indicating that model 1 provides better separation between the two classes in different thresholds (Fig. 3B).

Determining optimal cut-off thresholds is crucial for obtaining maximum sensitivity and specificity in medical diagnostic systems. Two common indices, ULI and YI, are used to achieve this. The ULI identifies the optimal cut-off threshold closest to the upper-left corner of the ROC curve, while YI, which ranges from 0 to 1 (with 1 being the best model), quantifies the greatest distance between the ROC curve and the chance level (where the probability of correctly diagnosing cases is 50%, represented by the dashed line in Fig. 3). The optimal cut-off value is the one that maximizes YI and minimizes ULI. Therefore, based on ULI and YI, model 1 outperformed model 2, indicating its ability to distinguish between positive and negative cases with higher performance (Fig. 3B). However, these indices have some limitations, including their failure to consider the costs of FPs and FNs, as well as more difficulty in their interpretations in comparison with other metrics (Bekkar et al., 2013; Koyama et al., 2016; Youden, 1950).

The GI was used originally in economics to measure inequality in income or wealth distribution, but it can also be used as an evaluation metric for binary classification models in AI studies. It is calculated based on the Lorenz curve (Mauguen and Begg, 2016) and ranges from 0 to 1 (1 belongs to the best model). The GI indicates how a model assigns a label to cases from different classes. If a specific label is more assigned to a specific class, the GI becomes higher and vice versa. Therefore, based on the higher GI value, model 1 indicated better classification performance (Fig. 3B). The GI is less sensitive to imbalanced datasets as it considers both true positives and false positives. In the end, ROC curve-based indices should be used alongside other metrics for a more comprehensive evaluation of model performance.



AUC	0.50-0.60	0.70-0.80	0.80-0.90	0.90-1.00
GI	0.20-0.40	0.40-0.60	0.60-0.80	0.80-1.00
Description	Poor	Fair	Good	Excellent

Fig. 3. ROC Curves and the related possible ranges for a classifier (A). The ROC curves of our models with the corresponding metrics values are shown in (B).

2.3. Confusion Matrix Based Features, Group 2

2.3.1. Matthews Correlation Coefficient

The Matthews correlation coefficient (MCC) is a measure to assess another aspect of the performance (correlation) of diagnostic models. It can be used to evaluate the quality of predictions in both binary and multi-class classification problems for both balanced and imbalanced datasets (Fig. 4A). The MCC uses TPs, TNs, FPs, and FNs obtained by a model and calculates a single numerical value. In other words, it turns a confusion matrix into a single value (Fig. 4B). The MCC is calculated by subtracting the product of the number of incorrectly diagnosed cases from the number of correctly diagnosed cases (FP×FN and TP×TN, respectively). Then, this value is divided by the square root of the product of the total number of positive (TP+FP) and negative predicted cases (TN+FN), as well as the total number of positive (TP+FN) and negative cases (TN+FP) (See supplementary material). To calculate the MCC for multi-class problems, we need to utilize the terms C, S, p_k, and t_k, which respectively represent the total number of correctly predicted cases (the diagonal of the confusion matrix), the total number of samples, the total number of the number of times class k was predicted by the model, and the number of times class k truly occurred (Grandini et al., 2020; Jurman et al., 2012; Matthews, 1975).

Interpretation

In binary classification problems, the MCC value ranges from -1 to 1. A binary classifier with an MCC equal to -1 indicates inverse predictions (assigning negative labels to all positive

ground-truth labels and vice versa). Hence, there is an inverse correlation between the predicted and ground-truth labels. On the other hand, if the MCC is zero, it means that the classifier has a performance similar to a random model (no correlation), and if it goes near 1, it shows that the model is performing well (positive correlation, Fig. 4C). Also, similar to the two-class classification, in multi-class problems, the range of MCC is from -1 to 1.

Our results indicate that model 1 outperformed model 2 according to the MCC values (0.60 vs. 0.47, respectively). This means that model 1 made predictions with higher performance in both classes. In other words, the correlation of model 1's predicted labels with true labels is higher than that of model 2.

Comparing the MCC with the balanced accuracy, the MCC provides a more informative insight into the models' overall performance. It can be concluded that a high MCC shows that the four values in the confusion matrix, sensitivity, specificity, precision, and negative predictive value (TN/(TN+FN)) all have high values. However, this cannot be concluded from the balanced accuracy. Balanced accuracy emphasizes overall accuracy while the MCC provides more detailed information considering all measures in the confusion matrix. Thereby, MCC is superior to balanced accuracy in evaluating the performance of a model when the positive and negative classes are at the same level of importance, and making predictions is as important as classifying the existing true labels. Otherwise, if classifying the existing true labels is more important, the balanced accuracy would be a better metric to assess the performance of the model. In addition, MCC is a better choice to rank and compare different models based on the same database, while balanced accuracy may lead to misinterpretation of models' performances. On the other hand, it may be hard to interpret the MCC of multi-class or even binary classifiers as there is no guideline to directly obtain the performance of the classifiers (Chicco et al., 2021a).

Comparing F1-Score and MCC, the F1-Score is used to obtain the balance between a classifier's precision and sensitivity while again the MCC has a more overall assessment over the classifier's predictions. The F1-Score is not a good measure to assess the performance of classifiers trained to solve imbalance problems in comparison with MCC. The MCC can overcome this problem and is a better evaluation criterion for imbalanced problems. In addition, as mentioned in the previous paragraph, the MCC of a model is high if the four rates of the model are high. Thereby, MCC reflects better information about the model's

performances. We suggest using MCC as a metric reference for models' ranking to prevent misleading results obtained by F1-Score (Chicco and Jurman, 2020; Wardhani et al., 2019).

Compared to the AUC, the MCC gives a summary of the overall performance of a model, while the AUC shows how well the model performs in differentiating and distinguishing the classes within the problem. Typically, both AUC and MCC indicate high levels of consistency in both balanced and imbalanced datasets. However, AUC is preferable due to its greater discriminatory power. MCC is better suited for situations where correct predictions and the importance of positive and negative classes are equal (Chicco and Jurman, 2023; Halimu et al., 2019).

2.3.2. Cohen's Kappa

Cohen's Kappa (CK) is another statistical measure used to assess the quality of predictions made by a binary classifier on only balanced datasets (Fig. 4A). CK is a tool used to investigate the level of agreement between two models or raters. As known in binary classification problems, there are two classes: negative and positive. According to CK, one of the raters is the classifier, and the other one is the ground-truth labels. CK is calculated using the observed agreement (p_o, the number of agreements between the two raters divided by the total number of observations) and expected agreement (p_e, the number of agreements obtained by chance divided by the number of total observations). To calculate CK, $p_e$is subtracted from p_o, and the obtained value is divided by the subtraction of the $p_e$due to chance from 1 (See supplementary material) (Cohen, 1960; Delgado and Tibau, 2019; Grandini et al., 2020).

Fleiss' kappa (FK) is another statistical measure very similar to CK, used to assess the level of agreement between different raters in a multi-class classification problem. The number of samples in the dataset, the number of raters, and the number of classes are used to calculate FK. In other words, it uses the distribution of ratings and calculates the proportion of raters that have an agreement for each subject. To calculate FK, the mean proportion of the total agreement of all raters and the mean agreement expected by chance are determined (observed agreement).

Then, they are used in the same formula as CK to calculate FK 1 (See supplementary material) (Fleiss, 1971; Nichols et al., 2010).

Interpretation

CK is a measure that helps to understand the agreement between two methods (raters). In other words, the CK can be used to measure the performance of a classifier (the first method) based on the degree of the agreement by chance between the classifiers and the ground-truth labels, which are defined by a gold standard method (for instance histopathologic examination in medicine). To calculate CK, the values of the confusion matrix are used (Fig. 4B). TPs and TNs show the number of cases in which the two raters agreed on the results, and the number of FPs and FNs are the cases in which the classifier did not agree with the ground-truth labels. CK calculates the agreement of the two raters using TPs and TNs, compared with the agreement that would be expected by chance. Then, it shows how much better the agreement of a classifier is compared to what could be achieved by chance. CK falls between -1 and 1 (Fig. 4C), and the strength of agreement can be interpreted as follows: $CK<0.00$: poor, $0.00<CK<0.20$: slight, $0.21<CK<0.40$: fair, $0.41<CK<0.60$: moderate, $0.61<CK<0.80$: substantial, and $0.81<CK<1.00$: almost perfect (Landis and Koch, 1977). Also, the range of FK is between -1 and 1. CK and FK are not usually good choices to assess the performance of classifiers that are used to solve classification problems with imbalanced datasets.

Our results proved that model 1 diagnosed benign and malignant breast lesions with a higher degree of agreement than model 2 (CK= 0.60 and 0.47, respectively). In other words, models 1 and 2 indicate substantial and moderate agreements, respectively, which prove the better performance of model 1.

Cohen's kappa is highly affected by the distribution of the predictions. The number of TPs and TNs (diagonal elements of the confusion matrix) can make a huge change in the value of the CK. For example, if a model has a high sensitivity and low specificity (indicating asymmetric confusion matrix, Fig. 4B), then the value of the CK will be abnormal and it may lead to wrong interpretations and misinformation about the performance of the model. As a result, CK is not a good choice for evaluating the performance of models trained for imbalanced problems, only of it leads to a confusion matrix with asymmetric off-diagonal elements. In other words, when the dataset is imbalanced, it is more likely to obtain a confusion matrix that is not diagonally symmetric, and a small mistake in the minority class may lead to a great loss in CK (Fig. 4B). In these cases, the CK may not be a good metric which leads to wrong

information about the model's performance. However, the MCC is a good choice to evaluate the performance of classifiers used to solve problems with imbalanced datasets as it is not much affected by the distribution of the values in the confusion matrix. In conclusion, it is better to use the MCC if the dataset is imbalanced. If the confusion matrix is diagonally symmetric, then the MCC will be equal to CK, and if there is a difference between the distribution of these values in the confusion matrix, then there will be a significant difference between the MCC and CK (Chicco et al., 2021b; Delgado and Tibau, 2019; Grandini et al., 2020).

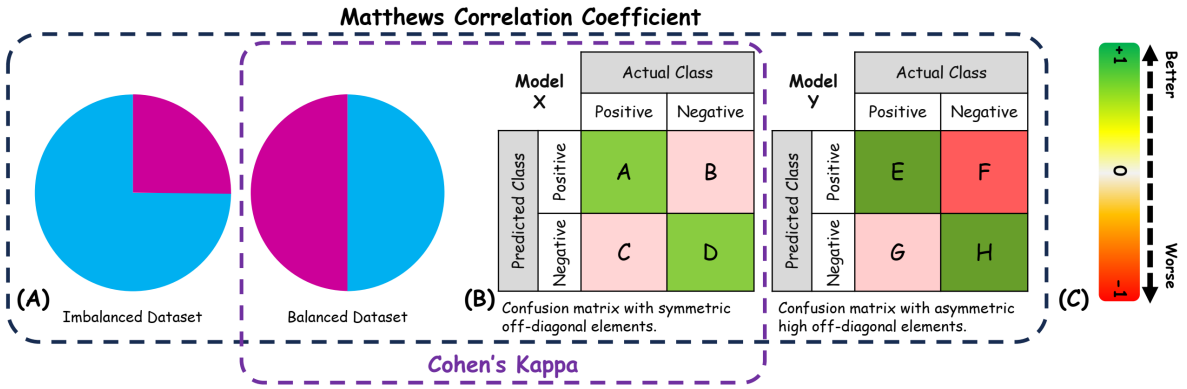


Fig. 4. The Matthews Correlation Coefficient (MCC) can be used for both balanced and imbalanced datasets (A), while Cohen's Kappa (CK) is preferably used for balanced datasets and confusion matrices with symmetric off-diagonal elements (B), The range of MCC and CK is from -1 to 1(C).

2.3.3. Top-K Accuracy

The top-k accuracy is measure that helps evaluate the performance of classifiers. It considers the top-k highest probabilities predicted by the classifier for the labels. It then checks if the true label for a given case is among those top-k predicted labels. If it is, the classifier's prediction is considered correct; otherwise, it is considered incorrect (Boyd et al., 2012; Kato and Hirohashi, 2019; Lapin et al., 2016). This measure is particularly useful for multi-class classification problems.

Interpretation

To understand the top-k accuracy better, let's look at its formula. When k is set to 1, the top-k accuracy becomes the same as classic accuracy. On the other hand, when k is set to the total number of classes in the classification problem, the top-k accuracy will always be 1, regardless of the classifier's correctness in predicting the labels with the highest probability (k=1). The

top-k accuracy ranges from 0 to 1. It's important to note that top-k accuracy is only applicable for evaluating the performance of classifiers in multi-class problems and cannot be used for binary classifiers. For binary classifiers, if k is set to 1, it is equivalent to classic accuracy, and if it is set to 2, the top-k accuracy will be 1. To compare classifiers effectively, comparing top-k accuracy to classic accuracy provides more realistic values. This ensures that small errors are not considered, making the comparison fairer between models.

In Fig. 5A, we can see the probabilities of five data points belonging to five different classes, sorted in decreasing order from left to right. The curve represents the relationship between k (on the x-axis) and the top-k accuracy (on the y-axis). At the beginning of the curve, when k is equal to 1, only the class with the highest probability is considered for the measure. In this case, there are only two correct predictions out of the five, resulting in a top-1 accuracy of 0.4. As 'k' increases to 2, the two predicted classes with the highest probabilities are considered, resulting in three correct predictions and a top-2 accuracy of 0.6. The same pattern applies to the fifth case (orange object), making the prediction of the classifier correct when 'k' is 3. Consequently, the top-3 accuracy increases to 0.8. Finally, when 'k' is 4 or 5, the top-k accuracy becomes 1. The key observation is that as the value of 'k' increases, the top-k accuracy also increases and eventually reaches 1 when k is equal to the number of classes.

A good model achieves an accuracy of 1.0 at lower values of 'k'. The best model is the one that achieves a top-k accuracy of 1 when k is equal to 1 (as shown in Fig. 5B). When comparing models based on their top-k accuracy, models that reach 1.0 sooner, like the green model, demonstrate better performance. Conversely, a model that achieves a top-k accuracy of 1.0 later exhibits weaker performance.

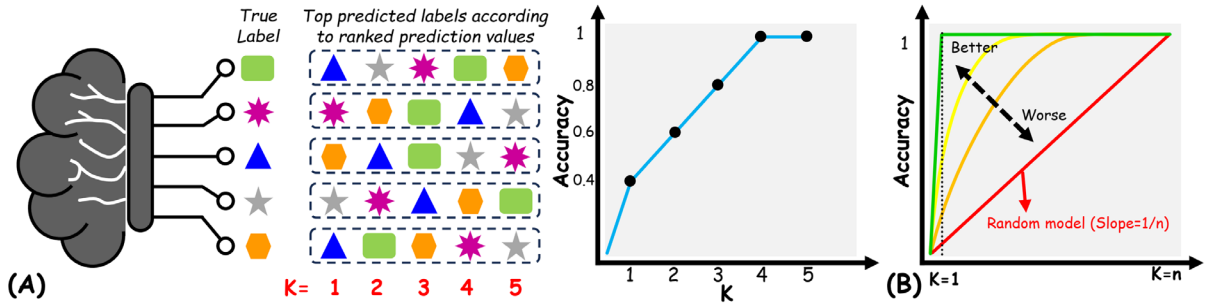


Fig. 5. Top-K Accuracy provides information about how a model could correctly diagnose cases according to the first to k-th highest probabilities (A). The best model is one that achieves an accuracy of 1 (100%) at a lower value of k (B).

2.4. Precision-Recall Curve

The precision-recall curve (PRC) is a graphical representation that depicts the relation between the precisions and recalls (sensitivity) of a binary classifier achieved at different decision thresholds (Fig. 6A). Compared to the ROC curve, the PRC is better for evaluating the performance of classifiers trained on imbalanced datasets. Moreover, it provides better insights into the important metrics, precision, and recall (sensitivity), which are used to describe a model's performance. On the other hand, interpreting a model based on PRC requires more effort than ROC. The average precision (AP) and the area under the PRC (AUPRC) are two metrics obtained from the PRC that are useful to assess a model's performance (Davis and Goadrich, 2006; Japkowicz and Shah, 2011; Saito and Rehmsmeier, 2015).

Interpretation

The AP presents the performance of a model in information retrieval and is a good metric for comparing binary classifiers. Although AP is a single numerical value, it is independent of the decision threshold as it is calculated based on averaging of precisions of a classifier at all different decision thresholds (Robertson, 2008; Su et al., 2015). The AUPRC is another metric to measure the overall performance of models in diagnosing positive cases. It calculates the area under the curve using the integral of the plotted curve concerning the model's precisions and recalls. The higher values of AUPRC and AP are desired, which indicates better performance in distinguishing between two classes in a binary classification problem. The AP and AUPRC are good metrics to compare models trained with imbalanced datasets as well (Davis and Goadrich, 2006; Japkowicz and Shah, 2011; Saito and Rehmsmeier, 2015).

Usually, the PRC starts from the top left corner of the plot, which shows a threshold with precision and recall of 1 and 0, respectively and finishes at a point with recall and precision of 1 and 0.5 (for balanced datasets), respectively. The latter point represents a high threshold, which has led to zero false negatives in the predictions, resulting in a recall of 1. On the other hand, the model will predict all negative cases incorrectly (a lot of false positives), which will result in a precision of 0.5 for a balanced dataset. The other points are obtained in the same way by calculating the precision and recall at different decision thresholds and connecting them (Fig 6A).

The baseline in the PRC represents the performance of a binary classifier that simply classifies all the points into the target (usually positive) class. This model has an AUPRC and AP of 0.5 (Fig. 6B). In general, the PRC of a classifier with higher AUPRC and AP, moves to the top right point of the plot, indicating better performance. The PRC of a model with ideal performance is shown in green, which has an AUPRC and AP of 1 (Fig. 6B).

Based on the AP and AUPRC values of our results, model 1 outperformed model 2 in classifying the breast lesions. Having higher AUPRC and AP values in model 1 means that it has higher precision at most of the thresholds compared to model 2, which translates to more TPs and fewer FPs, enabling it to detect malignant breast lesions better (Fig. 6C).

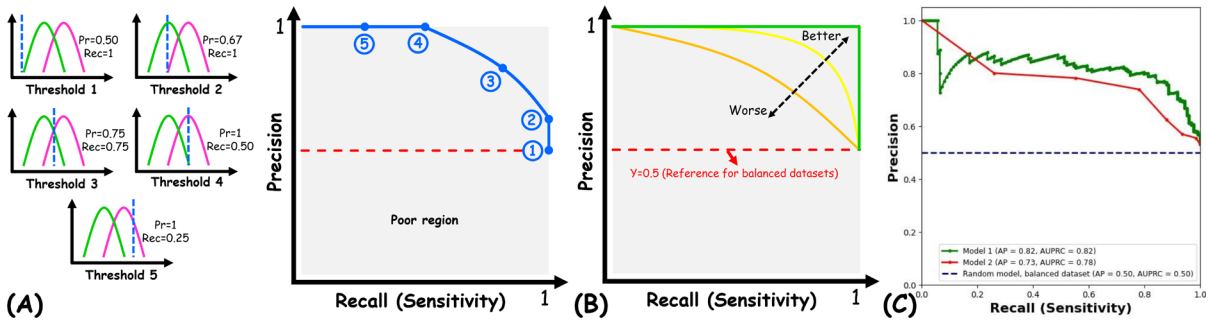


Fig. 6. The Precision-Recall Curve is depicted according to the different decision thresholds (A). The curve of a model with higher diagnostic performance moves to the top right point of the plot (B). The Precision-Recall curves of our models with the corresponding AUPRC and AP values are shown in (C).

2.5. Cumulative Gains Curve

The cumulative gains curve (CGC) is a graphical representation that illustrates the number of observations that a binary classifier can realistically obtain with the highest probability of belonging to the target (usually positive) class. The CGC plot shows the gain (sensitivity) on the y-axis and the proportion of the samples on the x-axis (Fig. 7A). The main metric obtained from the CGC is cumulative gains (CG) which is equal to the area under the CGC. The CGC is originally plotted for models used to solve binary classification problems. However, it does not provide any information about the classification threshold of the model (Bekkar et al., 2013; Japkowicz and Shah, 2011; Tufféry, 2011).

Interpretation

To plot the CGC of a model the first step is to calculate the gains for each decile. To do this, two different lists of data are used: the ground-truth labels of our cases, and the predicted

probabilities from the model. The predicted probability of the model is sorted in a decreasing order with their ground-truth labels where the highest probabilities appear at the top of this list. Afterward, the whole dataset is split into ten sub-samples, each containing 10% of the dataset. The first sub-sample (first decile) includes the points with the highest probability of belonging to the target class. The second sub-sample (second decile) contains the next 10% of cases with the highest probability. To calculate the gain of each decile, the total number of true predicted cases up to that decile with the target label is counted and divided by the total number of cases with the target label in the dataset. This procedure is continued until the gain of all deciles is calculated. Then, gains are plotted against the deciles, completing the CGC. Although we explained how to plot CGC for ten sub-samples, it can be plotted for any size of sub-sample.

Each point on the CGC has a gain on the y-axis and a percentage of the sample up to that point on the x-axis. In other words, each point on the CGC indicates what proportion of the total positive points lay in the sample of the size up to that point. For example, in Fig. 7A, the point showed that if we feed 30% of the cases with the highest probability of belonging to the target class into the model, it could correctly diagnose 90% of the total number of cases of the target class. In other words, by testing 30% of cases with the highest probability of belonging to the target class, we can expect to find 90% of the target cases. In Fig. 7B, the red line shows the CGC of a random model, which makes random predictions with a CG of 0.5. Moreover, the CGC of a model with perfect predictions and 100% sensitivity will look like the green curve. The CGC of this model should be a straight line without fluctuations starting from the gain of zero and moving to 1 because all the cases belonging to the target class are ordered at the top of the list and the plot moves up with a constant intercept and remains at 1 until the end of the plot. If the number of cases of the target class is high in the dataset, the plot will have a lower intercept; on the other hand, if they are few, it will have a high intercept.

The higher CG (area under the curve) reflects the better performance of the model. Based on CG values, model 1 outperformed model 2, indicating better identification of positive cases with fewer false negatives. In the sample size of 15%, it can be seen that model 2 has a higher CG value compared to model 1. This shows that model 2 has made more accurate predictions in the 15% of the data that had the highest probability of belonging to the malignant class. However, in an overall look, we can see that model 1 has a higher CG value compared to model 2 and outperformed it by having more accurate probabilities for each case belonging to the malignant class (Fig. 7C). Therefore, the CGC makes it possible to do comparisons between the performance of models.

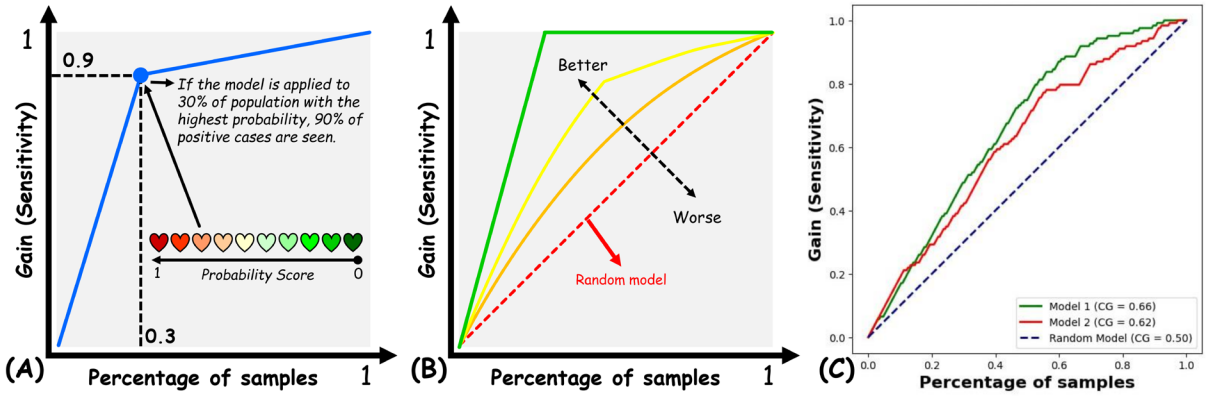


Fig. 7. The Cumulative Gains Curve provides information about how well a model could diagnose target cases based on the proportion of the selected samples (A). The curve of a model with higher diagnostic performance moves to the top left point of the plot (B). The Cumulative Gain curves of our models with the corresponding Cumulative Gain (CG) values are shown in (C).

2.6. Lift Curve

The Lift Curve (LC) is another graphical representation that can be used to evaluate the performance of a binary classifier. The LC is obtained by plotting the lift of a model against the proportion of the samples (Fig. 8A). The area under the lift curve (AUL) and the maximum lift point (MLP) are two metrics that can be used to compare the performances of different models based on LC. Like CGC, LC is also originally used to evaluate binary classifiers. Although both can be used to evaluate the performance of multi-class classifiers, the plot would be less informative. Moreover, the interpretation of LC for different models can be very challenging due to the fluctuations that appear in them (Bekkar et al., 2013; Japkowicz and Shah, 2011; Tufféry, 2011).

Interpretation

To plot an LC, first, the predicted probabilities by the model should be sorted in decreasing order. Then, the ground-truth label of each case should be stored in the same order. A subset of cases with high probabilities is determined. Afterward, the lift of the samples should be calculated for that subset as follows: the number of sorted target labels is divided by the number of all cases in that subset. Then, this value should be divided by the ratio of the target label in the whole dataset. In other words, to determine LC, the ratio of target labels in a subset of the dataset should be divided by the ratio of the target class in the whole dataset. The outcome of

this is the lift value for that subset. To calculate the lift value of different subset sizes, the ratio of the target label up to that point should be considered. According to this, usually, the LC starts from a point with the highest lift, and as it moves forward (the subset size increases), the lift value decreases until it reaches 1 (Fig. 8A).

Each point on the LC represents a sample (subset) size and the corresponding lift value. The lift value shows that the chosen subset has a lift times the number of target patients compared to the average number of target patients in the whole dataset. For example, in Fig. 8A the point with the lift value of 3 and the sample size of 0.2 shows that the subset with the 20% of the cases with the highest probability includes 3 times more target cases compared to the mean target cases ratio in the whole dataset. For example, if 10% of the cases in the whole dataset belong to the target class, it can be concluded that 60% (lift × subset size) of them lie in the sample which contains the top 20% of cases with the highest probabilities belonging to the target class.

A good model sorts the target probabilities in a decreasing order such that all the labels of the target class appear first followed by the labels of the other class. So, as shown in Fig. 8B this model indicates an LC which is constant at first and then starts to fall until reaching to 1, when the sample size is equal to the whole dataset. Models with better performances (higher AUL and MLP values) move to the top right point of the LC plot (Fig. 8B). Our results indicate that although the MLP value of the two models is equal, based on the AUL, model 1 outperformed model 2, indicating that the model 1 performed better in identifying positive (malignant) cases. In addition, we can say that both models performed better than a random model (y=1, Fig 8C).

Similar to the CGC, it can be seen that model 2 has a higher AUL in the sample size of 15% and provides better prediction probabilities for the target cases. But when looking at the whole plot, model 1 indicates better performance in predicting the probabilities of the malignant cases and outperformed model 2 (Fig. 8C).

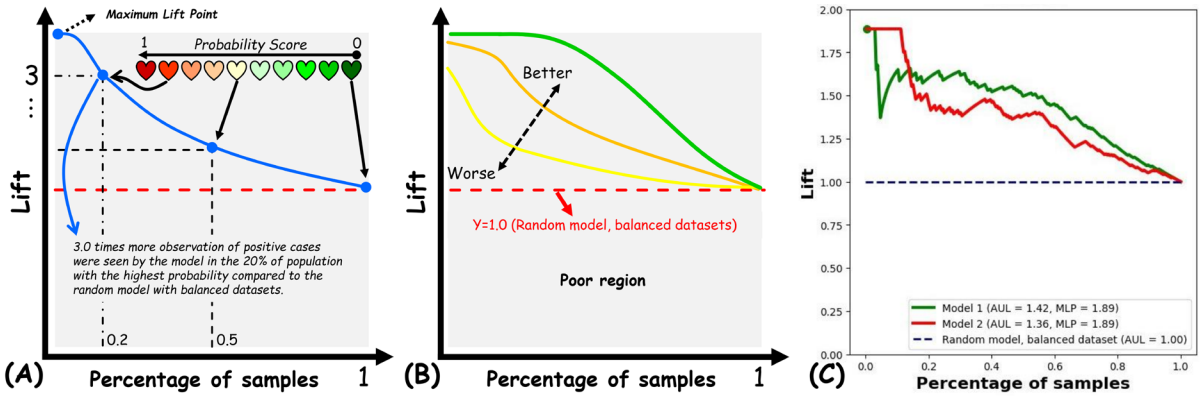


Fig. 8. The Lift Curve provides information about how a model could diagnose target cases based on the proportion of the selected samples compared with a random model (A). The curve of a model with higher diagnostic performance (higher AUL and MLP values) moves to the top right point of the plot (B). The Lift Curves of our models with the corresponding AUL and MLP values is shown in (C).

2.7. Decision Curve

The decision curve (DC) is a curve that is plotted to provide a better insight into the performance of a classifier, especially in medical binary classification problems. The DC is achieved by plotting the calculated net benefit of a classifier at different threshold probabilities. (Fig. 9A). The net benefit provides a comparison between the benefit of TPs and FPs in a model's classification outcomes and shows whether the benefit of TPs outweighs the harm of FPs or not (See supplementary material). The DC can be used to assess and compare the performance of the clinical model regarding the trade-off between TPs and FPs. The interpretation of the DC is straightforward, and unlike many other statistical measures, it considers the real-life consequences of decisions made by the classifier and how these decisions affect real-life situations. The DC analysis is done in a way that simplifies the assumptions, and this may lead to a lack of some information on the complexity of decision-making processes, making it less directly applicable in health centers (Rousson and Zumbrunn, 2011; Sande et al., 2021; Van Calster et al., 2018; Vickers and Elkin, 2006; Vickers et al., 2019).

Interpretation

Analyzing the DC will help find the range of optimal decision thresholds that can be used to make decisions about the treatment of the cases. The curve provides information about the thresholds at which the classifier performs better compared to other treatment strategies, and helps to compare classifiers and choose the best strategy for treating the patients. In Fig. 9A,

the green curve shows the maximum net benefit a classifier can achieve, where a classifier correctly classifies all target (positive) cases without any error (FP=0). Under this condition, the net benefit is equal to the proportion of truly predicted positive cases to the whole dataset. Generally, all classifiers predict all cases as patients when the decision threshold is set to zero. When the threshold starts to increase, the curve may still stay at the maximum point if there are no FPs in the predictions. Otherwise, the number of FPs affects the value of the net benefit, and lower net benefits will be obtained, which leads to a decrease in the curve. The blue line in Fig. 9A shows the DC of the "treat all" strategy, which acts as a classifier that classifies all cases as patients (positive). The red line on the other hand shows the DC of a classifier that classifies all cases as non-patients (negative) in which none of the cases will be treated. Thresholds that have a net benefit less than these two curves (fall below them) do not perform well and cannot be used in clinical scenarios. A curve that is closer to the maximum net benefit indicates better performance (Fig. 9B). The optimal range for the threshold of a classifier is the range in which the DC of the model is higher than the decision curves of the treat all and treat non-strategies.

Comparing the DC of model 1 with model 2, it can be seen that they first follow a similar pattern in their net benefit changes. However, after the threshold passes 0.2, model 1 constantly has a higher net benefit than model 2, which means the benefits of this model compared to its harm are more than model 2. In this case, we can conclude that model 1 can make better predictions. Moreover, the thresholds that lead to the least ULI in the ROC were found 0.54 and 0.67 for models 1 and 2, respectively, which led to a net benefit of 0.29 and 0.13 (Fig. 9C).

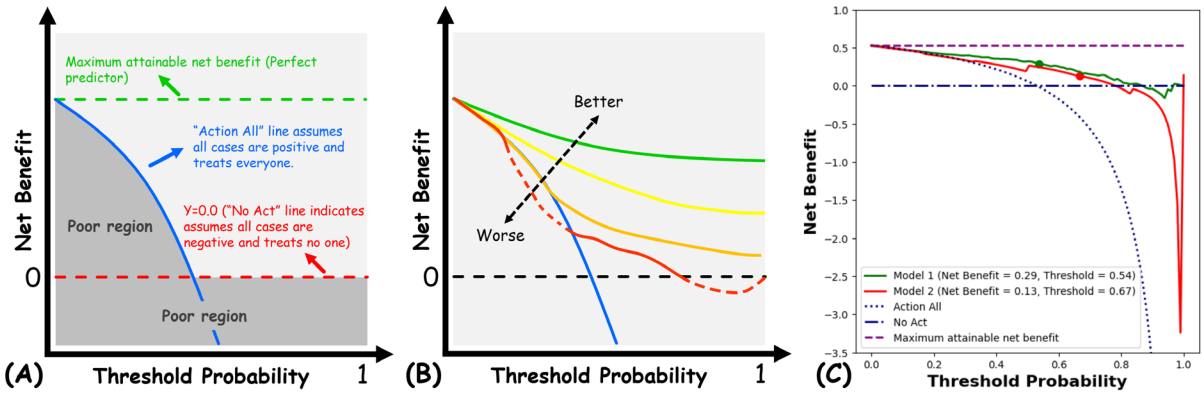


Fig. 9. The Decision Curve provides information about the net benefit of a model in diagnosing target cases based on different threshold probabilities (A). The curve of a model with a higher net benefit is

closer to the perfect model (B). The Decision curves of our models with the corresponding net benefit and threshold probability values are shown in (C).

2.8. Calibration Curve

A calibration curve (CC) is a curve that depicts the relationship between the predicted probability of a classifier and the outcomes, helping evaluate the classifiers' reliability. It is important to note that calibrating a classifier does not guarantee that its performance will be improved. The interpretation of CC may be challenging when comparing classifiers and adds complexity to making comparisons. A model with better CC is sensitive to outliers and data points considered as noise, which may lead to classifier overfitting with small datasets (Austin et al., 2020; Austin and Steyerberg, 2019; Cohen and Goldszmidt, 2004; Van Calster et al., 2019; Vuk and Curk, 2006).

Interpretation

All classification models return their output as predicted values. In binary classifiers, usually, the predicted value is between 0 and 1. The closer the predicted value of a data point is to one, the more likely that point belongs to the target class. The predicted value and predicted probability are different from each other, but under some conditions, the predicted value can be used as representative of the predicted probability. To interpret the predicted value as the predicted probability, the calibration of the model needs to be checked. This is done using the CC. In other words, CC is used to check if the predicted values of a model could be used as probabilities. To obtain the CC, the dataset is divided into desired bins (subsets, for example, five). Then, for each bin, the observed fraction of positive cases (y-axis) is calculated and plotted against the mean predicted probabilities of the classifier (x-axis). These two values fall within the interval of 0 and 1 (Fig. 10A). If the average point of a bin falls above the reference line, that means the classifier is under-predicting the true probabilities in that bin. On the other hand, if it falls below the reference line, it is over-predicting the true probabilities in that bin. These show that the classifier is not perfectly calibrated in these bins. However, if the average point falls on the reference line, it can be concluded that the classifier is perfectly calibrated in that bin (Fig. 10A).

CC provides valuable insights into the overall performance of a classifier, offers information about the bins that lead to uncalibrations in the classifier, and enables the

calibration of the classifier. A perfectly calibrated model has a CC like the green line (y = x) in Fig. 10B. If the CC of a classifier has fluctuations around this line, that classifier is not considered calibrated. The closer the CC of a classifier is to the reference line, the more calibrated it is (Fig. 10B).

Brier score is a numerical measure used to assess how well a classifier is calibrated. it is achieved by calculating the mean of the difference between the predicted value and the label of the points (cases) in the CC powered by two (See supplementary material). The lower Brier Score shows that the predicted values of the classifier are closer to the ground-truth values. Thereby, the classifier is more calibrated, and the predicted values can be used instead of probabilities. The Brier score is a good metric for comparing the calibration of classifiers (Cohen and Goldszmidt, 2004).

When comparing model 1 and model 2, the points on model 1 are closer to the reference line and have a lower Brier score compared to model 2, indicating that this classifier is more calibrated and provides more reliable results (Fig. 10C).

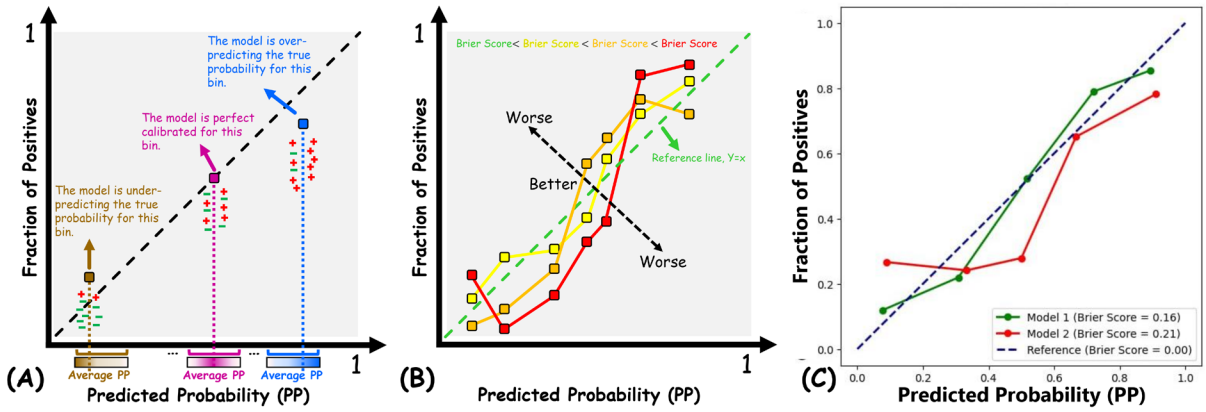


Fig. 10. The Calibration Curve provides information about how the outputs (predicted values) of a model are fit to the probability (A). The curve of better-calibrated model (lower Brier score value) moves closer to the reference line (B). The calibration curves of our models with the corresponding Brier score values are shown in (C).

3. Deep Learning Models

Deep-learning (DL) models are enhanced neural networks that have multiple layers. Each layer in the network tries to capture a more detailed representation of the input data. Many types of layers in the architecture of DL models make them able to extract and learn complex patterns from raw datasets, which can be used to undertake recognition tasks (LeCun et al., 1998). ML

techniques are algorithms that are used to automatically extract hidden patterns and useful information from training datasets. ML models can make reliable decisions and predictions using the data from previous computations. DL models are advanced neural networks that have a deeply nested network in their architecture. This characteristic makes these types of models able to extract and find patterns among raw data, which is the main difference and advantage between these models and classic ML techniques. Thereby, the DL technique makes them a better choice for solving problems with large and high-dimensional data. Although they are very powerful and useful, the interpretation of DL models is harder compared to ML models. ML models are still a better choice for solving problems with small datasets (Janiesch et al., 2021).

DL techniques are used in a variety of applications in which big datasets are involved, such as audio processing and speech recognition, natural language processing, image analysis, and object detection and recognition (Shinde and Shah, 2018). In the field of medical imaging, which is the focus of this paper, image segmentation like lesion contouring, object detection like tumor detection (localization), image enhancement like noise reduction and image super-resolution, and object classification are the most common applications of DL models (Azad et al., 2024; Currie et al., 2019). Image segmentation is a task in which an image is divided into different segments, which include pixels that share common features such as color, texture, and other visual attributes. In object detection, the models identify and localize the objects according to the similar attributes they share by gathering pixels (Montagnon et al., 2020). Finally, in object classification, the DL model extracted deep features from images and found the most informative features automatically to differentiate between the classes.

3.1. Convolutional Neural Network

Convolutional neural networks (CNN) are a type of DL model that is used for image analysis, object detection, segmentation, and recognition, which are based on the animals' visual cortex. In CNN models, each layer in the hidden layers is connected only to the nodes in the previous layer. CNNs are made of convolutional layers, pooling layers, and fully connected (or dense) layers. The architecture of CNNs leads to hierarchical feature extraction from images. In a CNN, the convolutional layer is the computational core, it applies filters to extract features from input data, generating 2-D activation maps. These maps capture relevant patterns using shared-weight neurons, which helps reduce network complexity. Pooling layers then downsample the activation maps, helping prevent overfitting while retaining important

information. Meanwhile, fully connected layers establish connections between all neurons in adjacent layers and resemble a traditional neural network to distill high-level features into probabilities for specific classes. Notably, some recent innovations, such as the 'Network In Network' (NIN) architecture, aim to enhance feature learning by replacing traditional fully connected layers (Lin et al., 2013; O'Shea and Nash, 2015).

Due to the fact that DL models such as CNN produce prediction values and classification results just like ML models, all the evaluation metrics explained for ML models can be used for DL models as well. Thereby, DL and ML models can be compared using the metrics that had been mentioned previously.

In this section, two DL models, ResNet-50 (model 1) and ResNet-101 (model 2) (He et al., 2016), were trained based on the transfer-learning method. The same breast ultrasound image used for ML models was considered for DL models to make a better comparison between the two groups (ML and DL).

In the following, similar to the ML models, the performance metrics of the two DL models were computed. The confusion matrix-based metrics are shown in Fig. 11. Other computed metrics and curves are available in the supplementary material. You can compare their performance by following the instructions explained in the ML section. Furthermore, there is an additional commonly used tool other than those mentioned previously that can be used to assess DL models: the Gradient-weighted Class Activation Mapping (Grad-CAM).

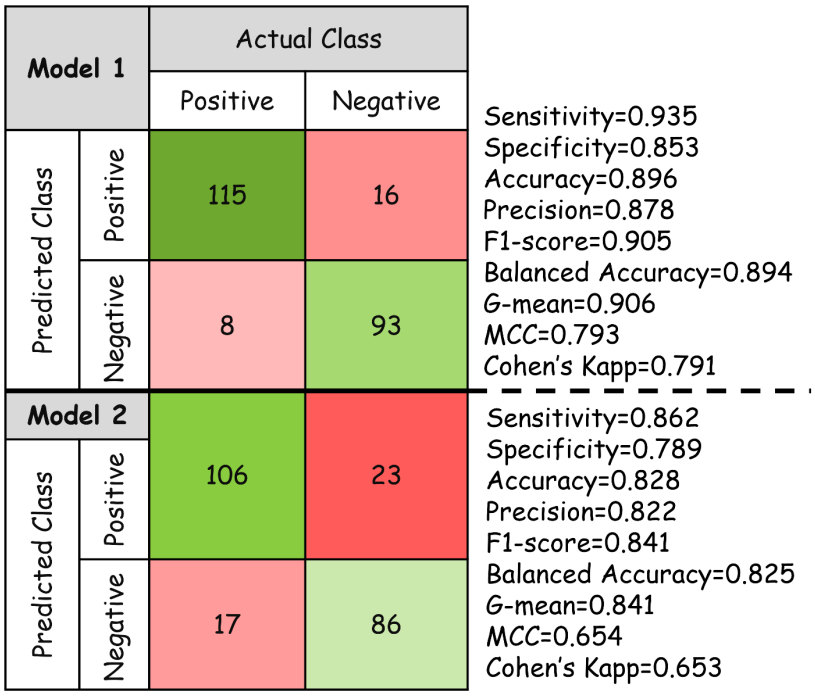


Fig. 11. Confusion matrices of the DL models 1 and 2 and their related evaluation metrics.

3.2. Grad-CAM

Grad-CAM is a technique employed for evaluating the extent of the contribution made by distinct image regions toward the predictions generated by a DL model. To assess the performance of a DL model's output, it is better to obtain the feature map from the last convolution layer, which is the layer before the dense layer (Fig. 12A). The Grad-CAM helps to determine whether the model can extract deep features from the target region in the image and use its features for classification or not. In other words, it highlights the most important and useful regions used for classification. Accordingly, the Grad-CAM is a map obtained from the original image in black and white regions (Fig. 12B). The black regions represent the areas in the image that are not used for classification. However, the lighter parts are used in making the classification of the image. The lighter and whiter a region is, the more it is used in making the final classification.

If we look at Fig. 12B, we can see that according to its Grad-CAM, it has detected the areas that contain the lesion and used the features form them in making the classification. Grad-CAMs were originally in gray-scale format. To get a better representation and insight into the Grad-CAM, a color map is used. To distinguish the regions within the Grad-CAM better in our work, we used the JET color map, which is the most commonly used color map for Grad-CAMs. Finally, to find the regions whose features have been used to make the classification, the color map is overlayed on the original image, and the regions with high importance are found more easily, which are shown in the last image on the right. In this case, it can be seen that the area in which the tumor is located was correctly detected by the model. However, this does not mean that the classification result of the model has to be correct. The model can make either correct or incorrect predictions in this situation (Selvaraju et al., 2017). We will highlight this point in the next section.

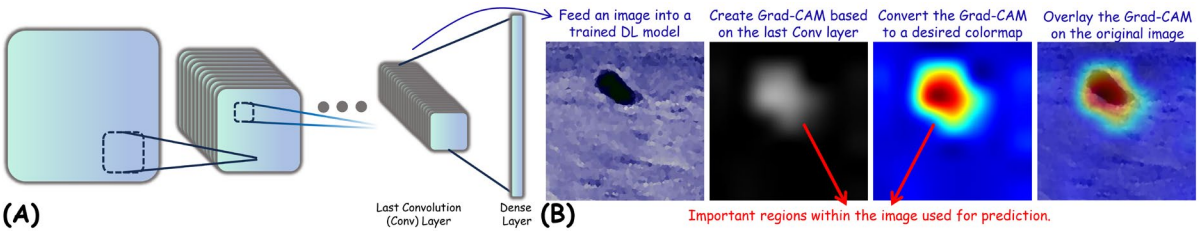


Fig. 12. Grad-CAM is obtained from the last convolution layer to probe the deep learning model performance (A). Visualization of lesion detection and classification using Grad-CAM (B).

Interpretation

Grad-CAM of the two DL models for six breast lesions is assessed. Referring to Fig. 13, the first column holds the B-mode ultrasound images of the lesions, the second column holds the ground-truth masks, and the third and fourth columns hold the Grad-CAM of the first and second DL models, respectively (model 1 has a better performance over model 2 and outperforms it in classifying the breast lesions, see supplementary material). Each lesion is represented by the label and prediction value (in parenthesis) that are assigned by the models.

Case A is a benign lesion. As can be seen, model 1 has classified the lesion as benign with a predicted value of 0.872, and model 2 has incorrectly classified it as malignant with a predicted value of 0.910. According to Grad-CAM, model 1 has detected the region of the lesion correctly and found the right areas to extract the features from. However, model 2 has only used a small region of the lesion for the classification and has also extracted the features of other regions of the image according to the colors in Grad-CAM. So, in this case, model 1 has correctly found the region and made a correct decision. Case B is a malignant lesion. Both models have misclassified this lesion. The Grad-CAM of the first model shows that it has correctly found the region which the features have to be extracted from, however, it has made a mistake with a prediction value of 0.566. However, model 2 has extracted the features used in the classification from regions that did not cover the lesion. Moreover, it misclassified the lesion with a higher prediction value of 0.879. Case C is a malignant lesion. Both models have correctly extracted important features from the region of the lesion, but despite this, both have wrongly classified the lesion as benign with a prediction score of 0.516 and 0.889, respectively. It can be seen that the predictive value of model 1 is better than model 2, while model 2 has a higher value, and a change in the decision threshold could have led to a correct classification in model 1.

Cases D and E have a similar interpretation. Case D shows the ultrasound of a malignant lesion. In the case of D, the regions used to extract the classification have been detected correctly by both models, and both models have correctly classified it as malignant. The only difference between the two models in this case is the prediction value, where model 1 has a higher prediction value compared to model 2 (0.997 and 0.867, respectively). Moreover, the Grad-CAM of model 1 is better compared to model 2, while model 2 has detected a region out of the lesion as part of the lesion's region. So, model 1 outperforms model 2 in this case. Case E has a similar situation, but in this case, the lesion is benign, and again, model 1 outperformed model 2 with a better Grad-CAM and prediction value. The last case, F, is a malignant lesion, which model 1 correctly classified and model 2 wrongly classified. In this case, both models found the regions of the lesion correctly according to their Grad-CAMs, but the second model also used some other regions as the lesion's region and made a mistake in classifying the case.

So, in general, we can conclude that in some situations, although a DL model extracts important features from the correct region (lesion's region), it can make a wrong decision. Therefore, Grad-CAM should not be considered alone for DL models' evaluation as it can lead to misunderstanding.

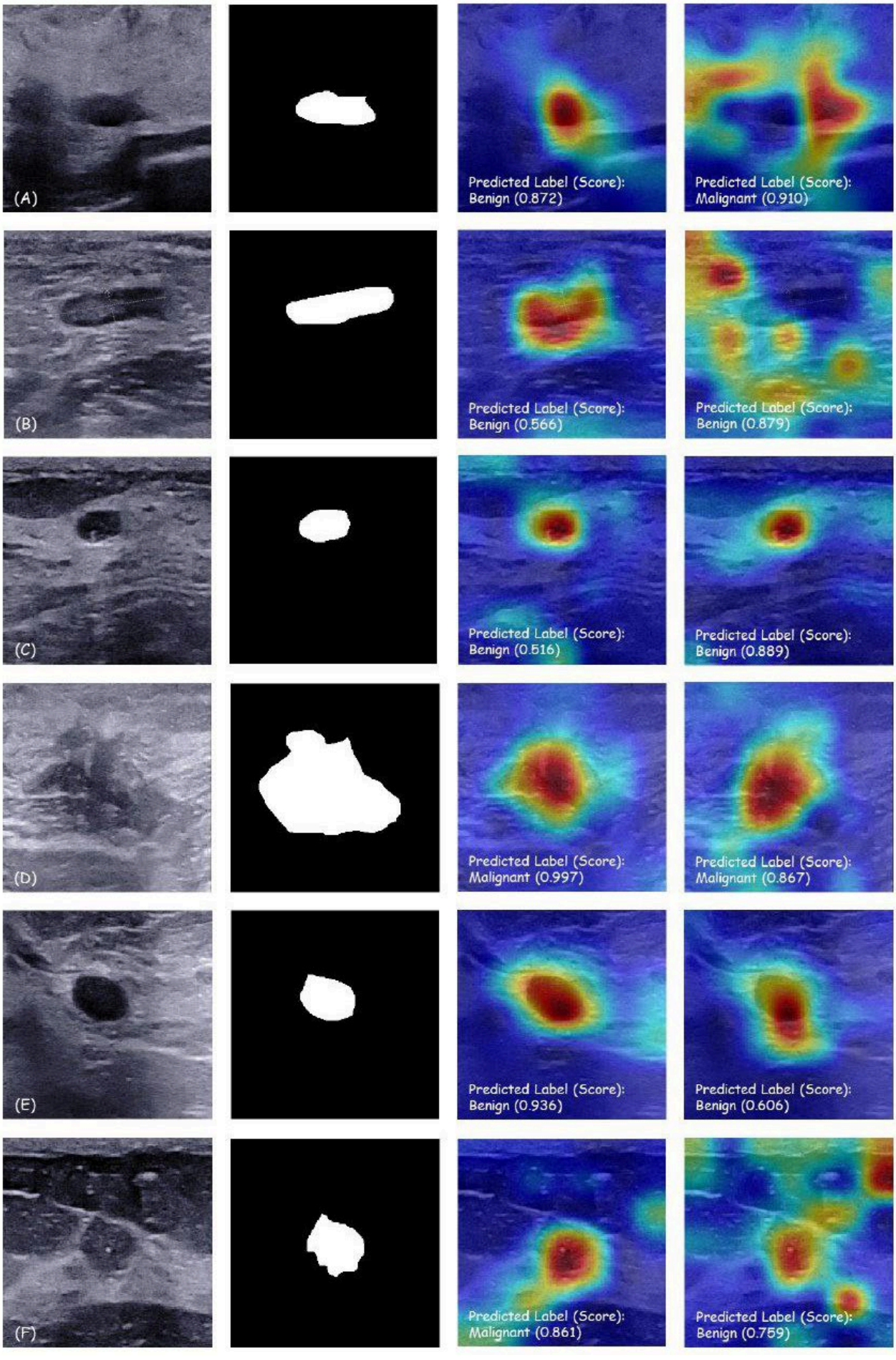

(A)
Predicted Label (Score):
Benign (0.872)
Predicted Label (Score):
Malignant (0.910)
(B)
Predicted Label (Score):
Benign (0.566)
Predicted Label (Score):
Benign (0.879)
(C)
Predicted Label (Score):
Benign (0.516)
Predicted Label (Score):
Benign (0.889)
(D)
Predicted Label (Score):
Malignant (0.997)
Predicted Label (Score):
Malignant (0.867)
(E)
Predicted Label (Score):
Benign (0.936)
Predicted Label (Score):
Benign (0.606)
(F)
Predicted Label (Score):
Malignant (0.861)
Predicted Label (Score):
Benign (0.759)

Fig. 13. The sample image of six breast lesions (A-F, first column), with the corresponding ground-truth (second column) and Grad-CAMs of DL model 1 (third column) and model 2 (forth column).

4. Discussion

In this paper, we reviewed twenty-two of the most common evaluation metrics used to assess the quality of the predictions made by medical classification models (Table 1). ML and DL techniques are mainly used to develop medical AI-aiding tools. To completely cover the metrics used to assess the performance of these models and provide real-world situations, we developed and used two ML and two DL models based on ultrasound images of patients with breast cancer. The metrics were explained and interpreted basically, and they were used to evaluate these four models and compare each pair of them to identify the best ones.

4.1 Impact of AI in Healthcare

The growth of AI applications in the healthcare industry has led to eye-catching improvements in the quality of physicians' work. AI models could help physicians in their daily routine practices in wide ranges from cancers and syndromes diagnosis to monitoring the status of patients after treatments. Carpal tunnel syndrome (CTS) is one of the most common peripheral neuropathies which is usually diagnosed using ultrasonography and clinical assessments such as electrodiagnostic tests. Besides, AI models developed for the diagnosis of CTS significantly improved physicians' performance and helped them avoid errors in their diagnosis (Faeghi et al., 2021; Mohammadi et al., 2023). The diagnosis of cancer, which is the leading cause of death, is another field in which the developed AI tools helped physicians increase the accuracy of their diagnosis. In this regard, the developed tools not only could predict whether a breast lesion is malignant or benign (Fan et al., 2023; Hamyoon et al., 2022; Wang et al., 2023), but also were capable of differentiating the molecular subtype of breast cancer (Ma et al., 2022). Moreover, AI models have gone further as they can monitor organs' function and predict whether the treatment was well done or not. For instance, AI models monitor renal function after allograft transplantation (Ardakani et al., 2017) and predict chronic kidney disease and its prognosis (Alnazer et al., 2021; Schena et al., 2022). Therefore, AI models have been proven to be effective and essential usage in healthcare. As a result, physicians need to improve their knowledge of AI to use them in their tasks effectively.

4.2 Interpretability and Evaluation of AI Models

Most of the AI models are like black boxes and the processes that happen in the core of them are not completely transparent. Therefore, having the knowledge and skills to interpret the outputs of AI models is a must for physicians who want to use them. In this paper, we equipped physicians and healthcare professionals who are going to join AI research projects or work with AI-powered machines, with the ability to understand deeply the way of evaluating the performance of AI systems. It is also important for AI models to provide uncertainty quantification in their predictions. Physicians need to understand the confidence levels and uncertainties in model diagnoses or recommendations. Providing uncertainty estimates helps physicians determine how much to trust the model's output and know when a second opinion may be needed. Quantifying prediction uncertainties is an active area of research and can help improve AI transparency (Seoni et al., 2023).

In addition, with the growing role of AI in healthcare, it is critical to develop these systems with strong ethical principles. Physicians should understand how to assess AI tools for fairness, accountability, and mitigation of unwanted biases. The models should be evaluated to ensure they do not negatively impact vulnerable patient populations or exacerbate existing healthcare disparities. Developing "ethical by design" AI aligned with medical ethics will be crucial for successful clinical adoption and maintaining public trust. The AI tools developed for the healthcare industry cannot replace physicians and are going to be used just to improve the quality and accuracy of their work. Many of these tools outperform or have at least very similar results to physicians in medical tasks such as disease diagnosis and symptom checking (Gottlieb et al., 2023; Gräf et al., 2022). However, clinicians will still need to combine AI outputs with their medical expertise and experience when making decisions about patient care.

4.3 Evaluation Metrics for Diagnostic Models

vProper evaluation is critical prior to the clinical deployment of AI models. Evaluation metrics reveal model weaknesses that can be addressed through optimization before implementation. The evaluation of the models is done by metrics which help us to gain information on the different aspects of their performance. To have a brief view, all metrics used in this review with their definitions, interpretations, and main limitations are listed in Table 1. The mentioned metrics can evaluate the performance of a diagnostic model from different aspects. It should be noted that each of them evaluates a different aspect of the model’s performance so they cannot be used alone to gain a complete insight into how well a model performs in making classifications and each of them has an advantage over the others. As a result, these metrics

need to be used together to make the physician able to assess the performance of a diagnostic model comprehensively. Accuracy is a metric used to measure the overall correctness of the predictions of a model while sensitivity and precision are used to monitor false negative and false positive predictions, respectively. Moreover, specificity aids sensitivity by providing information about the correctly identified negative cases. The F1-Score provides a balance between precision and sensitivity. When it comes to imbalance datasets, MCC, G-mean, F1-Score, balanced accuracy, AP, and AUPRC are some of the metrics, that can be used to provide a good assessment. The ROC curve is a curve that can help to obtain the TPR and FPR at different decision thresholds and identify the optimal decision threshold with the most balanced TPR and FPR. On the other hand, when it comes to imbalance datasets, PRC provides a better visual assessment of the models' performances and can be used to get a more accurate evaluation of the model's predictions. The CGC and LC help to assess the predicted values for the target cases. The CC shows how near the predicted values are to the real probabilities and how well-calibrated the model is. The DC and the net benefit show us how well the right predictions of our model outweigh the wrongs, while the MCC and Cohen's Kappa help to assess the correlation and agreement between the predicted and true labels, respectively.

The evaluation metrics can guide model optimization by highlighting areas for improvement. Additionally, they offer insights into the dataset used for training. For example, low sensitivity may indicate that more positive training examples are needed. High FPs could suggest overfitting requiring techniques like regularization. Iteratively evaluating models and evaluation metrics is essential for developing high-performing AI. While the evaluation metrics quantify model performance on given datasets, clinical deployment requires assessing generalizability to new data reflective of real-world diversity. Rigorous evaluation on multiple independent datasets, preferably multicenter and prospective, is advised. Continued evaluation is also needed to ensure models adapt to evolving clinical environments and populations over time.

Table 2. Overview, explanation, interpretation, and limitations of the metrics described in this review at a glance.

Group	Name	Definition	Trend and Interpretation	Limitation
Confusion Matrix	Sensitivity (Recall)	The percentage of positive cases that are correctly diagnosed.	Better when high, less false negative predictions.	It does not give information about false positives and focuses only on the target class.
	Specificity	The percentage of negative cases that are correctly diagnosed.	Better when high, less false positives.	It does not give information about false negatives and focuses only on the negative (untarget) class.
	Accuracy	The percentage of all cases that are correctly predicted.	Better when high, high means more correct predictions for both positive and negative cases.	It is not a good choice for problems with imbalanced datasets.
	Precision	Number of truly predicted target (positive) cases to the whole number of cases predicted as the target class.	Better when high, less false positive predictions.	It does not give information about false negatives and focuses on the predicted target class labels.
	F1-Score	It considers both the precision and sensitivity of a model and evaluates the balance between them. It is a good metric for models trained with imbalanced datasets.	Better when high, less false predictions.	Although it provides an overall assessment of precision and sensitivity, it focuses more on the target class.
	Balanced Accuracy	It evaluates the mean of sensitivity and specificity and is a good metric for models trained with imbalanced datasets.	Better when high, with more correct predictions.	It does not consider false positive or false negative rates and does not represent the trade-off between sensitivity and specificity.
	Geometric Mean (G-mean)	It evaluates the geometric mean of the specificity and sensitivity. It is a good metric for models trained with imbalanced datasets.	Better when high, with more correct predictions.	It will equal zero if either of the specificity or sensitivity is zero. It does not represent the trade-off between sensitivity and specificity.
	Matthews Correlation Coefficient (MCC)	It assesses and returns a measure for the difference between the predicted labels and true ones. It is a good metric for models trained with imbalanced datasets, which provides more detailed information than F1-Score, Geometric Mean, and Balanced Accuracy.	Better when high, shows a better correlation between the predicted labels and real ones.	It may be hard to interpret the MCC for multi-class classifiers even in binary problems. It will be uncalculatable when the model only predicts one class.
	Cohen's kappa (CK)	It evaluates and returns a measure for the agreement	Better when high, shows higher agreement	It is not a good choice for imbalanced datasets.

		between the predicted labels of a model and the true ones.	between the predicted classes and true labels.	
	Top-K Accuracy	It evaluates the predictions of models according to the top K classes which have a higher chance to be chosen as the data points label.	Better when high, shows the higher accuracy obtained by the model where the true label is among the top k true labels predicted by the model or not.	It cannot be used for binary classifiers.
ROC Curve	Area Under the ROC Curve (AUC)	It summarizes and evaluates the overall performance of a model at all possible thresholds.	Better when high, good performance independent from the decision threshold.	It does not take into account true negative or false negative rates. It is not a good choice when the consequences (costs) of changes in sensitivity and specificity are different.
	Upper-left Index (ULI)	It evaluates how near the ROC curve is to the top left corner, which indicates a better performance of the model.	Better when low, it has the minimum value for the optimal cut-off threshold.	It does not take into account true negative or false negative rates; it is not a good choice when the consequences (costs) of changes in sensitivity and specificity are different.
	Youden's Index (YI)	It evaluates the vertical distance of the ROC curve from the reference line (random model).	Better when the high, greater distance between the ROC curve and the chance level line.	It does not take into account true negative or false negative rates. It is not a good choice when the consequences (costs) of changes in sensitivity and specificity are different.
	Gini Index (GI)	It evaluates the overall performance of a model by comparing the ROC curve with the reference line (random model).	Better when high, and better performance than random guessing.	It does not consider true negative or false negative rates; it is not a good choice when the consequences (costs) of changes in sensitivity and specificity are different.
Precision Recall Curve	Average Precision (AP)	Evaluates the mean of the precision at different decision thresholds. It is a good metric for models trained with imbalanced datasets.	Better when high, shows fewer mean false positive predictions at different thresholds.	It does not provide information about the true and false negative rates and overall accuracy.
	Area Under the Precision-Recall Curve (AUPRC)	It determines how well the model could diagnose all positive cases correctly while not labeling negative cases incorrectly at the different decision thresholds. It is a good metric for models trained with imbalanced datasets.	Better when high, shows better trade-off between precision and recall and the model is better than random guessing.	It does not consider true negative or false negative rates and overall accuracy.
Cumulative Gain Curve	Cumulative Gain (CG)	It evaluates the probabilities assigned to the data points belonging to the target class over different sample sizes.	Better when high, shows more accurate probabilities for cases belonging to the target class.	It does not provide information about the decision threshold and overall accuracy. It only focuses on positive data, not negative ones.
Lift Curve	Area Under the Lift Curve (AUL)	It evaluates the overall lift (Number of times that more positive cases are seen by a model compared to the random model) at different sample sizes.	Better when high, shows more accurate probabilities for cases belonging to the target class.	It does not provide information about the decision threshold and overall accuracy. It only focuses on positive data, not negative ones.

	Maximum Lift Point (MLP)	It evaluates the maximum lift achieved by a model	Better when high, shows more accurate probabilities for cases belonging to the target class.	It does not provide information about the decision threshold and overall accuracy. It only focuses on positive data, not negative ones.
Decision Curve	Net Benefit	It evaluates how well the correct predictions of a model can benefit over the false ones at each decision threshold.	Better when high, shows more net benefit of a model's correct predictions over the harm of false ones.	It may lead to a lack of information on the complexity of decision-making processes, degree of calibration of the model's predictions, and overall accuracy.
Calibration Curve	Brier Score	It evaluates the difference between the predicted values and the real labels.	Better when low, shows more similarities between the model's predicted values and real labels. A more calibrated model represents a lower value.	It is sensitive to outliers and data points considered as noise.
Activation Mapping	Grad-CAM	It evaluates the extent of the contribution made by distinct image regions towards the predictions generated by a DL model.	Pixels with low (black) and high (wight) values within the map respectively highlight the most and least important regions that are considered in making the final classification.	It should not be considered alone as it can lead to misunderstanding; in some conditions, although a DL model extracts important features from the region of interest, it makes a mistake in classifying the case.

4.4 Limitations of the Review

In this section, the limitations of this review are mentioned and explained. Firstly, many evaluation techniques can be used to assess the performance of models, which were not mentioned. However, we did our best to gather the most commonly used metrics to gain an insight into how well the performance of a model is. According to the fact that most of the AI models used in the healthcare industry focus on classification tasks and the AI tools for other problems such as segmentation and image restoration are still at the research level and are not used in routine clinical tasks, we focused only on the evaluation of diagnostic models. So, the second limitation can be addressed as our review was only done for the classification problem. The third limitation appeared according to the dataset that was used in our review, which had a binary target that was going to be predicted. So, the evaluation metrics (except Top-K Accuracy) that were introduced and explained in our work are a good choice to be used for the evaluation of two-class (binary) classification problems. However, although some metrics can be used directly to evaluate AI models on multi-class classification problems, all binary metrics can be used for this aim as well. The main key is converting multi-class classification problems to two-class classification problems. In this regard, two well-known approaches can be used: one vs. one, and one vs. all (or one vs. rest). In these techniques, the data sets are divided into

subsets, which include the data of each class and the data of the rest of the classes. In the one vs one technique, the data sets belonging to each class are used to turn the multi-class problems into binary ones to solve problems such as classifying normal wrists, apparent fractures and occult fractures (Singh et al., 2023). For instance, we can consider groups two by two: normal wrists vs. apparent fractures, apparent fractures vs. occult fractures and so on. On the other hand, in the one vs approach, all the data of each class and the rest of the classes are considered as the two subsets to solve the problem and evaluate the performance of AI models such as classifying normal, novel coronavirus pneumonia (NCP), and common pneumonia (CP) groups (Zhang et al., 2020). For instance, we can consider groups as follows: NCP vs. all (Normal and CP), or CP vs. all (Normal and NCP). If these techniques are applied to solve multi-class classification problems, then the introduced binary metrics can be used to assess the performance of the developed AI models.

As can be seen in Fig. 14A, the confusion matrix of the multi-class (here three) problems can be turned into a two-class classification problem using the one vs. one technique. For example, in class one vs. class three, the number of truly predicted class one cases, A, will be considered as the true positive predictions, the number of truly predicted class three, I, will be the true negatives, and the number of wrong predictions of class one, C, and class three, G, will be the false positives and the false negatives, respectively. The confusion matrix of the one vs. two classes can be obtained in the same way. In Fig. 14B, to convert the three-class classification problem confusion matrix of the model and to the two-class problems (class one vs all classes) using the one vs. all method, the truly predicted data points in class one, A, will be considered as the true positive predictions. The sum of the non-class one predicted and non-class one actual predictions (E, F, H, and I) will be considered as the true negative predictions because no matter if they were predicted truly or not in their class in the binary problems made using the one vs. all method. The only important thing is that they were not classified as the target class (class one). The sum of the cases that were predicted not to be in class one but belonged to it, D and G, are the false negatives, and those that were predicted as class one but were not class one, B and C, are considered false positives in the new confusion matrix. The confusion matrix of the other combinations such as the class three vs. all classes, which is also shown in Fig. 14B can be achieved in the same way. So, using these approaches, multi-class classification problems can be turned into binary ones and the metrics introduced in our review can be used to evaluate these models as well.

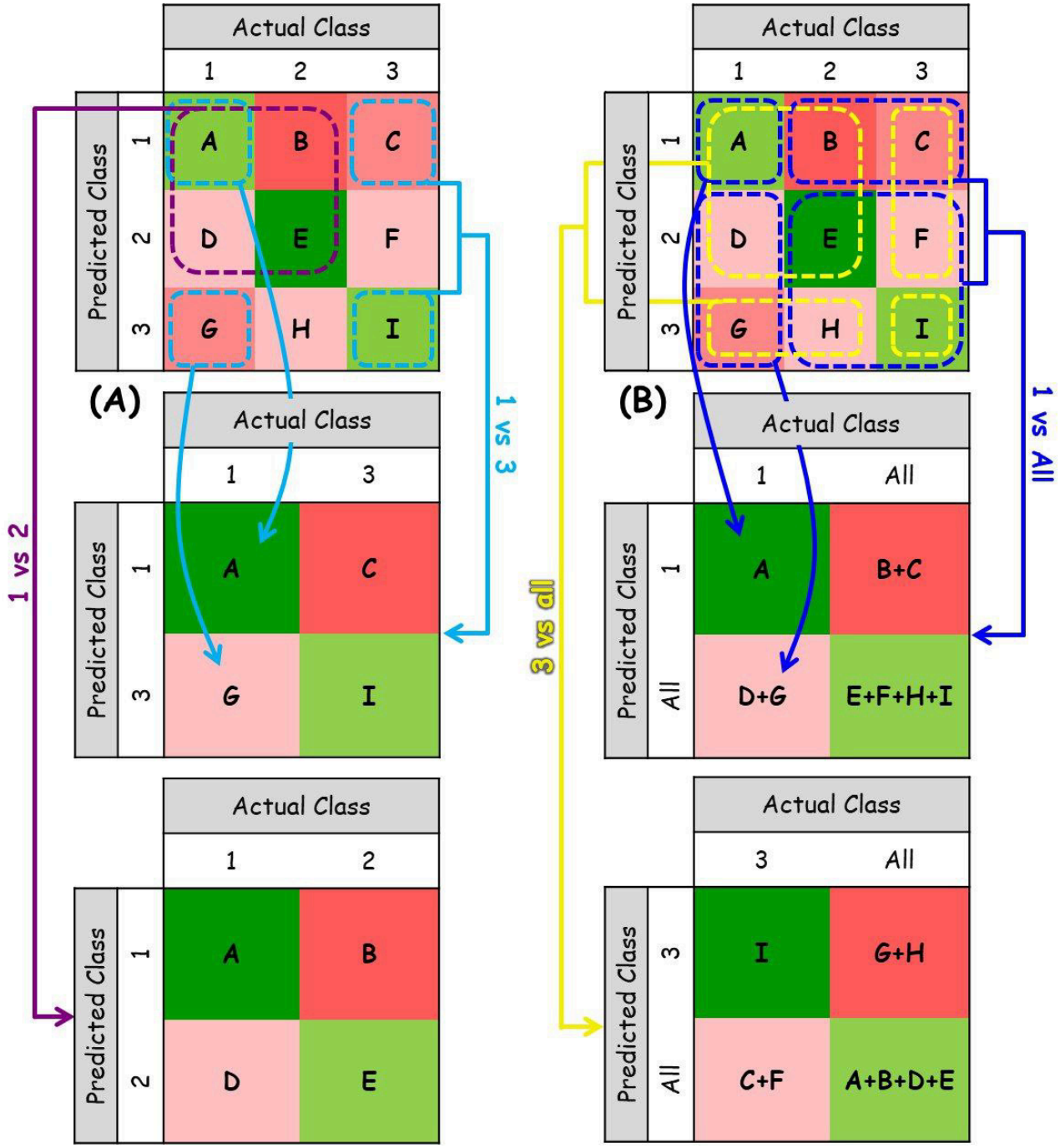


Fig. 14. Converting the confusion matrix of three-class classification problems into two-class classification problems using one vs. one (A) and one vs. all (B) methods.

5. Conclusion

In this paper, we have provided a comprehensive explanation of twenty-two commonly used metrics for the evaluation and interpretation of AI classification models. To facilitate a thorough understanding of these metrics, we developed and evaluated two ML models and two DL

models using a real-world breast cancer database. By applying these metrics, we assessed and compared the performance of the models. The findings presented in this paper contribute to enhancing physicians' understanding of AI classifiers' outputs and the evaluation process.

Authors' Contributions

Ali Abbasian Ardakani: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Supervision, Validation, Visualization, Writing - original draft, and Writing - review & editing. **Omid Airom:** Data curation, Formal analysis, Investigation, Validation, Visualization, Writing - original draft, and Writing - review & editing. **Hamid Khorshidi:** Data curation, Formal analysis, Investigation, Validation, Visualization, Writing - original draft, and Writing - review & editing. **Nathalie J Bureau:** Investigation, Methodology, Validation, Visualization, Writing - original draft, and Writing - review & editing. **Massimo Salvi:** Investigation, Validation, Visualization, Writing - original draft, and Writing - review & editing. **Filippo Molinari:** Investigation, Validation, Visualization, Writing - original draft, and Writing - review & editing. **U Rajendra Acharya:** Conceptualization, Data curation, Methodology, Supervision, Validation, Visualization, Writing - original draft, and Writing - review & editing.

Acknowledgments

Institutional Review Board (IRB) approval was obtained for this study by Research Ethics Committees of ViceChancellor in Research Affairs – Shahid Beheshti University of Medical Sciences, Tehran, Iran: # IR.SBMU.RETECH.REC.1402.372.

Formatting of funding sources

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

6. References

Abbasian Ardakani, A., Bureau, N.J., Ciaccio, E.J., Acharya, U.R., 2022. Interpretation of radiomics features–A pictorial review. Computer Methods and Programs in Biomedicine 215, 106609.

Abbasian Ardakani, A., Mohammadi, A., Mirza-Aghazadeh-Attari, M., Acharya, U.R., 2023. An open-access breast lesion ultrasound image database: Applicable in artificial intelligence studies. Computers in Biology and Medicine 152, 106438.

Allen, G., 2020. Understanding AI technology. Joint Artificial Intelligence Center (JAIC) The Pentagon United States.

Alnazer, I., Bourdon, P., Urruty, T., Falou, O., Khalil, M., Shahin, A., Fernandez-Maloigne, C., 2021. Recent advances in medical image processing for the evaluation of chronic kidney disease. Medical Image Analysis 69, 101960.

Ardakani, A.A., Mohammadi, A., Najafabad, B.K., Abolghasemi, J., 2017. Assessment of Kidney Function After Allograft Transplantation by Texture Analysis. Iranian journal of kidney diseases 11.
Austin, P.C., Harrell Jr, F.E., van Klaveren, D., 2020. Graphical calibration curves and the integrated calibration index (ICI) for survival models. Statistics in Medicine 39, 2714-2742.
Austin, P.C., Steyerberg, E.W., 2019. The Integrated Calibration Index (ICI) and related metrics for quantifying the calibration of logistic regression models. Statistics in Medicine 38, 4051-4065.
Azad, R., Kazerouni, A., Heidari, M., Aghdam, E.K., Molaei, A., Jia, Y., Jose, A., Roy, R., Merhof, D., 2024. Advances in medical image analysis with vision Transformers: A comprehensive review. Medical Image Analysis 91, 103000.
Bekkar, M., Djemaa, H.K., Alitouche, T.A., 2013. Evaluation measures for models assessment over imbalanced data sets. J Inf Eng Appl 3.
Boyd, S., Cortes, C., Mohri, M., Radovanovic, A., 2012. Accuracy at the top. Advances in neural information processing systems 25.
Bradley, A.P., 1997. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognition 30, 1145-1159.
Brodersen, K.H., Ong, C.S., Stephan, K.E., Buhmann, J.M., 2010. The balanced accuracy and its posterior distribution, 2010 20th international conference on pattern recognition. IEEE, pp. 3121-3124.
Bush, W.S., Edwards, T.L., Dudek, S.M., McKinney, B.A., Ritchie, M.D., 2008. Alternative contingency table measures improve the power and detection of multifactor dimensionality reduction. BMC Bioinformatics 9, 238.
Chicco, D., Jurman, G., 2020. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics 21, 6.
Chicco, D., Jurman, G., 2023. The Matthews correlation coefficient (MCC) should replace the ROC AUC as the standard metric for assessing binary classification. BioData Mining 16, 1-23.
Chicco, D., Tötsch, N., Jurman, G., 2021a. The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation. BioData Mining 14, 13.
Chicco, D., Warrens, M.J., Jurman, G., 2021b. The Matthews Correlation Coefficient (MCC) is More Informative Than Cohen's Kappa and Brier Score in Binary Classification Assessment. IEEE Access 9, 78368-78381.
Cohen, I., Goldszmidt, M., 2004. Properties and Benefits of Calibrated Classifiers, In: Boulicaut, J.-F., Esposito, F., Giannotti, F., Pedreschi, D. (Eds.), Knowledge Discovery in Databases: PKDD 2004. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 125-136.
Cohen, J., 1960. A coefficient of agreement for nominal scales. Educational and psychological measurement 20, 37-46.
Currie, G., Hawk, K.E., Rohren, E., Vial, A., Klein, R., 2019. Machine Learning and Deep Learning in Medical Imaging: Intelligent Imaging. Journal of Medical Imaging and Radiation Sciences 50, 477-487.
Davis, J., Goadrich, M., 2006. The relationship between Precision-Recall and ROC curves, Proceedings of the 23rd international conference on Machine learning, pp. 233-240.
Delgado, R., Tibau, X.-A., 2019. Why Cohen's Kappa should be avoided as performance measure in classification. PLOS ONE 14, e0222916.
Engler, R., Guisan, A., Rechsteiner, L., 2004. An improved approach for predicting the distribution of rare and endangered species from occurrence and pseudo-absence data. Journal of applied ecology 41, 263-274.
Faeghi, F., Ardakani, A.A., Acharya, U.R., Mirza-Aghazadeh-Attari, M., Abolghasemi, J., Ejtehadifar, S., Mohammadi, A., 2021. Accurate automated diagnosis of carpal tunnel syndrome using radiomics features with ultrasound images: A comparison with radiologists' assessment. European Journal of Radiology 136, 109518.

Fan, Z., Gong, P., Tang, S., Lee, C.U., Zhang, X., Song, P., Chen, S., Li, H., 2023. Joint localization and classification of breast masses on ultrasound images using an auxiliary attention-based framework. Medical Image Analysis 90, 102960.
Fawcett, T., 2006. An introduction to ROC analysis. Pattern recognition letters 27, 861-874.
Fleiss, J.L., 1971. Measuring nominal scale agreement among many raters. Psychological bulletin 76, 378.
Gottlieb, M., Patel, D., Viars, M., Tsintolas, J., Peksa, G.D., Bailitz, J., 2023. Comparison of artificial intelligence versus real-time physician assessment of pulmonary edema with lung ultrasound. The American Journal of Emergency Medicine.
Gräf, M., Knitza, J., Leipe, J., Krusche, M., Welcker, M., Kuhn, S., Mucke, J., Hueber, A.J., Hornig, J., Klemm, P., 2022. Comparison of physician and artificial intelligence-based symptom checker diagnostic accuracy. Rheumatology International 42, 2167-2176.
Graham, J., 2016. Artificial Intelligence, Machine Learning, And The FDA.
Grandini, M., Bagli, E., Visani, G., 2020. Metrics for multi-class classification: an overview. arXiv preprint arXiv:2008.05756.
Halimu, C., Kasem, A., Newaz, S.S., 2019. Empirical comparison of area under ROC curve (AUC) and Mathew correlation coefficient (MCC) for evaluating machine learning algorithms on imbalanced datasets for binary classification, Proceedings of the 3rd international conference on machine learning and soft computing, pp. 1-6.
Hamyoon, H., Yee Chan, W., Mohammadi, A., Yusuf Kuzan, T., Mirza-Aghazadeh-Attari, M., Leong, W.L., Murzoglu Altintoprak, K., Vijayananthan, A., Rahmat, K., Ab Mumin, N., Sam Leong, S., Ejtehadifar, S., Faeghi, F., Abolghasemi, J., Ciaccio, E.J., Rajendra Acharya, U., Abbasian Ardakani, A., 2022. Artificial intelligence, BI-RADS evaluation and morphometry: A novel combination to diagnose breast cancer using ultrasonography, results from multi-center cohorts. European Journal of Radiology 157, 110591.
Hanley, J.A., McNeil, B.J., 1982. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 143, 29-36.
Harvey, H.B., Gowda, V., 2020. How the FDA regulates AI. Academic radiology 27, 58-61.
He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770-778.
Janiesch, C., Zschech, P., Heinrich, K., 2021. Machine learning and deep learning. Electronic Markets 31, 685-695.
Japkowicz, N., Shah, M., 2011. Evaluating learning algorithms: a classification perspective. Cambridge University Press.
Jiang, F., Jiang, Y., Zhi, H., Dong, Y., Li, H., Ma, S., Wang, Y., Dong, Q., Shen, H., Wang, Y., 2017. Artificial intelligence in healthcare: past, present and future. Stroke and vascular neurology 2.
Jurman, G., Riccadonna, S., Furlanello, C., 2012. A Comparison of MCC and CEN Error Measures in Multi-Class Prediction. PLOS ONE 7, e41882.
Kato, T., Hirohashi, Y., 2019. Learning Weighted Top-$ k $ Support Vector Machine, Asian Conference on Machine Learning. PMLR, pp. 774-789.
Koyama, T., Hamada, H., Nishida, M., Naess, P.A., Gaarder, C., Sakamoto, T., 2016. Defining the optimal cut-off values for liver enzymes in diagnosing blunt liver injury. BMC Research Notes 9, 41.
Kubat, M., Holte, R.C., Matwin, S., 1998. Machine learning for the detection of oil spills in satellite radar images. Machine learning 30, 195-215.
Landis, J.R., Koch, G.G., 1977. The measurement of observer agreement for categorical data. biometrics, 159-174.
Lapin, M., Hein, M., Schiele, B., 2016. Loss functions for top-k error: Analysis and insights, Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1468-1477.
LeCun, Y., Bottou, L., Bengio, Y., Haffner, P., 1998. Gradient-based learning applied to document recognition. Proceedings of the IEEE 86, 2278-2324.
Lin, M., Chen, Q., Yan, S., 2013. Network in network. arXiv preprint arXiv:1312.4400.

Lötsch, J., Kringel, D., Ultsch, A., 2021. Explainable artificial intelligence (XAI) in biomedicine: Making AI decisions trustworthy for physicians and patients. BioMedInformatics 2, 1-17.
Ma, M., Liu, R., Wen, C., Xu, W., Xu, Z., Wang, S., Wu, J., Pan, D., Zheng, B., Qin, G., Chen, W., 2022. Predicting the molecular subtype of breast cancer and identifying interpretable imaging features using machine learning algorithms. Eur Radiol 32, 1652-1662.
Maldonado, S., López, J., 2014. Imbalanced data classification using second-order cone programming support vector machines. Pattern Recognition 47, 2070-2079.
Manne, R., Kantheti, S.C., 2021. Application of artificial intelligence in healthcare: chances and challenges. Current Journal of Applied Science and Technology 40, 78-89.
Matthews, B.W., 1975. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochimica et Biophysica Acta (BBA) - Protein Structure 405, 442-451.
Mauguen, A., Begg, C.B., 2016. Using the Lorenz Curve to Characterize Risk Predictiveness and Etiologic Heterogeneity. Epidemiology 27, 531-537.
McCarthy, J., 2007. What is artificial intelligence.
Miller, D.D., 2019. The medical AI insurgency: what physicians must know about data to practice with intelligent machines. NPJ digital medicine 2, 62.
Mohammadi, A., Torres-Cuenca, T., Mirza-Aghazadeh-Attari, M., Faeghi, F., Acharya, U.R., Abbasian Ardakani, A., 2023. Deep Radiomics Features of Median Nerves for Automated Diagnosis of Carpal Tunnel Syndrome With Ultrasound Images: A Multi-Center Study. Journal of Ultrasound in Medicine 42, 2257-2268.
Montagnon, E., Cerny, M., Cadrin-Chênevert, A., Hamilton, V., Derennes, T., Ilinca, A., Vandenbroucke-Menu, F., Turcotte, S., Kadoury, S., Tang, A., 2020. Deep learning workflow in radiology: a primer. Insights into imaging 11, 1-15.
Neri, E., de Souza, N., Brady, A., Bayarri, A.A., Becker, C.D., Coppola, F., Visser, J., European Society of, R., 2019. What the radiologist should know about artificial intelligence – an ESR white paper. Insights into Imaging 10, 44.
Nichols, T.R., Wisner, P.M., Cripe, G., Gulabchand, L., 2010. Putting the kappa statistic to use. The Quality Assurance Journal 13, 57-61.
O'Shea, K., Nash, R., 2015. An introduction to convolutional neural networks. arXiv preprint arXiv:1511.08458.
Paranjape, K., Schinkel, M., Panday, R.N., Car, J., Nanayakkara, P., 2019. Introducing artificial intelligence training in medical education. JMIR medical education 5, e16048.
Powers, D.M., 2020. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. arXiv preprint arXiv:2010.16061.
Robertson, S., 2008. A new interpretation of average precision, Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, pp. 689-690.
Rousson, V., Zumbrunn, T., 2011. Decision curve analysis revisited: overall net benefit, relationships to ROC curve analysis, and application to case-control studies. BMC Medical Informatics and Decision Making 11, 45.
Saito, T., Rehmsmeier, M., 2015. The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. PLOS ONE 10, e0118432.
Sande, S.Z., Seng, L., Li, J., D'AGOSTINO, R., 2021. Statistical Learning in Medical Research with Decision Threshold and Accuracy Evaluation. Journal of Data Science 19.
Schena, F.P., Anelli, V.W., Abbrescia, D.I., Di Noia, T., 2022. Prediction of chronic kidney disease and its progression by artificial intelligence algorithms. Journal of Nephrology 35, 1953-1971.
Secinaro, S., Calandra, D., Secinaro, A., Muthurangu, V., Biancone, P., 2021. The role of artificial intelligence in healthcare: a structured literature review. BMC Medical Informatics and Decision Making 21, 125.

Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization, Proceedings of the IEEE international conference on computer vision, pp. 618-626.
Seoni, S., Jahmunah, V., Salvi, M., Barua, P.D., Molinari, F., Acharya, U.R., 2023. Application of uncertainty quantification to artificial intelligence in healthcare: A review of last decade (2013–2023). Computers in Biology and Medicine 165, 107441.
Shinde, P.P., Shah, S., 2018. A Review of Machine Learning and Deep Learning Applications, 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), pp. 1-6.
Singh, A., Ardakani, A.A., Loh, H.W., Anamika, P., Acharya, U.R., Kamath, S., Bhat, A.K., 2023. Automated detection of scaphoid fractures using deep neural networks in radiographs. Engineering Applications of Artificial Intelligence 122, 106165.
Su, W., Yuan, Y., Zhu, M., 2015. A relationship between the average precision and the area under the ROC curve, Proceedings of the 2015 international conference on the theory of information retrieval, pp. 349-352.
Tufféry, S., 2011. Data mining and statistics for decision making. John Wiley & Sons.
Van Calster, B., McLernon, D.J., van Smeden, M., Wynants, L., Steyerberg, E.W., Bossuyt, P., Collins, G.S., Macaskill, P., McLernon, D.J., Moons, K.G.M., Steyerberg, E.W., Van Calster, B., van Smeden, M., Vickers, Andrew J., On behalf of Topic Group 'Evaluating diagnostic, t., prediction models' of the, S.i., 2019. Calibration: the Achilles heel of predictive analytics. BMC Medicine 17, 230.
Van Calster, B., Wynants, L., Verbeek, J.F.M., Verbakel, J.Y., Christodoulou, E., Vickers, A.J., Roobol, M.J., Steyerberg, E.W., 2018. Reporting and Interpreting Decision Curve Analysis: A Guide for Investigators. European Urology 74, 796-804.
Vickers, A.J., Elkin, E.B., 2006. Decision Curve Analysis: A Novel Method for Evaluating Prediction Models. Medical Decision Making 26, 565-574.
Vickers, A.J., van Calster, B., Steyerberg, E.W., 2019. A simple, step-by-step guide to interpreting decision curve analysis. Diagnostic and Prognostic Research 3, 18.
Vuk, M., Curk, T., 2006. ROC curve, lift chart and calibration plot. Advances in methodology and Statistics 3, 89–108-189–108.
Wang, J., Zheng, Y., Ma, J., Li, X., Wang, C., Gee, J., Wang, H., Huang, W., 2023. Information bottleneck-based interpretable multitask network for breast cancer classification and segmentation. Medical Image Analysis 83, 102687.
Wardhani, N.W.S., Rochayani, M.Y., Iriany, A., Sulistyono, A.D., Lestantyo, P., 2019. Cross-validation Metrics for Evaluating Classification Performance on Imbalanced Data, 2019 International Conference on Computer, Control, Informatics and its Applications (IC3INA), pp. 14-18.
World Health, O., 2021. Ethics and governance of artificial intelligence for health: WHO guidance. World Health Organization, Geneva.
Youden, W.J., 1950. Index for rating diagnostic tests. Cancer 3, 32-35.
Yu, K.-H., Beam, A.L., Kohane, I.S., 2018. Artificial intelligence in healthcare. Nature biomedical engineering 2, 719-731.
Zhang, K., Liu, X., Shen, J., Li, Z., Sang, Y., Wu, X., Zha, Y., Liang, W., Wang, C., Wang, K., Ye, L., Gao, M., Zhou, Z., Li, L., Wang, J., Yang, Z., Cai, H., Xu, J., Yang, L., Cai, W., Xu, W., Wu, S., Zhang, W., Jiang, S., Zheng, L., Zhang, X., Wang, L., Lu, L., Li, J., Yin, H., Wang, W., Li, O., Zhang, C., Liang, L., Wu, T., Deng, R., Wei, K., Zhou, Y., Chen, T., Lau, J.Y.-N., Fok, M., He, J., Lin, T., Li, W., Wang, G., 2020. Clinically Applicable AI System for Accurate Diagnosis, Quantitative Measurements, and Prognosis of COVID-19 Pneumonia Using Computed Tomography. Cell 181, 1423-1433.e1411.