Online-Journals.org

iJOE International Journal of Online and Biomedical Engineering

iJOE | eISSN: 2626-8493 | Vol. 20 No. 12 (2024) | OPEN ACCESS

https://doi.org/10.3991/ijoe.v20i12.49739

PAPER

The Study of the Effectiveness and Efficiency of Multiple DCNN Models for Breast Cancer Diagnosis Using a Small Mammography Dataset

Nourane Laaffat[1](✉), Ahmad Outfarouin[2], Walid Bouarifi[1], Abdelilah Jraifi[1]

[1]Mathematics Computer Science and Communication Systems Research Laboratory, Cadi Ayyad University, Safi, Morocco

[2]Management and Decision Support Laboratory, Ibn Zohr University, Dakhla, Morocco

nourane.laaffat@ced.uca.ma

ABSTRACT

Breast cancer (BC), the most prevalent cancer worldwide, poses a significant threat to women's health, often resulting in mortality. Early intervention is crucial for reducing mortality rates and improving recovery. Mammography plays a pivotal role in early detection through high-resolution imaging. Various classification techniques, including classical and deep learning (DL) methods, assist in diagnosing BC. Convolutional neural networks (CNN)-based classification with transfer learning enhances efficiency and accuracy, especially with limited datasets. This study evaluates the performance of different pretrained deep CNN architectures in classifying pathological mammography scans from the Mini-MIAS dataset. The results show that Xception, VGG16, VGG19, and MobileNetV2 achieve the highest accuracy (97%), with VGG19 demonstrating the fastest prediction speed (0.53 s).

KEYWORDS

breast cancer (BC), Xception, VGG16, VGG19, ResNet50, MobileNetV2, InceptionResNetV2, InceptionV3, DenseNet121, deep convolutional neural networks (DCNNs), classification, transfer learning, effectiveness, efficiency

1 INTRODUCTION

Breast cancer (BC) [1], [2], [3] ranks as the most prevalent cancer globally and stands as the leading factor in female mortality. In the year 2022, a total of 2.3 million cases of BC were identified among women, leading to 670,000 fatalities on a global scale, according to the WHO [3].

Achieving a reduction in the mortality rate associated with this particular form of cancer, along with enhancing the prospects of recovery, can solely be accomplished through vigilant management of the tumor, starting from its initial stages of development. Mammography [3] is recognized as the primary method for early

Laaffat, N., Outfarouin, A., Bouarifi, W., Jraifi, A. (2024). The Study of the Effectiveness and Efficiency of Multiple DCNN Models for Breast Cancer Diagnosis Using a Small Mammography Dataset. *International Journal of Online and Biomedical Engineering (iJOE)*, 20(12), pp. 72–89. https://doi.org/10.3991/ijoe.v20i12.49739

Article submitted 2024-04-19. Revision uploaded 2024-06-09. Final acceptance 2024-06-13.

identification of breast irregularities, providing high-resolution images of breast tissue. In response to the rise in mammogram utilization over the past few decades, numerous study studies are endeavoring to automatically identify breast abnormalities using computer-aided detection (CAD) [4], [5], [6], [7], [8], [9], [10], [11], or to automatically interpret mammograms through computer-aided diagnosis (CADx) [12], [13], [14], [15], [16], [17]. These study works are oriented, depending on the problem to be treated, in two main areas:

- Computer-aided detection [18], which aims to identify anomalies and classify regions of a mammogram as regions of interest (suspicious).
- Computer-aided diagnosis [18] entails the classification of detected anomalies, determining whether they are benign or malignant, or whether they fall within the categories of abnormality or normalcy.

Numerous algorithms exist for classifying and predicting BC outcomes [19], [20], [21], [22], and [23]. Deep learning (DL) models have greatly improved classification tasks, driven primarily by the adoption of transfer learning (TL) techniques and the growing availability of diverse image datasets. These methods have effectively tackled numerous challenges, notably addressing data scarcity, particularly prevalent in fields such as medical imaging where obtaining extensive datasets is arduous. Consequently, deep convolutional neural networks (DCNNs) often struggle to effectively learn from small datasets, leading to overfitting issues. TL offers a solution by leveraging pre-trained DCNNs in various ways. It involves either extracting essential features from the pre-trained model and transferring them to a classification model or making specific adjustments to the model to achieve optimal results. This last way is more sophisticated and will be our adopted way.

The present paper gives a performance comparison between eight classifiers: Xception, VGG16, VGG19, ResNet50, MobileNetV2, InceptionResNetV2, InceptionV3, and DenseNet121. These 8 different types of DCNNs were pretrained on a natural image dataset (ImageNet), whose weight is used in TL experiments. To ascertain the category of mammography scans of the female breast, the Softmax function is implemented at the final layer (FC) to generate the predicted probabilities.

The key novelty of this work lies in its comprehensive evaluation of multiple state-of-the-art pretrained CNN architectures on a small yet diverse mammography dataset. This study not only benchmarks the performance of these models in terms of accuracy, sensitivity, specificity, precision, and prediction speed but also highlights the practical implications of using TL in scenarios with limited labeled data. By demonstrating the effectiveness of sophisticated TL techniques and providing a detailed performance comparison, this study offers valuable insights that can guide future applications and study in the field of BC detection.

The main contribution of our approach to detailing the dynamics of the system to the literature lies in several key aspects:

Comprehensive Evaluation of Pretrained CNN Architectures: We conduct a thorough evaluation of eight state-of-the-art pretrained CNN architectures (Xception, VGG16, VGG19, ResNet50, MobileNetV2, InceptionResNetV2, InceptionV3, and DenseNet121) on a small yet diverse mammography dataset. This comprehensive benchmarking provides insights into the relative strengths and weaknesses of each model in the context of BC detection, a contribution that is valuable for studied and practitioners aiming to select the most suitable model for their specific needs.

Application of TL in Medical Imaging: Our study highlights the practical implications of using TL to address data scarcity issues prevalent in medical imaging. By leveraging pretrained models on a large natural image dataset (ImageNet) and fine-tuning them on mammography images, we demonstrate how TL can significantly enhance model performance even with limited labeled data. This underscores the potential of TL for similar medical imaging applications, thereby advancing the methodology in the literature.

Performance Metrics and Practical Implications: We evaluate the models not only in terms of traditional metrics such as accuracy, sensitivity, specificity, and precision but also in terms of prediction speed, Kappa statistic, and other error metrics. This multifaceted evaluation offers a holistic view of the models' performance, addressing both their effectiveness and efficiency. Such a detailed analysis helps in understanding the trade-offs involved and aids in making informed decisions regarding model deployment in real-world clinical settings.

Detailed Performance Comparison and Insights: Our study provides a detailed performance comparison of the eight CNN models, offering valuable insights into which models perform best under specific conditions. This comparative analysis contributes to the literature by identifying the most effective models for mammography classification, thereby guiding future study and development efforts.

Guidance for Future Applications and Research: By demonstrating the effectiveness of sophisticated TL techniques and providing a detailed performance comparison, our study offers actionable insights that can inform future applications and study in the field of BC detection. This contribution helps bridge the gap between theoretical study and practical application, facilitating the translation of advanced ML techniques into clinical practice.

The objective of this study is to evaluate how efficiently and effectively these algorithms perform through various metrics. The sections of this document are organized as follows: Section 2 explores some related works. Section 3 presents the methodology. Section 4 elucidates the experimental context. Section 5 examines the comparative analysis of the experiments. Section 6 deliberates on the acquired experimental results. Lastly, Section 7 presents the paper's concluding remarks.

2 RELATED WORK

Classification is an essential function of computer vision systems applied to medical imaging; it is involved in the final phase of diagnosis called decision-making. The categorization process holds a significant and indispensable position within the realms of machine learning (ML) and DL. Numerous study efforts are directed towards the application of ML and DL in the medical imaging domain for the purpose of disease classification and prevention, particularly in cases such as BC. This involves a meticulous evaluation of patient conditions through the analysis and categorization of images according to their respective pathological stages. The DCNN, categorized as a form of DL architecture, has exhibited notable efficacy in the examination and categorization of medical imagery. This phenomenon is particularly evident within healthcare infrastructures for the identification of various types of tumors, such as those found in the brain, lungs, breasts, and skin. The strategy of TL is implemented to enhance the structure of CNNs, thereby culminating in the development of a resilient model.

Alantari and Kim [24] present a system specifically developed for identifying and categorizing breast irregularities. Their objective is achieved using the YOLO

detector, resulting in F1 scores of 0.992 and 0.9802, respectively, on the Digital Database for Screening Mammography (DDSM) and INbreast datasets. Subsequently, categorization is conducted using three DL architectures: standard feedforward CNN, ResNet50, and InceptionResNetV2. The initial model achieves a 94.5% accuracy rate on DDSM and 88.7% on INbreast. ResNet-50 achieves an accuracy of 95.8% on DDSM and 92.5% on INbreast, while InceptionResNetV2 attains 97.5% and 95.3%, respectively, on DDSM and INbreast. The authors in the study [25] conducted a comparative analysis between VGG16 and ResNet50 to ascertain the best BC detector using the IRMA dataset. VGG16 demonstrates itself as the most precise classifier, achieving an accuracy rate of 94%, whereas ResNet50 attains a recognition accuracy of 91.7%.

In [26], the authors introduce a classification model for BC masses by leveraging CNNs. An evaluation is conducted to analyze the efficacy of two different structures, specifically AlexNet and GoogleNet, which demonstrate discrepancies in both architecture and hyper-parameters. This evaluation aims to identify the optimal classifier through the utilization of image data sourced from the Curated Breast Imaging Subset of DDSM (CBIS-DDSM), INbreast, the Mammographic Image Analysis Society (MIAS), and the Egyptian NCI. For CBIS-DDSM, AlexNet achieves 100% accuracy, while GoogleNet achieves 98.46%. In the case of the INbreast database, the models achieve 100% and 92.5%, respectively. Moreover, for NCI images, the models achieve 97.89% and 91.58%, respectively. Similarly, for the MIAS database, the models achieve accuracies of 98.53% and 88.24%, respectively. Consequently, the AlexNet model demonstrates superior accuracy and performance as a classifier. In the study by the authors cited as references [27], a novel framework is introduced for the segmentation and categorization of images depicting. Various DL architectures, such as InceptionV3, ResNet50, DenseNet121, MobileNetV2, and VGG16, were employed to differentiate benign from malignant instances. These models were applied to datasets including MIAS, DDSM, and CBIS-DDSM. The proposed methodology, particularly when employing InceptionV3 on the DDSM dataset during the classification phase, demonstrated superior performance. This was evidenced by achieving an impressive accuracy rate of 98.87%, an area under the curve (AUC) of 98.88%, sensitivity of 98.98%, precision of 98.79%, an F1 score of 97.99%, and a computational time of 1.2134 seconds.

3 METHODOLOGY

Our study methodology consists of several steps aimed at contributing to the development of BC diagnostic support systems. The first step involves preprocessing. Input scans are normalized, their sizes are adjusted according to each model's input, and they undergo various transformation phases (rotation, flipping, zooming, etc.) to augment them. The dataset is subsequently divided into three distinct subsets, enabling the training, testing, and validation of each classification model with the objective of identifying the most efficient and effective model. The classification process is conducted through the fine-tuning of pretrained DCNN to suit the current classification task: determining whether the patient is diagnosed with a normal or abnormal anomaly. The methodology approach is illustrated in Figure 1.

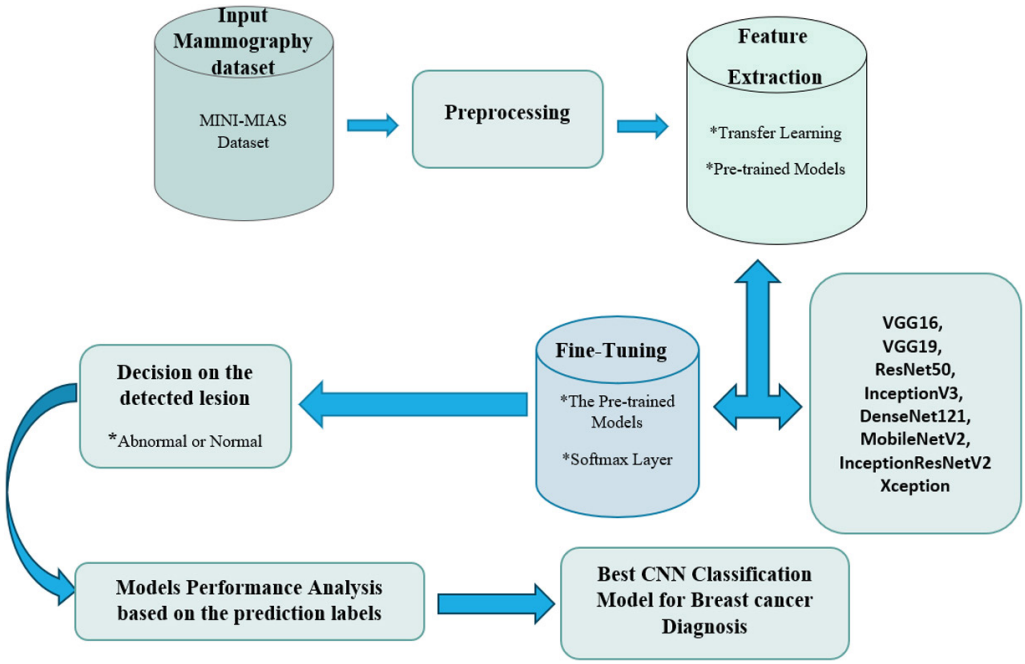


Fig. 1. Our methodology

4 EXPERIMENT

To compare the behaviors of VGG16, VGG19, ResNet50, InceptionV3, DenseNet121, MobileNetV2, InceptionResNetV2, and Xception, we carried out an experiment that was centered on evaluating both the effectiveness and efficiency of these networks. In particular, the study inquiries guiding the experiment are: Which of the models demonstrates greater efficiency? Which model provides greater accuracy? Which model is the fastest in its prediction?

4.1 Experiment environment

The experiment is conducted using Keras, which is one of the most important and famous libraries approved for building, developing, applying, and evaluating DL network models in a fast way. The idea behind the development of the Keras library, which is a high-level interface that uses Tensorflow and Tiano in the background and is characterized by its flexible and understandable application programming, is to facilitate experiments through rapid prototyping and to provide application models that are very easy to use and apply with minimal effort; the ability to go from idea to result with minimal delay is the key to good research. We carry out our experiment using Google Collaboratory, which is an environment that allows us to easily use a GPU accelerator to speed up the process of the operations carried out.

4.2 Breast imaging dataset

We used the Mini-MIAS dataset provided by the Mammographic Image Analysis Society to detect breast abnormalities earlier. It is a consortium of British research

collectives focused on the study of mammograms, having established a repository of digital mammograms. The Pilot European Image Processing Archive at Essex University provides access to mammographic scans. This dataset consists of 322 mammographic scans, among which 113 exhibit abnormalities that could potentially be benign or malignant, while 209 are classified as "Normal," indicating the absence of any abnormalities.

4.3 Preprocessing

The preprocessing initially begins with dataset partitioning, as detailed in the methodology section. Preparation of data input for all CNN architectures entails a mandatory step of normalization and then resizing images: preprocessing (224, 224, 3)-pixel-sized input images for VGG16, VGG19, ResNet50, DenseNet121, MobileNetV2, and InceptionResNetV2, and preprocessing (299, 299, 3)-pixel-sized input images for InceptionV3 and Xception. Adjustment of the original pretrained model is necessary to align with our specific requirements; the final fully connected output layer should facilitate binary classification (two classes) rather than the default 1000 classes.

4.4 Data augmentation

In general, enhanced performance in DL models is often observed with an increase in the volume of available data, thereby providing a richer source of information for extraction and facilitating heightened learning capabilities. Nevertheless, instances may arise where access to extensive datasets is limited. In such scenarios, a viable solution involves the application of image transformations, which encompass operations such as rotation, flipping, and brightness adjustments, among others, aimed at expanding the image inventory derived from the initial dataset. This is similar to our case, where we are using a small dataset. To optimize performance outcomes, the ImageDataGenerator API within Keras was employed to conduct image augmentation. This technique serves the purpose of artificially expanding the pool of training images sourced from the dataset by executing a range of diverse transformation methodologies.

4.5 Transfer learning

Owing to the constraints posed by scarce data and the essential requirement of significant computational resources, the employment of TL emerges as a viable approach for effectively training a DCNN. This method enables the enhancement of model efficiency by leveraging the existing architecture of a pretrained model, enabling the model to acquire knowledge pertaining to novel tasks through the utilization of parameters acquired during prior training on the ImageNet dataset, as opposed to commencing training with randomly initialized parameters.

4.6 Fine-tuning

It involves the process of modifying a pre-existing model by adjusting its para meters during the training phase on a novel dataset in order to enhance its versatility

and effectiveness on this specific dataset. This methodology is implemented on pretrained models to enhance their suitability for the classification mission related to BC. The final layers of the pretrained networks are initially set up to handle a classification task involving 1,000 different categories. Subsequently, we eliminate the ultimate layer and substitute it with a novel classifier. The primary function of this classifier is to categorize input images by utilizing the activations received from the feature extraction step of each convolutional neural network (CNN).

We first start by loading the pretrained model on the ImageNet dataset. The uppermost layer of this model, typically designated for ImageNet classification, is excluded. Then, each layer of the model is set as non-trainable, freezing the weights of the base model to prevent them from being updated during training. After that, a new model architecture is created by stacking the pre-trained model with a dense layer utilizing ReLU activation, alongside a dropout layer for the purpose of regularization. Ultimately, a dense output layer is incorporated into the model, featuring softmax activation, enabling binary classification (two classes). In general, the pretrained architecture serves as a base and is adapted to the specific task of binary classification of BC by adjusting the weights of the added layers during training.

5 EXPERIMENTAL RESULTS

In this particular section, the outcomes of the image examination conducted on our Mini-MIAS dataset are presented. In order to employ our classifiers and assess their performance, we partitioned our initial dataset into training, testing, and validation subsets. The test subset comprises a total of 32 scans.

We apply some DL models for the classification of BC while using TL and data augmentation techniques that evolve the performance of the models used and prevent overfitting. During the process of learning, various employed hyperparameters are adjusted. Table 1 gathers them.

Table 1. The fixed values of the hyperparameters

Hyperparameter	Value
Optimizer	Adam
Learning rate	$5 * 10^{-5}$
Batch size	32
Epochs	10

We assess the models implemented based on the criteria utilized to assess the efficacy and efficiency of the classification models. Accuracy, precision, recall, and sensitivity are widely utilized for classifying BC [28].

5.1 Effectiveness

In this particular section, we assess the performance of all utilized architectures based on various metrics such as model prediction speed, correctly classified

instances, misclassified instances, and accuracy. The findings are summarized in Table 2 and illustrated in Figure 2.

Additionally, simulation errors are considered to provide a comprehensive assessment of the classifiers' performance. This study also evaluates the effectiveness of our classifiers regarding:

- The Kappa Statistic (KS)
- The Root Mean Squared Error (RMSE)
- The Mean Absolute Error (MAE)
- The Root Relative Squared Error (RRSE)
- The Relative Absolute Error (RAE)

KS, RMSE, and MAE are expressed in numerical form, while RRSE and RAE are represented as percentages. These findings can be observed in Table 3 and Figure 3.

The choice of these criteria to assess the model's effectiveness relies on their ability to provide a comprehensive evaluation of its performance from various perspectives. The prediction speed is crucial as it measures how quickly the model can generate predictions, which is often a determining factor in real-time or large-scale applications. The correctly classified and misclassified instances provide a direct indication of the model's ability to make accurate predictions. Special attention is given to accuracy, as it represents the total proportion of instances correctly classified by the model, thus offering a holistic measure of its performance.

The Kappa statistic (KS) [18] is a measure of the model's classification reliability, taking into account both correct and incorrect predictions adjusted for what could be predicted by chance alone.

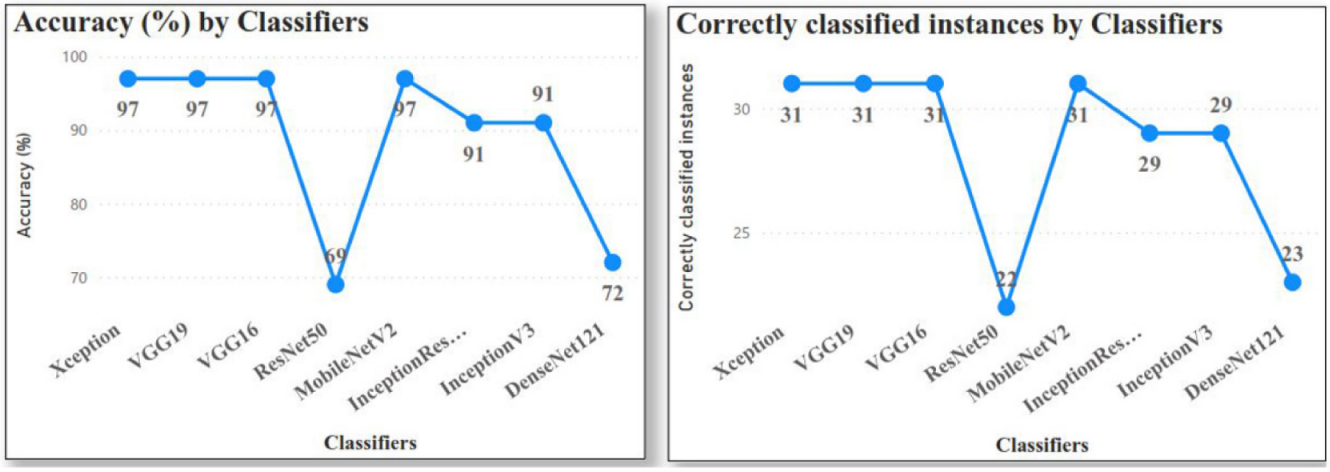
Error metrics such as root mean squared error (RMSE), mean absolute error (MAE), root relative squared error (RRSE), and relative absolute error (RAE) evaluate the precision of the model's predictions by measuring the difference between predicted and actual values. These metrics [18] provide insights into the model's ability to accurately and reliably estimate outcomes, which is crucial in many application domains, such as prediction and forecasting.

By combining these criteria, the assessment of the model's effectiveness becomes comprehensive, thereby providing an in-depth understanding of its performance across different aspects, from reliability to its ability to generate accurate predictions.

Table 2. Performance of the classifiers

Evaluation Criteria	Classifiers							
	Xception	VGG19	VGG16	ResNet50	MobileNetV2	InceptionResNetV2	InceptionV3	DenseNet121
Speed of prediction (s)	0.65	0.53	0.59	0.79	0.68	2.24	0.93	1.28
Correctly classified instances	31	31	31	22	31	29	29	23
Incorrectly classified instances	1	1	1	10	1	3	3	9
Accuracy (%)	97	97	97	69	97	91	91	72

a)



b)

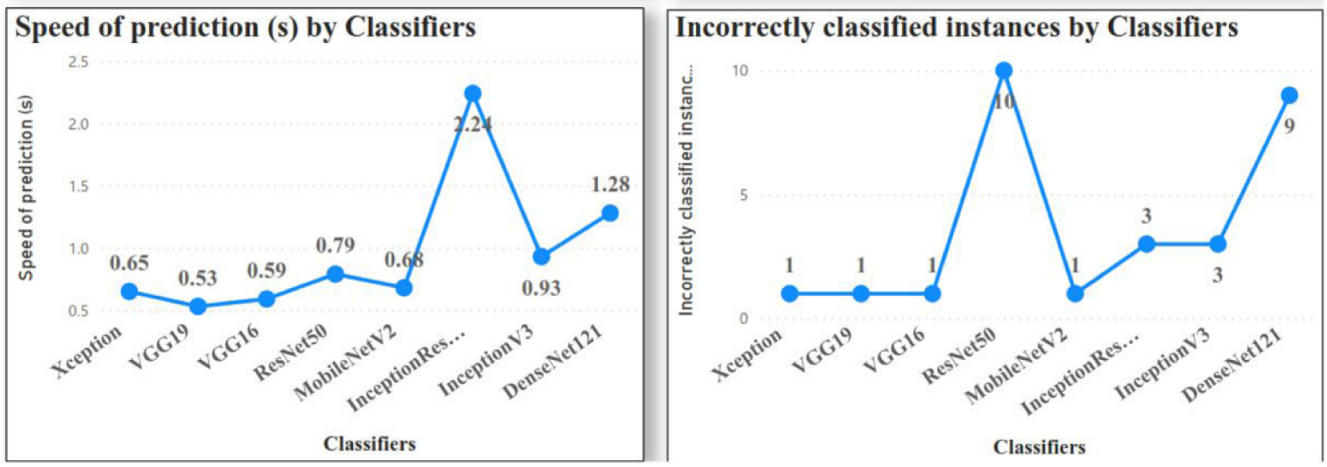


Fig. 2. Comparative graph of different classifiers (a) and (b)

5.2 Efficiency

We evaluate the efficiency of the adopted networks regarding precision, recall, TP rate (TPR, also called sensitivity), TN rate (TNR, also called specificity), FP rate (FPR), and FN rate (FNR) for all classifiers once the model used for prediction is built. Precision [28] [18] assesses the proportion of true positive predictions among all positive predictions made by the model, providing an indication of its ability to avoid false positives. Recall [28] and [18], on the other hand, measure the proportion of true positive predictions among all actual positive instances, highlighting the model's ability to avoid false negatives. TPR and TNR [18] specifically evaluate the sensitivity and specificity of the model, respectively, by quantifying its ability to correctly identify positive and negative instances. Similarly, FPR and FNR metrics [18] are crucial for understanding the model's classification errors, measuring the proportion of negative instances incorrectly classified as positive and positive instances incorrectly classified as negative, respectively. These metrics collectively provide a comprehensive assessment of the model's performance in terms of its ability to accurately classify instances into their respective classes. By considering these metrics, the evaluation of the model's efficiency becomes thorough, capturing its performance from multiple perspectives and enabling a more nuanced

understanding of its classification capabilities. Tables 4 and 5 encapsulate all the outcomes. Additionally, Table 6 showcases the confusion matrices, which serve as a valuable tool for assessing the classifier's performance. Rows correspond to rates pertaining to a specific class, and columns illustrate predictions. Figure 4 depicts the ROC curve of our classifiers to enhance comprehension of efficiency, specifically highlighting the precision of each classifier. The graphical representation provided by the ROC curve effectively showcases the performance of various classifiers. This plot enables the straightforward identification of optimal models and the elimination of less effective ones for improved classification.

Table 3. Simulation error

Evaluation Criteria	Classifiers							
	Xception	VGG19	VGG16	ResNet50	MobileNetV2	InceptionResNetV2	InceptionV3	DenseNet121
KS	0.93	0.93	0.93	0.12	0.93	0.78	0.78	0.3
MAE	0.03	0.03	0.03	0.31	0.03	0.09	0.09	0.28
RMSE	0.18	0.18	0.18	0.56	0.18	0.3	0.3	0.53
RAE (%)	6.92	6.92	6.92	69.26	6.92	20.78	20.78	62.33
RRQE (%)	37.22	37.22	37.22	117.69	37.22	64.46	64.46	111.65

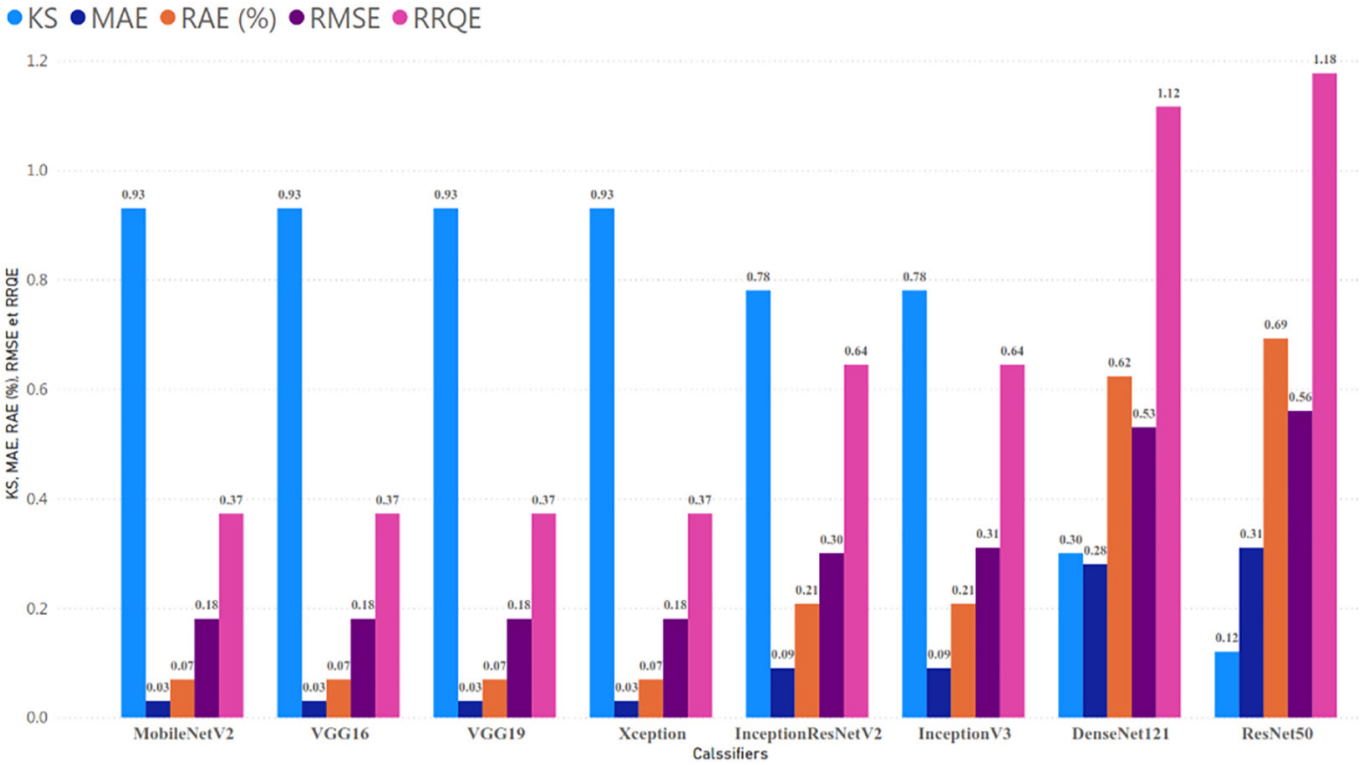


Fig. 3. Comparative diagram of the classifiers based on evaluation metrics: KS, MAE, RMSE, RAE, RRQE

Table 4. Rates of TP, TN, FP, FN for Xception, VGG19, VGG16, ResNet50, MobileNetV2, InceptionResNetV2, InceptionV3, DenseNet121

Classifiers	Precision	Recall	F1score	Class	AUC
Xception	1	0.91	0.95	Abnormal	0.95
	0.95	1	0.98	Normal	
VGG19	1	0.91	0.95	Abnormal	0.95
	0.95	1	0.98	Normal	
VGG16	1	0.91	0.95	Abnormal	0.95
	0.95	1	0.98	Normal	
ResNet50	1	0.09	0.17	Abnormal	0.54
	0.68	1	0.81	Normal	
MobileNetV2	1	0.91	0.95	Abnormal	0.95
	0.95	1	0.98	Normal	
InceptionResNetV2	0.9	0.82	0.86	Abnormal	0.88
	0.91	0.95	0.93	Normal	
InceptionV3	0.9	0.82	0.86	Abnormal	0.88
	0.91	0.95	0.93	Normal	
DenseNet121	0.67	0.36	0.47	Abnormal	0.63
	0.73	0.9	0.81	Normal	

Table 5. Comparison of accuracy measures for Xception, VGG19, VGG16, ResNet50, MobileNetV2, InceptionResNetV2, InceptionV3, DenseNet121

Performance Metrics / Classifiers	TP	TN	FP	FN
Xception	1	0.95	0.04	0
VGG19	1	0.95	0.04	0
VGG16	1	0.95	0.04	0
ResNet50	1	0.68	0.32	0
MobileNetV2	1	0.95	0.04	0
InceptionResNetV2	0.9	0.91	0.09	0.1
InceptionV3	0.9	0.91	0.09	0.1
DenseNet121	0.66	0.73	0.27	0.22

Table 6. Confusion matrices

Classifier	Actual Class		
	Abnormal	Normal	Predicted Class
Xception	10	1	Abnormal
	0	21	Normal
VGG19	10	1	Abnormal
	0	21	Normal

(Continued)

Table 6. Confusion matrices *(Continued)*

Classifier	Actual Class		
	Abnormal	Normal	Predicted Class
VGG16	10	1	Abnormal
	0	21	Normal
ResNet50	1	10	Abnormal
	0	21	Normal
MobileNetV2	10	1	Abnormal
	0	21	Normal
InceptionResNetV2	9	2	Abnormal
	1	20	Normal
InceptionV3	9	2	Abnormal
	1	20	Normal
DenseNet121	4	7	Abnormal
	2	19	Normal

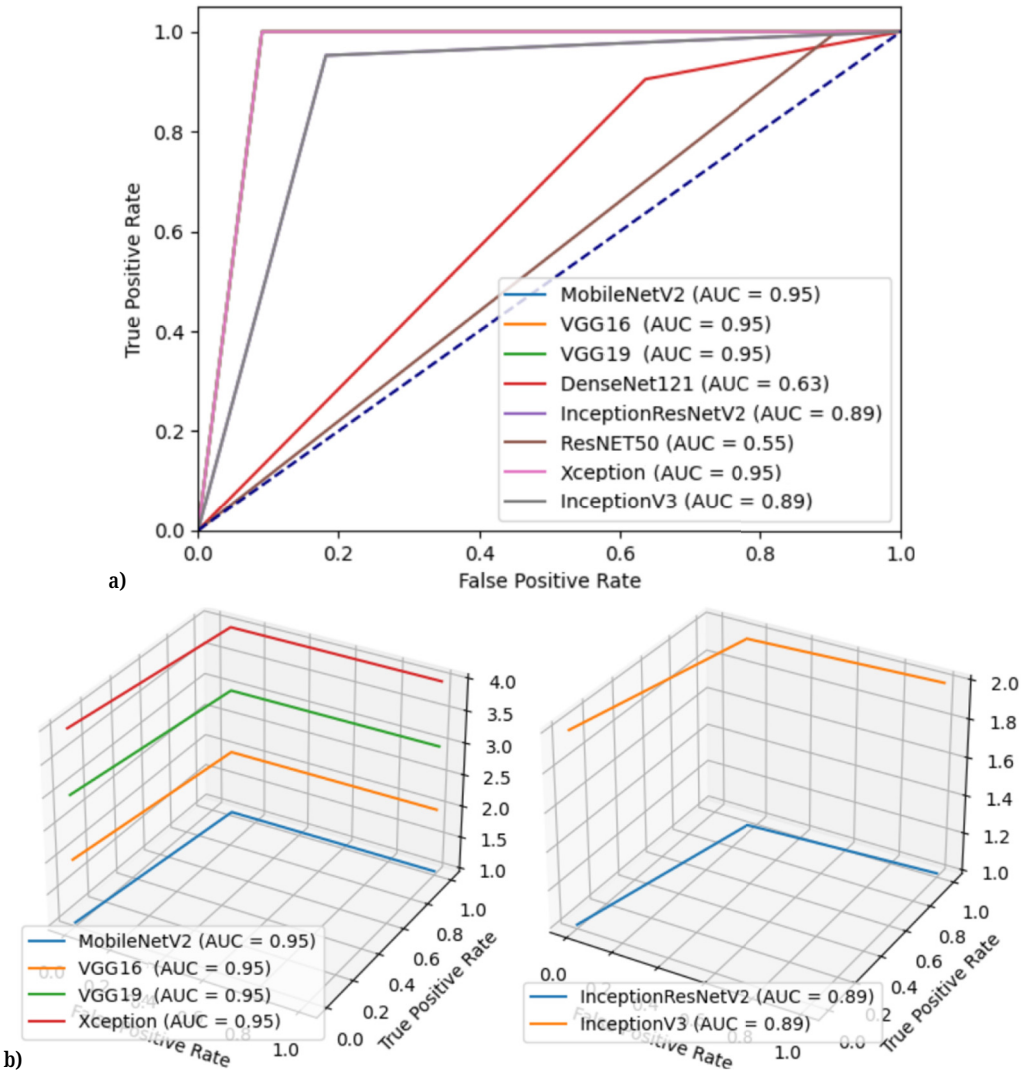


Fig. 4. Roc Curve (a) and (b)

Table 7. Comparison with other approaches using MIAS dataset for breast cancer classification

Approach	Accuracy %	Precision %	Recall %	F1-score %	Specificity %	AUC %	Speed of Prediction (s)
GoogleNet [26]	88.24	–	–	–	–	94.65	–
LSTM-RNN, CNN, random forest and boosting techniques [29]	95	–	97	98	97	94–97	–
InceptionResNetV2 + Random Forest [2]	88	88.6	88	–	87	84	0.0155
VGG19 + Random Forest [1]	91	90.7	90		91	88	0.02 s
DenseNet121 + SVM [2]	94	93.7	100		91	91	0.00107 s
NasNetLarge, Fine-Tuning [18]	94	100	82	90	100	91	–
CNN model from scratch [30]	95.3	–	–	–	92.3	97.4	–
Proposed Approach. The metrics values belong to the top Classifier (VGG19)	**97**	**96.6**	**100**	**96.5**	**95**	**95**	**0.53 s**

6 DISCUSSION

In essence, it's crucial to have models capable of swift and precise predictions. When developing such systems, we aim to fine-tune various neural network parameters —such as the number of layers, neurons, learning rates, and regularization factors— to minimize both prediction errors and processing time.

It's evident that VGG19 requires approximately 0.53 seconds for prediction on the test dataset, contrasting with InceptionResNetV2, which takes 2.24 seconds. On the other hand, the accuracy obtained by VGG16, VGG19, Xception, and MobileNetV2 (97%) is better than the accuracy obtained by InceptionV3 and InceptionResNetV2 (91% for both), which is better than that obtained by ResNet50 and DenseNet121 (69% and 72%, respectively). It can also be easily seen that VGG16, VGG19, Xception, and MobileNetV2, among all classifiers, demonstrate superior performance in correctly classifying instances and minimizing misclassifications (see Figure 2).

Table 3 highlights that VGG16, VGG19, Xception, and MobileNetV2 demonstrate the highest probability of achieving optimal categorization (93%) while maintaining the lowest margin of error (0.03). We can also notice that VGG16, VGG19, Xception, and MobileNetV2 exhibit the optimal balance between the reliability and validity of the collected data.

ResNet50 and DenseNet121 exhibit the highest error rates, as illustrated in Figure 3, contributing significantly to misclassifications across each architecture (out of 32 images, ResNet50 had 10 instances incorrectly classified, while DenseNet121 had 9 misclassified instances) (see Figure 2).

We can consistently ascertain the efficacy of a model through evaluation. In order to achieve this, it is imperative for the (TPR) and (TNR) to exhibit high values, while the (FPR) and (FNR) should be minimized as much as feasible. TPR denotes the sensitivity, representing the capacity to accurately detect individuals with the anomaly, whereas TNR signifies the specificity, denoting the ability to accurately detect individuals without the anomaly. Table 4 illustrates that VGG16, VGG19, Xception, and MobileNetV2 have reached peak levels of TPR and TNR, respectively, at 100% and 95%, while exhibiting minimal levels of FPR and FNR, respectively, at 0.04 and 0. In contrast, DenseNet121 has recorded the minimal values of TP and TN

rates, respectively, at 66% and 73%, while demonstrating the most important values of FP and FN rates at 0.27 and 0.22, respectively. From these results, we can understand why VGG16, VGG19, Xception, and MobileNetV2 have outperformed other classifiers.

The Area Under the Curve (AUC) is a metric that gauges a classifier's efficacy in discriminating between classes. A greater AUC signifies superior performance in distinguishing between positive and negative classes. From Table 5, we can see that VGG16, VGG19, Xception, and MobileNetV2 have the highest value of AUC (0,95), unlike ResNet50, which has the lowest value (0,64).

The ROC curve serves as a valuable tool for enhancing the comprehension of the efficacy of a given ML algorithm. It is evident from the analysis presented in Figure 4 that MobileNetV2, VGG16, VGG19, and Xception exhibit exemplary classification performance, characterized by a trajectory starting from the left corner's bottom, progressing vertically towards left corner's top, and then advancing to the right corner's top (with 100% sensitivity and 95% specificity). Subsequently, the performance of other algorithms such as InceptionResNetV2, InceptionV3, DenseNet121, and ResNet50 is observed. Given that MobileNetV2, VGG16, VGG19, and Xception all have an AUC of 0.95, their ROC curves overlap. Similarly, InceptionResNetV2 and InceptionV3 have an AUC of 0.89. To better visualize these results, we represented the ROC curves in 3D. In general, two models with the same AUC will have overlapping ROC (receiver operating characteristic) curves, indicating comparable performance regarding sensitivity and specificity for different classification threshold values.

Table 6 presents the confusion matrix, from which a comparison is made between the actual class and the predicted results acquired. VGG16, VGG19, Xception, and MobileNetV2 demonstrate accurate prediction results for 31 out of 32 instances. These instances consist of 10 abnormal cases correctly identified as abnormal, and 20 normal cases accurately classified as normal. However, there is one instance where the prediction was inaccurate; specifically, one normal case was erroneously predicted as abnormal, while no abnormal cases were misclassified as normal. This elucidates the rationale behind the superior performance in accuracy of VGG16, VGG19, Xception, and MobileNetV2 compared to other classifiers, attributed to their lower error rates.

To sum up, VGG16, VGG19, Xception, and MobileNetV2 demonstrated their effectiveness and efficiency through accuracy, sensitivity, specificity, and precision. Referring back to the prediction speed as indicated in Table 2, we can reevaluate the models based on the fastest rate of issuing predictions. We can thus see that VGG19 outperforms all other classifiers.

Table 7 presents a comparison with other approaches using the MIAS dataset for BC classification. The proposed approach demonstrates several advantages and improvements over existing methods, particularly with VGG19 being the top-performing classifier. It achieves an accuracy of 97%, surpassing GoogleNet (88.24%) [26], InceptionResNetV2 + Random Forest (88%) [2], and the fine-tuned NasNetLarge (94%) [18], and is comparable to LSTM-RNN, CNN, random forest, and boosting techniques (95%) [29] and the CNN model from scratch (95.3%) [30]. With a precision of 96.6%, it outperforms many methods, including VGG19 + Random Forest (90.7%) [1] and DenseNet121 + SVM [2] (93.7%), though the fine-tuned NasNetLarge [18] (100%) is higher. The proposed approach achieves a perfect recall of 100%, matching DenseNet121 + SVM [2] and surpassing others such as the fine-tuned NasNetLarge [18] (82%). Its F1-score of 96.5% indicates a balanced and effective model, higher than VGG19 + Random Forest [1] (91%) and the fine-tuned NasNetLarge [18] (90%). With a specificity of 95%, it ensures fewer false positives, outperforming many models and being on par with the CNN model from scratch

[30] (97.4%). The AUC of 95% reflects excellent classification performance, surpassing InceptionResNetV2 + Random Forest [2] (84%) and comparable to the fine-tuned NasNetLarge [18] (91%). Although the prediction speed of 0.53 seconds is slower than some methods, it remains practical for real-world applications. Overall, the proposed approach excels in accuracy, precision, recall, F1-score, specificity, and AUC, making it a superior and reliable solution for BC classification.

This work comprehensively evaluates eight pretrained CNN architectures, comparing their performance in mammography classification. Using models pretrained on the ImageNet dataset, the study shows how TL enhances performance with limited labeled data, crucial in medical imaging. The models are assessed for accuracy, sensitivity, specificity, prediction speed, and additional metrics such as Kappa, RMSE, MAE, RRSE, and RAE, providing a holistic performance view. The findings offer insights that will guide future BC detection applications and study. However, although the study uses a small yet diverse mammography dataset, the relatively limited size of this dataset might constrain the generalizability of the results. The sophistication of the TL techniques and the DCNN architectures used may require substantial computational resources and expertise, which could be a barrier for some practitioners. Additionally, while the study provides a detailed comparison of pretrained models, it does not explore custom or hybrid architectures that might offer additional benefits. These slight limitations are nonetheless offset by the improved performance achieved and the perspectives offered for future study.

7 CONCLUSION

To examine biomedical images for anomaly detection and classification, various tools and models in DL and ML are available. However, the persistent challenge remains to build a resilient model that is both robust, consistent, and precise. This work represents a comparative analysis of the classification performance of eight pretrained DCNN on the Mini-Mias dataset, which concerns female breast imaging, in terms of effectiveness and efficiency. Experience shows that the VGG16, VGG19, MobileNetV2, and Xception models outperform all other models with 97% accuracy. And based on the speed of prediction, we find that VGG19 (0.53 s) outperforms all models, followed by VGG16, Xception, and MobileNetV2.

8 REFERENCES

[1] N. Laaffat, A. Outfarouin, W. Bouarifi, and A. Jraifi, "A deep convolutional neural networks for the detection of breast cancer using mammography images," in *The International Conference on Artificial Intelligence and Smart Environment*, vol. 635, 2023, pp. 33–41. https://doi.org/10.1007/978-3-031-26254-8_5

[2] N. Laaffat, A. Outfarouin, W. Bouarifi, and A. Jraifi, "A deep learning model for breast cancer diagnosis using mammography images classification," in *International Conference on Computing, Intelligence and Data Analytics*, vol. 643, 2023, pp. 411–422. https://doi.org/10.1007/978-3-031-27099-4_32

[3] WHO "Breast Cancer." [Online]. Available: https://www.who.int/newsroom/factsheets/detail/breast-cancer

[4] J. Lacombe, A. Mange, A.-C. Bougnoux, I. Prassas, and J. Solassol, "A multiparametric serum marker panel as a complementary test to mammography for the diagnosis of node-negative early-stage breast cancer and DCIS in young women," *Cancer Epidemiol. Biomarkers Prev.*, vol. 23, no. 9, pp. 1834–1842, 2014. https://doi.org/10.1158/1055-9965.EPI-14-0267

[5] K. Kobayashi-Taguchi *et al.*, "Computer-aided detection of quantitative signatures for breast fibroepithelial tumors using label-free multi-photon imaging," *Molecules*, vol. 27, no. 10, 2022. https://doi.org/10.3390/molecules27103340

[6] M. A. Al-antari, M. A. Al-masni, and T.-S. Kim, "Deep learning computer-aided diagnosis for breast lesion in digital mammogram," in *Deep Learning in Medical Image Analysis: Challenges and Applications*, vol. 1213, 2020, pp. 59–72. https://doi.org/10.1007/978-3-030-33128-3_4

[7] F. Ayatollahi, S. B. Shokouhi, R. M. Mann, and J. Teuwen, "Automatic breast lesion detection in ultrafast DCE-MRI using deep learning," *Med. Phys.*, vol. 48, no. 10, pp. 5897–5907, 2021. https://doi.org/10.1002/mp.15156

[8] M. a. Z. Sousa, B. R. N. Matheus, and H. Schiabel, "Development of a structured breast phantom for evaluating CADe/Dx schemes applied on 2D mammography," *Biomed. Phys. Eng. Express*, vol. 4, p. 045018, 2018. https://doi.org/10.1088/2057-1976/aac2f2

[9] N. Vallez *et al.*, "Breast density classification to reduce false positives in CADe systems," *Comput. Meth. Programs Biomed.*, vol. 113, no. 2, pp. 569–584, 2014. https://doi.org/10.1016/j.cmpb.2013.10.004

[10] I. Reiser, R. M. Nishikawa, M. L. Giger, J. M. Boone, K. K. Lindfors, and K. Yang, "Automated detection of mass lesions in dedicated breast CT: A preliminary study," *Med. Phys.*, vol. 39, no. 2, pp. 866–873, 2012. https://doi.org/10.1118/1.3678991

[11] A. S. Chawla, R. S. Saunders, S. Singh, J. Y. Lo, and E. Samei, "Towards optimized acquisition scheme for multiprojection correlation imaging of breast cancer," *Acad. Radiol.*, vol. 16, no. 4, pp. 456–463, 2009. https://doi.org/10.1016/j.acra.2008.09.013

[12] V. Goreke, "A novel deep-learning-based CADx architecture for classification of thyroid nodules using ultrasound images," *Interdiscip. Sci.*, vol. 15, pp. 360–373, 2023. https://doi.org/10.1007/s12539-023-00560-4

[13] O. Attallah and M. Sharkas, "GASTRO-CADx: A three stages framework for diagnosing gastrointestinal diseases," *PeerJ Comput. Sci.*, vol. 7, p. e423, 2021. https://doi.org/10.7717/peerj-cs.423

[14] A. R. Jamieson, M. L. Giger, K. Drukker, and L. L. Pesce, "Enhancement of breast CADx with unlabeled data", *Med. Phys.*, vol. 37, no. 8, pp.3915–4522, 2010. https://doi.org/10.1118/1.3455704

[15] K. Drukker and M. L. Giger, "Computerized self-assessment of automated lesion segmentation in breast ultrasound: Implication for CADx applied to findings in the axilla – art. no. 69150G," in *Medical Imaging 2008: Computer-Aided Diagnosis, PTS 1 and 2*, in Proceedings of the Society of Photo-Optical Instrumentation Engineers (SPIE), M. L. Giger and N. Karssemeijer, Eds., Bellingham: Spie-Int Soc Optical Engineering, vol. 6915, 2008, pp. G9150–G9150. https://doi.org/10.1117/12.772029

[16] K. Drukker, C. A. Sennett, and M. L. Giger, "The effect of image quality on the appearance of lesions on breast ultrasound – Implications for CADx," in *Medical Imaging 2007: Computer-Aided Diagnosis, PTS 1 and 2*, M. L. Giger and N. Karssemeijer, Eds., Bellingham: Spie-Int Soc Optical Engineering, 2007, p. 65141E. https://doi.org/10.1117/12.707743

[17] M. Giger *et al.*, "Progress in breast CADx," in *2007 4th IEEE International Symposium on Biomedical Imaging: Macro* 2007, pp. 508–511. https://doi.org/10.1109/ISBI.2007.356900

[18] N. Laaffat, A. Outfarouin, W. Bouarifi, and A. Jraifi, "Breast cancer diagnosis with an ensemble deep neural network," *International Journal of Advances in Soft Computing & Its Applications*, vol. 15, no. 3, 2023.

[19] A. Jafari, "Machine-learning methods in detecting breast cancer and related therapeutic issues: A review," *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization,* vol. 12, no. 1, 2024. https://doi.org/10.1080/21681163.2023.2299093

[20] S. B. Mukadam and H. Y. Patil, "Machine learning and computer vision based methods for cancer classification: A systematic review," *Arch. Comput. Method Eng.*, vol. 31, pp. 3015–3050, 2024. https://doi.org/10.1007/s11831-024-10065-y

[21] L. Alkhathlan and A. K. J. Saudagar, "Machine learning techniques for breast cancer analysis: A systematic literature review," *Int. J. Comput. Sci. Netw. Secur.*, vol. 20, no. 6, pp. 83–90, 2020.

[22] N. Caballe-Cervigon, J. L. Castillo-Sequera, J. A. Gomez-Pulido, J. M. Gomez-Pulido, and M. L. Polo-Luque, "Machine learning applied to diagnosis of human diseases: A systematic review," *Applied Sciences*, vol. 10, no. 15, p. 5135, 2020. https://doi.org/10.3390/app10155135

[23] M. Hosni, I. Abnane, A. Idri, J. M. Carrillo de Gea, and J. L. Fernandez Aleman, "Reviewing ensemble classification methods in breast cancer," *Comput. Meth. Programs Biomed.*, vol. 177, pp. 89–112, 2019. https://doi.org/10.1016/j.cmpb.2019.05.019

[24] M. A. Al-antari, S.-M. Han, and T.-S. Kim, "Evaluation of deep learning detection and classification towards computer-aided diagnosis of breast lesions in digital X-ray mammograms," *Computer Methods and Programs in Biomedicine*, vol. 196, p. 105584, 2020. https://doi.org/10.1016/j.cmpb.2020.105584

[25] N. S. Ismail and C. Sovuthy, "Breast cancer detection based on deep learning technique," in *2019 International UNIMAS STEM 12th Engineering Conference (EnCon)*, 2019, pp. 89–92. https://doi.org/10.1109/EnCon.2019.8861256

[26] S. A. Hassan, M. S. Sayed, M. I. Abdalla, and M. A. Rashwan, "Breast cancer masses classification using deep convolutional neural networks and transfer learning." *Multimed Tools Appl*, vol. 79, pp. 30735–30768, 2020. https://doi.org/10.1007/s11042-020-09518-w

[27] W. M. Salama and M. H. Aly, "Deep learning in mammography images segmentation and classification: Automated CNN approach," *Alexandria Engineering Journal*, vol. 60, no. 5, pp. 4701–4709, 2021. https://doi.org/10.1016/j.aej.2021.03.048

[28] G. Murtaza *et al.*, "Deep learning-based breast cancer classification through medical imaging modalities: State of the art and research challenges," *Artif. Intell. Rev.*, vol. 53, pp. 1655–1720, 2020. https://doi.org/10.1007/s10462-019-09716-5

[29] S. J. Malebary and A. Hashmi, "Automated breast mass classification system using deep learning and ensemble learning in digital mammogram," *IEEE Access*, vol. 9, pp. 55312–55328, 2021. https://doi.org/10.1109/ACCESS.2021.3071297

[30] E. M. F. El Houby and N. I. R. Yassin, "Malignant and nonmalignant classification of breast lesions in mammograms using convolutional neural networks," *Biomedical Signal Processing and Control*, vol. 70, p. 102954, 2021. https://doi.org/10.1016/j.bspc.2021.102954

9 AUTHORS

Nourane Laaffat is a Moroccan Ph.D. student at Cadi Ayyad University, affiliated with the Mathematics, Computer Science, and Communication Systems Research Laboratory. She has received a bachelor's degree in mathematics and computer science from Cadi Ayyad University in 2018 and a master's degree in business intelligence from Sultan Moulay Slimane University in 2020. She is working in the medical field and decision making, focusing on medical image classification (E-mail: nourane.laaffat@ced.uca.ma).

Ahmad Outfarouin is a Moroccan professor in the Department of Business and Management at the National School of Business and Management, Dakhla, Ibn Zohr University, Morocco. Previously, he taught at the Private University of Marrakesh, the Faculty of Sciences and Techniques of Marrakesh, the National Superior School of Marrakesh, and the Faculty of Medicine and Pharmacy of Marrakesh (2014–2018). He is a member of the AI and Data Science for Economic and Environmental Development team in the Management and Decision Support Laboratory, working

on data science and decision support systems applied to various fields such as finance, healthcare, agriculture, and logistics (E-mail: a.outfarouin@uiz.ac.ma).

Walid Bouarifi is a Moroccan academic researcher at Cadi Ayyad University, specializing in artificial intelligence (AI) with a focus on computer vision and machine learning. His primary expertise includes facial recognition systems, person detection in surveillance footage, and identifying violent behavior in videos. He actively explores AI algorithms, particularly deep learning techniques, in medical imaging analysis. His work also includes developing innovative approaches for detecting individuals in enclosed spaces advancing security and surveillance technologies (E-mail: w.bouarifi@uca.ma).

Abdelilah Jraifi is a Moroccan academic researcher affiliated with Cadi Ayyad University, specializing in stochastic processes, finance, biomathematics, and deep learning. His research in finance focuses on modeling stochastic processes to analyze market behaviors, risk management, and derivatives pricing, advancing quantitative finance, and aiding financial decision-making. In biomathematics, he uses mathematical modeling to study biological phenomena, enhancing interdisciplinary collaboration in areas like population dynamics and epidemiology. Additionally, he applies deep learning methodologies to finance, utilizing neural networks to analyze data, forecast trends, and develop innovative investment and portfolio management solutions (E-mail: jraifi.abdelilah@gmail.com).