Neural Computing and Applications (2024) 36:6467–6496
https://doi.org/10.1007/s00521-023-09364-5

ORIGINAL ARTICLE

YOLO-based CAD framework with ViT transformer for breast mass detection and classification in CESM and FFDM images

Nada M. Hassan[1] · Safwat Hamad[2,3] · Khaled Mahar[4]

Received: 7 December 2022 / Accepted: 7 December 2023 / Published online: 16 January 2024

Abstract

Breast cancer detection is considered a challenging task for the average experienced radiologist due to the variation of the lesions' size and shape, especially with the existence of high fibro-glandular tissues. The revolution of deep learning and computer vision contributes recently in introducing systems that can provide an automated diagnosis for breast cancer that can act as a second opinion for doctors/radiologists. The most of previously proposed deep learning-based Computer-Aided Diagnosis (CAD) systems mainly utilized Convolutional Neural Networks (CNN) that focuses on local features. Recently, vision transformers (ViT) have shown great potential in image classification tasks due to its ability in learning the local and global spatial features. This paper proposes a fully automated CAD framework based on YOLOv4 network and ViT transformers for mass detection and classification of Contrast Enhanced Spectral Mammography (CESM) images. CESM is an evolution type of Full Field Digital Mammography (FFDM) images that provides enhanced visualization for breast tissues. Different experiments were conducted to evaluate the proposed framework on two different datasets that are INbreast and CDD-CESM that provides both FFDM and CESM images. The model achieved at mass detection a mean Average Precision (mAP) score of 98.69%, 81.52%, and 71.65% and mass classification accuracy of 95.65%, 97.61%, and 80% for INbreast, CE-CESM, and DM-CESM, respectively. The proposed framework showed competitive results regarding the state-of-the-art models in INbreast. It outperformed the previous work in the literature in terms of the F1-score by almost 5% for mass detection in CESM. Moreover, the experiments showed that the CESM could provide more morphological features that can be more informative, especially with the highly dense breast tissues.

Keywords CESM · Vision transformer · Mass detection · Mass classification

Abbreviations

ACR-BIRADS	American College of Radiology Breast Imaging Reporting and Data System
ACS	American Cancer Society
AUC	Area Under Curve
CAD	Computer-Aided (Diagnoses/Detection)
CBIS-DDSM	Curated Breast Imaging Subset of DDSM
CC	Carnio Caudal
CDD-CESM	Categorized Digital Database for Low energy and Subtracted Contrast Enhanced Spectral Mammography
CE	Contrast Enhanced
CESM	Contrast Enhanced Spectral Mammography
CL	Clip limit
CLAHE	Clip limit adaptive histogram equalization
CNN	Convolutional Neural Networks
CSP-Darknet	Cross Stage Partial Darknet
CSV	Comma-Separated Values
DEiT	Data Efficient Transformer
DM	Digital Mammography
ELM	Extreme Learning Machine
Faster-RCNN	Faster Region-Convolutional Neural Network
FFDM	Full Field Digital Mammography
FN	False Negative

✉ Nada M. Hassan
nadamahmoud@aast.edu

[1] College of Computing and Information Technology, Arab Academy for Science and Technology, Cairo, Egypt

[2] Faculty of Computer and Information Sciences, Ain Shams University, Cairo, Egypt

[3] Saint Mary's College of California, Moraga, CA 94575, USA

[4] College of Computing and Information Technology, Arab Academy for Science and Technology, Alexandria, Egypt

FNR	False Positive Rate
FP	False Positive
FPR	False Positive Rate
FRCN	Full Resolution Convolutional Network
GLCM	Gray Level Co-occurrence Matrix
IoU	Intersection over Union
LWT	Lifting Wavelet Transform
mAP	Mean Average Precision
MIAS	Ammographic Image Analysis Society
MLO	Medio Lateral Oblique
MLP	Multilayer Perceptron
NLP	Nature Language Processing
ROC	Receiver Operating Characteristics
ROI	Region of Interest
SFM	Screen Film Mammography
SVM	Support Vector Machine
TN	True Negative
TP	True Positive
TPR	True Positive Rate
ViT	Vision transformer
YOLO	You Look Only Once

1 Introduction

Cancer diseases are considered one of the most common diseases worldwide; breast cancer is one of the cancers that hits women hugely. Breast cancer develops in the breast tissues and occurs when the cells grow abnormally, forming a mass or lump, as these cells divide themselves rapidly compared to the other healthy ones.

The American Cancer Society (ACS) published an estimation for the breast cancer cases among females in the USA only (2022), and it stated that about 290,560 new cases would be diagnosed with breast cancer and 43,780 would die. Most cases mainly occur between 45 and 62 years old [1]. The substantial support for awareness about the risks of breast cancer helped a lot in the early diagnosis of this disease. This happens through encouraging regular check-ups, which can be done via different approaches, such as regular screening. There are various modalities for breast screening, but mammogram is the most widely used and effective one. Although screening cannot prevent breast cancer, it can help a lot in detecting the lesions early, even before the symptoms start to develop.

Mammograms are X-ray pictures of the breast with two different views, Medio Lateral Oblique (MLO) and Carnio Caudal (CC), for each breast side. Various abnormalities can be detected through mammogram interpretation; these abnormalities can be classified into masses and calcifications. Those abnormalities are diagnosed as benign or malignant according to their appearance and morphological features such as shape and pattern. Benign masses have an oval or circular shape with well-defined edges, while malignant masses look like it has spikes out from their center.

The accurate interpretation of the mammogram leads to precise diagnosis; Radiologists spend a lot of time and effort in the interpretation process. Interpretation of many cases can cause inaccurate diagnosing for some cases, especially with radiologists with fewer years of experience [2].

With the rapid development in machine learning and computer vision techniques, and because of the above problems, Computer-Aided Diagnosis (CAD) systems appeared and developed in the past decades. These systems are developed to process different forms of data such as medical images, clinical data, and genetic data. Recently, deep learning models have made these systems work more consistently, especially now there are a variety of neural network architectures that can be used in different ways depending on the data type in order to get the most out of them. For example, CNN models fit more the medical imaging tasks [3], while RNN [4, 5] models seem to be more effective with the sequential data such as genetic data [6]. Therefore, the data type is critical in determining which architecture will be used to design the CAD system.

This work mainly focuses on the mammographic CAD systems. Mammographic CAD systems are divided into Computer-Aided Diagnosis and Computer-Aided Detection systems; the first one mainly focuses on interpreting the predefined abnormality (benign or malignant). On the other hand, the second type of these systems aims to detect and localize the abnormality within mammography. Different factors make detecting and classifying the masses very challenging for the CAD system, such as shape, mass size variation ranging from too small to very large, and the nature of the breast tissues that sometimes mask the masses, especially with the highly dense breast tissues.

Mammograms were generated using systems based on phosphorescent screen-film technology until the US Food and Drug Administration approved the Full Field Digital Mammography (FFDM) systems. The FFDM systems have significantly improved the quality of mammographic images and the sensitivity of breast cancer detection. Even the models that were trained on FFDM images provided better results than the ones provided through the Screen Film Mammography technique (SFM) [7]. However, the contrast of the FFDM images depends on the differences between breast tissues; those images can provide just structural information. One of the limitations of this type of

mammographic image is that the highly dense breast tissues can mask the mass, as most of the fibro-glandular tissues have the same image gray levels of the lesions. This can lower the sensitivity in detecting tumors when breast density increases [8]. Recently, Contrast Enhanced Spectral Mammography (CESM) was introduced in 2011 as a new image technique for mammogram screening. CESM provides improved visualization for mammographic images by combining low and high-energy breast images. Figure 1 illustrates the difference between the FFDM and CESM images and how the CESM can provide more morphological features that can enhance breast cancer detection, especially with masked masses in highly dense breasts. Although the crucial information that these images can provide clinically, few studies have been proposed for developing deep learning-based CAD systems in CESM images. Almost all of the previously introduced studies by researchers used FFDM images or SFM images [9, 10].

These various studies proposed different models for mass detection and classification based on different deep learning techniques and with the use of transfer learning concept. As deep learning has great capability to automatically extract the deep features from the mammograms with no need for hand-crafted features, in addition to that, transfer learning reduces the computational cost and the need of large datasets for training.

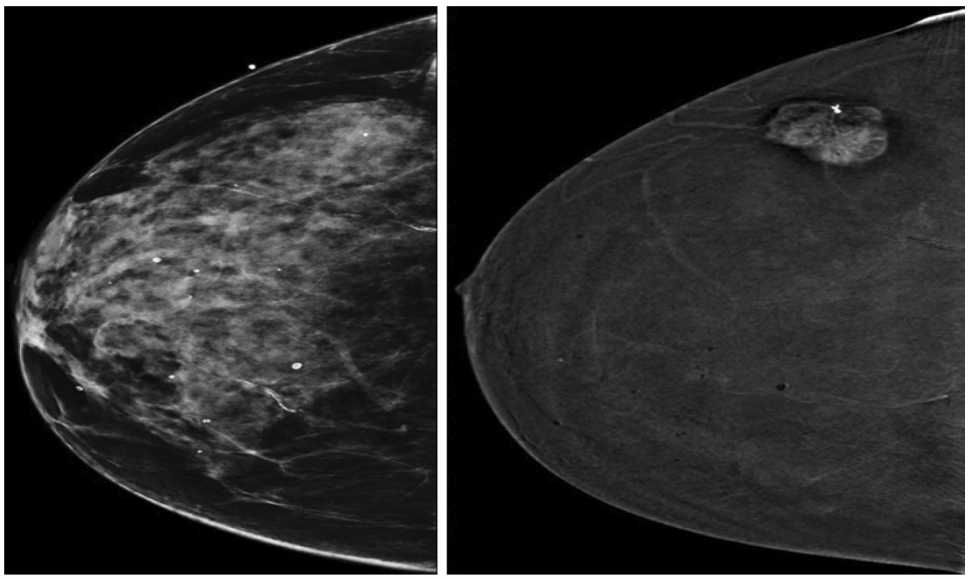
Convolutional Neural Networks (CNN) ruled the detection and classification tasks in medical imaging diagnosis through the last few years in the most of these studies, relying on the idea of the dependency on the immediate neighboring pixels which represent the local features of the image (such as color, contrast, etc.). This allows the model to learn and extract the essential features and edges only without learning the details of each pixel and the global context of the features. However, learning the entire image, rather than the parts that the filter extracts, may increase the chances of obtaining better performance from the model.

Recently, vision transformers have shown competitive performance compared to CNNs; vision transformers are built on the attention mechanism, which focuses on the local and global spatial features [11]. Few studies adopted the ViT transformers in medical imaging tasks [12–14] especially in breast cancer diagnosis. Furthermore, few works proposed a fully automated model for detecting and classifying the masses. Based on the aforementioned points, this work proposes a fully automated framework in an end-to-end training fashion for mass detection and classification based on YOLOv4 and ViT transformers in CESM mammographic images and FFDM images. This work leverages the ViT transformer to extract the global features alongside with the local features to enhance the accuracy of the diagnosis.

Moreover, the experiments are designed to explore the efficacy of the automated interpretation of CESM and how it can increase the sensitivity of breast cancer detection more than FFDM.

The novelty and contributions of this work can be summarized in the following points:

1. Integrate the YOLOv4 with ViT transformers for mass detection and classification, to utilize the capability of ViT in learning the local and global spatial features in the mass classification instead of the CNN models.
2. To the best of our knowledge, this is the first fully automated CAD framework for mass detection and



(a) (b)

Fig. 1 **a** Low-energy FFDM image shows negative findings; **b** recombination CESM image that clearly shows the existence of mass [8]

classification in CESM images, specifically in the CDD-CESM dataset [15].
3. Evaluate how CESM can have the potential to enhance the diagnostic accuracy comparing to Digital Mammography (DM).
4. Assess the performance of YOLOv4, YOLOv7, and YOLOv8 in mass detection on CDD-CESM and INbreast dataset based on a comprehensive evaluation through different experiments.
5. Conduct different experiments to evaluate and compare the performance of different vision transformer models and different CNN models at mass classification in mammograms.
6. Utilize the newly introduced CDD-CESM dataset [15] for mass detection and classification to evaluate the performance of the proposed model.

This paper is organized as follows: Sect. 2 presents the literature survey, while Sect. 3 demonstrates the methods and materials employed in this work. Section 4 shows the experimental design and results, and Sect. 5 discusses the results. Finally, Sect. 6 presents the conclusion, while Sect. 7 discusses the advantages, limitations, and directions for future work.

2 Related work

The rapid development of deep learning techniques hugely affected the researchers' contribution in developing more accurate CAD systems. Over the past years, many attempts have been made to introduce reliable systems for breast cancer diagnosis and prognosis, especially with the dependency on using mammograms as a first tool for the initial diagnosis. The proposed techniques in this literature mainly focused on one or more of three tasks; detection, segmentation, and classification.

Detection can be described as the process of localizing the abnormal area or spots within the mammogram images, while segmentation mainly targets the pixel-by-pixel annotation of the abnormal findings; finally, the classification is considered as the process of classifying the findings into (Normal/Abnormal) or (Benign/Malignant). Some studies focused on observing the morphological features (texture, color, brightness, etc.) in their works to extract the ROIs and then classify them into benign or malignant using feature-based/conventional machine learning techniques [16–20].

However, feature-based techniques have been used for a long time. Still, they have some drawbacks as those techniques mainly depend on classical feature engineering that is affected by different factors such as the subject knowledge of the developer, his intuition, and his skills in the mathematical models. This process is considered a time-consuming process; moreover, this may not capture all the relevant features in the image as those techniques cannot automatically learn the most discriminant features [21]. Furthermore, feature-based techniques are often designed to capture specific characteristics or patterns through the training phase. Accordingly, this may not generalize well to unseen data, especially since those techniques struggle with large and high-dimensional feature space datasets; and this can lead to more computational overhead.

On the other side, with the appearance of deep learning, CNNs replaced the traditional hand engineering approaches for feature extraction, as the CNNs can automatically extract and learn complex features in more detail and in a more efficient way that fits the required task [22, 23]. The initial convolutional layers in the deep CNN can effectively extract the low-level features, while subsequent layers propagate these features to extract more complex and abstract features. Through the training process, the filters and pooling operations automatically select the most discriminant and informative features. Using deep learning in feature extraction provides benefits that can overcome the problems of feature-based techniques. Automating the feature extraction process through deep learning saves the time and effort needed to extract hand-crafted features. The deep learning techniques capture the features at multiple levels of abstraction, and this hierarchal representation allows more informative features to be extracted. Moreover, transfer learning reduces the computational cost of learning the features from scratch; as the pre-trained models can be used as a feature extractor for related tasks [24].

Different studies adopted various architectures in their developed systems that integrates CNN with conventional machine learning techniques. In [25], they proposed a system based on CNN and SVM for mass and microcalcification segmentation and classification that achieved a classification accuracy of 80.5% and 87.2% on DDSM and CBIS-DDSM, respectively. One of the points that needs to be investigated in this work is employing a deep learning model for the segmentation phase to allow the model to learn more discriminant features about masses and calcifications; and accordingly, this may enhance the results.

In [26], the authors also proposed a CAD system for whole mammogram classification based on deep CNNs for feature extraction and SVM for classification. They conducted their experiments on two different datasets, MIAS and INbreast, that is composed of FFDM images; their approach achieved an accuracy of 97.93% and 96.64%, respectively.

In 2015, a giant leap occurred in object detection techniques, especially with the appearance of new deep learning-based models that can detect multiple objects within

one image. Those models are categorized into one-shot and two-shot detectors; the most well known among them is You Look Only Once (YOLO) [27] because of its performance at both accuracy and computational time levels. Al-Mansi et al. [28] were one of the pioneers who exploited YOLO in their work; they introduced a YOLO-based CAD system that achieved an accuracy of 85.2% for detection. In [29], they proposed a CAD system for mass detection, segmentation, and classification; they utilized YOLO for detection, then segmented the detected masses using Full Resolution Convolutional Network (FRCN) and AlexNet architecture-based classifier for classification. Their approach achieved a detection accuracy of 97.2%, segmentation accuracy of 92.97%, and classification accuracy of 95.3%. However, their model was straggling with small mass detection, moreover, one of the drawbacks in their model is the manual elimination of the false localized masses before segmentation phase, which is impractical for automated diagnosis.

In [30], the authors developed a YOLO-based CAD system to detect the tumors and classify them into masses and calcification; their experiments were conducted on two different datasets, INbreast and CBIS-DDSM. The model achieved 98.1% and 95.7%, respectively. Their model isn't providing a diagnosis about the malignancy of the detected masses. Additionally, the model has high inference rate/image relative to the other recent similar studies.

Also, Faster-RCNN [31] was one of the object detection models that showed promising performance in some studies. In [32], Ribli et al. also utilized the two-shot detector Faster-RCNN in their developed system for mass detection and classification. Their system detected 90% of the malignant masses with a classification accuracy of 95%. Agarwal et al. [33] also proposed a Faster-RCNN-based model for mass detection; the model achieved a sensitivity of 95–71% and a specificity of 70%. Comparing to the results of the models that adopted the YOLO, these models have lower detection sensitivity. Additionally, Faster-RCNN consumes more time at detection than YOLO.

Cao et al. [34] developed a novel model that detects breast masses based on anchor-free technique named FSAF, an enhanced model of RetinaNet [35]. Moreover, they proposed a new augmentation method to increase the size of the dataset. This augmentation technique enhanced their results; however, it has more computational cost rather than the traditional augmentation techniques. The model attained 0.495 False Positive Rate (FPR)/image for INbreast and 0.599 FPR/image for DDSM.

Shen et al. [36] proposed a framework for mass detection with an attempt to automate the process of mass annotation in mammographic images. The model mainly depended on adversarial learning; the experiments were done over a private dataset and INbreast; it achieved an AUC of 0.9083 and 0.8522 for each dataset, respectively. However, their training strategy addressed the problem of oscillation and the limitation of small batch size, this approach needs to be experimented on more different medical imaging datasets.

2.1 Insights from related work

The mammograms' quality affected breast cancer detection sensitivity; most of the proposed work was done over mammographic datasets of SFM or FFDM mammographic images. FFDM images provide better quality, so this type of image has replaced the SFM images in recent years; however, one of the problems still exists is the masked masses in cases with high breast density.

In FFDM, the gray levels of the glandular tissues have similar values as the masses. Accordingly, this makes the mass hide within the dense tissues and decreases the visibility of the masses. CESM is a relatively new technique for obtaining mammographic images; this technique depends on getting a new mammographic image by subtracting the low-energy image from the high-energy image.

Studies showed that the CESM could provide more morphological features and higher sensitivity in detecting lesions than the FFDM and the SFM [8, 37] especially for cases with highly dense tissues. Song et al. [38] proposed a deep-information bottleneck-based network for classifying the CESM images; their model aimed to learn the relevant features between the images. Their approach achieved an accuracy of 97.2%; they used the CDD-CESM dataset in their experiments. Other studies [39–41] introduced different classification models based on deep learning and conventional machine learning for CESM images. From Table 1, it can be noticeable that most of the proposed work was done mainly on the FFDM images, few studies only have been proposed CAD systems for CESM. There is not enough evaluation of the performance of CAD systems on CESM images, specifically in mass detection.

Moreover, based on the previously reviewed work, the proposed techniques in the literature mainly did not yet explore the potential of the transformers in learning the global context of the pixels in the classification task of abnormal findings in mammographic images. Therefore, a

Table 1 Summary of recent studies for both deep learning and feature-based techniques

Reference	Methodology	Feature extraction	Dataset	Image type	Detection	Classification	Results
[28]	YOLOv1	Deep learning	DDSM	FM	✔	✔	Detection Acc: 96.33% Classification Acc: 85.52%
[29]	YOLO + FRCN + Inception ResNet v2	Deep learning	INbreast	FFDM	✔	✔	Detection Acc: 97.27% Classification Acc: 95.32%
[30]	YOLO v3	Deep learning	INbreast CBIS-DDSM	FFDM FM	✔	—	Detection Acc: 98.1%, 95.7%
[33]	Faster-RCNN	Deep learning	INbreast OPTIMAM	FFDM FFDM	✔	—	Detection TPR @ FPI: 0.99 ± 0.03 at 1.17 for malignant and 0.85 ± 0.08 at 1.0 for benign (INbreast), 0.91 ± 0.06 at 1.69 (OPTIMAM)
[32]	Faster-RCNN	Deep learning	DDSM INbreast	FM FFDM	✔	✔	Detection of 90% of the malignant masses
[25]	AlexNet-based CNN lv + SVM	Deep learning	DDSM CBIS-DDSM	FM	—	✔	Classification Acc: 80.5%, 87.2%
[38]	Multi-feature deep-information bottleneck	Deep learning	CDD-CESM	CESM	—	✔	Classification Acc: 97.2%
[39]	Multilayer Perceptron (MLP)	Feature-based	Private dataset	CESM	—	✔	Classification acc: 84.8%
[41]	AlexNet and RawNet	Deep learning	Private dataset	CESM	—	✔	Classification sensitivity: 100% Classification specificity: 66%
[42]	Three different features extraction methods (Statistical + GLCM features, two methods based on Wavelet Transformer WT) + SVM	Feature-based	DDSM	FM	—	✔	Classification Acc: 98.69%
[18]	Extract intensity, texture, and morphological features using mathematical expressions + SVM for classification	Feature-based	DDSM MIAS	FM SFM	—	✔	Classification Acc: 96.6% Classification Acc: 97.5%
[43]	Lifting Wavelet Transform (LWT) for feature extraction from ROIs + Extreme Learning Machine (ELM) with Moth Flame Optimization (MFO) for classification	Feature-based	MIAS DDSM	SFM FM	—	✔	Classification Acc: 99.94% Classification Acc: 99.68%

comprehensive assessment has been done between CNN models and vision transformers models to exploit the ability of the vision transformers in learning the long-range dependencies and the relationship between image pixels in mass classification.

Consequently, this work proposes a fully automated framework for mass detection and classification in CESM images by integrating the power of YOLO with the ViT transformer. Furthermore, the performance of the model was also evaluated on FFDM images. Also, the work

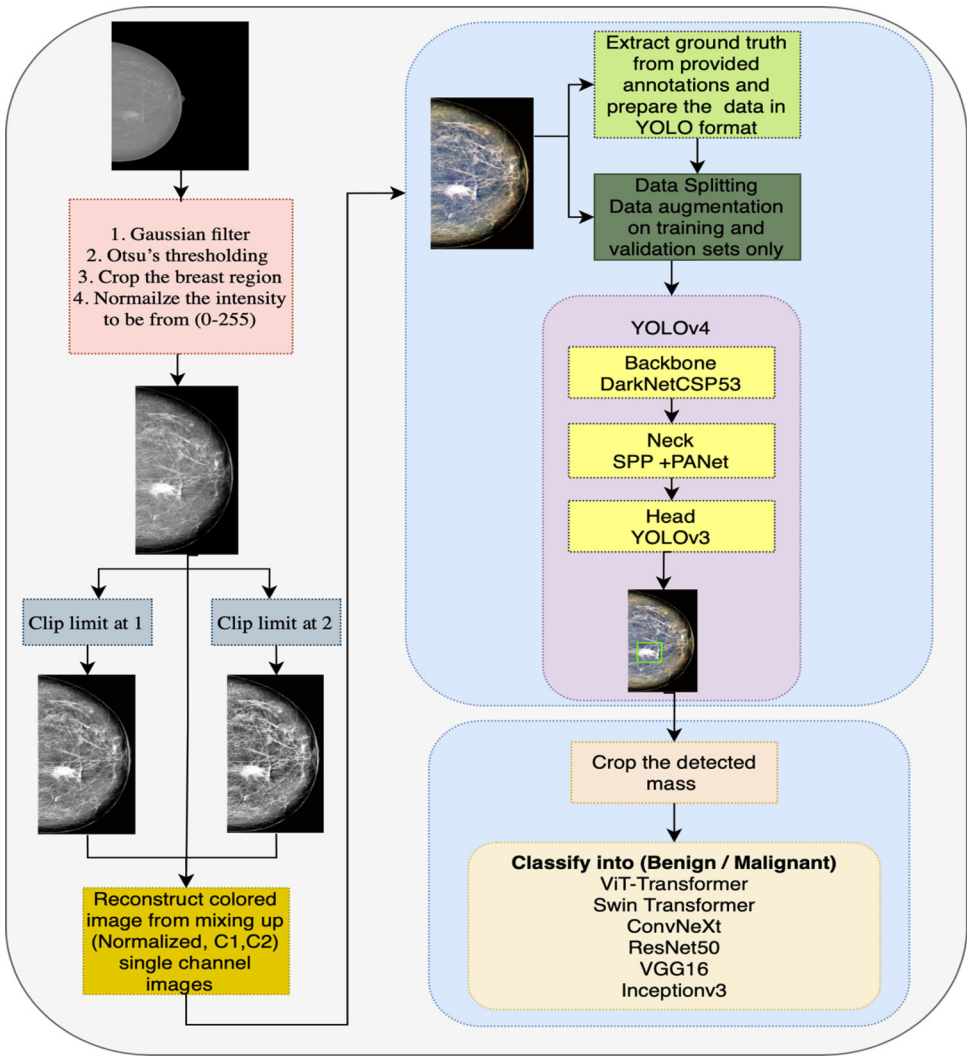


Fig. 2 Flow diagram for the steps of the proposed framework for mass detection and classification

provides a comprehensive study of the potential of CESM vs. FFDM in mass detection and classification. Table 1 summarizes recent studies in FFDM and CESM images for mass detection and/or (mass/whole image) classification.

3 Methods and materials

This work proposes a fully automated framework for mass detection and classification in end-to-end training strategy in CESM and FFDM images; deep learning models are adopted in both phases. The proposed CAD system can be divided into three parts, as shown in Fig. 2; pre-processing, detection, and classification. The pre-processing steps were inspired by the technique that was used by [34]; three methods were applied at this phase Gaussian Filter [44], Otsu's thresholding [45] and clip limit adaptive histogram equalization (CLAHE) [46] as shown in Fig. 2. YOLOv4 architecture was adopted for mass detection while the classification network of YOLO was replaced with the ViT-transformer network; Fig. 3 illustrates the proposed framework.

3.1 Pre-processing phase

3.1.1 Gaussian filter and Otsu's thresholding

Firstly, the Gaussian filter was used to reduce the noise and blurring of the images. Gaussian filter is a linear filter with a symmetric kernel with an odd size that passes through each pixel in the image. The values inside the kernels are calculated as shown in Eq. 1, where (x, y) represents the pixel coordinate, σ is the standard deviation of the Gaussian distribution.

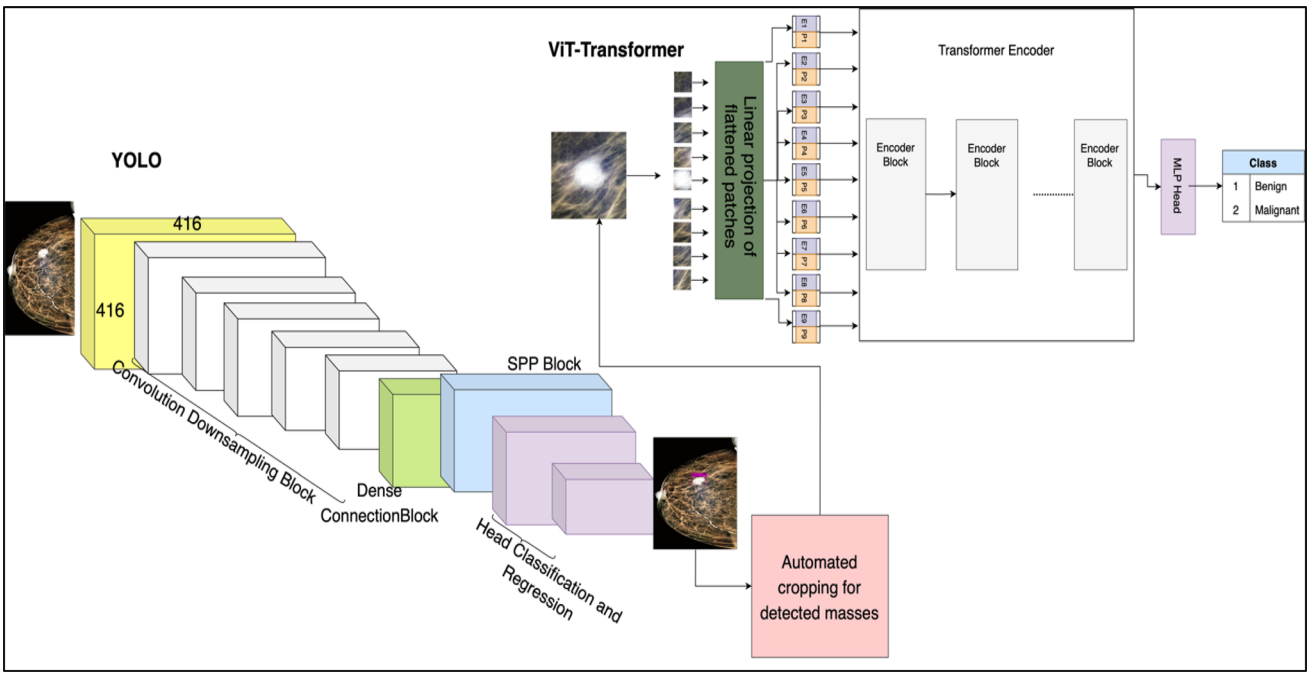


Fig. 3 A proposed integrated framework of YOLO and ViT transformer for mass detection and classification

$$G(x,\ y)\ =\ \frac{1}{2\pi\sigma^2}e^{-\frac{x^2+y^2}{2\sigma^2}} \tag{1}$$

Otsu thresholding was also used to find the suitable threshold for separating the foreground pixels and the background pixels to minimize the area of the background to crop the Region of Interest (RoI) of the breast.

3.1.2 Clip limit adaptive histogram equalization (CLAHE)

CLAHE was used to enhance the contrast of the mammographic image; CLAHE mainly enhances the local contrast of the image. It works at small tiles of the image rather than the whole image. This algorithm is used to enhance the contrast of the medical images to improve the visual appearance of the mammogram. The clip limit (CL) is an essential parameter for CLAHE, as this parameter controls the image's brightness level. In the pre-processing phase, two clip limits were used, as shown in Fig. 2.

3.2 Detection phase

At this phase, You Look Only Once (YOLO) model has been selected for mass detection. YOLO is a well-known object detection architecture known as a single-shot detector, as the image is processed in one shot to detect multiple objects within it. One of the advantages of YOLO is looking at the complete image, which means less information loss, and this is considered one of the crucial points that affect the interpretation of medical images. Furthermore, the detection and classification are done simultaneously, making the YOLO faster than other detectors. Figure 4 illustrates the architecture of YOLO.

There are different official versions of YOLO, which are YOLOv1, YOLOv2 [47], YOLOv3 [48], YOLOv4 [49], YOLOv7 [50], and YOLOv8 [51]. Different studies [7, 52, 53] show that YOLOv4 performs better than the other elder versions, YOLOv1, YOLOv2, and YOLOv3. The improvements that were introduced in YOLOv4 enhanced the accuracy and detection time. However, the recent versions, YOLOv7 and YOLOv8, were introduced with some improvements to enhance the trade-off between accuracy and time.

3.2.1 YOLOv4 vs. YOLOv7 and YOLOv8

An experiment was conducted to select the most suitable version of YOLO that fit the used datasets in this work among the recently introduced YOLO versions. The experiments use the same datasets for YOLOv4, YOLOv7, and YOLOv8 to evaluate the performance of these recent models on medical images, specifically mammograms. According to the experiments, YOLOv4 outperformed

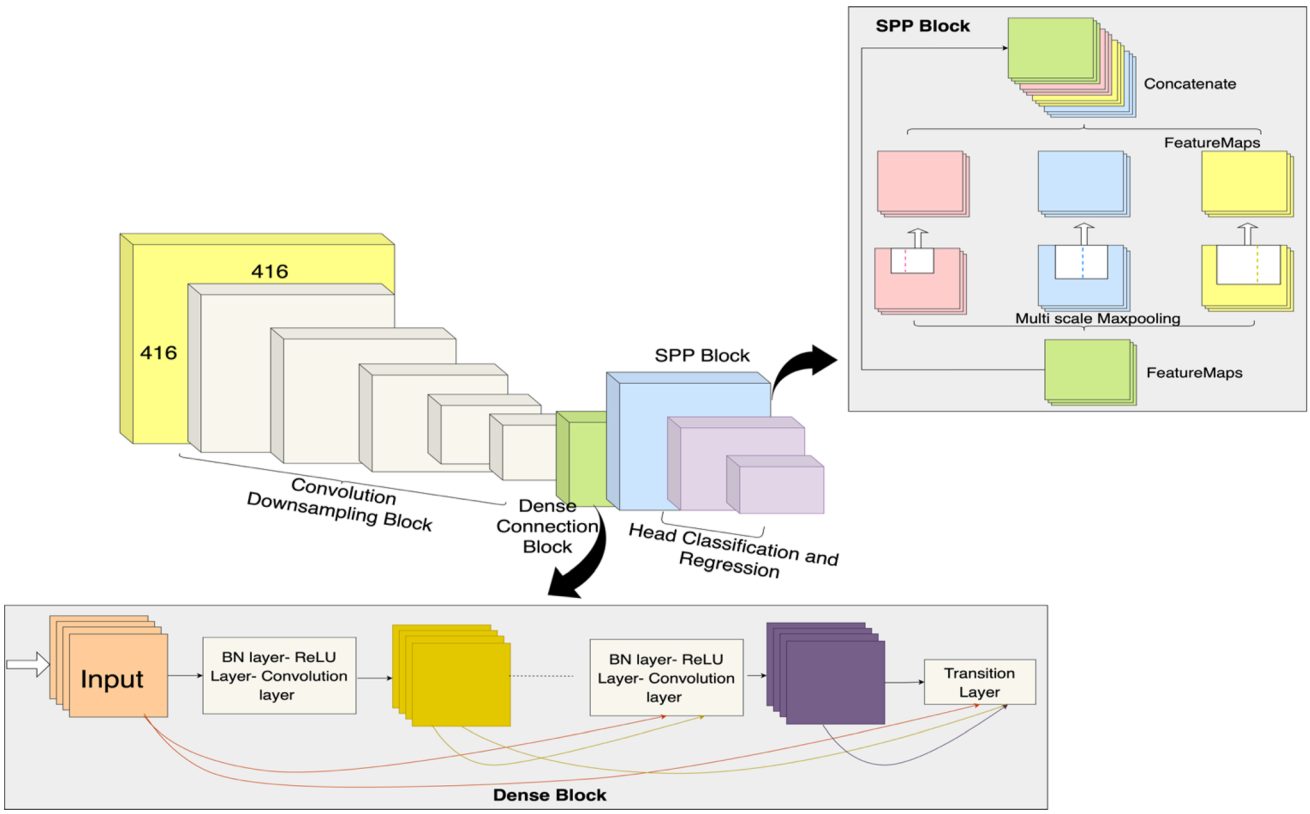


Fig. 4 YOLO v4 Architecture [33]

Table 2 Detection results of YOLOv4, YOLOv7, and YOLOv8 on INbreast and CESM-CDD datasets

Dataset	Model	mAP(%)@ IoU = 0.5	Recall (%)	Precision (%)	Inference/image
INbreast	YOLOv4	98.96%	100	92	0.0892 s
	YOLOv7	97.60%	95.60	95.70	0.0377 s
	YOLOv8	89.3%	85.7	86.8	0.0545 s
CE-CDD	YOLOv4	81.52%	79	77	0.0833 s
	YOLOv7	61.70%	64.10	66.30	0.0469 s
	YOLOv8	50.60%	54.80	58.50	0.0578 s
DM-CDD	YOLOv4	71.65%	71	68	0.0144 s
	YOLOv7	58.10%	54.10	67.30	0.0297 s
	YOLOv8	46.20%	45.20	50.70	0.0758 s

YOLOv7 and YOLOv8 regarding mAP, recall, and precision; however, YOLOv7 showed competitive results on the INbreast dataset. On the other side, YOLOv7 and YOLOv8 provide faster performance than YOLOv4, as shown in Table 2.

Based on the conducted experiments, YOLOv7 struggles in small mass detection, especially with crowded mammographic scenes, whether false mass detection or missed mass detection, especially with CESM-CDD dataset. It is not performed well in detecting masses at different scales, as it struggles with masses that are very large or very small regarding the other masses in the mammographic image. In addition, lighting changes can cause significant variations in the appearance of masses; YOLOv7 is adversely affected by the changes in lighting [54], which makes it inconvenient for mass detection, especially in mammographic scanning, where lighting variations are common.

Regarding YOLOv8, the main change is adopting a new anchor-free detection mechanism [55]. Also, the model is

built based on a new modified version of CSP-DarkNet-53 as the backbone [51] such as YOLOv4. On the other hand, the model is still under construction and development, so its performance is not stable yet. Also, the model struggles with small object detection in complex scenes [56], especially with the probability of the overlapping between the small-size objects and other size objects that may partially block its appearance.

Due to the consequences of the above reasons, at the detection phase, YOLOv4 was used to mainly detect the masses existing in the mammograms. It splits the input mammographic image into grid cells (s × s) cells; if the mass falls within the cell, it is considered responsible for detecting this mass. A fixed number of bounding boxes is predicted for each cell with their confidence score; each box's confidence score represents the probability of containing a mass multiplied by the Intersection over Union (IoU) between the ground truth and the predicted box.

Many different configurations were introduced for YOLO; these configurations are set up according to the application domain and the used datasets in the experiments. The most important and effective step before training the YOLO is adjusting the anchor boxes according to the dataset and the resolution of the input image; accordingly, the K-Means clustering algorithm was used for that in all conducted experiments. The anchors were generated in these experiments for each dataset separately and based on the different resolutions used.

3.3 Classification phase

3.3.1 Vision transformers

The transformer is considered a de facto architecture for Natural Language Processing (NLP) tasks. The transformers generally are built on the self-attention mechanism that allows the model to learn the global dependencies between the inputs and outputs. This mechanism mainly lets the inputs interact with each other to know which features the model should pay more attention to.

The promising results achieved with transformers in NLP tasks, especially with the transfer learning on downstream tasks, opened the door to introducing the transformers to the computer vision tasks. Many versions of vision transformers have been introduced recently; Dosovitskiy et al. [57] were the first ones who introduced the ViT transformer. After that, many were proposed, such as DeiT (Data Efficient Transformer) [58], Swin–Transformer [59], and ConvNeXt [60]. Furthermore, different studies recently utilized transformers in various medical imaging tasks such as classification, segmentation, and detection. In [61], the authors proposed a model for predicting breast tumor malignancy using a convNeXt transformer over ultrasound images. Van et al. [62] also introduced a model that utilized transformers to build a cross-view transformer model that was tested on multi-view medical images from two datasets CBIS-DDSM for breast cancer mammography and CheXpert for chest X-rays. In addition, different models based on transformers were proposed for COVID-19 diagnosis [63, 64]. Accordingly, and due to its success in prediction and classification, one of the main objectives of this work is to utilize the transformers in the detected masses classification.

3.3.2 Why transformers?

The kernels of the convolutional networks focus only on the local texture, which represents a local subset of pixels from the image, and that enforces the network to ignore the global context of the features as the network fails to encode the relative position of the features; accordingly, different studies were proposed recently to overcome this problem by utilizing attention mechanisms and pyramid networks.

Transformers are one of the recent architectures that built on the concept of the self-attention mechanism. The significant advantage of transformers that can be exploited in medical image diagnosis is the model's ability to understand the pixels' global context through learning the long-term dependencies between the data. For example, in this work, the surrounding area of the mass provides more information that can help the model observe more discriminant features about the global context and the correlation between the masses and its surrounding tissues; this can provide a more accurate diagnosis.

3.3.3 How vision transformer (ViT) works

The ViT Transformer is used for classification in this work; as shown in Fig. 5, at the classification phase, there are three main components of the ViT network; patch embedding, feature extraction using stacked transformer encoders, and the classification head that is built with Multilayer Perceptron (MLP).

The mammography image is reshaped into a sequence of patches. Then, the patches are flattened and mapped to dimensions with linear projection, as a constant latent

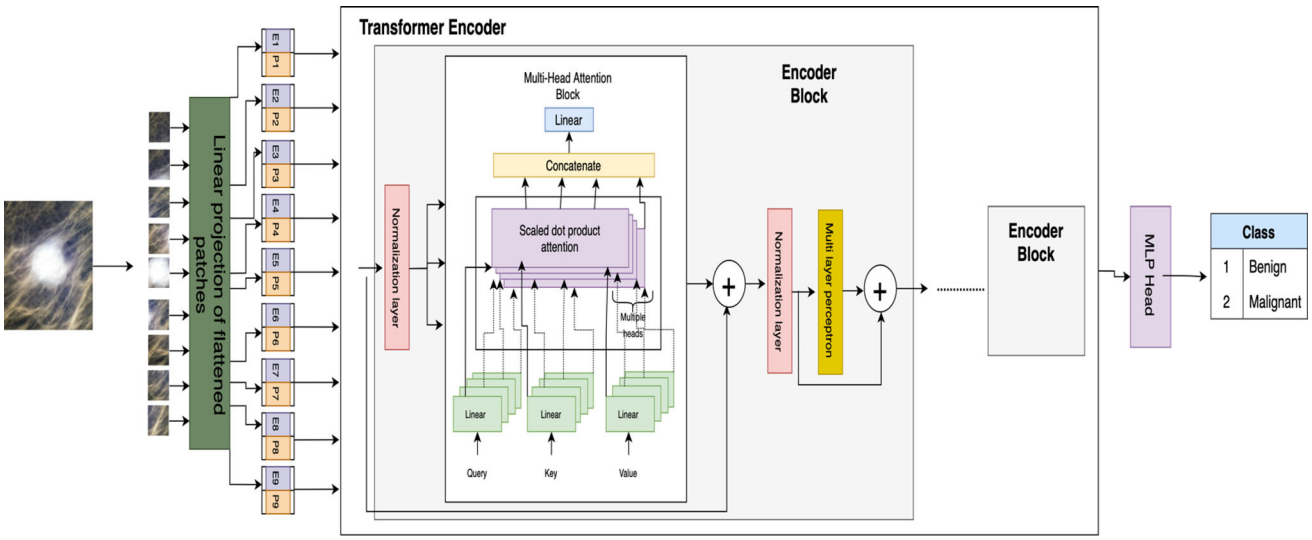


Fig. 5 Modified ViT transformer for mass classification

vector is used through all the layers of the transformer network. Each image $\mathbf{x} \in \mathbb{R}^{H \times W \times C}$ (where H, W, and C are representing height, width, and number of channels of the image) is split into N non-overlapped patches, each of size 16×16. These patches are reshaped into this form $\mathbf{x}_p \in \mathbb{R}^{N \times \left(P^2 \cdot C\right)}$ to be in sequence of 2D patches in n line vectors of the shape of $(1, P^2 \cdot C)$; where P is the resolution of each patch, while N represents the input sequence length (number of patches) that will be fed to the transformer after applying linear projection according to Eq. 2.

$$\mathbf{z}_0 = \left[\mathbf{x}_{\text{class}};\ \mathbf{x}_{\text{p}}^1\mathbf{E};\ \mathbf{x}_{\text{p}}^2\mathbf{E};\cdots;\ \mathbf{x}_{\text{p}}^N\mathbf{E}\right] + \mathbf{E}_{\text{pos}}, \quad \mathbf{E} \in \mathbb{R}^{\left(P^2 \cdot C\right) \times D},\ \mathbf{E}_{\text{pos}} \in \mathbb{R}^{(N+1) \times D} \tag{2}$$

At the patch embedding phase and after splitting the image into n^2 patches of shape (P, P, C), the flattened patches are multiplied by a trainable embedding tensor which is in a shape of $(P^2 \cdot C, d)$ to learn how to project each patch linearly into d dimension (where d is a constant value with the network architecture). The embedded patches are aggregated with the positional embeddings generated through $\mathbf{E}_{\text{pos}}$, a trainable positional embedding tensor. The resultant $\mathbf{z}_0$ is fed into the transformer encoders; the encoders learn the global features from the embedded patches of the mass through multi-headed self-attention layers. The encoders generate a sequence of tokens. Those tokens are fed into the MLP to generate the prediction of the class label if it is a benign mass or a malignant mass.

3.4 Transfer learning

This work adopted the transfer learning concept due to the lack of publicly available mammograms and to minimize the training time. Pre-trained weights for different networks were used to initialize the parameters of these networks through the training phase.

The YOLO network was fine-tuned for detection, and the pre-trained weights on the COCO dataset were transferred to initialize the network weights for training on the two mammographic datasets. On the other side, the models that were used for classification were trained on ImageNet and fine-tuned to predict the class of each mass.

In the conducted experiments for classification, six classification models were used; three of them are transformer based, which are ViT, SWIN, and ConvNeXt, while the other three are CNN-based, which are VGG16, ResNet50, and Inception v3.

The proposed model used ViT transformer; this model was pre-trained on ImageNet21K. The input image size of the model is 224×224, and the input images in this model are divided into 16×16 patches. The Adam optimizer was used through training. A linear layer was added on the top of the pre-trained encoder to downstream the model for the mass classification task. Additionally, the MLP head was modified to just generate two outputs that are malignant and benign.

Furthermore, for the other classification models, some modifications have been done to transfer the pre-trained weights of those models into the required task. The base

model of SWIN was used; the model was pre-trained on ImageNet1K, and the model took an input image size of 224 × 224. Also, the Adam optimizer was used with weight decay to avoid the overfitting problem, and the MLP head was also modified to output malignant or benign.

ConvNeXt architecture was adopted in the experiments to explore the potential of this model, as the architecture of this model integrates the ConvNeXt based on ResNet50 with the design of the training approaches of the vision transformer, the network was pre-trained on ImageNet22K, with an input image size of 224 × 224.

ResNet50 was modified; the last dense layer was excluded from the feature extraction layers. Moreover, the layers of ResNet were frozen to use the pre-trained weights during the training phase. A fully connected layer was added to obtain the final prediction. Binary cross-entropy loss function and Adam optimizer were used.

VGG16 was also modified as the classification layer was removed and replaced by a fully connected layer to classify the masses into malignant and benign. Furthermore, the base layers of VGG16 were frozen to restore the pre-trained weights of the ImageNet.

The final tested model was Inception v3, used in the classification phase; the model took input images of size 224 × 224. The base layers were set to be not trainable, and the last layer was also replaced with a fully connected layer to fit the mass classification task. For ResNet50, VGG16, and Inception v3, the Adam optimizer was used with binary cross-entropy loss function; also, the three models were pre-trained on ImageNet.

4 Experimental design and results

The experiments were conducted through three main phases; the first phase is the pre-processing of the datasets, followed by the mass detection phase, and finally, the mass classification phase. For mass detection, the YOLOv4 model was trained to detect the masses regardless of their type; then, the detected masses are fed into a model based on the vision transformers architecture to classify those masses into benign and malignant. Furthermore, different classification models were used through the experiments to evaluate the performance of the transformer-based models versus the CNN-based models on mammographic diagnosing.

4.1 Dataset

This work used two datasets: INbreast for FFDM images and CDD-CESM for both FFDM and CESM images. Each dataset was split into training, validation, and testing sets using the splitting ratio of 70%–10% and 20%, respectively. During the splitting process, the images for the same patients were included in the same set to prevent the results from being biased. The same sets were used across both detection and classification tasks to guarantee fair evaluation for the whole model.

4.1.1 INbreast

This dataset is composed of mammographic images that were produced using the FFDM technique. It includes 410 images in DICOM format for 115 cases with both views, Medio Lateral Oblique (MLO) and Carnio Caudal (CC). However, the dataset did not provide both views for each patient. Ninety cases of this dataset were diagnosed with cancer; 107 images were diagnosed with benign and malignant masses, and those are included in this work's experiments. The masses were classified into benign and malignant based on the BI-RADS score that is provided with the dataset, where 1, 2, and 3 are considered benign; on the other hand, 3, 4, and 5 are counted as malignant. The mass annotations were extracted from the XML files that are provided with the dataset, as those annotations were converted to be in the accepted form for YOLO annotation. As each image should be attached with its corresponding txt file that has the bounding boxes of the existing masses, where each box is represented by (C_x, C_y, w, h); (C_x, C_y) represents the coordinates of the center point of the bounding box, w is the width, and h is the height of this box. Table 3 illustrates the distribution of the dataset over training, validation, and testing sets that have been used for both detection and classification.

4.1.2 CDD-CESM

CDD-CESM is a newly introduced publicly available dataset that provides images in two types FDDM and CESM. It is the first publicly available dataset that contains CESM images. The dataset includes 2006 images divided into 1003 low-energy images (FDDM) and 1003 subtracted images representing the (CESM). The dataset has MLO and CC view images; the images were acquired using two different scanners that are G.E. Healthcare Senographe DS

Table 3 INbreast dataset splitting distribution for the experiments

	Training	Validation	Testing
Benign	25	6	4
Malignant	51	3	18
Total	76	9	22

and Hologic Selenia Dimensions Mammography Systems. The average resolution of the images is 2355 × 1315; the images were obtained from 326 patients, all of whom are females between 18 and 90 years old. The dataset provides manual segmentation annotations for the existing abnormalities in the mammographic images, which are provided according to the ACR-BIRADS lexicon. The images are available in JPEG format with a CSV file for the annotations; the medical reports are also attached to the dataset. The dataset includes 310 FFDM images with masses and 333 CESM images with masses; in the conducted experiments, we used 310 images for the experiments on FFDM and CESM. Table 4 shows the data distribution of CESM in the training, validation, and testing sets. The corresponding images (cases) to those used in the CESM modality were included in the experiments conducted on the FFDM modality to guarantee a fair evaluation of each modality's efficacy in the automated breast cancer diagnosis process.

As the provided annotations in this dataset were done for segmentation, there are three forms for those annotations, polygon, circle, and ellipse. Accordingly, some pre-processing was done to convert those annotations into bounding boxes; the polygon points for X and Y were used to get $X_{\min}$, $X_{\max}$, $Y_{\min}$, and $Y_{\max}$ from those lists, and those coordinates were used as coordinates for obtaining the bounding boxes.

For the circle annotations, the dataset provides center x (cx), center y (cy), and the radius r; So Eqs. (3) and (4) were used to get, x_1, and, y_1 of the box; to get the width and the height the r multiplied by 2 then the width were added to x_1 to get, x_2 and height to y_1 to get y_2.

$$x_1 = cx - r \tag{3}$$

$$y_1 = cy - r \tag{4}$$

To get the bounding box of the ellipse segmentation annotation, the provided values for cx, cy, rx, and ry were used to get x_1, y_1, x_2, and, y_2 according to Eqs. (5), (6), (7), and (8).

$$x_1 = cx - rx \tag{5}$$

$$x_2 = cx + rx \tag{6}$$

Table 4 CDD-CESM dataset splitting distribution for experiments

	Training	Validation	Testing
Benign	50	9	13
Malignant	173	16	49
Total	223	25	62

$$y_1 = Cy - ry \tag{7}$$

$$y_2 = Cy + ry \tag{8}$$

After converting the segmentation annotations into bounding boxes, those boxes' coordinates are altered to fit the format of the YOLO annotations, as mentioned before. Table 4 illustrates the distribution of the dataset over training, validation, and testing sets that have been used for both detection and classification.

4.2 Implementation environment

The experiments were conducted on a single machine with NVIDIA GeForce RTX3080Ti GPU with 12 vRAM, Intel® Core i7-11700 k processor with 3.200 GHz frequency, and 32 GB RAM. C + +, python 3.8, and TensorFlow were used to implement the proposed system on Windows 10 operating system.

4.3 Implementation set-up

The proposed framework was implemented over different phases: pre-processing, augmentation, detection, and classification. The following subsections demonstrate the implementation set-up of each phase.

4.3.1 Pre-processing and augmentation

For INbreast images, the images were converted from DICOM images into JPEG images; however, this step was skipped for CDD-CESM as the images are already in the JPEG format. The breast region was cropped to minimize the background area. The images were normalized to make the pixel intensity distribution in the range from (0–255). Then CLAHE was used at two different clip limits (1 and 2), and the generated images were mixed up with the normalized image to reconstruct new colored images. This step was done for both datasets to obtain new images with enhanced visibility for the mammographic images in addition to the original images in the training phase. The images were augmented to increase the number of images in the training phase, especially with the existence of imbalanced distribution between malignant and benign cases; four augmentation techniques were applied (vertical flip. Horizontal flip, multiplicative noise, and random rotation) for both images the original and the newly reconstructed one. At the detection phase, one more augmentation technique (mosaic) was used and selected based on previous experiments by [65], as it showed better performance in detection than other techniques that are used with YOLO. Algorithms 1 and 2 show the pre-processing algorithmic steps for CDD-CESM and INbreast.

Algorithm 1 Pre-processing CDD- CESM mammograms algorithm

Input:
- Patient ID and associated image files.
- CSV files containing mass descriptions and annotations for the images.

Output:
- Transformed and augmented images and their corresponding annotations in YOLO format.
- Text files containing bounding box information in YOLO format.
- Preprocessed images for each patient.

Algorithmic Steps:

1. **Import Libraries:** Load libraries like cv2, pandas, numpy, etc., for image processing and data handling.
2. **Define Helper Functions:**
   - get polygon formatted: Extracts X and Y coordinates from a given set of points, filtering out specific characters.
   - bounding_box: Calculates the bounding box for a given set of points.
   - mass type: Determines the type of mass (benign or malignant) for a given patient ID from a CSV file.
   - MinMaxNormalise: Performs min-max normalization on the given image.
   - clahe: Enhances the image using the Contrast Limited Adaptive Histogram Equalization (CLAHE) method.
   - synthetized_images: Creates a 3-channel image composed of the normalized and contrast-enhanced images at *clip= 1 and clip =2*.
3. **Read and Process Data:**
   - Read a CSV file containing mass annotations and convert it to a panda DataFrame.
   - Iterate over image files in a specified directory.
4. **Image Preprocessing:**
   - For each image file, read the image and optionally apply image preprocessing techniques like normalization, contrast enhancement, etc.
5. **Generate YOLO Annotations:**
   - For each image, determine its shape and mass type.
   - Iterate over the DataFrame rows to extract and process annotations (circle, ellipse, polygon).
   - Calculate bounding box coordinates for different shapes using these equations:

     If *Circle* use *center x* (cx) and $center\ y\ (cy)$:

     $$x_1 = cx - r$$
     $$y_1 = cy - r$$
     $$x_2 = x_1 + (2 * r)$$
     $$y_2 = y_1 + (2 * r)$$

     If *Ellipse* use *center x (cx), center (cy), radius (rx)* and *radius y (ry):*

     $$x_1 = cx - rx$$
     $$x_2 = cx + rx$$
     $$y_1 = cy - ry$$
     $$y_2 = cy + ry$$

     If *Polygon* use X and Y list of pointsto get Xmin, Ymin, Xmax and Ymax:

     $$x_{\text{scale}} = 1024/image\_width$$
     $$y_{\text{scale}} = 1024/image\_height$$
     $$xmin = X1[0] * y_{\text{scale}}$$
     $$xmax = (\text{int}\ (Y1[0])) * y_{\text{scale}}$$
     $$ymin = (\text{int}\ (X1[1])) * x_{\text{scale}}$$
     $$ymax = (\text{int}\ (Y1[1])) * x_{\text{scale}}$$
   - Convert these coordinates to the YOLO format (normalizing by image dimensions) and write them to a text file.
6. **Process and Augment Data:**
   - Run a loop to process files in a specified directory, applying image transformations, and saving both the transformed images and new annotations (apply the augmentation on training and validation sets).

Algorithm 2 Pre-processing INbreast mammograms algorithm

Input:
- DICOM images and XML files containing mammogram data and annotations.
- Directory paths for DICOM images *(DCM_PATH)* and XML annotations *(XML_PATH)*.
- *Patient ID* for processing specific patient data.

Output:
- Transformed and augmented images and their corresponding annotations in YOLO format.
- Text files containing bounding box information in YOLO format.
- Preprocessed images for each patient.

Algorithmic Steps:

1. **Import Libraries:** Load necessary libraries for image processing, data manipulation, and file handling.
2. **Define Paths:** Set the directory paths for DICOM images and XML annotation files.
3. **Image Transformation Definition:**
   - Define a function getTransform to create various image transformations (flipping, rotating, etc.) using the albumentations library.
4. **Load and Process Masks:**
   - Implement function named load_inbreast_mask to read XML files and create image masks indicating regions of interest in mammograms.
5. **Convert Masks to YOLO Format:**
   - Implement a function mask_to_yolo to transform image masks into bounding box annotations compatible with the YOLO model.
   - For each mask get mass bound box coordinates and convert it into YOLO coordinates using these equations:

$$x = |x|/\text{ image\_width}$$
$$y = |y|/\text{ image\_height}$$
$$h = |h|/\text{ image\_height}$$
$$w = |w|/\text{ image width}$$

6. **Generate Annotation Text Files:**
   - Convert bounding boxes into text annotations in YOLO format using a function bbox to txt..
7. **Extract Mass Type:**
   - Get the type of mass (benign or malignant) using *patient ID* and a reference CSV file to add it the YOLO txt file for the annotations using the mass_type function.
8. **Image Preprocessing Functions:**
   - Implement functions like crop, truncation normalization, clahe, and synthetized_images for various image preprocessing tasks (e.g., cropping, normalization, contrast enhancement) to Creates a 3-channel image composed of the normalized and contrast-enhanced images at *clip= 1 and clip =2*.
9. **Process and Augment Data:**
   - Run a loop to process files in a specified directory, applying image transformations, and saving both the transformed images and new annotations (apply the augmentation on training and validation sets).

4.3.2 Mass detection

The YOLOv4 is used in this work for the detection task; Darknet with CSP53 is used as a backbone for the network. Some network parameters were set up and modified at the network layers based on the domain and the used datasets in the experiments.

Table 5 shows the set-up configurations that were used in this work for YOLOv4 layers. In this work, different experiments were conducted to select the suitable input image size for the network, so the experiments were done

Table 5 Set up configurations for YOLOv4 network layers on 416 × 416 and 640 × 640-input sizes

Image size	(416 × 416), (640 × 640)
Nom. of anchors	9
Batches	64
Subdivisions	16 for 416 image size, 64 for 640 image size
Number of classes	1, 2
Learning rate (LR)	0.001
Scales	0.1, 0.1
Steps	3200, 3600
Max. batches	4000
No of filters	(18), (21)

for two input sizes (416 × 416) and (640 × 640). Moreover, the model performance was evaluated by two different scenarios; the first one was to detect the existence of the masses regardless of their type (benign/malignant), while the other one was designed to detect the benign masses and malignant masses separately. Accordingly, the number of classes in some experiments was 1 class, and in other experiments was 2. This consequently affects the number of filters to be 18 with the 1 class detection and 21 with the 2 classes detection based on the equation mentioned in [10]. The learning rate was selected to be 0.001 based on different experiments. Two scales, 0.1 and 0.1, were used to change the learning rate at two different steps that are 3200 and 3600. Finally, the max batches were set up to be 4000 as recommended in [27] (max batches = number of classes × 2000). Also, K-means clustering was used to select the anchors' sizes based on the dataset; nine anchors were used for those experiments.

4.3.3 Mass classification

Six experiments were conducted for classification with the same sets for training, validation, and testing that were used in detection for each dataset. The experiments were done to evaluate the performance of transformers versus the CNN-based networks. The selected CNN-based networks are VGG16, ResNet50, and Inception v3, as those networks showed promising results in other studies for mass classification. This work uses three transformer-based networks: ViT transformer, SWIN, and ConvNeXt. The input size for all networks was 224 × 224, Adam optimizer was used with binary cross-entropy loss function, and all the models were modified to transfer their pre-trained weights to fit the mass classification task. Furthermore, weight decay was used for all transformer models to avoid the overfitting problem. Algorithm 3 shows the algorithmic steps for the mass detection and classification process.

Algorithm 3 Mass detection and classification algorithm

Input:
- Transformed and augmented images and their corresponding annotations in YOLO format.
- Text files containing bounding box information in YOLO format.
- Preprocessed images for each patient.

Output:
- Detected masses and its pathology (Benign / Malignant)

Algorithmic Steps:
1. **Data Splitting:** Split the dataset into training, validation and testing with ratio of 70, 10 and 20 respectively.
2. **Perform mass detection with pre-trained YOLOv4 model:**
   - Define anchor boxes according image size using K means clustering.
   - Set values of (image size, number of batches, subdivisions, number of classes, Learning Rate, scales, steps, max-batches and number of filters).
   - Enable mosaic augmentation.
   - Train YOLOv4 using training set and validation set for fine tuning through training.
   - Evaluate the model on the testing set in terms of *mAP, precision, recall, F1-score* and *FNR.*
   - Save the coordinates of detected masses in a CSV file that has *patient_id, top_x, top_y, width* and *height.*
3. **Crop Detected Masses and masses of training and validation sets to be used in classification:**
   - Load and read the file of the coordinates of detected masses from YOLOv4.
   - Load and read the file of the coordinates of masses in training and validation sets
   - Run nested loops to process images in a specified directory to crop the masses (one to iterate over the records of the coordinates file and the other to find the image that will be processed).
   - Save the cropped images into folders based on whether the cropped mass is benign or malignant.
4. **Perform classification with ViT Transformer network:**
   - Load pretrained ViT Feature Extractor model (*ViT-base16*) for feature extraction (act as a tokenizer).
   - Load training, validation and testing data.
   - Apply tensor Transforms on (training, validation and testing)
   - Apply Data loader on the training set
   - Define the model for classification (pretrained model -*ViT-base16*)
   - Add a linear layer on the top of the pre-trained encoder to downstream the model for the mass classification.
   - Replace the network classification head (fully connected layer) with a new one to map the features of masses for prediction score for Benign and Malignant classes.
   - Specify training parameters (*LR, epochs, batches, training set, validation set, metrics, … etc.*).
   - Define optimizer with (*ADAM optimizer*) and loss function with (*Binary cross entropy*).
   - Load the evaluation metric and train the model.
   - Get validation accuracy and evaluate on testing set.
   - Get classification report in terms of *accuracy, sensitivity, specificity,* and *F1-score*.
   - Calculate confusion matrix.
   - Calculate the *AUC* and plot *ROC*.

4.4 Evaluation metrics

Different evaluation metrics were used to evaluate the performance of the proposed model. The Intersection over Union (IoU), mean Average Precision (mAP), F1-score, precision, and recall are used for detection. Moreover, the True Positive Rate (TPR) and False Negative Rate (FNR) were calculated using True Positive (TP), False Negative (FN), and False Positive (FP). The performance was evaluated for classification through accuracy, sensitivity, specificity, confusion matrix, Area Under the Curve AUC, and Receiver Operating Characteristics (ROC).

IoU was used in evaluating the detection; IoU represents the overlapping between the predicted bounding box and

the ground truth at a specific threshold; the threshold used in the proposed model is 0.5.

For detection, only TP, FP, and FN were defined as the detected mass is considered to be TP if the IoU $> = 0.5$ and considered to be FP if the IoU is < 0.5. If the model fails to detect an existing mass, this is counted as FN.

The mAP is also calculated to estimate the mean of all average precisions over all classes in the used dataset, where mAP is calculated using the following Equation.

$$\text{mAp} = \frac{1}{N}\sum_{i=1}^{N}(AP)_i, \quad \text{where } N \text{ is the number of classes} \tag{9}$$

Also, the precision was measured to define the percentage of the truly detected masses regarding the total number of actual existing masses, as shown in Eq. (10).

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{10}$$

Furthermore, the recall (sensitivity) and FNR were calculated as shown in Eqs. (11) and (13).

The dataset sets suffer from an imbalanced class problem, as shown from the data distribution in Tables 3 and 4 in Sects. 4.1.1 and 4.1.2. Accordingly, F1-score was calculated as it is used as one of the useful evaluation matrices in this case, as shown in Eq. (12).

$$\text{Recall(senstivity/TPR)} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{11}$$

$$F1 - \text{score} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \tag{12}$$

$$\text{FNR} = \frac{\text{FN}}{\text{TP} + \text{FN}} \tag{13}$$

Different classification metrics were used for evaluation, and the accuracy score of the model was calculated for both validation and testing according to Eq. (14). The confusion matrix was used to represent mainly four attributes based on the classification results that are:

True Positive (TP): this represents the number of benign masses that were classified correctly as benign.

True Negative (TN): this indicates the number of malignant masses that were classified as malignant.

False Positive (FP): this represents the number of the misclassified masses as benign, and they are malignant.

False Negative (FN): this represents the number of the misclassified masses as malignant, and they are benign.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TN} + \text{TP} + \text{FP} + \text{FN}} \times 100 \tag{14}$$

Moreover, specificity was calculated as shown in Eq. (15), also FPR and TPR were calculated to plot the ROC curve and calculate the AUC score. The ROC curve represents the trade-off between FPR and TPR; the higher the AUC score, the better the ability of the model to differentiate between benign and malignant masses.

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{15}$$

$$\text{FPR} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{16}$$

4.5 Mass detection results

This section illustrates the detection results on both datasets INbreast and CDD-CESM, with its two image types (DM-CESM). Table 6 shows the Benign vs. malignant mass detection results on INbreast. In contrast, Tables 7 and 8 show the results of mass detection regardless of its type on INbreast and CESM, respectively, at different resolution input sizes. Moreover, Fig. 6 presents false detection on a mammographic image from the INbreast dataset. At the same time, Fig. 7 illustrates detection results on some images on FFDM vs. its corresponding CESM images from the CDD-CESM dataset.

Tables 6 and 7 show that the performance of YOLO is better at detecting the existence of masses regardless its type. Also, Tables 7 and 8 show that the higher input image size improves the detection accuracy as the best mAP is achieved at input size of 640. The 640-input size achieved mAP of 98.96% in INbreast dataset, 81.52% on CESM images and 71.65% on DM images from CDD_CESM dataset. Moreover, the experimental results showed that DM lower detection sensitivity than CESM.

It can be seen from Fig. 6 an example of a false detection mass of a mammographic image from the INbreast dataset. The ground truth based on the provided annotations has only one mass for this case with patient_id (22,614,236), while the model detected two masses. One of them is truly detected with confidence score of 94% and the other one is a false detected mass.

For further demonstration of the impact of mass detection on CESM, Fig. 7 illustrates the effectiveness of using CESM in enhancing the sensitivity of mass detection rather than FFDM images. The figure provides samples of images from CDD-CESM dataset in forms of CESM and their corresponding DM to illustrate the performance of the model at mass detection in these images regarding their ground truth.

4.6 Mass classification results

This section provides the results of the mass classification using six modified models; three of them are CNN based, and the other three are vision transformer-based models.

Table 6 Detection results on INbreast for Benign masses vs. Malignant masses in terms of mAP, F1-score, TP, FP, FN, recall, and precision

Dataset	Input image size	mAP(%)@ IoU = 0.5	F1-score (%)	TP	FP	FN	Recall (%)	Precision (%)	B-AP (%)	M-AP (%)	TP-FP (B)	TP-FP (M)
INbreast	416 × 416	84.43%	86	19	3	3	86	86	75.00	93.86	3–2	16–1

Table 7 INbreast mass detection results regardless of its type (Benign/Malignant) at different input image sizes in terms of mAP, F1-score, TP, FP, FN, recall, precision, and FNR

#	Input image size	Normal cases	mAP(%)@IoU = 0.5	F1-score (%)	TP	FP	FN	Recall (%)	Precision (%)	FNR
Trial 1	416 × 416	Without normal images	97.78%	96	22	1	1	96	96	0.0454
Trial 2	416 × 416	With normal images	95.27%	94	22	2	1	96	92	0.0454
Trial 3	640 × 640	Without normal images	**98.96%**	96	**23**	2	0	**100**	92	0

The best results that were obtained during the experiments are indicated by bold

Table 8 Detection results for CDD-CESM for FDDM and CESM images at different input image sizes in terms of mAP, F1-score, TP, F.P., F.N., recall, precision, and FNR

#	Input image size	Image type	mAP(%)@IoU = 0.5	F1-score (%)	TP	FP	F.N	Recall (%)	Precision (%)	FNR
Trial 1	416 × 416	DM	70.27%	72	73	24	33	69	75	0.311
Trial 2	640 × 640	DM	**71.65%**	69	**75**	34	31	**71**	**68**	0.292
Trial 3	416 × 416	CESM	73.71%	74	71	16	35	67	82	0.330
Trial 4	640 × 640	CESM	**81.52%**	78	**84**	25	22	**79**	**77**	0.207

The best results that were obtained during the experiments are indicated by bold

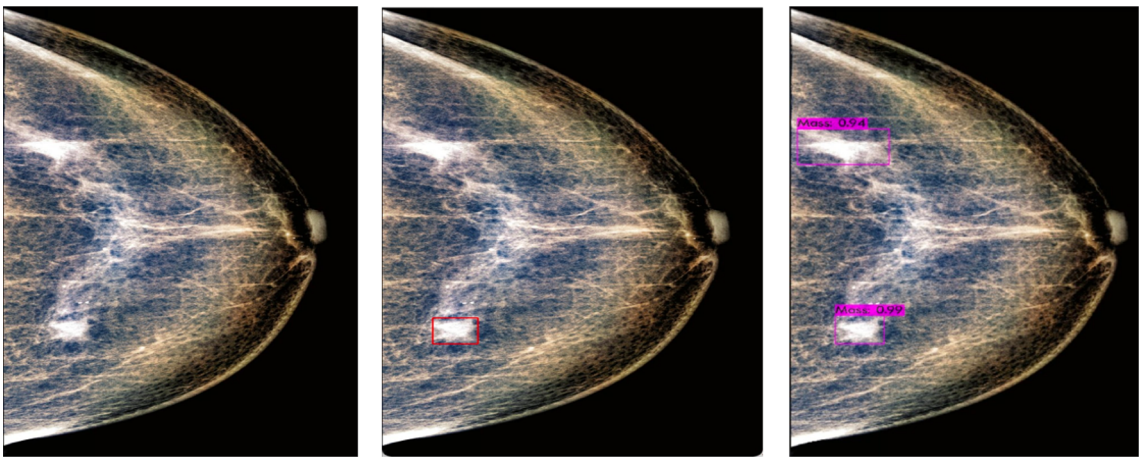


Fig. 6 False detection for a mammographic image from the INbreast dataset on the left, followed by its corresponding ground truth and then its detection result (The purple box indicates the detected mass while the red box indicates the ground truth)

Table 9 shows each network's results in different evaluation metrics. Moreover, Figs. 8, 9, 10, 11, 12, and 13 show the ROC and Confusion matrix for the results on each dataset; INbreast, CE-CESM, and CDD-CESM, respectively.

Table 9 shows that ViT transformer outperforms the other networks, specifically the CNN-based ones. It can be

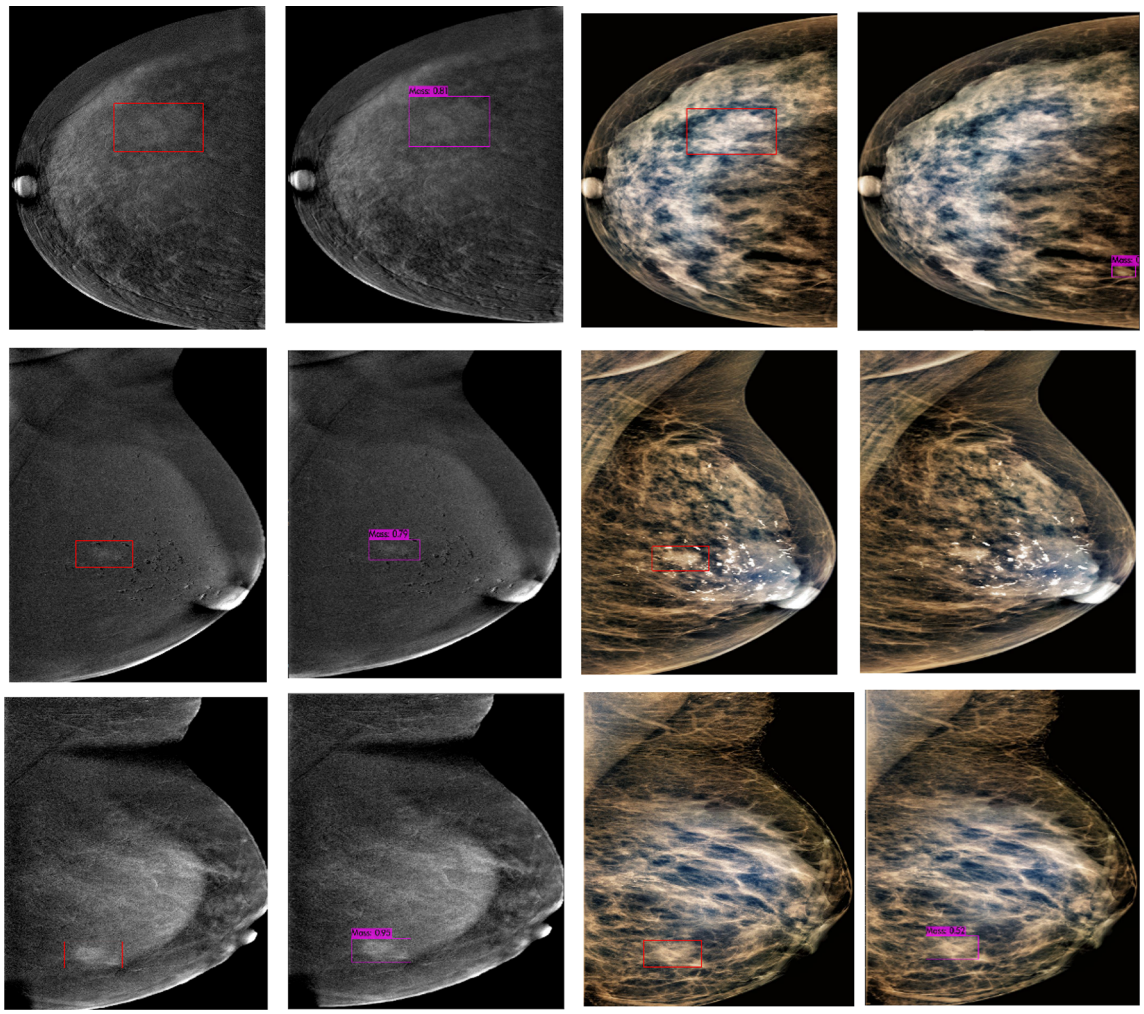


Fig. 7 Detection results for CESM images with their corresponding ground truth on the left. While on the right the same images in FFDM with its detection results and their corresponding ground truth. (Purple box indicates to the detected mass while red box indicates to the ground truth)

noticed that model achieved competitive accuracy scores on the testing set with 95.65% in INbreast, 97.61% in CESM, and 80% in DM. Also, it provides the highest AUC score among the other experimented networks, it achieved 88%, 90%, and 70% in INbreast CESM and DM, respectively. Figures 8, 10, and 12 show that ViT has the ability to differentiate better between benign and malignant masses comparing to the other networks. However, SWIN transformer and ConvNext also provide higher AUC scores than the CNN-based models (ResNet50, VGG16, and Inceptionv3). Furthermore, Confusion matrix in Figs. 9, 11, and 13 demonstrates that most of the misclassified masses is benign masses and this may have happened because the datasets lacked sufficient samples of benign masses.

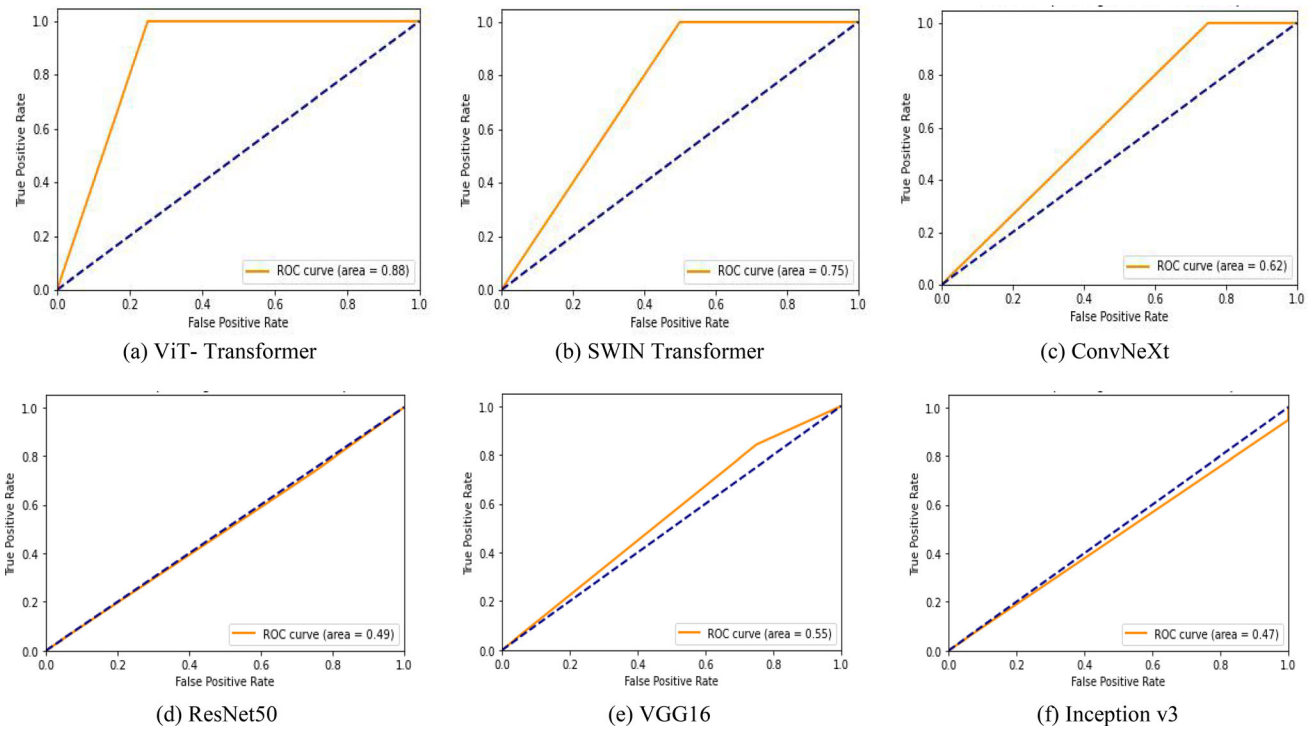
Table 9 Mass classification results for INbreast, DM-CESM, and CE-CESM in terms of validation accuracy, testing accuracy, sensitivity, specificity, precision, F1-score, and AUC

Dataset	Network	Validation accuracy (%)	Testing accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)	F1-score (%)	AUC (%)
INbreast	VIT transformer	78.50	**95.65**	**100**	**75**	**95**	**97.43**	**88**
	SWIN-Transformer	82.80	91.30	100	50	90.47	95	75
	ConvNeXt	85.70	86.90	100	25	86.36	92.68	62
	ReseNet50	97.50	90.87	73.68	25	82.35	77.77	49
	VGG16	67.14	73.91	84.21	25	84.21	84.21	55
	Inception v3	42.85	86.96	94.73	0	81.81	87.80	47
CE-CDD	VIT transformer	95.55	**97.61**	**100**	**80**	**97.36**	**98.66**	**90**
	SWIN-Transformer	95.87	94.04	98.64	60	94.80	96.68	79
	ConvNeXt	96.34	94.04	97.29	70	82.81	85.48	84
	ReseNet50	93.33	88.09	100	10	89.15	94.26	50
	VGG16	89.36	92.85	97.29	60	94.73	95.99	79
	Inception v3	92.70	84.52	94.59	10	88.60	91.50	52
DM-CDD	VIT transformer	72.83	**80.00**	**86.66**	**53.33**	**88.13**	**87.39**	**70**
	SWIN-Transformer	76.41	77.33	83.33	53.33	87.71	85.47	68
	ConvNeXt	71.69	76.00	88.33	26.66	82.81	85.48	57
	ReseNet50	60.38	**82.66**	**100**	13.33	82.19	**90.22**	57
	VGG16	66.98	70.66	78.33	40	83.92	81.03	59
	Inception v3	68.68	76	76.66	20	79.31	77.96	48

The best results that were obtained during the experiments are indicated by bold

4.6.1 INbreast mass classification results in terms of ROC and confusion matrix

See Figs. 8, 9 here.



(a) ViT- Transformer (b) SWIN Transformer (c) ConvNeXt

(d) ResNet50 (e) VGG16 (f) Inception v3

Fig. 8 ROC curve for the classification models on INbreast

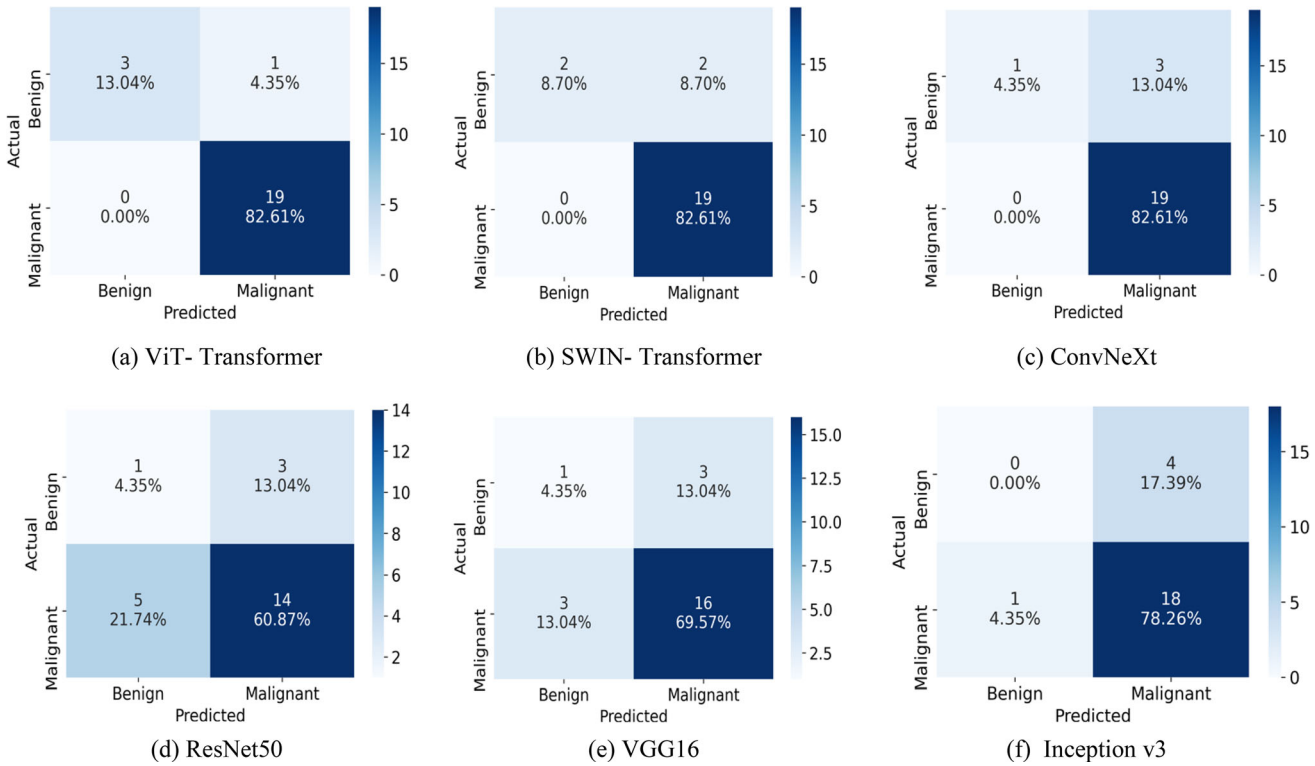


Fig. 9 Confusion Matrix (CM) for the classification models on INbreast

4.6.2 CESM-CDD mass classification results in terms of ROC and confusion matrix

See Figs. 10, 11 here.

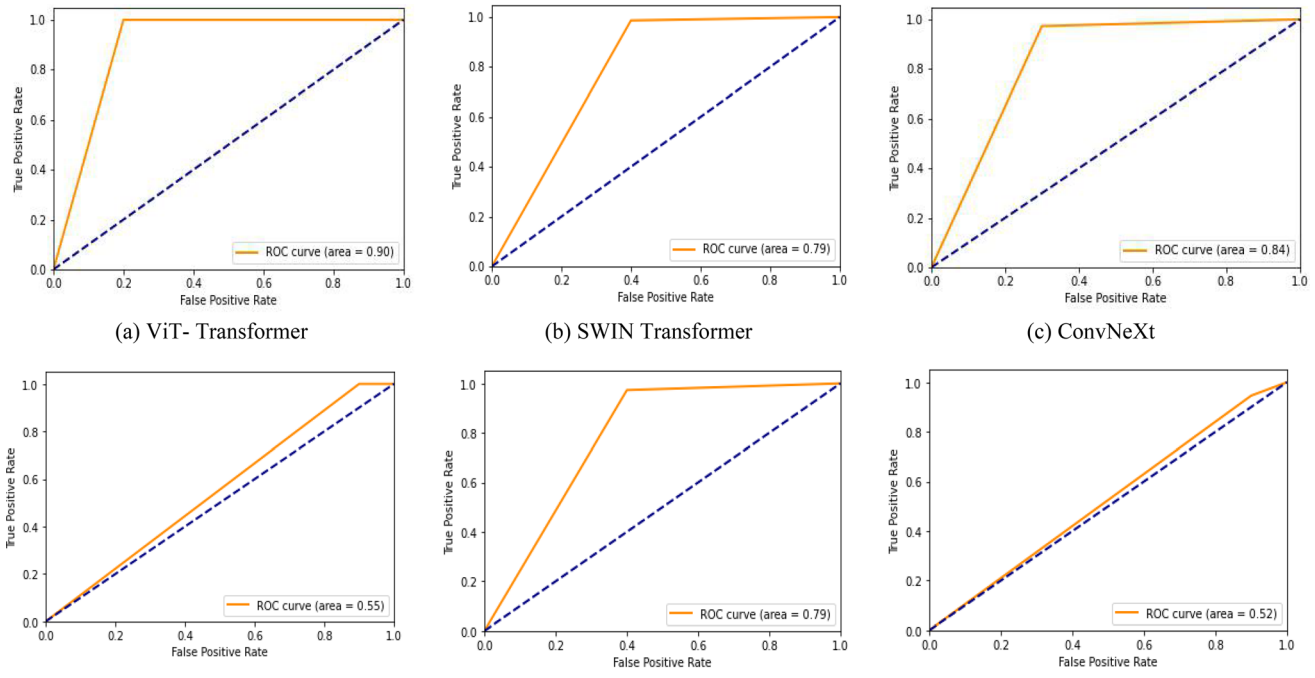


Fig. 10 ROC curve for the classification models on CE images from CDD-CESM

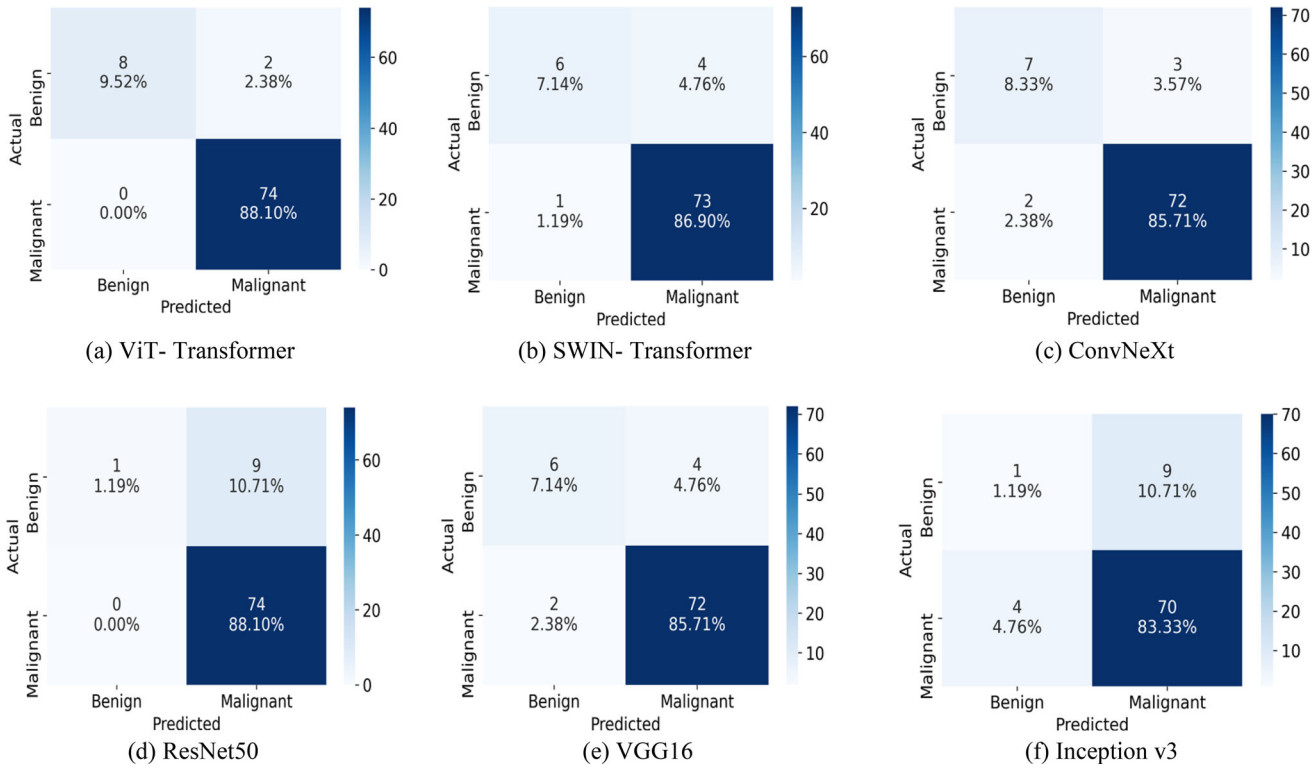


Fig. 11 Confusion Matrix (CM) for the classification models on CE images from CDD-CESM

4.6.3 DM-CDD mass classification results in terms of ROC and confusion matrix

See Figs. 12, 13 here.

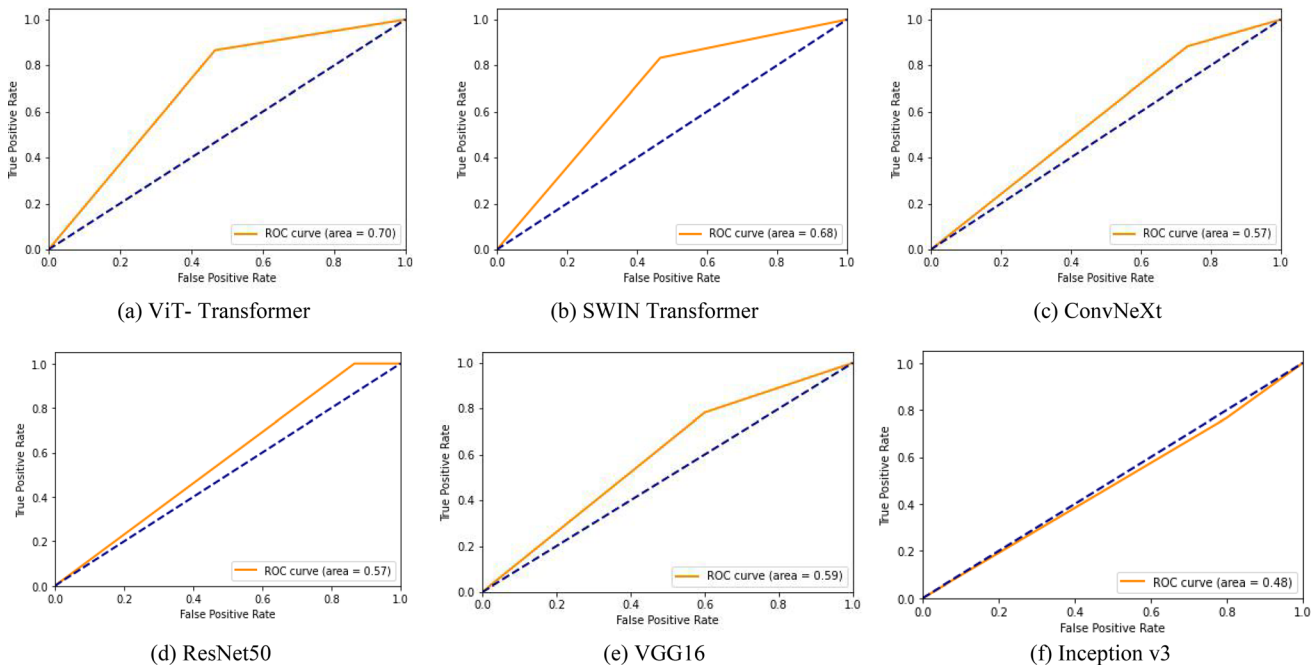


Fig. 12 ROC curve for the classification models on DM images in CDD-CESM

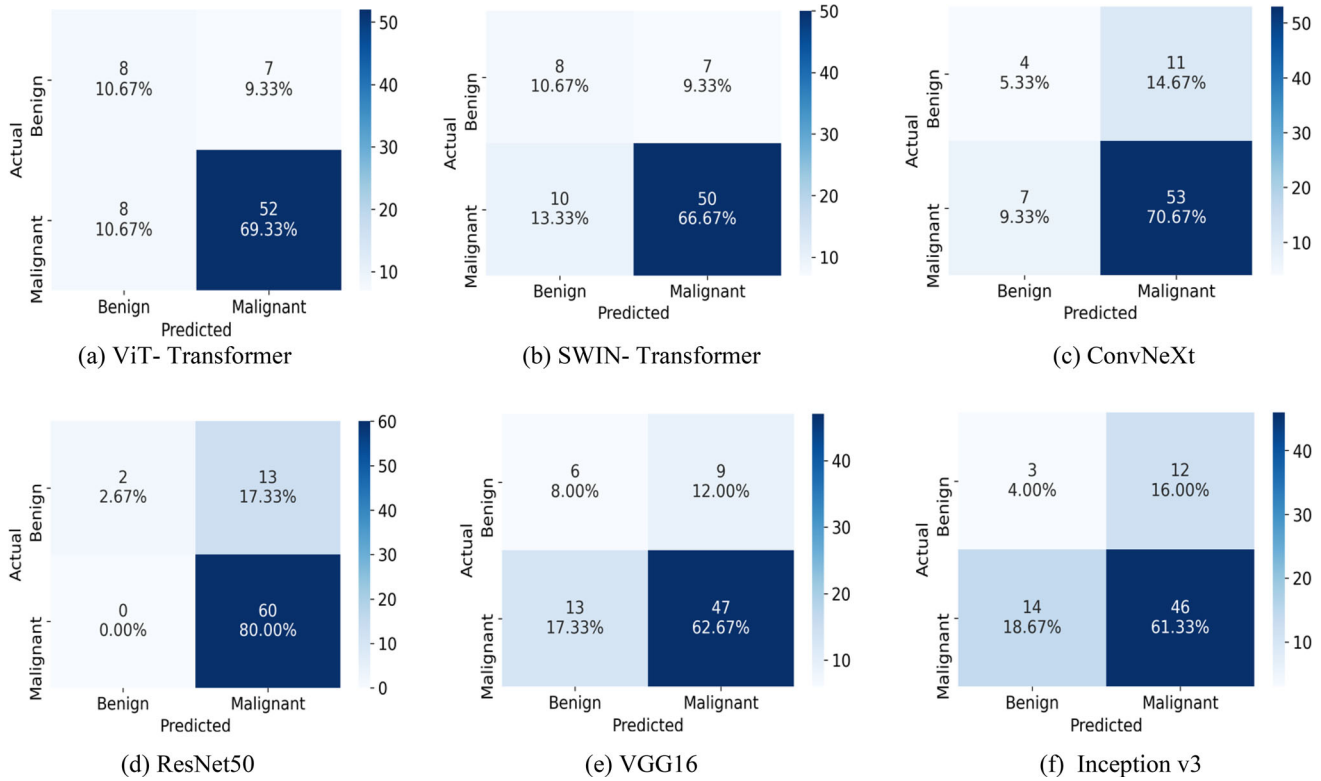


Fig. 13 Confusion Matrix (CM) for the classification models on DM images in CDD-CESM

5 Discussion

This section provides a discussion of the results. This work conducted several experiments to evaluate the model at mass detection and classification. For mass detection, various experiments were carried on to explore the significance of the higher-resolution input size of YOLO. In addition, evaluating the network's performance at detecting the existence of the masses versus detecting benign and malignant masses in both CE and FFDM images. At the classification phase, the classification network of YOLO was replaced by different Networks. Furthermore, the experiments aimed to explore the potential of the transformers regarding the CNN-based models that were used for mass classification.

5.1 Mass detection

Four experiments are conducted on INbreast. The first two experiments were done to evaluate the model in two cases; the first was to detect the benign and malignant masses, while the other aimed to detect the existence of the mass regardless it was benign or malignant. Considering the results of Tables 6 and 7, it can be deduced that the model showed better results at detecting the masses regardless of its type. The performance enhanced significantly by almost $\simeq$ 13% in terms of mAP. Moreover, this approach improved the number of the true detected masses from 19 to 22 with less FP and FN, as shown in Tables 6 and 7.

Table 7 also shows that the input image size affected the model's performance as the higher resolution showed better mAP. The experiments are done for an input image size of 416 $\times$ 416 and 640 $\times$ 640; the proposed model in Trial 3 detected all 23 existing masses in the testing set; however, there were 2 FP masses. This model achieved 98.96% mAP with a sensitivity of 100%. In Trial 1, the model achieved mAP of 97.78%, as the model failed to detect one of the existing masses, and there was one falsely detected mass. Higher resolution means less information loss, and this affected the detection, especially with the existence of small masses and high fibro-glandular tissues. Accordingly, this can clarify why the model performed better in Trial 3 than in Trial 1. Table 8 also showed that the higher input resolution provides higher mAP with more TP and less FN for both trials on DM and CE images from CDD-CESM. However, the higher resolution needs more time through training as this affects the batch size and subdivision values to allow enough memory through training.

The model in Trial 2 from Table 7 was trained with some normal images within the training set; it showed lower mean Average Precision than in Trial 1. However,

the model succeeded in detecting 22 masses out of the 23 from the testing set, but it detected 1 extra false detection more than in Trial 1. This may explain why the mAP became lower, as this false detection was in one of the normal images.

The CESM is considered a relatively recent technique for screening mammograms; based on the literature, no CAD system model for mass detection in CESM images has been introduced yet, specifically on the CDD-CESM dataset. Accordingly, in this work, we conducted experiments to explore the potential of using those images in detecting the masses. Moreover, the experiments evaluate the detection in contrast-enhanced mammography and its equivalent digital mammography for the same dataset. The results of Table 8 showed that mass detection in CESM has promising results and outperformed the model's performance on the FFDM images. In Table 8, Trials 2 and 4 showed that the mAP of the detection model was improved by 3.4%–9.8% on the CESM images for an input image size of 416 and 640, respectively. Furthermore, the model succeeded in enhancing the true mass detections, sensitivity, and precision.

Moreover, Fig. 7 illustrates the results on FFDM images and their corresponding CESM images, and it can be noticed that according to ground truth, CESM provides more accurate detection results. Apparently, from those conducted experiments, this can show that the CESM can reveal more morphological features rather than FFDM images. And accordingly, this helped the model to learn more discriminant features during the training phase. These improvements can be demonstrated from even the trials that were done on the images of size 416 × 416.

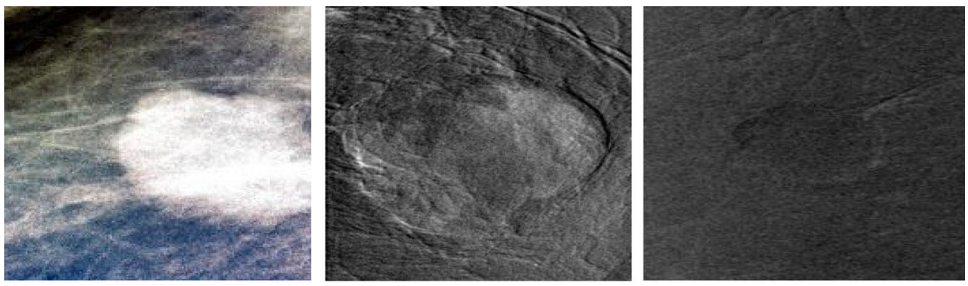
5.2 Mass classification

For classification, six models were experimented on each dataset. From Table 9, it can be deduced that the vision transformer (ViT) outperformed the other classifiers in classifying the detected masses. As shown in Table 9, the vision transformers achieved the highest accuracy for INbreast, DM images, and CE images from the CDD-CESM dataset. The model achieved a classification accuracy of 95.65% for INbreast, 97.61% for CE images, and 80% for FFDM. For INbreast and CE images, the results of ViT outperformed the other classifiers in all terms (accuracy, sensitivity, specificity, and AUC).

The ResNet 50 achieved the highest accuracy score for the D.M. images with 82%; however, it showed a lower AUC score than ViT. The ViT achieved 80% on DM images, but its performance is still considered the best among the other classifiers regarding the ROC and AUC values. The model achieved the highest score of 70%, as shown in Fig. 12. This means that the model can differentiate between benign and malignant masses more than the other classifiers. On the other hand, ResNet50 achieved an AUC of 57% and sensitivity of 100%, which means that the model is useless because it considered most of the cases as malignant. It showed bad performance at predicting the benign masses rightly, as shown in the confusion matrix from Fig. 13. Moreover, the results show that the ViT model can be generalized; based on the validation accuracy and testing accuracy scores in all trials, as the model provided a higher performance on the testing set.

SWIN and ConvNeXt performed relatively better than the CNN-based classifiers, especially in AUC scores. The transformer interprets the images as a matrix of patches, not a matrix of pixels, and this allowed the model to preserve long global relationships between the patches and obtain more semantic information rather than CNN-based models. However, the SWIN transformer provides some improvements on the regular vision transformer (ViT); it showed lower performance than the ViT. The architecture of ViT is mainly based on observing the relationships between each patch (image token) and all of the rest patches of the input image. On the other hand, SWIN is built on the idea of the shifted window design; accordingly, it looks only at the relationships between the patch and only the other patches in the windowed area. The main task for the classification part of this work is to classify the cropped mass from the detection phase into benign or malignant. In this case, the input image is the masses itself, not the whole mammographic images. Accordingly, the relationship between each patch and all other patches for the mass matters; it is not only about the relation of the patch and its neighboring patches of the same window. And so, this can clarify why ViT outperformed the SWIN in the conducted experiments. Maybe the approach of SWIN can increase the computation efficiency of the model; however, this was not so influential in these experiments as the images were not with high resolution, especially with the fact that they represent a cropped part of the original images. ConvNeXt architecture is mainly based on ResNet architecture with the same training approaches as the basic vision transformer. As shown in Table 9, ResNet provides low performance compared to the other transformer-based classifiers; which means that the architecture of the ResNet was not performed well with the mass classification task, and this can justify why the performance of ConvNeXt is lower than ViT and SIWN. However, it showed better performance than ResNet50.

From Fig. 9, it can be deduced that only one benign mass was misclassified with ViT in INbreast Dataset; the model succeeded in classifying all the malignant masses. Also, from Fig. 11 for CE images, the proposed model rightly predicted all the malignant masses. Only two benign



(a) (b) (c)

Fig. 14 Misclassified masses **a** INbreast **b**, **c** CE-CDD (truth: Benign, prediction: Malignant)

masses were misclassified; Fig. 14 shows the misclassified masses in INbreast and CE images from CESM-CDD.

It can be noticed from Fig. 13 that the most significant ratio of the misclassified masses was the benign ones compared to the ratio of the misclassified malignant masses. This may happen because the existing datasets do not have enough cases with benign masses compared to the number of existing malignant masses, and accordingly, the model trained in a better way to classify the malignant masses. However, augmentation techniques were used to overcome this problem, but those techniques do not provide too much realistic transformation for the images. And this can be considered one of the limitations facing breast cancer CAD systems.

ViT transformers utilize the idea of parallel processing, making the transformers provide more computational efficiency than CNN-based models. Table 10 shows that the inference time that the proposed framework took per image is less than the time YOLOv4 took before replacing the classification layers of YOLO with the ViT.

Tables 11 and 12 show a comparison between the proposed work and other recent studies on INbreast and CDD-CESM. The results in those tables illustrate that the proposed work shows promising and competitive results regarding the previous work. The proposed model achieved a detection accuracy of 98.96% and a classification accuracy of 95.64% in INbreast, which is almost the same result provided by the model introduced by [66]. Moreover, Table 12 shows that the proposed model outperformed the proposed model by [15] in terms of F1-score for mass detection in CE images by almost 5%; however, it achieved the same score as DM images.

Table 10 Inference time per image for the proposed framework

Method	Inference time/image (s)
YOLOv4 + ViT transformer	0.0378
YOLOv4	0.0652

6 Conclusion

Vision transformers are considerably revolutionizing computer vision tasks, especially image classification. Utilizing the power of transformers in medical image interpretation can help in enhancing the performance of CAD systems. This work proposed a novel framework for mass detection and classification based on integrating the YOLOV4 with the basic architecture of the vision transformer (ViT).

CESM images are a relatively new type of mammographic images that need more investigation in the direction of developing CAD systems that can utilize the morphological features of these images. Accordingly, this work introduces the first automated CAD system for mass detection and classification in CESM images. Furthermore, the model also was evaluated on FFDM images. The INbreast and the newly introduced CDD-CESM datasets were used in the experiments. The conducted experiments showed that the CESM images could improve the CAD system's performance at both detection and classification levels, as they showed better results than FFDM images. The proposed model achieved detection accuracy of 98.96% and 81.52%; moreover, it achieved a classification accuracy of 95.65% and 97.61% for INbreast and CESM, respectively.

The experiments also showed that the image size affected the detection results specifically for the CDD-CESM dataset as the image size of 640 enhanced the mAP for DM images by 3.4% and 9.8% for CE images compared to the image size of 416.

The proposed model utilized the potential of the vision transformers with mammographic images in classifying the masses detected using YOLOv4. Integrating the ViT architecture into our model has not only boosted its performance, but also revealed its potential in learning global and semantic features, crucial for the task at hand. ViT transformer showed very promising results compared to the other experimented models; it shows the best AUC score for INbreast and CDD-CESM datasets. Vit achieved AUC scores of 88%, 90%, 70%, and F1-score of 97.43%,

Table 11 Comparison of detection and classification results between proposed work and related studies on INbreast with YOLO

Reference	Method	Dataset	Splitting ratio (training-validation-testing) (%)	Detection accuracy (%)	Classification accuracy (%)
[7]	YOLOv3 + Inception v3	INbreast, DDSM	80–20	89.5	95.5
[66]	YOLO	INbreast	75–6.25–18.75	98.96	95.6
[30]	YOLO-based Fusion models	INbreast, CBIS-DDSM	70–10–20	98.1, 95.7	–
[67]	YOLO-based ROI Classifier	CBIS-DDSM	–	90	93.5
[68]	YOLO + InceptionResNetv2	INbreast	70–10–20	97.27	97.50
Proposed model	YOLOv4 + ViT transformer	INbreast	70–10–20	98.96	95.65
		CE-CESM		81.52	97.61
		DM-CESM		71.65	80.00

Table 12 Comparison of detection and classification results between proposed work and related studies on CDD-CESM

Reference	Method	Dataset	Segmentation/detection (F1-score)	Classification accuracy
[15]	EffecientNetB0 + GradCAM	CDD-CESM	DM: 72% C.E.: 73%	_
Proposed model	YOLOv4 + ViT transformer	INbreast	96%	95.65%
		CE-CESM	78%	97.61%
		DM-CESM	72%	80.00%

98.66%, 87.39% for INbreast, CE-CESM, and DM-CESM, respectively.

7 Advantages, limitations, and future work

Based on the conducted experiments, transformers showed better results than the CNN-based models; the transformers showed great capability at learning the global features along with the semantic features, especially with the existence of the attention mechanism that can help the model to learn which features to pay attention to at the classification task. Moreover, these experiments showed the efficacy of CESM images in developing breast cancer CAD systems.

However, some limitations need to be considered in future work; the available datasets suffer from the imbalance class problem even with using augmentation techniques as they did not provide natural transformations. Furthermore, the size of the publicly available datasets is relatively small. Moreover, the proposed model's computational time can be considered another limitation that can be enhanced in the future regarding the performance of the recent versions of YOLO.

Based on the results, this work can be extended in different directions; CESM images can be used along with FDDM for introducing a multimodal CAD system that can utilize both of these types to improve the performance of the CAD systems for breast cancer diagnosis. Also, the newly introduced YOLO models configurations need more investigations, especially with the significant improvement in the computational time and performance, so this work can be extended to investigate the potential of using these models on medical images, especially with the benefit of the ability to use them for detection or segmentation.

Author contribution NMH was involved in conceptualization, methodology, software, data curation, writing—original draft. SH helped in conceptualization, methodology, writing—reviewing and editing, supervision. Khaled Mahar contributed to conceptualization, validation, writing—reviewing and editing, supervision.

Funding Open access funding provided by The Science, Technology & Innovation Funding Authority (STDF) in cooperation with The Egyptian Knowledge Bank (EKB).

Data availability The datasets analyzed during the current study are available in the Cancer Imaging archive repository, https://wiki.cancerimagingarchive.net/pages/viewpage.action?pageId=109379611, and in the Kaggle repository, https://www.kaggle.com/datasets/ramanathansp20/inbreast-dataset.

Declarations

Competing interests The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Giaquinto AN, Sung H, Miller KD et al (2022) Breast cancer statistics, 2022. CA Cancer J Clin 72:524–541. https://doi.org/10.3322/CAAC.21754
2. Miglioretti DL, Smith-Bindman R, Abraham L et al (2007) Radiologist characteristics associated with interpretive performance of diagnostic mammography. J Natl Cancer Inst 99:1854–1863. https://doi.org/10.1093/JNCI/DJM238
3. Alzubaidi L, Zhang J, Humaidi AJ et al (2021) Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. J Big Data 8:53. https://doi.org/10.1186/s40537-021-00444-8
4. Kumar R (2023) Memory recurrent elman neural network-based identification of time-delayed nonlinear dynamical system. IEEE Trans Syst Man Cybern Syst 53:753–762. https://doi.org/10.1109/TSMC.2022.3186610
5. Sherstinsky A (2020) Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. Phys D Nonlinear Phenom 404:132306
6. Nasser M, Yusof UK (2023) Deep learning based methods for breast cancer diagnosis: a systematic review and future direction. Diagnostics 13:161. https://doi.org/10.3390/DIAGNOSTICS13010161
7. Aly GH, Marey M, El-Sayed SA, Tolba MF (2021) YOLO based breast masses detection and classification in full-field digital mammograms. Comput Methods Programs Biomed 200:105823. https://doi.org/10.1016/J.CMPB.2020.105823
8. Sensakovic WF, Carnahan MB, Czaplicki CD et al (2021) Contrast-enhanced mammography: how does it work? Radiographics 41:829–839. https://doi.org/10.1148/RG.2021200167/ASSET/IMAGES/LARGE/RG.2021200167.TBL2.JPEG
9. Wei J, Hadjiiski LM, Sahiner B et al (2007) Computer-aided detection systems for breast masses: comparison of performances on full-field digital mammograms and digitized screen-film mammograms. Acad Radiol 14:659–669. https://doi.org/10.1016/J.ACRA.2007.02.017
10. Hassan NM, Hamad S, Mahar K (2022) Mammogram breast cancer CAD systems for mass detection and classification: a review. Multimed Tools Appl 81:20043–20075. https://doi.org/10.1007/S11042-022-12332-1/FIGURES/5
11. Raghu M, Unterthiner T, Kornblith S et al (2021) Do vision transformers see like convolutional neural networks? Neural Inf Process Syst 34:12116–12128
12. He K, Gan C, Li Z et al (2022) Transformers in medical image analysis: a review. Intell Med. https://doi.org/10.1016/J.IMED.2022.07.002
13. Gheflati B, Rivaz H (2022) Vision transformers for classification of breast ultrasound images. In: Proceedings of the annual international conference of the IEEE engineering in medicine and biology society, EMBS 2022-July, pp 480–483. https://doi.org/10.1109/EMBC48229.2022.9871809
14. Shamshad F, Khan S, Zamir SW et al (2023) Transformers in medical imaging: a survey. Med Image Anal 88:102802. https://doi.org/10.1016/j.media.2023.102802
15. Khaled R, Helal M, Alfarghaly O et al (2022) Categorized contrast enhanced mammography dataset for diagnostic and artificial intelligence research. Sci Data 9:122. https://doi.org/10.1038/S41597-022-01238-0
16. Suhail Z, Denton ERE, Zwiggelaar R (2018) Classification of micro-calcification in mammograms using scalable linear Fisher discriminant analysis. Med Biol Eng Comput 56:1475–1485. https://doi.org/10.1007/S11517-017-1774-Z/TABLES/2
17. Punitha S, Amuthan A, Joseph KS (2018) Benign and malignant breast cancer segmentation using optimized region growing technique. Future Comput Inform J 3:348–358. https://doi.org/10.1016/J.FCIJ.2018.10.005
18. Mughal B, Sharif M, Muhammad N (2017) Bi-model processing for early detection of breast tumor in CAD system. Eur Phys J Plus 132:1–14. https://doi.org/10.1140/EPJP/I2017-11523-8
19. Rouhi R, Jafari M, Kasaei S, Keshavarzian P (2015) Benign and malignant breast tumors classification based on region growing and CNN segmentation. Expert Syst Appl 42:990–1002. https://doi.org/10.1016/J.ESWA.2014.09.020
20. Dong M, Lu X, Ma Y et al (2015) An efficient approach for automated mass segmentation and classification in mammograms. J Digit Imaging 28:613–625. https://doi.org/10.1007/S10278-015-9778-4
21. Montenegro L, Abreu M, Fred A, Machado JM (2022) Human-assisted vs. deep learning feature extraction: an evaluation of ECG features extraction methods for arrhythmia classification using machine learning. Appl Sci (Switzerland) 12:7404. https://doi.org/10.3390/app12157404
22. Dara S, Tumma P (2018) Feature extraction by using deep learning: a survey. In: Proceedings of the 2nd international conference on electronics, communication and aerospace technology, ICECA 2018, pp 1795–1801. https://doi.org/10.1109/ICECA.2018.8474912
23. LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521:436–444. https://doi.org/10.1038/nature14539
24. Bengio Y, Courville A, Vincent P (2012) Representation learning: a review and new perspectives. IEEE Trans Pattern Anal Mach Intell 35(8):1798–1828
25. Ragab DA, Sharkas M, Marshall S, Ren J (2019) Breast cancer detection using deep convolutional neural networks and support vector machines. PeerJ 7:e6201. https://doi.org/10.7717/PEERJ.6201
26. Sannasi Chakravarthy SR, Bharanidharan N, Rajaguru H (2022) Multi-deep CNN based experimentations for early diagnosis of breast cancer. IETE J Res. https://doi.org/10.1080/03772063.2022.2028584
27. Redmon J, Divvala S, Girshick R, Farhadi A (2016) You only look once: unified, real-time object detection
28. Al-Masni MA, Al-Antari MA, Park JM et al (2017) Detection and classification of the breast abnormalities in digital mammograms via regional convolutional neural network. In: Annual international conference IEEE engineering medicine and biology society, pp 1230–1233. https://doi.org/10.1109/EMBC.2017.8037053
29. Al-antari MA, Al-masni MA, Kim TS (2020) Deep learning computer-aided diagnosis for breast lesion in digital

mammogram. Adv Exp Med Biol 1213:59–72. https://doi.org/10.1007/978-3-030-33128-3_4
30. Baccouche A, Garcia-Zapirain B, Olea CC, Elmaghraby AS (2021) Breast lesions detection and classification via YOLO-based fusion models. Comput Mater Contin 69:1407–1425. https://doi.org/10.32604/CMC.2021.018461
31. Ren S, He K, Girshick R, Sun J (2017) Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans Pattern Anal Mach Intell 39:1137–1149. https://doi.org/10.1109/TPAMI.2016.2577031
32. Ribli D, Horváth A, Unger Z et al (2018) Detecting and classifying lesions in mammograms with deep learning. Sci Rep 8:1–7. https://doi.org/10.1038/s41598-018-22437-z
33. Agarwal R, Díaz O, Yap MH et al (2020) Deep learning for mass detection in full field digital mammograms. Comput Biol Med 121:103774. https://doi.org/10.1016/J.COMPBIOMED.2020.103774
34. Cao H, Pu S, Tan W, Tong J (2021) Breast mass detection in digital mammography based on anchor-free architecture. Comput Methods Programs Biomed 205:106033. https://doi.org/10.1016/J.CMPB.2021.106033
35. Zhu C, He Y, Savvides M (2019) Feature selective anchor-free module for single-shot object detection
36. Shen R, Yao J, Yan K et al (2020) Unsupervised domain adaptation with adversarial learning for mass detection in mammogram. Neurocomputing 393:27–37. https://doi.org/10.1016/j.neucom.2020.01.099
37. Mohamed SAS, Moftah SG, Chalabi NAEM, Salem MAAW (2021) Added value of contrast-enhanced spectral mammography in symptomatic patients with dense breasts. Egypt J Radiol Nuclear Med 52:1–10. https://doi.org/10.1186/S43055-020-00372-2/FIGURES/4
38. Song J, Zheng Y, Wang J et al (2022) Multi-feature deep information bottleneck network for breast cancer classification in contrast enhanced spectral mammography. Pattern Recognit 131:108858. https://doi.org/10.1016/J.PATCOG.2022.108858
39. Danala G, Patel B, Aghaei F et al (2018) Classification of breast masses using a computer-aided diagnosis scheme of contrast enhanced digital mammograms. Ann Biomed Eng 46:1419–1431. https://doi.org/10.1007/s10439-018-2044-4
40. Gao F, Wu T, Li J et al (2018) SD-CNN: a shallow-deep CNN for improved breast cancer diagnosis. Comput Med Imaging Graph 70:53–62. https://doi.org/10.1016/J.COMPMEDIMAG.2018.09.004
41. Perek S, Kiryati N, Zimmerman-Moreno G et al (2019) Classification of contrast-enhanced spectral mammography (CESM) images. Int J Comput Assist Radiol Surg 14:249–257. https://doi.org/10.1007/s11548-018-1876-6
42. Berbar MA (2018) Hybrid methods for feature extraction for breast masses classification. Egypt Inform J 19:63–73. https://doi.org/10.1016/j.eij.2017.08.001
43. Muduli D, Dash R, Majhi B (2020) Automated breast cancer detection in digital mammograms: a moth flame optimization based ELM approach. Biomed Signal Process Control 59:10192. https://doi.org/10.1016/j.bspc.2020.101912
44. D'Haeyer JPF (1989) Gaussian filtering of images: a regularization approach. Signal Process 18:169–181. https://doi.org/10.1016/0165-1684(89)90048-0
45. Otsu N (1979) Threshold selection method from gray-level histograms. IEEE Trans Syst Man Cybern SMC 9:62–66. https://doi.org/10.1109/TSMC.1979.4310076
46. Pizer SM, Amburn EP, Austin JD et al (1987) Adaptive histogram equalization and its variations. Comput Vis Graph Image Process 39:355–368. https://doi.org/10.1016/S0734-189X(87)80186-X
47. Redmon J, Farhadi A (2016) YOLO9000: better, faster, stronger. In: Proceedings–30th IEEE conference on computer vision and pattern recognition, CVPR 2017-January, pp 6517–6525. https://doi.org/10.1109/CVPR.2017.690
48. Redmon J, Farhadi A (2018) YOLOv3: an incremental improvement
49. Bochkovskiy A, Wang C-Y, Liao H-YM (2020) YOLOv4: optimal speed and accuracy of object detection
50. Wang C-Y, Bochkovskiy A, Liao H-YM (2022) YOLOv7: trainable bag-of-freebies sets new state-of-the-art for real-time object detectors
51. Reis D, Kupec J, Hong J, Daoudi A (2023) Real-time flying object detection with YOLOv8
52. Nepal U, Eslamiat H (2022) Comparing YOLOv3, YOLOv4 and YOLOv5 for autonomous landing spot detection in faulty UAVs. Sensors 22:464. https://doi.org/10.3390/S22020464
53. Ismail A, Mehri M, Sahbani A et al (2021) Performance benchmarking of YOLO architectures for vehicle license plate detection from real-time videos captured by a mobile robot. Sorbonne University, Paris
54. Zhou S, Cai K, Feng Y et al (2023) An accurate detection model of Takifugu rubripes using an improved YOLO-V7 network. J Mar Sci Eng 11:1051. https://doi.org/10.3390/jmse11051051
55. Tian Z, Shen C, Chen H, He T (2019) FCOS: fully convolutional one-stage object detection
56. Lou H, Duan X, Guo J et al (2023) DC-YOLOv8: small-size object detection algorithm based on camera sensor. Electronics (Switzerland) 12:2323. https://doi.org/10.3390/electronics12102323
57. Dosovitskiy A, Beyer L, Kolesnikov A et al (2022) An image is worth 16 × 16 words: transformers for image recognition at scale
58. Touvron H, Cord M, Douze M et al (2021) Training data-efficient image transformers & distillation through attention
59. Liu Z, Lin Y, Cao Y et al (2021) Swin transformer: hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE international conference on computer vision 9992–10002. https://doi.org/10.48550/arxiv.2103.14030
60. Liu Z, Mao H, Wu C-Y et al (2022) A ConvNet for the 2020s
61. Hassanien MA, Singh VK, Puig D, Abdel-Nasser M (2022) Predicting breast tumor malignancy using deep ConvNeXt radiomics and quality-based score pooling in ultrasound sequences. Diagnostics (Basel) 12:1053. https://doi.org/10.3390/DIAGNOSTICS12051053
62. van Tulder G, Tong Y, Marchiori E (2021) Multi-view analysis of unregistered medical images using cross-view transformers. Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics) 12903 LNCS. Springer, Cham, pp 104–113
63. Fan X, Feng X, Dong Y, Hou H (2022) COVID-19 CT image recognition algorithm based on transformer and CNN. Displays 72:102150. https://doi.org/10.1016/J.DISPLA.2022.102150
64. Al-Rahhal MM, Bazi Y, Jomaa RM et al (2022) COVID-19 detection in CT/X-ray imagery using vision transformers. J Pers Med 12:310. https://doi.org/10.3390/JPM1202031
65. Hassan NM, Hamad S, Mahar K (2022) A deep learning model for mammography mass detection using mosaic and reconstructed multichannel images. Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics). Springer, Cham, pp 544–559
66. Al-antari MA, Al-masni MA, Choi MT et al (2018) A fully integrated computer-aided diagnosis system for digital X-ray mammograms via deep learning detection, segmentation, and classification. Int J Med Inform 117:44–54. https://doi.org/10.1016/J.IJMEDINF.2018.06.003
67. Platania R, Zhang J, Shams S et al (2017) Automated breast cancer diagnosis using deep learning and region of interest

detection (BC-DROID). In: ACM-BCB 2017–proceedings of the 8th ACM international conference on bioinformatics, computational biology, and health informatics, pp 536–543. https://doi.org/10.1145/3107411.3107484

68. Al-antari MA, Han SM, Kim TS (2020) Evaluation of deep learning detection and classification towards computer-aided diagnosis of breast lesions in digital X-ray mammograms. Comput Methods Programs Biomed 196:105584. https://doi.org/10.1016/J.CMPB.2020.105584