Turkish Journal of Electrical Engineering and Computer Sciences

Volume 31 | Number 7 Article 10

11-30-2023

A comparative study of YOLO models and a transformer-based YOLOv5 model for mass detection in mammograms

DAMLA COŞKUN

DERVİŞ KARABOĞA

ALPER BAŞTÜRK

BAHRİYE AKAY

ÖZKAN UFUK NALBANTOĞLU

See next page for additional authors

A comparative study of YOLO models and a transformer-based YOLOv5 model for mass detection in mammograms

Authors
DAMLA COŞKUN, DERVİŞ KARABOĞA, ALPER BAŞTÜRK, BAHRİYE AKAY, ÖZKAN UFUK NALBANTOĞLU, SERAP DOĞAN, İSHAK PAÇAL, and MERYEM ALTIN KARAGÖZ

Turkish Journal of Electrical Engineering & Computer Sciences

http://journals.tubitak.gov.tr/elektrik/

Research Article

Turk J Elec Eng & Comp Sci
(2023) 31: 1294 – 1313

doi:10.55730/1300-0632.4048

A comparative study of YOLO models and a transformer-based YOLOv5 model for mass detection in mammograms

Damla COŞKUN[1,2,*], Derviş KARABOĞA[1,2], Alper BAŞTÜRK[1,2], Bahriye AKAY[1,2]
Özkan Ufuk NALBANTOĞLU[1,2], Serap DOĞAN[3], İshak PAÇAL[2,4], Meryem Altin KARAGÖZ[1,2,5]
[1]Department of Computer Engineering, Faculty of Engineering, Erciyes University, Kayseri, Turkiye
[2]Artificial Intelligence and Big Data Application and Research Center, Erciyes University, Kayseri, Turkiye
[3]Department of Radiology, Erciyes University Medical Faculty, Kayseri, Turkiye
[4]Computer Engineering Department, Engineering Faculty, Iğdır University, Iğdır, Turkiye
[5]Department of Computer Engineering, Sivas Cumhuriyet University, Sivas, Turkiye

Received: 31.03.2023 • **Accepted/Published Online:** 12.10.2023 • **Final Version:** 30.11.2023

Abstract: Breast cancer is a prevalent form of cancer across the globe, and if it is not diagnosed at an early stage it can be life-threatening. In order to aid in its diagnosis, detection, and classification, computer-aided detection (CAD) systems are employed. You Only Look Once (YOLO)-based CAD algorithms have become very popular owing to their highly accurate results for object detection tasks in recent years. Therefore, the most popular YOLO models are implemented to compare the performance in mass detection with various experiments on the INbreast dataset. In addition, a YOLO model with an integrated Swin Transformer in its backbone is proposed for mass detection in mammography images within the study. The performance of YOLOv5 models and a transformer-based YOLO model is compared to that of each other and YOLOv3 and YOLOv4 models using images with different sizes on the INbreast dataset. The best results are obtained by the transformer-based YOLO model of YOLOv5 for 832×832 image size. In another experiment, we compared the default anchors against the anchors provided by the YOLOv5 autoanchor function before training and saw that the anchors generated by the YOLOv5 autoanchor increased the success rates. Furthermore, various experiments were conducted to observe how data augmentation affects performance. Although a small amount of data was used in the study, high performance was obtained by YOLO algorithms, which are promising tools for cancer detection.

Key words: Breast cancer, deep learning, YOLO, computer-aided detection, transformer-based YOLO, data augmentation

1. Introduction

Breast cancer ranks among the most prevalent types of cancer worldwide. Globally, breast cancer was diagnosed in 2.3 million women and 684,996 deaths occurred in 2020 [1]. In 2023, it is estimated by the American Cancer Society that there will be approximately 300,590 (297,790 cases in women and 2800 cases in men) new breast cancer cases and 43,700 deaths from breast cancer among the US population [2]. Although breast cancer sometimes occurs after the onset of symptoms, many breast cancer cases do not show symptoms. Early detection of breast cancer is crucial and regular screening plays a vital role in achieving it [3]. Various medical imaging techniques are employed to assist in the detection of breast cancer, such as mammography (MG) [4], magnetic resonance imaging (MRI), positron emission tomography (PET), computed tomography (CT) [5],

*Correspondence: damlacoskun@erciyes.edu.tr

breast thermography (BT) [6], histopathology (HP) [7], and ultrasound (US) [8, 9]. Among them, MG with a low-dose X-ray is the most prevalent [3]. A mammogram scan takes two images for each breast from two views: craniocaudal (CC) and medial-lateral oblique (MLO) [10]. The cancer level is evaluated by radiologists utilizing the Breast Imaging Reporting & Data System (BI-RADS). This system classifies results numbered from 0 to 6 [11]. However, the large number of evaluations performed by radiologists and the complexity of MG can lead to misdiagnosis [12].

Computer-aided detection (CAD) systems have helped clinicians in diagnosis, detection, and classifying results. Nevertheless, the biggest challenge when using CAD to detect MG anomalies is the high false-positive rate. False-positive results cause anxiety in patients, extra exposure to radiation, redundant biopsies, high medical costs, high call-back rates, and further evaluations. Thus, the current CAD systems should be improved for more accurate and robust detection by avoiding these limitations.

Recent advances in computer technology, machine learning, and imaging technologies and the widespread adoption of digital MG imaging have contributed to solving the complex problem of early detection of breast cancer by deep learning (DL) techniques. In addition, DL methods, especially convolutional neural networks (CNNs), have received significant interest in CAD for digital MG imaging due to their ability to overcome the limitations of CAD models. CNNs lead to better detection performance than CAD models and aid radiologists in identifying and diagnosing potentially malign lesions. CNN models aim to enhance the capability of radiologists to identify breast cancer in its early stages, even in small lesions, and to notify them for additional analysis [12]. Technologies for detecting objects using DL methods are at the core of the state-of-the-art CAD systems for the prospective generation. Acting as decision-making assistants, object detection algorithms are expected to mark/emphasize the lesions and abnormalities in MG images. Since false discoveries and misses pose a great risk in diagnostic processes, it is crucial to maximize the detection and classification performances for MG images. The You Only Look Once (YOLO) class of object detection algorithms has become a very popular approach not only in general object detection tasks but also in the field of mass detection in mammograms, due to their accuracy and flexibility [13]. Different versions and settings of YOLO algorithms might have substantial variations in detection performance; hence, it is critical to assess the performance of these models in mass detection.

In the present study, we used the INbreast dataset for training and testing, and we did not do any preprocessing except for converting images from DICOM to JPEG format. Then we utilized YOLOv3, YOLOv4, and YOLOv5 models and the different variations of YOLOv5 models (YOLOv5s, YOLOv5m, YOLOv5l, and YOLOv5x) and compared the performance of these models. Since Aly et al. [13] indicated that YOLOv3 outperformed YOLOv1 and YOLOv2 models, no separate comparisons were conducted with YOLOv1 and YOLOv2. The backbone of YOLOv5 has been modified by integrating the Swin Transformer and has also been employed for the first time in the detection of masses from MG images. We also performed another experiment to observe the effect of anchor production using the K-means clustering algorithm, according to the trained data. In our last experiment, we observed the performance effect of data augmentation. The present article is structured as follows: Section 2 presents an overview of related research on the detection of breast cancer; Section 3 details the methodology used in the present study; Section 4 outlines the experiments and presents the results obtained; and, finally, the conclusions are given in Section 5.

2. Related work

Various neural network models have been investigated by researchers in recent years for detecting and classifying malign and benign lesions associated with breast cancer. Studies utilizing different imaging methods such as HP and US are also present in the literature. However, in the present study, the focus of the literature review is on MG since mammogram data are used.

Ribli et al. [14] created a CAD system utilizing the Faster R-CNN model for the detection and classification of breast lesions within the INbreast dataset. The lesions were categorized as either malign or benign. Peng et al. [15] proposed a method for mass detection by integrating the Faster R-CNN architecture with a multiscale-feature pyramid network.

Al-Masni et al. [16] suggested a CAD system that utilizes YOLO for the automatic detection and classification of masses using mammograms. The system was trained and tested on images selected from the Digital Database for Screening Mammography (DDSM) dataset, with equal representation of benign and malign classes. Al-antari et al. employed YOLO for mass detection in mammograms, as referenced in [17, 18]. In their 2018 study [17], they presented a CAD system. YOLO was utilized for detecting and localizing masses. Subsequently, a full resolution convolutional network (FRCN) was employed for segmenting the detected masses, followed by the classification of segmented masses into benign and malign categories using a pretrained CNN based on the AlexNet architecture. In a later work [19], Al-antari et al. enhanced the model introduced in [17] with refinements in the classification and segmentation phases. In another study [20], they conducted similar research. Firstly, a YOLO detector is employed and assessed for the detection of breast lesions in complete mammograms. Subsequently, regular feedforward CNN, ResNet-50, and InceptionResNet-V2 are adapted and evaluated for the task of classifying breast lesions. The performance of the proposed system was tested on two databases: DDSM and INbreast.

Djebbar et al. [21] introduce an innovative CAD system based on the YOLOv3 architecture. This YOLO-based CAD system is capable of simultaneously conducting detection and classification within a unified framework. Despite its slightly larger size compared to previous models, this system exhibits improved accuracy. Aly et al. [13] suggested an end-to-end YOLO-based model for detecting masses and categorizing benign and malign lesions. YOLOv1, YOLOv2, and YOLOv3 were employed to automatically detect masses and compare their performance on the INbreast dataset. Furthermore, they reported that utilizing K-means to create anchors improved the detection performance of their model. Finally, the classification performance was compared between YOLO, ResNet, and Inception networks on the INbreast dataset for various resolutions. Additionally, YOLO-v3 integrated with K-means clustering for anchors and data augmentation techniques improved the performance even on challenging samples.

Baccouche et al. [22] present an end-to-end system that utilizes the YOLO model to achieve simultaneous detection and classification of breast lesions in mammograms. The initial stage of the proposed system involves preprocessing the images, after which it proceeds to detect anomalous regions suggestive of breast lesions and subsequently categorizes them based on their pathology, classifying them as either masses or calcifications. The performance of this model was evaluated on three datasets: CBIS-DDSM, INbreast, and a privately collected dataset containing 487 mammograms. In another study, Baccouche et al. [23] introduced a YOLO-based fusion model for detecting and classifying breast lesions in current mammograms. They expanded this model's application by retrospectively implementing it on synthetic mammograms. These synthetic mammograms were generated through image-to-image translation models like CycleGAN and Pix2Pix, aiming for early cancer prediction based on previous mammograms. The evaluation results demonstrated the effectiveness of the proposed

methodology in detecting and classifying breast lesions on current mammograms.

Hamed et al. [24] introduced a CAD system based on YOLOv4, employing a 2-path detection approach for masses in both full and cropped mammograms. The masses were subsequently classified as benign or malign. Zhao et al. [25] suggested a YOLOv3-based CAD system for mammogram detection consisting of three main steps: preprocessing, YOLOv3-based DL model, and model evaluation. They combined detecting the position of masses and classification tasks for microcalcification, mass, benign, malign, and other classes. Their proposed model has three models for training on the CBIS-DDSM dataset: general used all images, mass used only mass samples, and microcalcification used only microcalcification samples. The YOLOv3-based CAD system achieved high performance and ensured a robust model for lesion detection and various classification tasks. Kolchev et al. [26] compare the performance of the YOLOv4-based CNN model and Nested Contours Algorithm (NCA) for lesion detection. They used the INbreast dataset and augmented the data with the mosaic method. In the training of YOLOv4 1080 images were used; besides 200 proven images were employed in testing models with the same number for each class: breast cancer and absence. Although the NCA performed better than YOLOv4, YOLOv4 has a lower false-positives rates. Therefore, they suggest a hybrid model with YOLOv4 and NCA to get more accurate and robust results.

Yasir et al. [27] employed cutting-edge object detection models, specifically YOLOv5 and Mask RCNN. The YOLOv5 model was utilized for detecting and categorizing masses as either benign or malign. On the other hand, Mask RCNN was applied to identify tumor edges that extend into the breast parenchyma, along with determining tumor sizes. The model was trained on the INbreast dataset using a combination of YOLOv5 and Mask RCNN. The performance of this proposed model was assessed by comparing it to the original version of YOLOv5. Hassan et al. [28] present a model based on YOLOv4 for mass detection and classification. Their study explores the performance of various augmentation techniques, including the newly introduced mosaic technique, within the YOLOv4 framework. Additionally, they enhance the detection accuracy by transforming the images into a multichannel format during the preprocessing phase, resulting in an approximate improvement of nearly 10%. The evaluation of the model involves experimenting with different combinations of augmentation techniques. The experiments are conducted using the INbreast and MIAS datasets. For the INbreast dataset, the results demonstrate that the combination of mosaic with YOLOv4 yields the best performance.

Zhang et al. [29] have proposed an anchor-free-YOLOv3 network for mass detection in mammograms. To mitigate the issues caused by using Mean Squared Error (MSE) as the box regression loss, a Generalized Intersection over Union (GIoU) loss has been adopted for training bounding box regression. For objectness prediction loss, "focal loss" has been used, which can prevent network deterioration caused by a large number of easy negatives. Additionally, a new feature fusion method called the summation method has been designed and employed in the top-down pathway. Comparative experiments were conducted on two databases, the publicly available INbreast dataset and the private TXMD dataset. Su et al. [30] introduced a YOLO-LOGO model to detect and segment masses by combining YOLO and LOGO (local-global) networks. The first step of the YOLO-LOGO model is implementing the YOLOV5L6 model for detecting mass locations and cropping masses from MG images. Afterward, the LOGO training method was adapted to separately train global and local transformer branches on both whole and cropped images. Then the segmentation decision was made by merging the two branches. Thus, LOGO enhanced the balance of the model performance for training and segmentation. The CBIS-DDSM and INbreast datasets were used to test the effectiveness of the proposed YOLO-LOGO model.

Kamran et al. [32] introduced novel U-net-shaped transformer-based architecture (SWIN-SFTNet) based on transformers that surpasses the performance of existing architectures in the segmentation of micromasses in breast MG. The CBIS-DDSM and INbreast datasets were utilized in the study. In 2022, Chen et al. [33] introduced a Multiview Vision Transformer architecture that can independently capture patch relationships among four mammograms obtained from two different views (CC/MLO) of breasts from both sides (right/left), using both local and global transformer blocks. The transformer-based model they proposed was tested on a private dataset. Betancourt Tarifa et al. [31] have developed and conducted experiments involving transformer-based models for the purpose of mass detection in mammograms. They utilize the Swin Transformer as a backbone multiscale feature extractor. The OMI-DB dataset was utilized in the study. Lu et al., Yang et al., and Vasanthi et al.[34–36] proposed a novel detection model that integrates the Swin Transformer and YOLO. In [34], Swin Transformer-YOLOv5 was proposed for the real-time detection of wine grape bunches. In [35], the model was proposed for surface defect detection. That study focuses on enhancing the capability of capturing long-range semantic information through the introduction of the Swin Transformer Block in the design of the C3STR module, built upon the foundation of YOLOv5. The utilization of the lightweight attention module, Coordinate Attention (CA), in the CAHead structure further contributes to the fusion of feature information. The combined benefits of the ST and CA module result in improved detection ability, particularly for small objects, leading to an overall enhancement in detection performance. In order to address the visual challenge of detecting small-scale objects, a YOLOv5X-transformer model was introduced [36]. In this architecture, the Multihead-Self-Attention module is employed to extract comprehensive details from the feature maps. Subsequently, the obtained feature maps are aggregated across five distinct scales through Spatial Pyramid Pooling-Faster (SPPF), enhancing the feature map's overall quality. For preserving spatial information and accurate pixel localization, the Path-Aggregated Network (PANet) is employed as the neck model.
The details of the summarized studies in this section are also presented separately in Table 1.

Table 1. Related work.

References	Year	Dataset	Methods	Performance
Al-Masni et al. [16]	2017	DDSM	YOLO	Detection accuracy: 96.33%, Classification accuracy: 85.52%
Ribli et al. [14]	2018	DDSM, INbreast, Private dataset	Faster R-CNN	AUC = 0.95 Malignant lesion detection: 90%, with only 0.3 FP marks per image in the INbreast dataset
Al-antari et al. [18]	2018	DDSM	FCNN, YOLO	Detection accuracy: 99.7% Classification accuracy: 97%
Al-antari et al. [17]	2018	INbreast	YOLO, FrCN, DCNN	Detection accuracy: 98.96% F1-Score: 99.24% Mass segmentation: Accuracy: 92.97% F1-Score: 92.69%
Djebbar et al. [21]	2019	DDSM	YOLO	Detection accuracy: 97%, Classification accuracy: 96.7%
Peng et al. [15]	2020	CBIS-DDSM, INbreast	Faster R-CNN, ResNeXt-101	CBIS-DDSM: TPR of 0.93 at 2.28 FP per image, INbreast: TPR of 0.95 at 0.3829 FP per image

Al-Antari et al. [19]	2020	INbreast	YOLO, FrCN, DCNN, Regular Feed-forward CNN, ResNet-50, InceptionResNetV2	Detection accuracy: 92.97%, Segmentation accuracy: 92.97%, Classification accuracy: CNN: 88.74%, ResNet-50: 92.56%, and InceptionResNet-V2: 95.32%
Al-Antari et al. [20]	2020	DDSM, INbreast	YOLO, feed-forward CNN, ResNet-50, InceptionResNetV2	Detection accuracy: DDSM: 99.17%, INbreast: 97.27%, F1-Score: DDSM: 99.28%, INbreast: 98.02%, Classification accuracy: DDSM: CNN: 94.50%, ResNet-50: 95.83%, InceptionResNet-V2: 97.50%, INbreast: CNN: 88.74%, ResNet-50: 92.55%, InceptionResNet-V2: 95.32%
Aly et al. [13]	2020	INbreast	YOLO, ResNet, InceptionNet	Detection accuracy: 89.4%, Classification: YOLO: Average precision for benign: 94.2%, for malign: 84.6%, ResNet: 91.0%, InceptionV3: 95.5%
Baccouche et al. [22]	2021	CBIS-DDSM, INbreast, Private dataset	YOLO	Mass detection accuracy: CBIS-DDSM: 95.7%, INbreast: 98.1%, Private dataset:98%
Hamed et al. [24]	2021	INbreast	YOLO, ResNet, VGG, Inception	Mass detection accuracy: 97.86%, Classification accuracy: 91%
Zhao et al. [25]	2022	CBIS-DDSM	YOLO	Mass detection accuracy: 97.77%, Mass classification accuracy: 98.12%
Hassan et al. [28]	2022	INbreast, MIAS	YOLO	Detection: INbreast: mAP: 99.5%, Precision: 98%, Recall: 94%, MIAS: mAP: 95.28%, Precision: 93%,Recall: 90%
Baccouche et al. [23]	2022	Private dataset	YOLO, CycleGAN	For mass: Accuracy: 94%, Precision: 94%, Recall: 94%, Sensitivity: 95%, AUC: 95
Kolchev et al. [26]	2022	INbreast, Private dataset	YOLO, Nested Contours Algorithm (NCA)	YOLOv4: Precision: 85%, Recall: 60%, F1-Score: 70%, NCA: Precision: 59%, Recall: 93%, F1-Score: 72%
Yassir et al. [27]	2022	INbreast	YOLO, Mask RCNN	Sensitivity: 95%, Specificity: 97%, Precision: 91.08%, mAP: 95.20% Accuracy: 98%, MCC:92.02%
Zhang et al. [29]	2022	INbreast, TXMD (Private dataset)	YOLO	INbreast: TPR: 0.95, FP: 1.7, TXMD: TPR: 0.94, FP: 5.75
Su et al. [30]	2022	CBIS-DDSM, INbreast	YOLO, Local-Global (LOGO) Networks	Detection mAP: CBIS-DDSM: 65, INbreast: 61.4 Segmentation F1: CBIS-DDSM: 74.52, INbreast: 69.37
Kamran et al. [32]	2022	CBIS-DDSM, INbreast	Swin-SFTNet	Segmentation Dice-score: CBIS-DDSM: 3.10%, INbreast: 3.81%, CBIS pretrained model tested on INbreast: 3.13%
Chen et al. [33]	2022	Private dataset	Multiview Vision Transformers	AUC = 0.818

Lu et al. [34]	2022	Wine Grape Dataset	Faster R-CNN, YOLO, Swin-T-YOLOv5	Faster R-CNN: mAP: 53.54%, F1-Score: 0.67, YOLOv3: mAP: 78.93%, F1-Score: 0.72, YOLOv4: mAP: 83.45%, F1-Score: 0.76, YOLOv5: mAP: 93.64%, F1-Score: 0.83, Swin-T-YOLOv5: mAP: 97.19%, F1-Score: 0.89
Yang et al. [35]	2023	NEU, DAGM2007, RSDDs	ST-CA YOLOv5	AP: on NEU: 43.3, on DAGM2007: 65.9, on RSDDs: 50.6
Vasanthi et al. [36]	2023	PASCAL, VOC	YOLO	PASCAL: mAP: 87.7%, Precision: 85.2%, Recall: 81.4%

3. Methodology

3.1. You Only Look Once (YOLO)

There are two approaches for detecting objects: a one-stage detector and a two-stage detector. The object detection models are built in such a way that they first extract the features of the input images through the backbone. Then the features are passed on to the object detector, which consists of the detector neck and detector head. The neck in object detection models serves as a feature aggregator, responsible for merging and blending the features generated by the backbone, in preparation for the detection step carried out by the head. The difference between these approaches is that the detection responsibility, including the localization and classification of each bounding box, is on the head. A two-stage detector performs these two tasks independently and merges the results (sparse detection), while a single-stage detector applies them simultaneously (dense detection). YOLO is a one-stage detector [37].

The goal of YOLO is to create a single model for all stages of a neural network. YOLO calculates all the features of the image and makes predictions for all objects simultaneously, in place of repeating the classification of disparate areas of the image. The basic idea behind YOLO is to apply S × S cells to the image. When an object is present in an image and the center of that object falls inside a particular grid cell, then it is the responsibility of that grid cell to detect that object [37]. Each grid cell is also responsible for class probabilities and predicting the bounding boxes (B) with the confidence scores for those boxes. Every bounding box is composed of 5 values of (x, y, w, h) (at the center of the bounding box, width, height) and confidence score. The confidence score, given in Equation 1, indicates whether there are objects in this box.

$$confidencescore = p(Object) \times IoU_{pred}^{truth} \tag{1}$$

$$IoU_B^A = \frac{A \cap B}{A \cup B} \; ; \; IoU_B^A \, \epsilon \, [0,1] \tag{2}$$

$p(Object)$ is the probability of having an object in the cell and IoU_{pred}^{truth} is the intersection over union (IoU) of the prediction box and ground truth box, given in Equation 2. Because $p(Object)$ is in the range [0,1], if there are no objects in the cell, $p(Object)$ will be zero. In this situation, the confidence score is equal to zero. When $p(Object)$ is equal to one, the confidence score will be equal to IoU_{pred}^{truth} [38]. After the input image passes through a single neural network of multiconvolutional networks, the system creates a prediction vector

for each object in the image. The detection process of the YOLO model is shown in Figure 1. Furthermore,

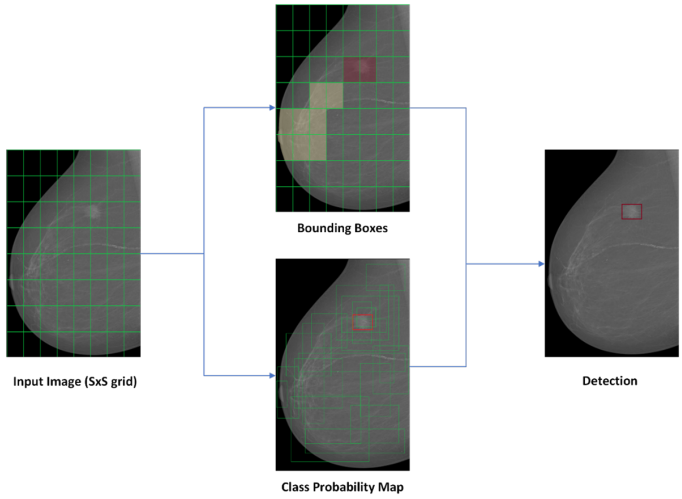

Figure 1. The detection process of the YOLO model.

YOLO adopts nonmaximum suppression (NMS) to eliminate bounding boxes that do not contain any objects or have the same object as other bounding boxes. NMS cleans all the overlapping bounding boxes with an IoU value greater than the threshold [37, 39].

YOLOv1 is the first version of YOLO and it has 24 convolutional and 2 fully connected layers [38]. The second version of YOLO, known as YOLOv2 or YOLO9000, consists of a network called Darknet-19 that includes 19 convolutional layers and 5 max-pooling layers [39]. Moreover, in YOLOv2, batch normalization has been added to all convolution layers. This method increases the performance of the network while reducing the training time [37, 39]. In YOLOv3, Darknet-19's feature extraction backbone, which has difficulty in detecting small objects, has been replaced with Darknet-53 to overcome this issue. The network is constructed with a bottleneck structure (1×1 and 3×3 convolution layers) within each residual block and a skip connection. YOLOv3 adds prediction layers aside instead of placing them in the last layers as before. Three different scale detectors use attributes of the last 3 residual blocks. Feature Pyramid Network (FPN) architecture is used in the neck of YOLOv3 [37, 40]. In April 2020, Bochkovskiy et al. published YOLOv4. YOLOv4 uses a novel backbone, CSPDarknet53 (CSP stands for Cross Stage Partial), and adds the Spatial Pyramid Pooling (SPP) block over the CSPDarknet53 and uses PANet, which is an advanced version of the FPN, instead of the FPN [41]. The fifth version of YOLO (YOLOv5) differs from other models in that it uses a CSPNet [42] as the model backbone and PANet [43] as the neck for feature aggregation. The model structure is given in Figure 2.

Through this progress, feature extraction is improved and the mAP score is significantly boosted [44]. The SPPF module used is the same as the SPP implemented in YOLOv3, but it is an optimized version that reduces floating point operations per second (FLOPs) and runs faster than SPP [45, 46]. The YOLOv5 models with different configurations and parameter sizes are YOLOv5s, YOLOv5m, YOLOv5l, and YOLOv5x. Due to its successful performance in object detection, the YOLOv5 algorithm has been primarily employed in the present study.

[1]Ultralytics, (2022). GitHub [Online]. Website https://github.com/ultralytics/yolov5/issues/6998 (Accessed 16 Feb. 2023)

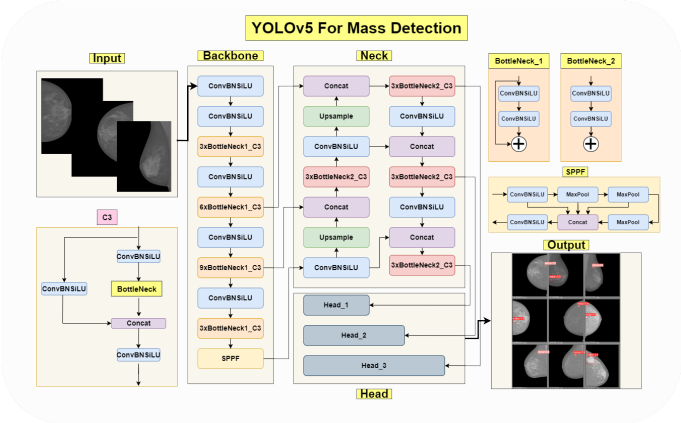

Figure 2. Overview of the structure of YOLOv5.[1]

3.2. Shifted Window (Swin) Transformer

In the realm of computer vision, Vision Transformers (ViTs) have recently emerged as a viable alternative to traditional CNNs. Notably, the Swin Transformer stands out as a versatile backbone that excels at learning attention-based hierarchical features, thereby attaining state-of-the-art performance across a wide range of vision tasks [31]. The Swin Transformer represents a hierarchical vision transformer with the ability to serve as a versatile foundational structure for various computer vision tasks [48]. Contrary to prior Vision Transformers like ViT [49], which relied on global self-attention between nonoverlapping, medium-sized image patches (e.g., 16×16 pixels) at a fixed scale, Swin adopts a distinct approach. It employs a window-based strategy alongside window-shifting at multiple scales. This design constrains self-attention calculations among small patches (4×4 pixels) within nonoverlapping windows, while still facilitating cross-window connections. Consequently, this yields linear computational complexity relative to image size and renders Swin suitable for dense vision tasks. To date, Swin and its variations have emerged as the foundational architectures underpinning cutting-edge methodologies in image classification [50], semantic segmentation [50, 51], instance segmentation [52], and object detection [50]. Considering all the aforementioned attributes, especially its aptitude for extracting hierarchical multiscale attention features leading to a top-tier performance in intricate computer vision tasks, Swin stands as the ideal cornerstone for our transformer-based mass detection framework.

3.3. Swin Transformer-based YOLOv5

Incorporating both YOLOv5 and the Swin Transformer, the two models were merged by substituting Swin Transformer encoder blocks for the final C3 layer in the underlying network of the YOLOv5. The Swin Transformer has the ability to capture distant relationships and maintain various local details [48]. While this merging process might lead to a slight decrease in YOLOv5's inference speed, it has the potential to improve detection accuracy. Hence, our suggested approach amalgamated YOLOv5s and the Swin Transformer to ensure the new architecture inherits their strengths while safeguarding global and local attributes. The structural diagram of the proposed model is presented in Figure 3.

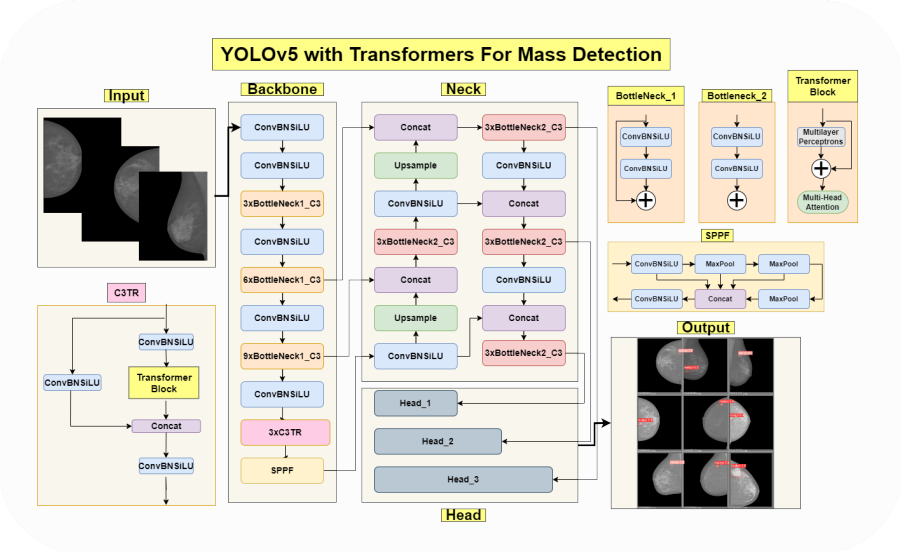

Figure 3. Structural diagram of the Swin Transformer-based YOLOv5.

4. Experiments and results

4.1. Experimental setup

The experiments in this manuscript were conducted using an Intel Core i5 processor, a single RTX 2080TI graphics card, and 16 GB DDR4 RAM hardware with Ubuntu 20.04 operating system. The IoU, number of classes, the confidence probability threshold, and epochs number used in the experiments are set to 0.5, 2 (benign and malign class), 0.25, and 6000, respectively. The value of the learning rate was set to 0.001 and adjusted with the scale of 0.1 at the 4000th and 5000th iterations.

4.2. Evaluation metrics

Various assessment measures are employed in studies related to detecting and classifying breast cancer including True Positives (TP), True Negatives (TN), False Positives (FP), False Negatives (FN), Precision, Recall, Average Precision (AP), mean Average Precision (mAP), and F1-Score [53].

4.3. Dataset and preprocessing

The INbreast dataset consists of digital mammograms obtained using the Siemens MammoNovation mammography system. The format of MG images is DICOM. Four mammograms were collected from 90 patients with both breasts affected (totaling 360 mammograms), while only two mammograms were gathered from each of the other 25 patients (totaling 50 mammograms) who had mastectomy in the dataset. Thus, a total of 410 mammograms with both MLO and CC views including normal, benign, and malign cases were gathered from 115 patients. There are 107 patients in total where breast mass is present in both MLO and CC views of the mammograms. There are some mammograms that have multiple tumors in different positions. Therefore, there are a total of 116 masses from 50 patients (41 benign and 75 malign masses from 18 and 32 patients, respectively) among 107 images. The dimensions of the mammograms are either 3328×4084 or 2560×3328 pixels and they have 14-bit contrast resolution [54, 55]. A total of 107 MG images with masses in both views (CC and MLO) in the dataset were selected to evaluate the models in the present study. These images' dataset

is used by dividing to contain 80% of its size for the training dataset and 20% for the test dataset. Despite the limited amount of data in the dataset, in order to ensure a fair comparison of results, the same amount of data as the reference article was used, and the data were partitioned for training and testing in the same manner as in the reference article [13]. The class distribution of images in the training and test datasets are given in Table 2.

Table 2. Class distribution of images in the training and test datasets.

Class	Images in test dataset	Images in training dataset	Total
Benign	7	28	35
Malign	14	58	72
Total	21	86	107

In the INbreast dataset, mammograms have a contrast resolution of 14 bits. Therefore, firstly, all DICOM files are processed by converting them to an 8-bit contrast resolution. Then the coordinates of the masses and their corresponding pathology classes are obtained from the XML file that lists the cases. After acquiring mass coordinates, they are normalized to adjust to different image sizes. This normalization process involves scaling the mass ground truth coordinates into a range of [0.0 to 1.0] relative to the width and height of the image. Finally, the mammograms are converted to a lower dimension to fit the model. We investigated different image sizes including 448 × 448, 608 × 608, and 832 × 832 as in [13].

4.4. Comparison results of YOLO models (YOLOv3, YOLOv4, and YOLOv5)

In this subsection, the performance of different YOLO models is compared on mass detection. The same parameter values weren selected for models for a fair comparison. In this experiment, YOLOv3, YOLOv4, and YOLOv5(x) were used for mass detection and classification on the INbreast dataset. The image size was chosen as 448×448 for this experiment. Table 3 summarizes the test results of the YOLOv3, YOLOv4, and YOLOv5 models for 448×448 image size. The best benign AP, malign AP, recall, mAP, and F1 values were obtained in the YOLOv5 model. The YOLOv3 model has higher results than YOLOv4 in this experiment.

Table 3. The results of the YOLOv3, YOLOv4, and YOLOv5 models for 448×448 image size.

	Benign AP (%)	Malign AP (%)	Precision	Recall	mAP	F1-Score
Aly et al. [13]	56.0	81.7	80.0	67.0	68.9	72.9
YOLOv3	75.0	78.6	86.7	59.4	76.8	70.5
YOLOv4	51.0	60.6	85.3	56.2	55.8	67.7
YOLOv5	83.0	90.0	83.5	87.0	88.0	85.2

4.5. Comparison results of YOLOv5 models (YOLOv5s, YOLOv5m, YOLOv5l, and YOLOv5x) and transformer-based YOLOv5 model

In the reference article [13], researchers used different image sizes for training, namely 448 × 448, 608 × 608, and 832 × 832, for the YOLOv3 experiment. In this experiment, YOLOv5 models were trained and tested

with these image sizes. The performance of different YOLOv5 models was compared with each other and reference article. For 448×448 image size, the best recall, benign AP, mAP, and F1 values were obtained in the YOLOv5x model. The best malign AP and precision values were obtained in the YOLOv5s and YOLOv5l models, respectively. Performance in all metrics is better than in the YOLOv3 model in [13]. For 608×608 image size, the best recall and F1 values were obtained in the YOLOv5x model. The best malign AP, mAP, and precision values were obtained in the YOLOv5s, YOLOv5m, and YOLOv5l models, respectively. For 832×832 image size, the YOLOv5x model has the best mAP value and malign AP, and YOLOv5m has the best precision value. In the experiments conducted, the best results among the YOLOv5 models were obtained using the YOLOv5x model with an image size of 832. Therefore, the transformer-based model was implemented solely in the architecture of the YOLOv5x model and for an image size of 832. As a result of this experiment, it was observed that the implementation of the transformer-based model led to an improvement in performance across nearly all metrics. The best Benign-AP, recall, mAP, and F1-Score values for the same image size were obtained with this model. The test results for 448×448, 608×608, and 832×832 image size are presented in Table 4.

Table 4. The results of the YOLOv5 models for 448×448, 608×608, and 832×832 image size.

Image size	Model	Benign			Malign			Metrics (%)			
		TP	FP	AP (%)	TP	FP	AP (%)	Precision	Recall	mAP	F1-Score
448	Aly et al. [13]	4	2	56.0	12	2	81.7	80.0	67.0	68.9	72.9
	YOLOv5s	5	9	50.0	16	1	93.0	64.9	82.8	71.0	72.8
	YOLOv5m	5	1	78.0	13	1	91.8	86.3	75.5	84.9	80.5
	YOLOv5l	5	0	80.0	15	3	92.0	89.4	79.8	86.0	84.3
	YOLOv5x	6	2	83.0	15	1	90.0	83.5	87.0	88.0	85.2
608	Aly et al. [13]	4	0	90.7	12	2	72.1	89.0	67.0	81.4	76.4
	YOLOv5s	5	2	63.0	16	0	99.0	84.5	83.4	81.0	83.9
	YOLOv5m	6	2	83.0	15	2	92.0	81.3	87.0	88.0	84.1
	YOLOv5l	4	0	71.5	15	1	95.4	92.6	73.6	83.4	82.0
	YOLOv5x	6	3	77.6	17	1	97.9	79.5	92.5	87.7	85.5
832	Aly et al. [13]	7	0	87.5	14	1	80.8	95.0	88.0	84.2	91.3
	YOLOv5s	6	2	82.0	16	2	96.0	81.8	87.7	89.5	84.6
	YOLOv5m	5	0	85.0	16	1	96.0	96.7	77.3	87.0	85.9
	YOLOv5l	4	0	69.0	15	1	96.0	95.2	73.3	82.8	82.8
	YOLOv5x	5	0	85.0	17	1	97.0	96.3	85.7	91.6	90.7
	Swin Transformer Based YOLOv5x	6	0	92.8	16	3	96.6	92.1	89.9	94.7	91.0

4.6. Using anchors generated by K-means

Anchor boxes are predetermined bounding boxes with fixed widths and heights used to identify the scale and aspect ratio of particular object classes for detection. These boxes are selected based on the sizes of objects in the training datasets and are designed to capture the essential features of the objects. During the detection process, predefined boxes are distributed over the image. The network calculates probabilities as much as the allowed number of anchor boxes for each grid and other properties. Calculations are used to improve each

individual anchor box. Several anchor boxes, each for a different object size can be defined [2].
The anchor boxes can be adjusted to the size of the object being detected using the final neural network outputs. The network should not predict the final width and height of the object, such as a breast mass in cancer detection. Instead, only the anchor box closest in size to the detected object should be adjusted to fit the object size [3]. In the original configuration file of the YOLOv3, v4, and v5 models, there are 9 anchors with different scales presented in the form of (width, height). In the studies given in the previous subsections, these original values were taken as the anchor values. However, reconstructing new anchors representing breast masses of the dataset instead of using random anchors may affect the success rate. This has been accomplished by grouping the masses into 9 relative anchors using bounding boxes and the K-means algorithm integrated with the YOLO-V3 model on the training set mammograms in [13]. Because the model should not have seen the test mammograms before, clustering is carried out on the data after excluding the test data. The composed anchors that are generated are reported as ((15,11), (24,17), (29,21)), ((32,28), (41,31), (63,36)), ((58,46), (75,65), (120,95)) in the reference article [13]. In Table 5 test results for the YOLOv5 model trained with these anchors are presented. The image size is chosen as 448×448 to compare the results of the reference article. The dataset has split 80% training set and 20% testing set.

Table 5. Results for YOLOv5 model trained using anchors given in [13] (448×448 image size).

	Benign			Malign			Metrics (%)			
	TP	FP	AP (%)	TP	FP	AP (%)	Precision	Recall	mAP	F1-Score
Aly et al. [13]	4	2	76.7	11	0	75.1	88.0	62.0	68.9	72.7
YOLOv5s	4	2	66.7	15	1	95.4	79.0	72.7	81.0	75.1
YOLOv5m	5	3	71.4	14	1	82.4	74.2	76.9	79.0	75.5
YOLOv5l	6	3	80.4	16	2	93.8	75.9	89.9	87.1	82.3
YOLOv5x	5	1	77.1	14	2	90.1	85.4	76.6	83.6	80.7

In another experiment, we utilize the autoanchor function in YOLOv5 to verify and generate anchors, before training starts. The autoanchor function analyzes anchors based on data and training settings, and if it determines that the initially presented anchors are not suitable or if the number of anchors is specified in the model file instead of the anchor values, it adjusts the anchors accordingly. To generate new anchors, the autoanchor function applies K-means clustering to the dataset labels, which have been scaled to the training image size. The resulting K-means centroids serve as initial conditions for a Genetic Evolution (GE) algorithm. This GE algorithm then evolves all anchors for 1000 generations using CIoU loss and Best Possible Recall (BPR) as the fitness function [4,5]. Since the approach in the reference paper was implemented with YOLOv3 using an image size of 448×448, we also conducted the experiment with YOLOv5 using the same image size for a fair comparison. The test results for the YOLOv5 model trained with anchors generated by the autoanchor function for 448×448 image size are presented in Table 6. The results of the YOLOv5 model trained with anchors generated by K-means clustering by autoanchor are more successful than those of the YOLOv5 model trained using the anchors given in [13].

[2] MathWorks Inc. (2022). Anchor Boxes for Object Detection [online]. Website https://www.mathworks.com/help/vision/ug/anchor-boxes-for-object-detection.html (Accessed 07 Nov. 2022)
[3] Bochkovskiy A. (2018). GitHub [online]. Website https://github.com/pjreddie/darknet/issues/568 (Accessed 07 Nov. 2022)
[4] Ultralytics, (2021). GitHub [online]. Website https://github.com/ultralytics/yolov5/issues/3482 (Accessed 06 Nov. 2022)
[5] Ultralytics, (2022). GitHub [online]. Website https://github.com/ultralytics/yolov5/issues/6838 (Accessed 07 Nov. 2022)

Table 6. Results for YOLOv5 models trained with anchors generated by K-means clustering in this study (448×448 image size).

	Benign			Malign			Metrics (%)			
	TP	FP	AP (%)	TP	FP	AP (%)	Precision	Recall	mAP	F1-Score
YOLOv5s	5	2	61.5	15	0	95.6	84.4	78.7	78.6	81.4
YOLOv5m	6	2	76.2	14	3	89.2	78.5	83.4	82.7	80.8
YOLOv5l	5	0	81.7	16	1	96.3	88.0	82.8	89.0	85.3
YOLOv5x	6	1	84.9	14	0	93.4	91.8	84.0	89.1	87.7

4.7. Data augmentation approach

Deep learning requires a large quantity of data to overcome overfitting issues, to be able to train appropriately, and to achieve better test performance. Data augmentation refers to the process of creating new training data by applying various techniques to the existing data. The main goal of data augmentation is to create different versions of training set images that the model may have already seen during the training process. Thus, data augmentation is used to solve the issue of limited data in small medical datasets. The primary aim of the experiments in this subsubsection is to assess the impact of data augmentation on performance. In the reference study [13], the training data were augmented through rotation and added to the training set, while the test data were augmented and used as validation data. No modifications were made to the test set.

Initially, we split the original dataset into two parts, where 80% was reserved for the training set and 20% for the test set as in the reference article [13]. In total, 107 images were divided into five folds each having 21 images and each fold was used as a test dataset. Then we applied data augmentation techniques to increase the size of the training set. The training dataset was created by augmenting the remaining 86 images with the data augmentation technique. Specifically, we rotated each image in the training set by three angles of 90°, 180°, and 270° and thus we created a manually augmented dataset. No validation dataset was used at first. Subsequently, the data obtained by rotating the mammograms in the test set were used as a validation dataset. In this experiment, more successful results were obtained than in the approach with no validation dataset. The results of testing the YOLOv5s model trained on an augmented training dataset without a validation dataset and trained on an augmented training dataset with a validation dataset are presented in Table 7. For fair comparison as in the reference article, this experiment was done with only 448×448 input image size and 5-fold cross-validation. Each fold has different results depending on the selected data for the training and test datasets. The precision value is above 70% and 80% for the first and second experiments for this subsection, respectively. However, in many folds, the mAP value is above 70% for these experiments in all folds. In addition to these experiments, we conducted a data augmentation experiment using the YOLOv5x model with an image size of 832, with which we previously achieved the best results, and the results are presented in Table 8.

In the table, classical data augmentation refers to training on the augmented dataset with recommended hyperparameters [6] such as scale, HSV, and mosaic in YOLOv5 were utilized. Hyperparameter values, which are specifically evolved for fine-tuning (COCO pretrained) YOLOv5 models on the VOC dataset, are used. The values indicate the likelihood of employing that data augmentation method. The hyperparameter values used are as follows: hsv-h: 0.0138, hsv-s: 0.664, hsv-v: 0.464, degrees: 0.373, scale: 0.898, shear: 0.602, translate: 0.245, flipud: 0.00856, fliplr: 0.5, perspective: 0.0, mosaic: 1.0, mixup: 0.243, copy-paste: 0.0

[6] Ultralytics, (2020). GitHub [Online]. Website https://github.com/ultralytics/yolov5/issues/852 (Accessed 16 Feb. 2023)

Table 7. Test results of the YOLOv5s model without augmentation, trained on an augmented training dataset without a validation dataset, and trained on an augmented training dataset with a validation dataset with 448×448 image size.

	Benign AP	Malign AP	Precision	Recall	mAP	F1-Score
Aly et al. [13]	91.7	64.0	83.0	79.0	77.8	80.9
Original dataset without augmentation	50.0	93.0	64.9	82.8	71.0	72.7
Augmented training dataset without validation dataset	69.0	98.0	96.7	77.4	84.0	85.9
Augmented training dataset with validation dataset	87.5	93.8	94.3	82.2	90.7	87.8

Table 8. Effect of different augmentation techniques on the training dataset with 832×832 image size YOLOv5x models.

	Benign AP	Malign AP	Precision	Recall	mAP	F1-Score
Classical data augmentation	71.4	87.4	93.8	62.6	79.4	75.1
Augmentation with recommended hyperparameter values	85.0	97.0	96.3	85.7	91.6	90.7
Classical data augmentation + augmentation with recommended hyperparameter values	92.8	99.2	94.7	92.9	96.0	93.8

5. Discussion and conclusions

Early and accurate diagnosis of cancer is extremely important. CAD has been used frequently in the field of medicine in recent decades. These models can help reduce exposure to radiation from X-rays and unnecessary biopsies. There are several computer-aided approaches employed in the diagnosis of breast cancer. Most of the various computer-assisted approaches used in breast cancer diagnosis use multistage object detection/image processing systems rather than a one-time model. Nevertheless, using a two-stage system for detection and classification can be challenging as the data require preparation for each stage. However, the use of single-stage detectors such as the YOLO may be more appropriate since this process is not required. YOLO algorithms are used in object detection and give successful results. Aly et al. [13] conducted various experiments with the YOLOv1, YOLOv2, and YOLOv3 algorithms in their study. In the present research, we extended the findings of the previous study using the YOLOv5 model for a variety of situations and compared the performance of recent YOLO models in mass detection through different experiments. The size of labeled breast lesion images is still limited. We used the YOLOv5 model in the INbreast dataset. The dataset has 107 mass mammograms and has 2 classes, benign and malign. Although limited data are available, the results of YOLO algorithms in breast mass detection are very promising. The performance of the models is assessed by utilizing precision, recall, mAP, and F1-Score metrics. For various image sizes in individual experiments, we get varying results. The best results are obtained by using the YOLOv5 model when the YOLOv3, YOLOv4, and YOLOv5 models are compared and the result of the YOLOv5x model has a higher score for many metrics than the other models for 448×448 image size. Moreover, YOLOv5 has better performance than YOLOv3 for other image sizes, especially for the mAP value. Using YOLOv3, Aly et al. [13] achieved 56%, 81.7%, 80%, 67%, 68.9%, and 72.9% success for benign AP, malign AP, precision, recall, mAP, and F1-Score, while we achieved 83%, 90%, 83.5%, 87%, 88%, and 85.2% success using YOLOv5(x), respectively. For 608×608 image size, the YOLOv5

model achieved very successful results compared to the YOLOv3 model for almost all metrics. The YOLOv3 model used in [13] performed slightly better for benign AP, recall, and F1-Score for 832×832 image size. For all that, the YOLOv5 model performed noticeably better for malign AP, precision, and mAP values. Although the mAP and F1-Scores for the 448×448 and 608×608 image sizes are similar, the highest performance was obtained with the largest image size (832×832) for YOLOv5(x). The utilization of YOLOv5 in conjunction with a transformer has proven to be a contributing factor to enhanced performance. In the proposed study, we implement YOLO with a transformer-based backbone for mammogram mass detection, leading to models that achieve better performance than the previous YOLO methods. Since the most successful results were achieved with an image size of 832 for YOLOv5x, the transformer-based model was also applied with this image size. Furthermore, YOLOv5 has an autoanchor function for anchor verification and generation before training. Using anchors produced with YOLOv5 improved the results because the anchors are generated according to the training data. Two more experiments were conducted to see how data augmentation affects performance during the training phase. Data augmentation improved the performance, and the most successful result was achieved using the augmented training and validation set and augmentation with optimized parameters during the training phase. In experiments, augmentation with recommended hyperparameter values was used. The impact of various augmentation techniques on the training dataset of YOLOv5x models with an image size of 832×832 is also demonstrated in the study. Apart from these, some of the challenges in mass detection studies for breast cancer include memory and GPU limitations, having a limited amount of data, and datasets with missing annotations. In addition, the small number of samples belonging to a class adversely affects the class results. Furthermore, creating a well-annotated dataset can be time-consuming, costly, and prone to errors. Semisupervised, self-supervised, or methods capable of generating synthetic data can assist in mitigating the issue arising from the scarcity of labeled data in experiments. Despite the limitation in the dataset size, applying transfer learning by leveraging the weights of a model trained on a larger dataset for training can contribute to an improvement in performance. If the dataset includes both CC and MLO mammograms of the same breast, utilizing these images together with multiview inputs can contribute to better information extraction, as the features would correspond to the same breast rather than treating these images as separate inputs. In recent times, transformer-based designs have demonstrated the capability to capture extensive spatial relationships through the replacement of convolutional operations, exploiting the self-attention mechanism inherent in the encoder–decoder architecture, and acquiring intricate and highly expressive representations. Different layers of YOLO models can also incorporate distinct transformer-based architectures. If transformers were to entirely replace convolutional operators in machine vision tasks, several challenges would arise, encompassing elevated memory consumption and computational expenses, but the amalgamation of Transformer and Convolutional Neural Network (CNN) architectures can potentially yield improved outcomes. In spite of the limitations, our results indicate that YOLO-based algorithms are promising tools for breast cancer diagnosis.

Acknowledgments

This study is supported by the Council of Higher Education 100/2000 doctoral program. Experimental calculations were conducted on the computer at Erciyes University Artificial Intelligence and Big Data Application and Research Center.

References

[1] Arnold M, Morgan E, Rumgay H, Mafra A, Singh D et al. Current and future burden of breast cancer: global statistics for 2020 and 2040. The Breast 2022; 66: 15-23. https://doi.org/10.1016/J.BREAST.2022.08.010

[2] Siegel RL, Miller KD, Wagle NS, Jemal A. Cancer statistics, 2023. CA: a cancer journal for clinicians 2023; 73 (1): 17-48. https://doi.org/10.3322/CAAC.21763

[3] Parsa P, Kandiah M, Abdul Rahman H, Mohd Zulkefli NA. Barriers for breast cancer screening among Asian women: a mini literature review. Asian Pacific Journal of Cancer Prevention 2006; 7 (4): 509-514.

[4] Gøtzsche PC, Jørgensen KJ. Screening for breast cancer with mammography. Cochrane database of systematic reviews. 2013; (6). https://doi.org/10.1002/14651858.CD001877.pub5

[5] Domingues I, Pereira G, Martins P, Duarte H, Santos J et al. Using deep learning techniques in medical imaging: a systematic review of applications on CT and PET. Artificial Intelligence Review 2020; 53: 4093-4160. https://doi.org/10.1007/S10462-019-09788-3

[6] Moghbel M, Mashohor S. A review of computer assisted detection/diagnosis (CAD) in breast thermography for breast cancer detection. Artificial Intelligence Review 2013; 39: 305-313. https://doi.org/10.1007/S10462-011-9274-2

[7] Veta M, Pluim JPW, van Diest PJ, Viergever MA. Breast cancer histopathology image analysis: a review. IEEE Trans Biomed Eng, 2014; 61 (5): 1400. https://doi.org/10.1109/TBME.2014.2303852

[8] Kozegar E, Soryani M, Behnam H, Salamati M, Tan T. Computer aided detection in automated 3-D breast ultrasound images: a survey. Artificial Intelligence Review 2020; 53: 1919-1941. https://doi.org/10.1007/S10462-019-09722-7

[9] Shah SM, Khan RA, Arif S, Sajid U. Artificial intelligence for breast cancer analysis: trends & directions. Computers in Biology and Medicine 2022; 105221. https://doi.org/10.1016/J.COMPBIOMED.2022.105221

[10] Li Y, Chen H, Cao L, Ma J. A survey of computer-aided detection of breast cancer with mammography. J Health Med Inf 2016; 4 (7): 1-6. https://doi.org/10.4172/2157-7420.1000238

[11] Liberman L, Menell JH. Breast imaging reporting and data system (BI-RADS). Radiologic Clinics 2002; 40 (3): 409-430. https://doi.org/10.1016/S0033-8389(01)00017-3

[12] Abdelhafiz D, Yang C, Ammar R, Nabavi S. Deep convolutional neural networks for mammography: advances, challenges and applications. BMC Bioinformatics 2019; 20: 1-20. https://doi.org/10.1186/s12859-019-2823-4

[13] Aly GH, Marey M, El-Sayed SA, Tolba MF. YOLO based breast masses detection and classification in full-field digital mammograms. Computer Methods and Programs in Biomedicine 2021; 200: 105823. https://doi.org/10.1016/j.cmpb.2020.105823

[14] Ribli D, Horváth A, Unger Z, Pollner P, Csabai I. Detecting and classifying lesions in mammograms with deep learning. Scientific Reports 2018; 8 (1): 4165.

[15] Peng J, Bao C, Hu C, Wang X, Jian W et al. Automated mammographic mass detection using deformable convolution and multiscale features. Medical & Biological Engineering & Computing 2020; 58: 1405-1417.

[16] Al-masni MA, Al-antari MA, Park JM, Gi G, Kim TY et al. Detection and classification of the breast abnormalities in digital mammograms via regional convolutional neural network. In: 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE, 2017; 1230-1233. https://doi.org/10.1109/EMBC.2017.8037053

[17] Al-Antari MA, Al-Masni MA, Choi MT, Han SM, Kim TS. A fully integrated computer-aided diagnosis system for digital X-ray mammograms via deep learning detection, segmentation, and classification. International Journal of Medical Informatics 2018; 117: 44-54.

[18] Al-Masni MA, Al-Antari MA, Park JM, Gi G, Kim TY et al. Simultaneous detection and classification of breast masses in digital mammograms via a deep learning YOLO-based CAD system. Computer Methods and Programs in Biomedicine 2018; 157: 85-94.

[19] Al-Antari MA, Al-Masni MA, Kim TS. Deep learning computer-aided diagnosis for breast lesion in digital mammogram. Deep Learning in Medical Image Analysis: Challenges and Applications 2020; 59-72.

[20] Al-Antari MA, Han SM, Kim TS. Evaluation of deep learning detection and classification towards computer-aided diagnosis of breast lesions in digital X-ray mammograms. Computer Methods and Programs in Biomedicine 2020; 196: 105584.

[21] Djebbar K, Mimi M, Berradja K, Taleb-Ahmed A. Deep convolutional neural networks for detection and classification of tumors in mammograms. In: IEEE 2019 6th International Conference on Image and Signal Processing and Their Applications (ISPA), 2019. pp. 1-7.

[22] Baccouche A, Garcia-Zapirain B, Olea CC, Elmaghraby AS. Breast lesions detection and classification via YOLO-based fusion models. Computers, Materials & Continua 2021; 69 (1).

[23] Baccouche A, Garcia-Zapirain B, Zheng Y, Elmaghraby AS. Early detection and classification of abnormality in prior mammograms using image-to-image translation and YOLO techniques. Computer Methods and Programs in Biomedicine 2022; 221: 106884.

[24] Hamed G, Marey M, Amin SE, Tolba MF. Automated breast cancer detection and classification in full field digital mammograms using two full and cropped detection paths approach. IEEE Access 2021; 9: 116898-116913.

[25] Zhao J, Chen T, Cai B. A computer-aided diagnostic system for mammograms based on YOLOv3. Multimedia Tools and Applications 2022; 1-25. https://doi.org/10.1007/s11042-021-10505-y

[26] Kolchev A, Pasynkov D, Egoshin I, Kliouchkin I, Pasynkova O et al. YOLOv4-based CNN model versus nested contours algorithm in the suspicious lesion detection on the mammography image: a direct comparison in the real clinical settings. Journal of Imaging 2022; 8 (4): 88.

[27] Yasir N, Anwar S, Khan MT. Machine vision-based intelligent breast cancer detection. Pakistan Journal of Engineering and Technology 2022; 5 (1): 1-10.

[28] Hassan NM, Hamad S, Mahar K. A deep learning model for mammography mass detection using mosaic and reconstructed multichannel images. In: International Conference on Computational Science and Its Applications, 2022, (pp. 544-559). Cham: Springer International Publishing.

[29] Zhang L, Li Y, Chen H, Wu W, Chen K et al. Anchor-free YOLOv3 for mass detection in mammogram. Expert Systems with Applications 2022; 191: 116273.

[30] Su Y, Liu Q, Xie W, Hu P. YOLO-LOGO: a transformer-based YOLO segmentation model for breast mass detection and segmentation in digital mammograms. Computer Methods and Programs in Biomedicine 2022; 221: 106903. https://doi.org/10.1016/j.cmpb.2022.106903

[31] Betancourt Tarifa AS, Marrocco C, Molinara M, Tortorella F, Bria A. Transformer-based mass detection in digital mammograms. Journal of Ambient Intelligence and Humanized Computing 2023; 14 (3): 2723-2737.

[32] Kamran SA, Hossain KF, Tavakkoli A, Bebis G, Baker S. Swin-sftnet: spatial feature expansion and aggregation using swin transformer for whole breast micro-mass segmentation, 2022. arXiv preprint arXiv:2211.08717.

[33] Chen X, Zhang K, Abdoli N, Gilley PW, Wang X et al. Transformers improve breast cancer diagnosis from unregistered multi-view mammograms. Diagnostics 2022; 12 (7): 1549.

[34] Lu S, Liu X, He Z, Zhang X, Liu W et al. Swin-Transformer-YOLOv5 for real-time wine grape bunch detection. Remote Sensing 2022; 14 (22): 5853.

[35] Yang W, Wu H, Tang C, Lv J. ST-CA YOLOv5: Improved YOLOv5 based on Swin Transformer and coordinate attention for surface defect detection. In: IEEE 2023 International Joint Conference on Neural Networks (IJCNN) 2023. pp. 1-8.

[36] Vasanthi P, Mohan L. Multi-Head-Self-Attention based YOLOv5X-transformer for multi-scale object detection. Multimedia Tools and Applications 2023; 1-27.

[37] Thuan D. Evolution of YOLO algorithm and YOLOv5: the state-of-the-art object detention algorithm. MA, Oulu University of Applied Sciences, Finland, 2021.

[38] Redmon J, Divvala S, Girshick R, Farhadi A. You Only Look Once: unified, real-time object detection. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 779-788. https://doi.org/10.1109/CVPR.2016.91

[39] Redmon J, Farhadi A. YOLO9000: better, faster, stronger. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 2017, pp. 6517-6525. https://doi.org/10.1109/CVPR.2017.690

[40] Redmon J, Farhadi A. Yolov3: an incremental improvement. arXiv preprint arXiv:1804.02767, 2018. https://doi.org/10.48550/arxiv.1804.02767

[41] Bochkovskiy A, Wang CY, Liao MH. YOLOv4: optimal speed and accuracy of object detection. arXiv 2020, arXiv:2004.10934. https://doi.org/10.48550/arXiv.2004.10934

[42] Wang CY, Mark Liao HY, Wu YH, Chen PY, Hsieh JW et al. CSPNet: a new backbone that can enhance learning capability of CNN. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Seattle, WA, USA, 2020, pp. 1571-1580. https://doi.org/10.1109/CVPRW50498.2020.00203

[43] Liu S, Qi L, Qin H, Shi J, Jia J. Path aggregation network for instance segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018; pp. 8759-8768. https://doi.org/10.48550/arXiv.1803.01534

[44] Dlužnevskij D, Stefanovič P, Ramanauskaite S. Investigation of YOLOv5 efficiency in iPhone supported systems. Baltic Journal of Modern Computing 2021; 9.3: 333-344. https://doi.org/10.22364/bjmc.2021.9.3.07

[45] Liu H, Sun F, Gu J, Deng L. SF-YOLOv5: A lightweight small object detection algorithm based on improved feature fusion mode. Sensors 2022; 22 (15): 5817. https://doi.org/10.3390/s22155817

[46] Xue Z, Lin H, Wang F. A small target forest fire detection model based on YOLOv5 improvement. Forests 2022; 13 (8): 1332. https://doi.org/10.3390/f13081332

[47] Chen PY, Hsieh JW, Gochoo M, Wang CY, Liao HYM. Smaller object detection for real-time embedded traffic flow estimation using fish-eye cameras. In: 2019 IEEE International Conference on Image Processing (ICIP). 2019. pp. 2956-2960.

[48] Liu Z, Lin Y, Cao Y, Hu H, Wei Y et al. Swin transformer: hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision 2021. pp. 10012-10022.

[49] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X et al. An image is worth 16x16 words: transformers for image recognition at scale, 2020. arXiv preprint arXiv:2010.11929.

[50] Liu Z, Hu H, Lin Y, Yao Z, Xie Z et al. Swin transformer v2: scaling up capacity and resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022. pp. 12009-12019.

[51] Wei Y, Hu H, Xie Z, Zhang Z, Cao Y et al. Contrastive learning rivals masked image modeling in fine-tuning via feature distillation, 2022. arXiv preprint arXiv:2205.14141

[52] Li F, Zhang H, Xu H, Liu S, Zhang L et al. Mask dino: towards a unified transformer-based framework for object detection and segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023. pp. 3041-3050.

[53] Hassan NM, Hamad S, Mahar K. Mammogram breast cancer CAD systems for mass detection and classification: a review. Multimedia Tools and Applications 2022; 81 (14): 20043-20075. https://doi.org/10.1007/s11042-022-12332-1

[54] Dhungel N, Carneiro G, Bradley AP. A deep learning approach for the analysis of masses in mammograms with minimal user intervention. Med Image Anal 2017; 37: 114-128. https://doi.org/10.1016/j.media.2017.01.009

[55] Moreira IC, Amaral I, Domingues I, Cardoso A, Cardoso MJ et al. INbreast: toward a full-field digital mammographic database. Acad Radiol 2012; 19 (2): 236-248. https://doi.org/10.1016/j.acra.2011.09.014