

5-20-2024

Deep learning-based breast cancer diagnosis with multiview of mammography screening to reduce false positive recall rate

Meryem Altın KARAGÖZ

Özkan Ufuk NALBANTOĞLU


Derviş KARABOĞA

Bahriye AKAY

Alper BAŞTÜRK

See next page for additional authors

Follow this and additional works at: <https://journals.tubitak.gov.tr/elektrik>

 Part of the [Computer Engineering Commons](#), [Computer Sciences Commons](#), and the [Electrical and Computer Engineering Commons](#)

Recommended Citation

KARAGÖZ, Meryem Altın; NALBANTOĞLU, Özkan Ufuk; KARABOĞA, Derviş; AKAY, Bahriye; BAŞTÜRK, Alper; ULUTABANCA, Halil; DOĞAN, Serap; COŞKUN, Damla; and DEMİR, Osman (2024) "Deep learning-based breast cancer diagnosis with multiview of mammography screening to reduce false positive recall rate," *Turkish Journal of Electrical Engineering and Computer Sciences*: Vol. 32: No. 3, Article 3.

<https://doi.org/10.55730/1300-0632.4076>

Available at: <https://journals.tubitak.gov.tr/elektrik/vol32/iss3/3>



This work is licensed under a [Creative Commons Attribution 4.0 International License](#).

This Article is brought to you for free and open access by TÜBİTAK Academic Journals. It has been accepted for inclusion in Turkish Journal of Electrical Engineering and Computer Sciences by an authorized editor of TÜBİTAK Academic Journals. For more information, please contact pinar.dundar@tubitak.gov.tr.

Deep learning-based breast cancer diagnosis with multiview of mammography screening to reduce false positive recall rate

Authors

Meryem Altın KARAGÖZ, Özkan Ufuk NALBANTOĞLU, Derviş KARABOĞA, Bahriye AKAY, Alper BAŞTÜRK, Halil ULUTABANCA, Serap DOĞAN, Damla COŞKUN, and Osman DEMİR

Deep learning-based breast cancer diagnosis with multiview of mammography screening to reduce false positive recall rate

Meryem Altın KARAGÖZ^{1,2*}, O. Ufuk NALBANTOĞLU^{2,3,4}, Derviş KARABOĞA^{2,3},
Bahriye AKAY^{2,3}, Alper BAŞTÜRK^{2,3}, Halil ULUTABANCA⁵,
Serap DOĞAN⁶, Damla COŞKUN^{2,3}, Osman DEMİR⁷

¹Department of Computer Engineering, Sivas Cumhuriyet University, Sivas, Türkiye

²Artificial Intelligence and Big Data Application and Research Center, Kayseri, Türkiye

³Department of Computer Engineering, Erciyes University, Kayseri, Türkiye

⁴Genome and Stem Cell Center (GenKök), Erciyes University, Kayseri, Türkiye

⁵Department of Neurosurgery, Medical Faculty, Erciyes University, Kayseri, Türkiye

⁶Department of Radiology, Medical Faculty, Erciyes University, Kayseri, Türkiye

⁷Republic of Türkiye Ministry of Health, Kayseri, Türkiye

Received: 15.11.2023

Accepted/Published Online: 27.04.2024

Final Version: 20.05.2024

Abstract: Breast cancer is the most prevalent and crucial cancer type that should be diagnosed early to reduce mortality. Therefore, mammography is essential for early diagnosis owing to high-resolution imaging and appropriate visualization. However, the major problem of mammography screening is the high false positive recall rate for breast cancer diagnosis. High false positive recall rates psychologically affect patients, leading to anxiety, depression, and stress. Moreover, false-positive recalls increase costs and create an unnecessary expert workload. Thus, this study proposes a deep learning-based breast cancer diagnosis model to reduce false positive and false negative rates. The proposed model has two steps: unsupervised feature extraction with Variational Autoencoder (VAE) and classification with CNN using extracted features by VAE. The proposed model is trained and evaluated on in-house anonymized and public mammography datasets. The proposed model provides efficient processing of multiview mammography by maintaining higher accuracy, efficiency, consistency, and faster than transfer learning-based models even on the imbalanced test set of the in-house dataset with obtaining 0.99 AUC, 95.05% accuracy, 97.85% precision, 95.05% recall, and 96.43% F1 score and an AUC of 0.98 on INbreast dataset. Furthermore, the proposed model significantly reduces the false positive recall rate, decreasing it from 6.13% to 2.61% compared to expert diagnosis while achieving an accuracy of 97.03% and AUC of 0.99. Overall, the proposed deep learning-based model enhances breast cancer diagnosis and reduces the false positive recall rate by obtaining high accuracy.

Key words: Breast cancer, computer-aided diagnosis, deep learning, mammography classification, false positive recall rate

1. Introduction

According to the World Health Organization (WHO), breast cancer is the most prevalent cancer type seen worldwide. An early diagnosis of breast cancer and its treatment is crucial for preventing breast cancer and reducing mortality¹. Currently, several screening technologies are used for the diagnosis of breast cancer,

*Correspondence: makaragoz@cumhuriyet.edu.tr

¹World Health Organization (2023). Breast cancer [online]. Website <https://www.who.int/news-room/fact-sheets/detail> [accessed 12 July 2023].

including mammography, computed tomography technology (CT), photoacoustic imaging, nuclear magnetic resonance imaging (MRI), microwave imaging, and other technologies. Routine mammography screening technology is the most reliable and effective screening for early diagnosis thanks to its high-resolution imaging and proper visualization of abnormalities [1, 2].

On the other hand, many breast cancer diagnoses with mammography result in high false positive rates. The outcomes are mostly benign when additional imaging or biopsy is recommended for certain diagnoses. Nevertheless, interpreting mammograms can be a complex and an error-prone task, with at least 25% of detectable cancers going unnoticed [2–7]. In general, a large number of mammograms, that is to say, 10%-15% of them require additional clarification with additional screening such as ultrasound. After that, only 10%-20% of additional imaging is recommended to refer to biopsy for certain diagnosis. Finally, only 20%-40% of biopsy operations end up with a cancer diagnosis [8, 9]. The false-positive recall rate is so high that most women will face at least one false-positive recall in 10 years of the annual screening period [10, 11]. Researchers have also focused on false-positive effects in several studies [8, 11, 12]. According to those studies, the psychology of women negatively affects the result of false positive recall, which leads to distress, anxiety, depression, etc. Moreover, a high rate of false positive recalls may also have a negative impact on the cost-effectiveness of mammography screening. The cost of false-positive recalls reached almost 4 billion dollars annually in the USA, reported by [13] in 2016. As a result, reducing the false positive recall rate in screening mammography is crucial for mitigating adverse effects.

Mammography images are usually read and interpreted by expert radiologists. However, radiologists' performances can vary subjectively due to different factors such as their experience, workloads, and complexity or quality of mammography screening cases. Moreover, the reading process of mammograms is an extremely labor-intensive process for radiologists due to the increased number of mammography imaging. Computer-aided detection (CAD) has been utilized for mammography screening to assist radiologists for two decades. However, studies reported that CAD does not sufficiently improve diagnostic performance [14–16]. Furthermore, traditional CAD programs require hand-crafted feature extraction and have a high false positive recall rate [17]. In recent years, deep learning-based CADs, particularly in Convolutional neural networks (CNN) [18–21], have emerged with encouraging results to enhance the performance of CADs owing to their end-to-end feature extraction mechanisms and learning ability for complex problems. Several deep learning studies have recently been developed for analyzing mammography images [22–26].

Deep learning has been a significant gain in reducing false positive recalls by obtaining high performance. Wu et al. [9] reported that their proposed end-to-end deep learning-based diagnosis model could achieve more accurate results than radiologists' decisions. Clancy et al. [10] employ a pretraining strategy for several deep learning models to distinguish false positive recall in mammography screening. Mayo et al. [17] compare the AI-based CAD algorithm with the conventional CAD algorithm. They report that AI-based CAD reduces false-positive rates by a significant reduction of 69%. Advanced pretraining techniques have a considerable improvement in determining recalled-benign mammography images. Bozkurt et al. [27] incorporate a handcrafted, deep, and fusion-based feature extraction framework (HANDEFU), allowing users to build models interactively for COVID-19 from chest X-ray images. The study of the proposed framework achieves superior performance with a 99.36% accuracy using the LBP+SVM model among various models. Kyono et al. [28] introduce MAMMO as a clinical decision support system with multi-task learning for the diagnosis with mammography images and patients' personal information. MAMMO provides automatic triaging of patients

to save scarce clinical resources and improve the diagnosis performance of radiologists. Aboutalib et al. [29] propose deep learning-based breast cancer diagnosis to reduce unnecessary recalls (recalled-benign). Kim et al. [30] develop an AI-based model for breast cancer diagnosis with mammography images. According to the study, the AI-based model performs notably better than radiologists with AUROC of 0.94–0.97 and reduces false positive recalls. Adedigba et al. [31] offer a discriminative fine-tuning and mixed-precision training model on public mammogram datasets. They also use data augmentation for rapid convergence and improved performance. Their proposed model obtains the highest accuracy of 0.998 owing to fine-tuning and data augmentation strategies. Maqsood et al. [32] present a transferable texture convolutional neural network (TTCNN) based on convolutional neural network models that are InceptionResNet-V2, Inception-V3, VGG-16, VGG-19, GoogLeNet, ResNet-18, ResNet-50, and ResNet-101. The proposed model employs public DDSM, INbreast, and MIAS datasets, reaching an average accuracy of 97.49%. Hamidinekoo et al. [33] offer a deep learning-based CAD model to enhance breast cancer identification results by leveraging the relationship between mammography and breast histopathology images. Thus, they aim to create a mapping of features/phenotypes between mammographic abnormalities and their histopathological representation. Ragap et al. [34] utilize a multi-DCNN model for classifying breast cancer lesions in mammograms. They apply four experiments to improve classification performance, including end-to-end fine tuning, deep feature extraction, fusion of features extracted by DCNNs, and using PCA to reduce the computational cost of the DCNN model on public CBIS-DDSM and MIAS datasets. Their proposed fusion model outperforms state-of-the-art models by reducing computational cost. Al-Mansour et al. [35] present a comprehensive analysis for multi-label classification based on two-view mammography images, including density, lesion types, and lesion states. Their proposed ConvNeXt-based CAD enables a thorough assessment of the patient's condition and the preparation of a detailed patient report by offering radiologists an in-depth analysis of mammograms. In summary, the studies in the literature show that deep learning models can offer promising results by distinguishing unnecessary recalls and improving radiologists' performance with high accuracy, efficiency, and consistency.

Although deep learning-based models have demonstrated remarkable attention, deep learning models face limited and imbalanced dataset problems in mammography identification. Deep learning models require large datasets to learn the numerous parameters involved effectively. Training deep learning models on a limited dataset can lead to an over-fitting problem, where the model performs well on the training data but struggles to generalize to new examples. On the other hand, training on an imbalanced dataset can result in an under-fitting problem, as the model may not sufficiently capture the patterns of the minority class, leading to the mislabeling of diverse data. Only a few publicly available datasets exist in a small size, such as INbreast [36], MIAS [37], Digital Database for Screening Mammography (DDSM) [38], CBIS-DDSM [39], BCDR [40], and WBCD [41]. Collecting and publishing publicly available large-scale data practically is challenging in the short term due to concerns about patient privacy, the need for expert interpretation, and the laborious, exhausting, and costly process. Previous studies applied various strategies, such as using transfer learning models, data augmentation, and multi-task learning, to mitigate over-fitting and under-fitting problems in deep networks. Even though these strategies address insufficient dataset problems, developing new and robust deep learning strategies is essential to resolve these problems.

In recent years, self-supervised learning models (SSL) have gathered remarkable attention to deal with insufficient dataset problems in computer vision tasks. SSL provides a powerful solution to the lack of datasets and allows the usage of unlabeled datasets owing to the pretext task step. Thus, SSL models extract features from unlabeled datasets in an unsupervised manner during the pretext task step. Then, the pretrained model

is implemented in the main task step via end-to-end fine-tuning or feature extraction [42]. Nevertheless, mammogram classification with deep learning still faces imbalanced and limited annotated dataset problems, and there is a lack of self-supervised model study in mammography. Additionally, conducting deep learning model research on clinical datasets is crucial to contribute reliability and relevance in healthcare and clinical applications. From this perspective, this study aims to develop a new self-supervised learning model to assess both clinical applications and public datasets in mammography. The proposed model offers two stages for breast cancer diagnosis using multiview mammography screening to diagnose and reduce the false positive recall rate and maintain high performance. The first stage ensures the usage of a large number of negative samples in an unsupervised manner for feature extraction with a Variational Autoencoder (VAE) as a pretext task. Thus, we aim to develop a robust deep learning model by allowing the usage of a large number of mammography samples in the first stage. The second stage is classification by conducting a lightweight Convolutional neural network that exploits extracted features (latent space of VAE) by the pretrained encoder of VAE. Thanks to utilizing low dimensional latent space and implementing lightweight classifier models, the proposed deep learning model provides a solution to prevent over-fitting problems on a limited mammogram dataset. In summary, the proposed model aims to develop a robust deep model and reduce the false positive and false negative recall rates for breast cancer diagnosis with two networks, which are pretext network for training VAE for feature extraction and classifier network using latent space of four image modalities extracted by the encoder. The proposed model compares with commonly used CNN models, which are Resnet50V2 [43], ResNet101 [43], DenseNet121 [44], and EfficientNetB1 [45]. We use an in-house mammography screening dataset to evaluate the proposed model. Furthermore, we retrospectively compare the performance of the proposed model with expert decisions. The main contributions of this study can be stated as follows:

- We propose a self-supervised learning-based breast cancer diagnosis model that enables the usage of multiview mammograms to reduce false positive and false negative rates. The proposed model provides high accuracy, efficiency, and consistency even on an imbalanced test set.
- The main problem with training deep learning-based breast cancer diagnosis models with mammography is the scarcity of positive samples. The proposed model allows the usage of a large number of samples in an unsupervised manner in the pretext task step. Then, the pretrained encoder extracts deep features for classification. Thus, the proposed model performs better than the conventional transfer learning models owing to the two-stage mechanism.
- The proposed model incorporates the utilization of multiview mammography images. Utilizing multiview mammograms yielded improved classification performance compared to single-view-based models.
- The proposed model is assessed on the public INbreast dataset and shows competitive performance with previous studies by demonstrating a high AUC of 0.98.
- The reduced false positive rate is crucial for decreasing costs and minimizing psychological effects on patients. The proposed model reduces the false positive recall rate from 6.13% to 2.61% compared to expert diagnosis.

This paper is organized as follows. Section 2 describes the proposed deep learning model for mammography classification. Section 3 gives information about study cohorts, datasets, and experimental setups. Section 4 presents the results of the proposed model and transfer learning-based models for each test dataset in detail.

Section 5 discusses the limitations of the proposed model and its future direction. Finally, Section 6 briefly concludes this paper.

2. The proposed deep learning model

We proposed a new self-supervised learning-based CAD model via Variational Autoencoder (VAE) in the pretext task step and a Convolutional Neural Network (CNN) in the classification step for breast cancer diagnosis. The proposed deep model has two stages that provide usage of all images with pretext task network in an unsupervised manner and classification network by using extracted latent features by the pretrained encoder of VAE. The proposed model is given in Figure 1.

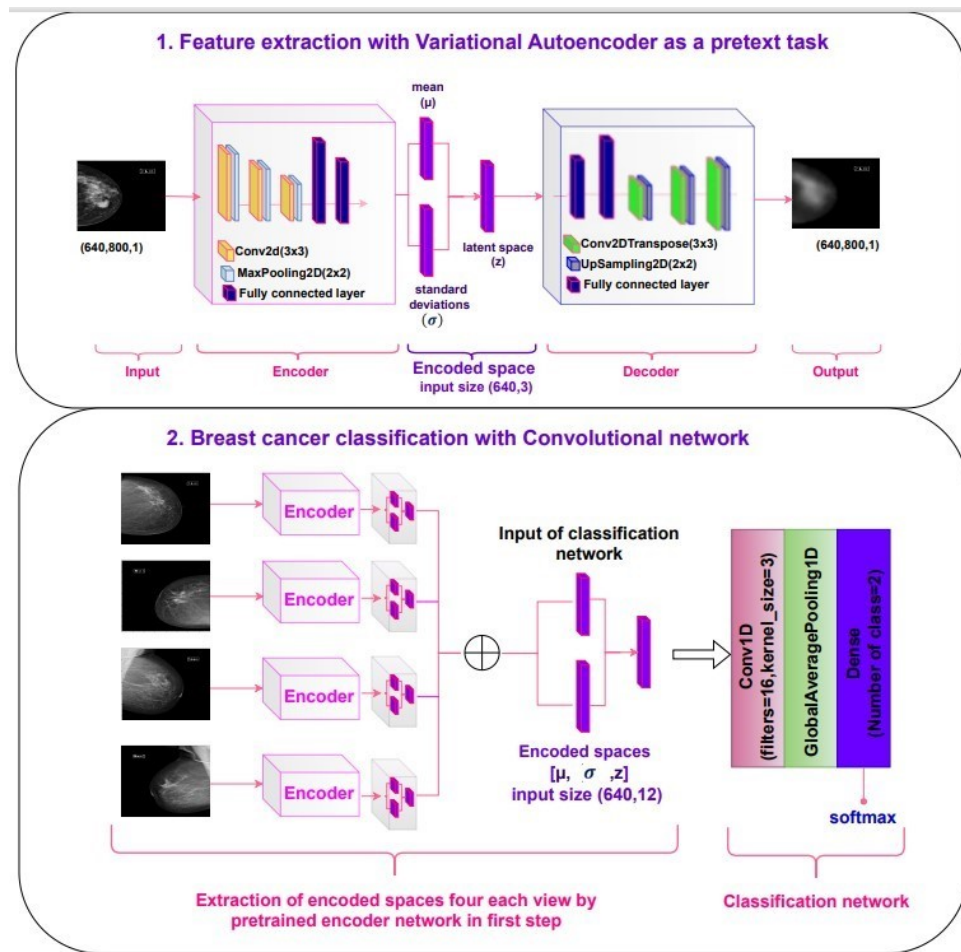


Figure 1. The multiview-based proposed model for classification of mammogram images.

2.1. Feature extraction with Variational Autoencoder as a pretext task

The first stage of the proposed model is feature extraction and dimensional reduction through VAE during the reconstruction of mammography images in the pretext task. Variational Autoencoder (VAE) was introduced by Kingma and Welling [46] as an unsupervised generative network. VAE differs from traditional autoencoders

by regularized latent space and calculation loss. VAE conducts dimensionality reduction by mapping high-dimensional input data to a lower-dimensional latent space. The regularized latent space of VAE generates more meaningful features for generative processes compared with other generative networks [47]. This reduction can facilitate more efficient processing and representation of datasets. Therefore, we choose VAE as a pretext task network to extract deep features and conduct dimensionality reduction from mammography images. Then, we aim to use the extracted latent representation in the classification step. VAE is composed of Bayesian inference-based probabilistic encoder and decoder networks. In addition, VAE contains latent space (z) that represents a distribution of input (x) encoded with a probabilistic encoder network by calculating an approximated posterior distribution $p_\theta(z|x)$. Latent space is described by mean (μ), standard deviations (σ) and random variable $\varepsilon \sim N(0, 1)$, providing an efficient input data compression and is calculated as follows:

$$z_i = (\mu_i + \sigma_i \cdot \varepsilon) \sim q_\phi(z|x) \quad (1)$$

Decoder network gets latent space (z) as input by maximizing the marginal log-likelihood $p_\theta(x'|z)$. VAE used reconstruction loss (L_{rec}) and by minimizing Kullback–Leibler divergence (L_{KL}) loss for calculation of reconstruction error between x (input) and x' (output) that is calculated as follows:

$$L_{rec} = -E_{q_\phi(z|x)}[\log p_\theta(x'|z)] \quad (2)$$

$$L_{KL} = D_{KL}(q_\phi(z|x) || p_\theta(x'|z)) \quad (3)$$

$$L = L_{rec} + L_{KL} \quad (4)$$

The encoder network of VAE is built up by three 2D convolution layers with 3x3 kernel size and ReLU activation, each followed by a max pooling layer of 2x2. Furthermore, the last max pooling layer of the encoder network is followed by a flattened layer, which is then fed into three fully connected layers with sizes of 800, 640, and 640, respectively, to yield the mean (μ) and log variance (σ) vectors. Then, a lambda layer computes the final latent vector (z) using (μ) and (σ) vectors. We used a grid search to determine latent space sizes among various sizes of 64, 128, 256, 512, 640, and 1024. Based on the outcomes of several experiments, it was observed that a small latent size is insufficient for reconstructing high-quality images. In contrast, a larger latent space size allows for more faithful reconstructions. To balance the need for faithful reconstructions with considerations of computational complexity and the risk of over-fitting, we selected a latent space size of 640 for aligning with the proper image size, ensuring that it is appropriately scaled to handle the complexity of the input images. Thus, the encoder network attempts to efficiently map high-dimensional mammograms of size (640x800) onto a low-dimensional latent vector of size 640. On the other hand, the decoder network aims to reconstruct the image to its original size through the latent space(z). Therefore, the decoder gets the latent vector (z) as an input and then is fed into a fully connected layer and a reshape layer to convert the 1D output to a 4D tensor. Thus, the fully connected layer and reshape layer help transform the latent vector into a format compatible with the decoder. Additionally, $L1 = 1e - 5$ and $L2 = 1e - 4$ regularization are applied to the fully connected layers in both the encoder and decoder, contributing to the model's generalization and mitigating over-fitting concerns. The decoder networks consist of three 2D transposed convolutional layers with 3x3 kernel size and ReLU activation, each following three up-sampling layers of 2x2. The last layer of the decoder network provides ultimate output via a 2D transposed convolutional layer with a 3x3 kernel size and sigmoid activation to reconstruct an image of size (640x800). While ReLU activation is used in the encoder

layers and decoder layers (excluding the last reconstruction layer) for its computational efficiency and ability to address gradient-related issues, sigmoid is utilized only in the output layer of the decoder for its suitability in generating probabilities and aligning with loss function of VAE. Furthermore, we used a small kernel size in each convolution layer and pooling layer for enhanced localized features, preservation of fine details particularly in small lesions, improved generalization ability, and reduced over-fitting risk. VAE is trained with a large number of healthy mammography images (86,894) with a fixed size (640x800) belonging to BIRADS 1-2 classes in an unsupervised manner. Thus, the proposed model provides a solution for using large imbalanced datasets to extract diverse and further information during reconstruction.

2.2. Breast cancer classification with convolutional network

The second stage is the classification of mammograms as a patient or healthy for breast cancer diagnosis, using latent space (z) with mean (μ) and standard deviations (σ) vectors. We utilized the CNN model in the classification step because CNNs are well-recognized models and enable the capture of robust, discriminative, and local features thanks to the convolutional mechanisms. The classification network is generated as a lightweight model to prevent over-fitting on a small number of training mammography datasets, containing solely a 1D convolutional layer, a global average pooling layer, and a fully connected layer. The 1D convolutional layer employs a kernel size of 3 and ReLU activation. Subsequently, the global average pooling layer enhances effective feature extraction and dimensionality reduction, creating a meaningful representation fed into the fully connected layer for classification. Consequently, the final classification layer is constructed with a specified number of class sizes (2 for patient and healthy classification) and utilizes softmax activation. The computational complexity of the proposed models for each network is presented in Table 1. The multiview-based proposed model offers low complexity and rapid processing with a depth of 3, 389.20k FLOPs (Floating-point Operations), 338 parameters, and 0.354 s execution time of training per epoch.

The classification network gets low dimensional latent space extracted by a pretrained encoder of VAE in the first stage. Firstly, the encoder of VAE gets each imaging modality separately and extracts a latent vector with (640,3) size. Then, each latent space of four image modalities is stacked vertically as a final input vector with (640,12) size for classification. This experimental design includes image-based positive samples in the patient class for training and testing. In some real-case scenarios, patients may have only one breast or may not have positive findings in all four images (RCC, RML, LCC, LML). If all four images of a patient are not available, the encoded spaces are padded by using copy augmentation via other images of the same patient. The classifier gets latent spaces of four image modalities (RCC, RML, LCC, LML) together as an input vector for patient-based classification in training and testing. In summary, our purpose is to build a robust model that can leverage a large dataset in an unsupervised manner during the pretext task step. Then, the latent spaces of multiview mammograms are utilized as input in the classification step.

3. Experimental study

3.1. Study cohorts and datasets

The anonymized data were collected from the field screening conducted by cancer early detection, scanning, and education centers (KETEM) of the Turkish Ministry of Health in Kayseri Province on “women older than 40 years old between the years 2015-2018. Out of 25,432 mammography screenings, 1611 were referred for additional imaging or biopsy. The unresolved or missing cases were excluded from the dataset, resulting in a total of 23,258 samples.

Table 1. The computational complexity of the proposed models is evaluated by including the number of depths, FLOPs, trainable parameters, and execution time (second) of training per epoch.

Network	subnetwork	depth	FLOPs(G)	parameters	execution time (second)
Pretext task network	encoder of VAE	10	0.163	52.22M	
	decoder of VAE	9	0.147	41.03M	
	VAE (total)	19	0.311	93.25M	28985.912
Classification network	CNN	3	0.000389 (389.20k)	338	0.354

In this study, we used 92,938 mammography images belonging to 23,258 women from the in-house dataset. The information on the in-house dataset is given in Table 2. The Breast Imaging Reporting and Data System (BIRADS) offers a standardized framework for organizing and reporting breast imaging findings. This structure ensures clarity and consistency in communication among healthcare professionals. BIRADS employs assessment categories, ranging from 0 to 6, each with defined criteria, to convey the likelihood of malignancy associated with specific imaging findings. This categorization facilitates a uniform and systematic interpretation and reporting of breast imaging. 1531 women were referred to additional imaging or biopsy for further examination. Only 111 of 1531 women had suspicious lesions, of which 73 were diagnosed with BIRADS 4 and BIRADS 5, most likely targeted as malignant. The mammography images have four views for each patient: right/left with craniocaudal (RCC/LCC) and right/left mediolateral oblique (RMLO/LMLO), as shown in Figure 2.

The annotated BIRADS 0-3-4-5 cases were included in the patient class. Thus, 206 biopsy-proven lesions on mammograms (belonging to 111 women) were labeled as a patient, 92,732 mammography images belonging to BIRADS 1-2, and benign calls from the cases referred to further examination were labeled as healthy. All samples of BIRADS 1-2 (86,894) were utilized for training the pretext task network. The difference between the healthy and patient dataset sizes is large, which can lead to over-fitting problems on deep models. Therefore, we used a balanced and relatively small dataset to train the classifier network. The classifier network was tested on four datasets to observe model robustness, including a different number of healthy samples (Table 3). The balanced test dataset has been split into 80% for training and 20% for testing. Other test datasets used a pretrained model with a balanced dataset to evaluate different test datasets. The unbalanced test dataset contains a sample size of the healthy class that is 25 times larger than that of the patient class. The proposed model was tested on patients who had undergone the biopsy test dataset to compare the model and physician performance. The human expert comparison test dataset has 20% of all healthy samples ($\sim 19,000$) to reflect a realistic scenario and measure model robustness.

The proposed model was evaluated on the public INbreast dataset to assess its effectiveness against previous studies. Meanwhile, BI-RADS 1 and 2 in the INbreast dataset were included as the benign class, and BI-RADS 4, 5, and 6 were targeted as malignant. We applied contrast-limited adaptive histogram equalization (CLAHE) to mammograms emphasizing the improvement of local contrast and the visibility of details in both bright and dark regions. The train set and test set were split into 80% and 20% for each class in the classification process. The number of images for each train and test dataset is presented in Table 3.

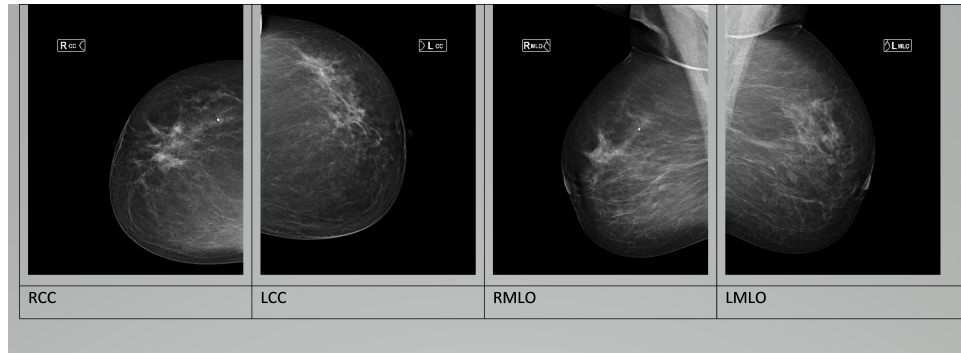


Figure 2. The multiview mammogram example of the in-house dataset.

Table 2. The information about the in-house dataset.

Category	Subcategory	Subcategory	Number of patients	Number of images
Healthy	BIRADS 1-2		21,727	86,894
Referred			1531	6044
	Biopsy-proven		111	206
		BIRADS 0	34	64
		BIRADS 3	4	6
		BIRADS 4	34	62
		BIRADS 5	39	74
Total			23,258	92,938

Table 3. The information about train and test dataset for each healthy (BIRADS 1-2 and recalled-healthy from the cases referred to biopsy) and patient (BIRADS 0-3-4-5 biopsy-proven) classes.

Dataset	Classes	Number of women	Number of images
Train dataset	Healthy	61	242
	Patient	89	166
Balanced test dataset	Healthy	17	60
	Patient	20	40
Unbalanced test dataset	Healthy	505	2001
	Patient	20	40
Undergone biopsy test dataset	Healthy	1420	5485
	Patient	20	40
Human expert comparison test dataset	Healthy	4742	18,753
	Patient	20	40
INbreast train dataset	benign	39	80
	malignant	70	228
INbreast test dataset	benign	10	20
	malignant	26	60

3.2. Experimental setup

The proposed deep models were built by Keras Library. GeForce RTX 2080 Ti GPU with Tensorflow-gpu library was used for training and testing. We consider several key points to set hyperparameters for VAE and classification models, such as model complexity and preventing over-fitting and vanishing gradients. We utilized small batch sizes and $L1 - L2$ regularization over latent layers of VAE to avoid over-fitting and achieve faster convergence and better generalization. Therefore, we implemented VAE with the following initial hyperparameters settings to train: batch size = 16, $L1 = 1e-5$, $L2 = 1e-4$, optimizer = RMSprop with learning rate = 0.001 for minimizing Kullback–Leibler divergence (L_{KL}) loss, and epochs=100. RMSprop optimizer was utilized to minimize VAE loss because of the adaptive learning rate mechanism and its effectiveness in handling sparse gradients. Furthermore, to optimize the learning process, we dynamically adjusted the learning rate of the VAE using cyclic learning rate (CLR). The CLR was configured with a maximum learning rate of 0.0001, 0.25 step size per epoch, and triangular mode. The large healthy samples of the in-house dataset split into training and validation sets in ranges of 75 and 25 for the VAE model, respectively. Finally, the best-performing model attained through VAE training and monitoring via validation loss was saved and utilized for feature extraction. The pretrained encoder of VAE, trained on the in-house dataset, is utilized for latent feature extraction in classification on both in-house and public dataset experiments. On the other hand, the classifier network was set up with a batch size of 16, the Adam optimizer with a learning rate of 0.001, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and a momentum of 0.99 for minimizing sparse categorical cross-entropy loss. The classifier model was trained with 10,000 epochs. Adam optimizer is implemented in the classifier model for adaptive optimization, faster convergence, and efficient training, especially in scenarios where the gradients of different parameters have varying scales. Furthermore, 5-fold cross-validation was implemented during training of the classifier model for robust model evaluation, hyperparameter tuning, and mitigating the impact of data variability on performance metrics. According to the minimum validation loss, the best-performing model of the classifier during training is utilized to evaluate mammography samples. Accuracy, Recall, Precision, and F1 evaluation metrics were used to evaluate the classifier models with a weighted average. Furthermore, the mean area under the curve (AUC) is calculated to evaluate the classifier's overall performance and generalization across multiple datasets or experimental conditions.

4. Results

In this study, we used a balanced dataset to train the proposed model. Then, the pretrained model training on the balanced dataset evaluates different imbalanced test datasets. The information on the dataset is given in Table 3, and the results on the balanced dataset are presented in Table 4. Additionally, the ROC curve and confusion matrix results are given in Figure 3 for the proposed models and Figure 4 for the transfer learning models. Firstly, the classifier was trained by considering a single whole image's latent space with (640,3) size, called a single-view-based model. The second proposed classifier model was trained by staked latent spaces of four image modalities with (640,12) size for classification, called a multiview-based model. We compared both proposed models with transfer learning models, which are Resnet50V2, ResNet101, DenseNet121, and EfficientNetB1. The single-view-based proposed model surpasses transfer learning models on balanced test datasets. The results of the single-view-based proposed model achieved an AUC score of 0.99, an accuracy of 93.00%, and an f1 score of 93.16%. On the other hand, the multiview-based proposed model indicates a higher sensitivity (true positive rate) for the patient class compared to the single-view-based model, with only a small number of healthy cases being misclassified as patients. The single- and multiview-

based models outperformed state-of-the-art transfer learning models by nearly doubling the performance on the balanced dataset. Additionally, the proposed models facilitate a fast training process with 0.354 s/epoch for classification (see Table 4). According to the confusion matrix results in Figure 3, the proposed models accurately identify patient and healthy mammogram samples. On the other hand, transfer learning models demonstrate difficulty in correctly determining the patient and healthy classes, indicating a notable biased tendency to classify mammograms as healthy (see Figure 4). Thus, the proposed models provide a promising solution by enhancing performance in accurately classifying mammograms, addressing the over-fitting in deep learning models, achieving fast training execution times due to the feature extraction mechanism in the pretext step, and using a lightweight classifier model via the extracted latent space.

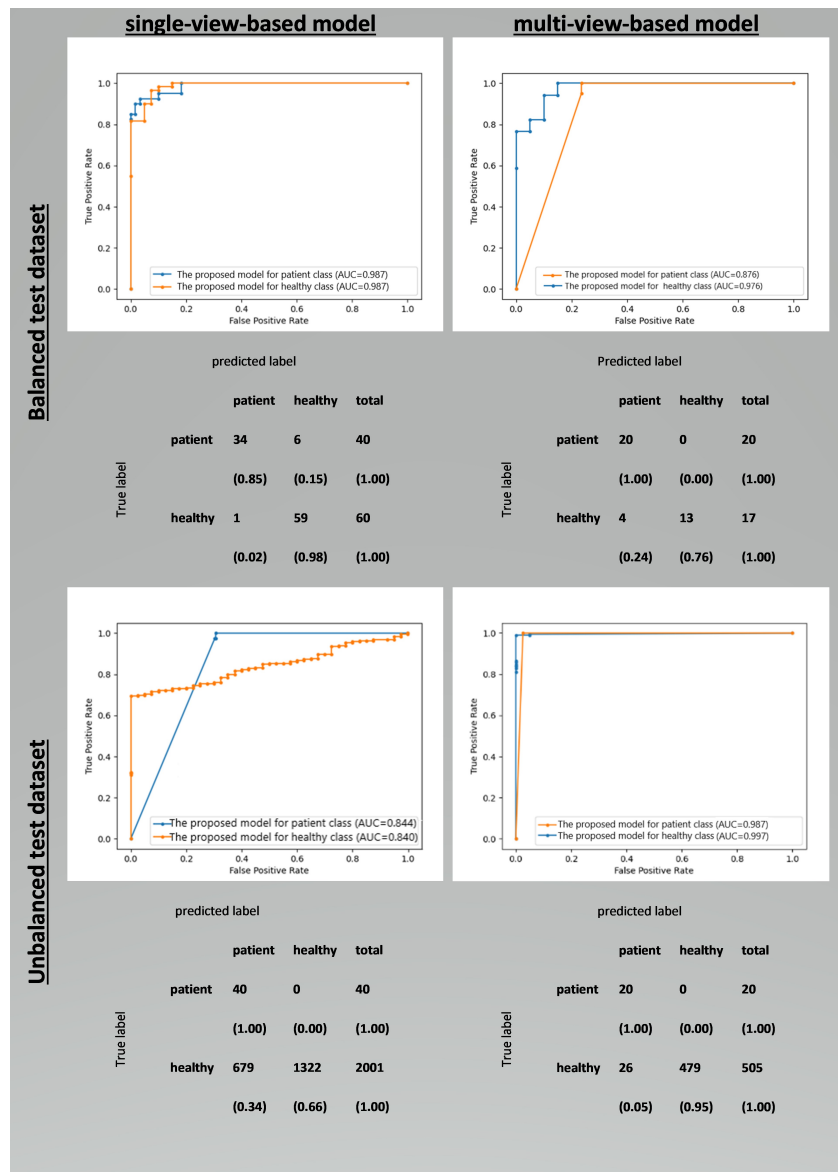


Figure 3. The ROC curve and confusion matrix results of single-view-based and multiview-based models for balanced and unbalanced test datasets.

Table 4. The accuracy, precision, recall, f1 with a weighted average, the mean of AUC results, and the execution time (second) of training per epoch on the balanced test dataset.

Model	AUC	accuracy	precision	recall	f1	execution time (second)
Resnet50V2	0.51	59.00	53.60	59.20	50.20	122.607
ResNet101	0.48	55.00	43.60	54.80	45.20	157.776
DenseNet121	0.39	58.00	45.40	58.20	45.80	139.386
EfficientNetB1	0.40	55.00	47.80	55.00	47.40	122.474
The single-view based proposed model	0.99	93.00	93.32	93.00	93.16	0.555
The multiview based proposed model	0.93	89.19	90.99	89.19	90.08	0.354

The results of models on the unbalanced test dataset are given in Table 5. The purpose of the involved unbalanced test dataset is to assess the robustness of the proposed model on an imbalanced test dataset. The sample size of the healthy class is 25 times larger than the sample size of the patient class in the imbalanced test dataset. One of the critical benefits of deep models is accurately identifying patients, which means avoiding false negatives (failing to classify a true patient as positive). The other important benefit is avoiding false positive recall rate, which involves not labeling healthy individuals as patients to reduce adverse effects such as anxiety, depression, cost-effect, etc. Ultimately, it is crucial to strike a balance between two critical factors. According to the confusion matrix results in Figure 3, both proposed models exhibit high sensitivity for patient samples, correctly identifying all of them. On the other hand, the multiview-based proposed model has a small number of false positive samples, with only 26 healthy cases in 505 cases being misclassified as patients on the unbalanced dataset. Furthermore, the multiview-based proposed model exhibits significantly higher AUC scores of 0.987 for patients and 0.997 for healthy instances than the transfer learning models. In addition, the confusion matrix results of transfer learning models are given in Figure 4. Although the precision, recall, f1, and accuracy values of transfer learning models are relatively high on an unbalanced dataset due to the large number of healthy samples calculated with a weighted average, the AUC values are almost half the proposed method on the unbalanced dataset. Furthermore, while the transfer learning models accurately classified most healthy samples, they misclassified most patient samples as the results of the balanced dataset. The transfer learning models might not be as effective as the proposed model in overall performance, especially in capturing the true discriminator between healthy and patient classes because of failure in patient identification. Overall, the multiview-based proposed model significantly improves performance for the patient class compared to transfer learning models and a single-view-based model when the test dataset becomes increasingly unbalanced. We execute Grad-CAM (Gradient-weighted Class Activation Mapping) [48] to compare the results of single-view and multiview-based proposed models. Thus, we aim to provide visual explanations for the decision-making process of the models by highlighting the regions in the mammogram image that contribute the most to a particular class decision. The example results of Grad-CAM for single- and multiview-based models are presented in Figure 5. The proposed model focuses on lesions and masses in mammograms for making model decisions during training (see Figure 5). While the single-view-based model leverages only significant patterns in a single mammogram, the multiview-based model provides a stronger, more effective, and robust solution by utilizing powerful patterns from multiple mammography views.

Table 5. The accuracy, precision, recall, f1 with a weighted average and the mean of AUC results on the unbalanced test dataset.

Model	AUC	accuracy	precision	recall	f1
Resnet50V2	0.50	92.31	96.14	92.35	94.22
ResNet101	0.50	93.68	96.12	93.24	95.16
DenseNet121	0.40	91.08	96.10	91.24	93.16
EfficientNetB1	0.51	86.48	96.12	86.47	91.24
The single-view-based proposed model	0.84	66.73	98.15	66.73	79.45
The multiview-based proposed model	0.99	95.05	97.85	95.05	96.43

We applied patients who had undergone a biopsy test dataset to evaluate all referred samples and compare the performance of the proposed models and physicians. The results of the multiview-based proposed model on patients who have undergone a biopsy test dataset and human expert comparison test dataset are given in Table 6. The proposed model accurately identifies all patient cases in the patient class and reduces the false positive recall rate in the healthy class. Additionally, the ROC curve and confusion matrix results of the multiview-based proposed model are given in Figure 6 for patients who have undergone the biopsy test dataset and human expert comparison test dataset. According to the confusion matrix results in Figure 6, the proposed multiview-based model has incorrectly classified 131 cases out of all cases referred to as patients. On the other hand, the physicians referred to all 1420 cases as suspicious, which suggests a higher rate of false positives (i.e. actually healthy cases targeted as patients) compared with the proposed model. Moreover, the proposed model has a lower false positive rate for identifying healthy cases as patients than physicians. As a result, the proposed model achieved an accuracy of 90.86%, AUC of 0.97, and f1 of 94.14% on the undergone biopsy test dataset.

Table 6. The accuracy, precision, recall, f1 with a weighted average, and the mean of AUC results for the multiview-based proposed model on patients who have undergone biopsy test dataset and human expert comparison test dataset. 5-fold cross-validation is implemented for the proposed model.

Dataset	AUC	accuracy	precision	recall	f1
Undergone biopsy test set	0.97 ± 0.002	90.86 ± 0.008	98.75 ± 0.000	90.86 ± 0.008	94.14 ± 0.005
Human expert comparison test set	0.99 ± 0.000	97.03 ± 0.003	99.63 ± 0.000	97.03 ± 0.003	98.17 ± 0.002

The proposed model utilized a stratified test dataset with 20% sampling of the groups to reflect a realistic scenario that could reflect the real-life scenario and measure the robustness of the model. Table 6 and Figure 6 illustrate the results of the proposed models on the multiview-based proposed model. The proposed model shows high robustness on the human expert comparison test dataset with an accuracy of 97.03%, AUC of 0.99 (according to ROC curve $\sim 99\%$ AUC for each class), and f1 of 98.17%. The proposed model was able to classify all patients accurately, and it also demonstrated a low false positive recall rate by identifying only 124 out of 4742 healthy cases as patients. On the other hand, a total of 1531 women out of 23,258 were recommended by physicians to undergo further examination with a biopsy. However, biopsy results showed that only 111 of these women were actually patients, and the majority of the referred women (1420) were found to be healthy. Thus, the proposed model was ensured to decrease the false positive recall rate to 2.61%. In contrast, the expert diagnosis had a false positive recall rate of 6.13% according to the confusion matrix for all samples of the in-house dataset (without being split into train and test) given in Figure 6. As a result, the proposed model improved the accuracy of patient diagnosis with a promising result and reduced false positive recalls, minimizing unnecessary biopsies and adverse effects such as financial burden and psychological stress.

		Models						
		<u>Resnet50V2</u>			<u>ResNet101</u>			
<u>Balanced test dataset</u>	Predicted label		patient	healthy	total	Predicted label		
	True label	patient				healthy	total	patient
	patient	4	36	40	patient	2	38	40
	healthy	5	55	60	healthy	7	53	60
<u>Unbalanced test dataset</u>	Predicted label		patient	healthy	total	Predicted label		
	True label	patient				healthy	total	patient
	patient	4	36	40	patient	2	38	40
	healthy	121	1880	2001	healthy	91	1910	2001
<u>Balanced test dataset</u>	Predicted label		patient	healthy	total	Predicted label		
	True label	patient				healthy	total	patient
	patient	1	39	40	patient	4	36	40
	healthy	3	57	60	healthy	9	51	60
<u>Unbalanced test dataset</u>	Predicted label		patient	healthy	total	Predicted label		
	True label	patient				healthy	total	patient
	patient	1	39	40	patient	4	36	40
	healthy	143	1858	2001	healthy	240	1761	2001

Figure 4. The confusion matrix results of transfer learning-based models for balanced test dataset and unbalanced test dataset.

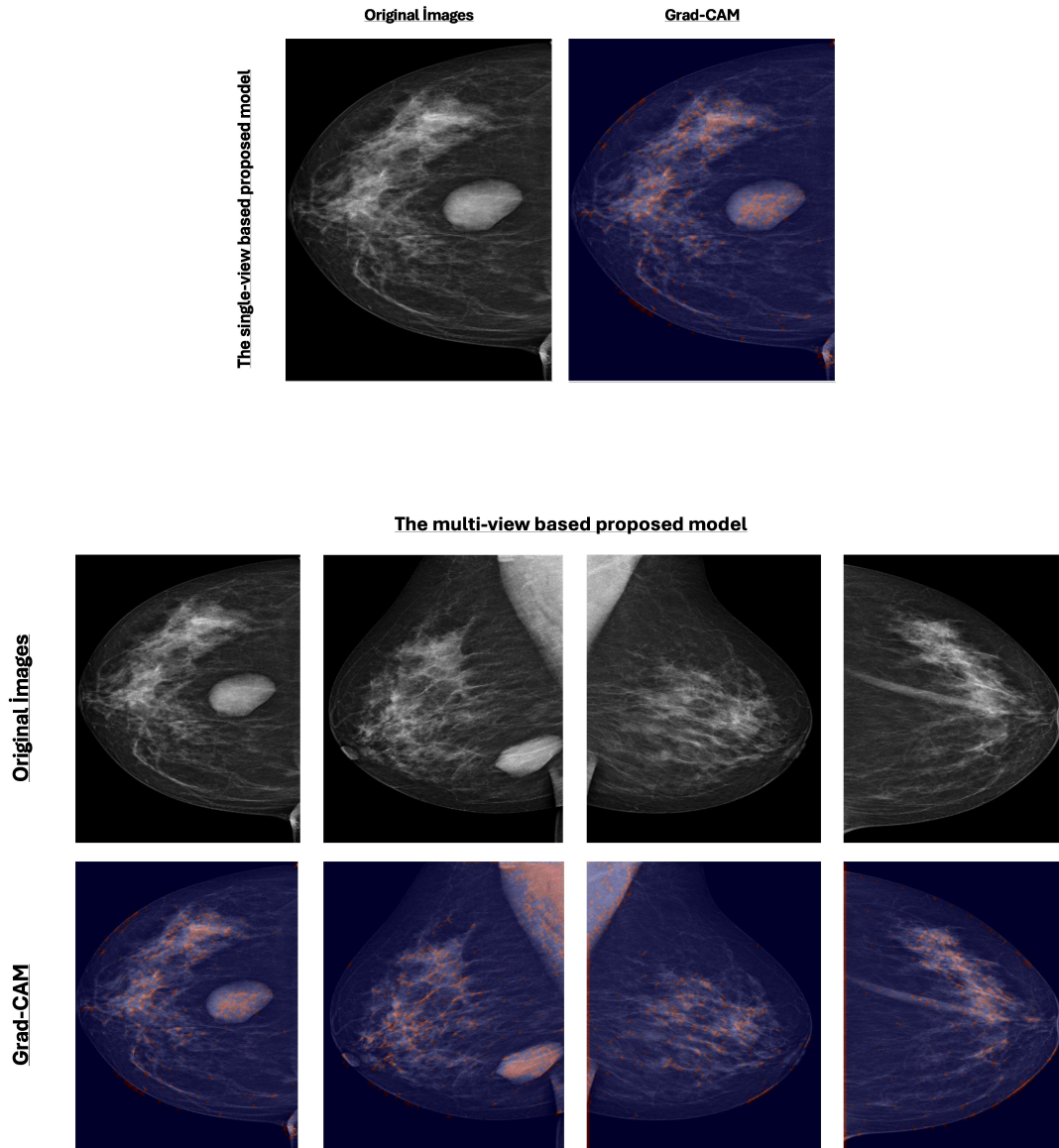


Figure 5. The results of Grad-CAM for single- and multiview-based models. The key and meaningful patterns in mammograms are marked using jet colormap, which plays a significant role in the model's decision-making process during classification.

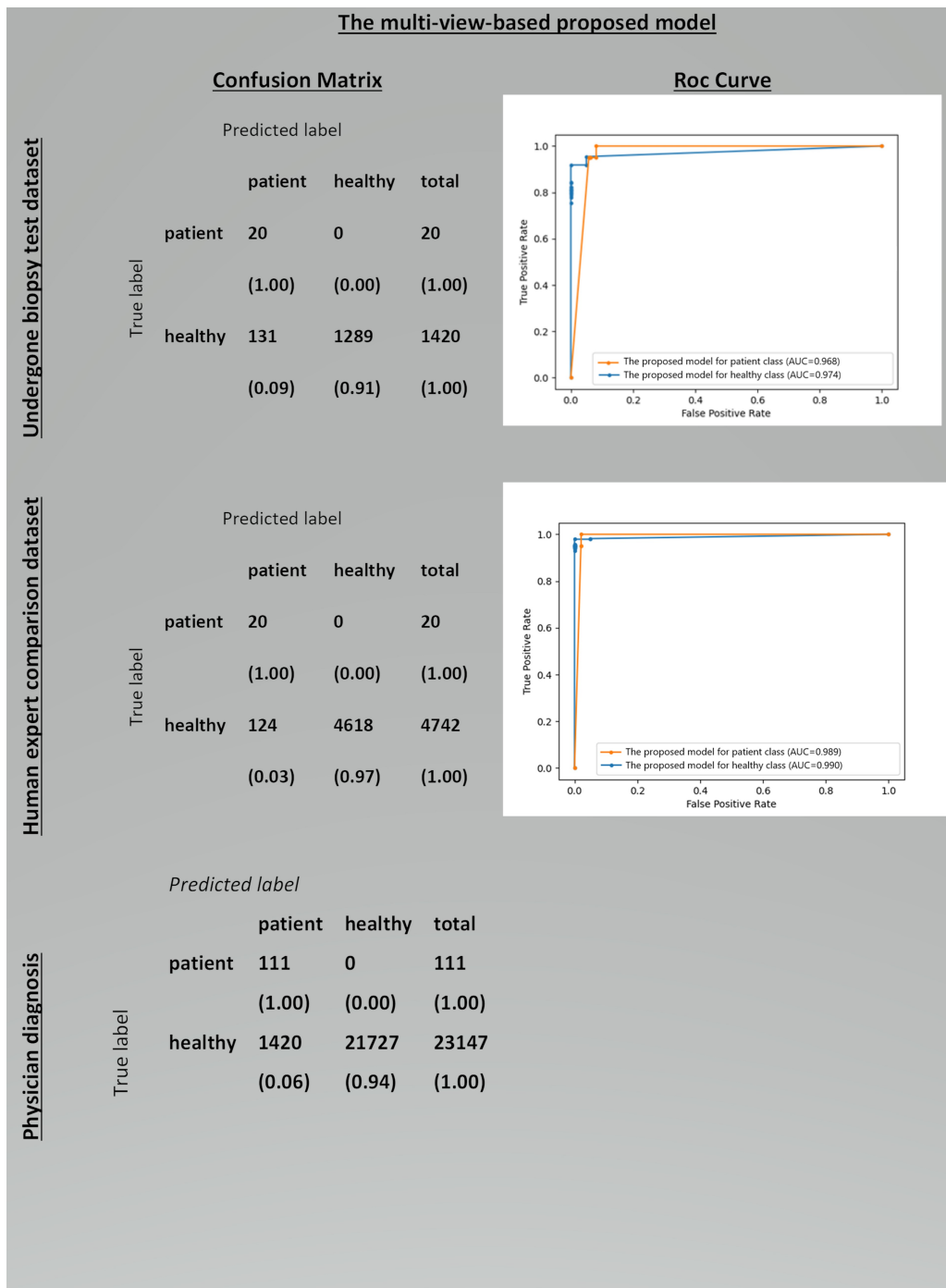


Figure 6. The ROC curve and confusion matrix results of the proposed model (multiview-based model) for patients who have undergone biopsy test dataset, human expert comparison test dataset, and physician diagnosis for 20% of all samples of the in-house dataset.

We assessed the performance of the proposed model in benign and malignant classification using the public INbreast dataset [36] for comparison with previous mammography studies. The proposed classification model was operated with single-view and multiviews (RCC, LCC, RMLO, LMLO) for mammography images. The proposed classification model was trained and tested on INbreast with extracted latent features by pretrained VAE in the pretext task. The latent features were extracted by the pretrained VAE on the in-house dataset and then fed into the classification model. Thus, we highlighted the transferability of the pretrained VAE on the in-house dataset to the INbreast dataset. Table 7 shows a comparative analysis of results between the proposed model and previous studies. According to the results, the multiview-based proposed model demonstrates competitive results against previous studies, achieving a high AUC of 0.98. Consequently, the proposed model shows high performance, generalization ability, and transferability of the encoded space between different datasets making it a valuable contribution to the field of breast cancer diagnosis.

Table 7. Comparison of the proposed models with previous studies on INbreast dataset for benign and malignant classification, considering accuracy, precision, recall, f1 with weighted average, and mean AUC.

dataset	model	AUC	accuracy	precision	recall	f1
INbreast	TSBN [49]	-	85.53	-	84.00	75.06
	CNN [50]	0.95	93.04	-	94.83	93.22
	Three-stage PAA [51]	0.96	-	-	-	-
	ECA-Net50 [52]	0.96	92.9	-	92.8	-
	The single-view based proposed model	0.96	92.50	92.87	92.50	92.62
	The multiview-based proposed model.	0.98	91.67	93.59	91.67	91.96

5. Discussion

In this study, we comprehensively analyze the proposed model examining the complexity of the model and regions' effect on decision-making, evaluating various test sets, including private and public datasets, and comparing their performance against state-of-the-art models and human experts. Our proposed model effectively addresses challenges associated with imbalanced and insufficient mammogram datasets, providing high, fast, and robust performance. However, it is noted that the limited number of positive samples (malignant, etc.) in mammograms still poses a challenge in assessing the deep models. The proposed model requires further training and testing on larger positive samples to address this limitation. Therefore, we will focus on developing synthesized mammography images, particularly for malignant cases, as data augmentation in future studies. Additionally, all mammogram views may not be available for patients. Therefore, image-to-image translation models can improve the model accuracy in cases where specific views are missing. Moreover, image translation tasks can potentially recognize and map corresponding structures and features across different mammogram views. Synthetic image generation techniques in future mammography studies have the potential to contribute significant benefits to overcome data limitations and enhance the overall efficacy of breast cancer detection and diagnosis. While the VAE exhibits a powerful approach in the representation and dimensionally reduction of images in a low latent space, Generative Adversarial Network (GAN)-based and diffusion models excel in synthetic image generation tasks. Therefore, in future research, we intend to leverage GAN and diffusion-based models for synthesizing mammography images, especially in cases where views are missing, incomplete, or imbalanced.

6. Conclusion

In this study, we propose an efficient and robust deep learning-based breast cancer diagnosis in two stages by overcoming over-fitting and underrepresentation problems. Firstly, VAE extracts encoded feature spaces by allowing the usage of large samples in an unsupervised manner, and then encoded spaces of four views were used by the simple classifier model. Thus, the proposed model achieves lower complexity by employing a low-dimensional encoded space vector (640,12) as input rather than utilizing high-dimensional mammography images. The proposed model incorporates a lightweight classifier model with only 338 trainable parameters, 389.20k FLOPs, and 0.354 s execution time for training for mammogram classification. Overall, the proposed model adopting a two-stage mechanism allows for the efficient processing of mammography data by maintaining higher accuracy, rapidity, and robustness even on the imbalanced test set than transfer learning-based models.

On the other hand, the first step in breast cancer diagnosis is mammography screening, as it often requires additional follow-up tests to arrive at a conclusive diagnosis. False positive recall occurs when a woman is recommended for additional imaging or biopsy, but most of them subsequently turn out benign. The high false positive recall rate is a significant concern, as it often leads to unnecessary follow-up tests, causing additional healthcare costs, anxiety, and stress for the patient. In this study, we observed a significant improvement in the proposed model in reducing the false positive recall rate for breast cancer diagnosis. The study shows that the proposed model decreased the false positive recall rate to less than half of the false positive recall rate with expert diagnosis. The reduced false positive recall rate would reduce costs and minimize psychological effects on patients. The significance of our results is that the experiments were conducted retrospectively in real-life data. This strongly encourages the idea that deep learning-based CAD and medical decision support systems might be highly productive and effective for the health economy and public health in mammography-based breast cancer screening. As a result, the proposed model has the potential to be a valuable tool in improving breast cancer diagnosis and reducing the negative impact on patients.

7. Declaration of competing interest

The authors declare no potential conflicts of interest regarding any financial support, research, authorship, or publication of this article.

8. Data availability

The authors do not have permission to share data. The source codes, models, and study results are available from the corresponding author upon reasonable request.

9. Acknowledgements

Computational experiments of this study were performed on the resources of the Artificial Intelligence and Big Data Application and Research Center at Erciyes University in Türkiye. Ethical clearance was obtained from Erciyes University in Türkiye, with the decision number 2019/64.

References

- [1] Li H, Zhuang S, Li Da, Zhao J, Ma Y. Benign and malignant classification of mammogram images based on deep learning. *Biomedical Signal Processing and Control* 2019; 51:347–354.
- [2] Ueda D, Yamamoto A, Onoda N, Takashima T, Noda S et al. Development and validation of a deep learning model for detection of breast cancers in mammography from multi-institutional datasets. *PLoS One* 2022; 17 (3):e0265751.
- [3] Bae MS, Moon WK, Chang JM, Koo HR, Kim WH et al. Breast cancer detected with screening US: reasons for nondetection at mammography. *Radiology* 2014;270 (2):369–77. Epub 2014/01/30. pmid:24471386.
- [4] Bird RE, Wallace TW, Yankaskas BC. Analysis of cancers missed at screening mammography. *Radiology* 1992;184 (3):613–7. Epub 1992/09/01. pmid:1509041.
- [5] Broeders MJ, Onland-Moret NC, Rijken HJ, Hendriks JH, Verbeek AL et al. Use of previous screening mammograms to identify features indicating cases that would have a possible gain in prognosis following earlier detection. *Eur J Cancer* 2003;39 (12):1770–5. Epub 2003/07/31. pmid:12888373.
- [6] Majid AS, de Paredes ES, Doherty RD, Sharma NR, Salvador X. Missed breast carcinoma: pitfalls and pearls. *Radiographics* 2003;23 (4):881–95. Epub 2003/07/11. pmid:12853663.
- [7] Weber RJ, van Bommel RM, Louwman MW, Nederend J, Voogd AC et al. Characteristics and prognosis of interval cancers after biennial screen-film or full-field digital screening mammography. *Breast Cancer Res Treat* 2016;158 (3):471–83. Epub 2016/07/10. pmid:27393617.
- [8] Duffy S, Tabar L, Smith R. The mammographic screening trials: commentary on the recent work by Olsen and Gotzsche. *Ca-a Cancer Journal for Clinicians* 2002; 52 (2):68–71.
- [9] Wu N, Phang J, Park J, Shen Y, Huang Z et al. Deep neural networks improve radiologists' performance in breast cancer screening. *IEEE transactions on medical imaging* 2019; 39 (4):1184–1194.
- [10] Clancy K, Aboutalib S, Mohamed A, Sumkin J, Wu S. Deep learning pre-training strategy for mammogram image classification: an evaluation study. *Journal of Digital Imaging* 2020; 33:1257–1265.
- [11] Hubbard R, Kerlikowske K, Flowers C, Yankaskas B, Zhu W et al. Cumulative probability of false-positive recall or biopsy recommendation after 10 years of screening mammography: a cohort study. *Annals of Internal Medicine* 2011; 155 (8):481–492.
- [12] Brodersen J, Thorsen H, Cockburn J. The adequacy of measurement of short and long-term consequences of false-positive screening mammography. *Journal of Medical Screening* 2004; 11 (1):39–44.
- [13] Siu A. US Preventive Services Task Force. Screening for breast cancer: US Preventive Services Task Force recommendation statement. *Annals of Internal Medicine* 2016; 164 (4):279–296.
- [14] Fenton J, Taplin S, Carney P, Abraham L, Sickles E et al. Influence of computer-aided detection on performance of screening mammography. *New England Journal of Medicine* 2007; 356 (14):1399–1409.
- [15] Fenton J, Abraham L, Taplin S, Geller B, Carney P et al. Breast Cancer Surveillance Consortium. Effectiveness of computer-aided detection in community mammography practice. *Journal of the National Cancer Institute* 2011; 103 (15):1152–1161.
- [16] Lehman C, Wellman R, Buist D, Kerlikowske K, Tosteson A et al. Breast Cancer Surveillance Consortium et al. Diagnostic accuracy of digital screening mammography with and without computer-aided detection. *JAMA Internal Medicine* 2015; 175 (11):1828–1837.
- [17] Mayo R, Kent D, Sen L, Kapoor M, Leung J et al. Reduction of false-positive markings on mammograms: a retrospective comparison study using an artificial intelligence-based CAD. *Journal of Digital Imaging* 2019; 32:618–624.
- [18] LeCun Y, Bengio Y, Hinton G. Deep learning. *nature* 2015; 521 (7553):436–444.

- [19] Karaman A, Pacal I, Basturk A, Akay B, Nalbantoglu U et al. Robust real-time polyp detection system design based on YOLO algorithms by optimizing activation functions and hyper-parameters with artificial bee colony (ABC). *Expert Systems with Applications* 2023; 221:119741.
- [20] Alici-Karaca D, Akay B, Yay A, Suna P, Nalbantoglu O et al. A new lightweight convolutional neural network for radiation-induced liver disease classification. *Biomedical Signal Processing and Control* 2022; 73:103463.
- [21] Karagoz M, Akay B, Basturk A, Karaboga D, Nalbantoglu O. An unsupervised transfer learning model based on convolutional auto encoder for non-alcoholic steatohepatitis activity scoring and fibrosis staging of liver histopathological images. *Neural Computing and Applications* 2023:1–15.
- [22] Becker A, Marcon M, Ghafoor S, Wurnig M, Frauenfelder T et al. Deep learning in mammography: diagnostic accuracy of a multipurpose image analysis software in the detection of breast cancer. *Investigative radiology* 2017; 52 (7):434–440.
- [23] Tang C, Cui X, Yu X, Yang F et al. Five Classifications of Mammography Images Based on Deep Cooperation Convolutional Neural Network. *American Scientific Research Journal of Engineering Technology and Sciences* 2019; 57:10–21.
- [24] Kooi T, Litjens G, Van Ginneken B, Gubern-Mérida A, Sánchez C et al. Large scale deep learning for computer aided detection of mammographic lesions. *Medical image analysis* 2017; 35:303–312.
- [25] Lotter W, Sorensen G, Cox DA. multi-scale CNN and curriculum learning strategy for mammogram classification. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: Third International Workshop, DLMIA 2017, and 7th International Workshop, ML-CDS 2017, Held in Conjunction with MICCAI 2017, Québec City, QC, Canada September 14, Proceedings 3 2017*; 169–177.
- [26] Shen L, Margolies L, Rothstein J, Fluder E, McBride R et al. Deep learning to improve breast cancer detection on screening mammography. *Scientific Reports* 2019; 9 (1):12495.
- [27] Bozkurt F. A deep and handcrafted features-based framework for diagnosis of COVID-19 from chest x-ray images. *Concurrency and Computation: Practice and Experience* 2022; 34 (5):e6725.
- [28] Kyono T, Gilbert F, Schaar M. MAMMO: A deep learning solution for facilitating radiologist-machine collaboration in breast cancer diagnosis. *arXiv preprint* 2018; arXiv:1811.02661.
- [29] Aboutalib S, Mohamed A, Berg W, Zuley M, Sumkin J et al. Deep learning to distinguish recalled but benign mammography images in breast cancer screening deep learning in mammography. *Clinical Cancer Research* 2018; 24 (23):5902–5909.
- [30] Kim HE, Kim H, Han BK, Kim K, Han K et al. Changes in cancer detection and false-positive recall in mammography using artificial intelligence: a retrospective, multireader study. *The Lancet Digital Health* 2020; 2 (3):e138–e148.
- [31] Adedigba AP, Adeshina SA, Aibinu AM. Performance Evaluation of Deep Learning Models on Mammogram Classification Using Small Dataset. *Bioengineering* 2022; 9 (4):161. <https://doi.org/10.3390/bioengineering9040161>
- [32] Maqsood S, Damasevicius R, Maskeliunas R. TTCNN: A Breast Cancer Detection and Classification towards Computer-Aided Diagnosis Using Digital Mammography in Early Stages. *Applied Sciences* 2022; 12 (7):3273. <https://doi.org/10.3390/app12073273>
- [33] Hamidinekoo A, Denton E, Rampun A, Honnor K, Zwigelaar R. Deep learning in mammography and breast histology, an overview and future trends. *Medical Image Analysis* 2018; 47:45–67.
- [34] Ragab D, Attallah O, Sharkas M, Ren J, Marshall S. A framework for breast cancer classification using multi-DCNNs. *Computers in Biology and Medicine* 2021; 131:104245.
- [35] Al-Mansour E, Hussain M, Aboalsamh H, Al-Ahmadi S. Comprehensive Analysis of Mammography Images Using Multi-Branch Attention Convolutional Neural Network. *Applied Sciences* 2023; 13 (24):12995.
- [36] Moreira I, Amaral I, Domingues I, Cardoso A, Cardoso M et al. Inbreast: toward a full-field digital mammographic database. *Academic radiology* 2012; 19 (2):236–248.

- [37] Suckling J, Parker J, Dance D, Astley S, Hutt I et al. Mammographic image analysis society (mias) database v1. 21. 2015.
- [38] Heath M, Bowyer K, Kopans D, Kegelmeyer Jr P, Moore R et al. Current status of the digital database for screening mammography. In *Digital Mammography*. Nijmegen 1998; pp. 457-460. Dordrecht: Springer Netherlands.
- [39] Lee R, Gimenez F, Hoogi A, Miyake K, Gorovoy M et al. A curated mammography data set for use in computer-aided detection and diagnosis research. *Scientific data* 2017; 4 (1):1–9.
- [40] Lopez M, Posada N, Moura D, Pollán R, Valiente J et al. BCDR: a breast cancer digital repository. In *15th International conference on experimental mechanics* 2012; 113–120.
- [41] Marcano-Cedeño A, Quintanilla-Domínguez J, Andina D. WBCD breast cancer database classification applying artificial metaplasticity neural network. *Expert Systems with Applications* 2011; 38 (8):9573–9579.
- [42] Huang SC, Pareek A, Jensen M, Lungren MP, Yeung S et al. Self-supervised learning for medical image classification: a systematic review and implementation guidelines. *npj Digital Medicine* 2023;6 (1). <https://doi.org/10.1038/s41746-023-00811-0>
- [43] He K, Zhang X, Ren S, Sun J. Identity Mappings in Deep Residual Networks. arXiv 2016.
- [44] Huang G, Liu Z, Maaten L, Weinberger K. Densely Connected Convolutional Networks. arXiv 2018.
- [45] Mingxing T, Quoc V. Le. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. arXiv 2020.
- [46] Kingma D, Welling M. Auto-encoding variational bayes. arXiv preprint 2013; arXiv:1312.6114.
- [47] Kazerouni A, Aghdam EK, Heidari M et al. Diffusion Models for Medical Image Analysis. A Comprehensive Survey. arXiv preprint 2022:1-28. <http://arxiv.org/abs/2211.07804>.
- [48] Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D et al. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision* 2017; 618-626.
- [49] Gong R, Lu Z, Shi J. Task-driven Self-supervised Bi-channel Networks Learning for Diagnosis of Breast Cancers with Mammography. arXiv preprint 2021;arXiv:2101.06228.
- [50] El Houby E, Yassin N. Malignant and nonmalignant classification of breast lesions in mammograms using convolutional neural networks. *Biomedical Signal Processing and Control* 2021; 70:102954.
- [51] Razali NF, Isa IS, Sulaiman SN, Karim NKA, Osman MK. Improvement of Breast Density Classifier based on CNN Features Extraction and SVM in Mammogram Images. *training* 2022; 7, 18.
- [52] Lou Q, Li Y, Qian Y, Lu F, Ma J. Mammogram classification based on a novel convolutional neural network with efficient channel attention. *Computers in Biology and Medicine* 2022; 150:106082.