Multi-View Attention-Based Late Fusion (MVALF) CADx System for Breast Cancer Using Deep Learning

Hina Iftikhar[1,2], Ahmad Raza Shahid[1,2], Basit Raza[1,2], Hasan Nasir Khan[1,2]

[1]*Medical Imaging and Diagnostics Laboratory (MID), National Centre of Artificial Intelligence (NCAI), Islamabad, Pakistan*

[2]*Department of Computer Science, COMSATS University Islamabad (CUI), Pakistan*

basit.raza@comsats.edu.pk

Abstract. Breast cancer is a leading cause of death among women. Early detection can significantly reduce the mortality rate among women and improve their prognosis. Mammography is the first line procedure for early diagnosis. In the early era, conventional Computer-Aided Diagnosis (CADx) systems for breast lesion diagnosis were based on just single view information. The last decade evidence the use of two views mammogram: Medio-Lateral Oblique (MLO) and Cranio-Caudal (CC) view for the CADx systems. Most recent studies show the effectiveness of four views of mammogram to train CADx system with feature fusion strategy for classification task. In this paper, we proposed an end-to-end Multi-View Attention-based Late Fusion (MVALF) CADx system that fused the obtained predictions of four view models, which is trained for each view separately. These separate models have different predictive ability for each class. The appropriate fusion of multi-view models can achieve better diagnosis performance. So, it is necessary to assign the proper weights to the multi-view classification models. To resolve this issue, attention-based weighting mechanism is adopted to assign the proper weights to trained models for fusion strategy. The proposed methodology is used for the classification of mammogram into normal, mass, calcification, malignant masses and benign masses. The publicly available datasets CBIS-DDSM and mini-MIAS are used for the experimentation. The results show that our proposed system achieved 0.996 AUC for normal vs. abnormal, 0.922 for mass vs. calcification and 0.896 for malignant vs. benign masses. Superior results are seen for the classification of malignant vs benign masses with our proposed approach, which is higher than the results using single view, two views and four views early fusion-based systems. The overall results of each level show the potential of multi-view late fusion with transfer learning in the diagnosis of breast cancer.

Key words: breast cancer, mammogram, four-view mammogram, information fusion, late fusion, transfer learning.

1. Introduction

Breast cancer is one of the most death-causing invasive diseases among women. In 2018, 2.1 million cases of breast cancer were recorded by the World Health Organization (WHO) and 627 000 women died of breast cancer, which is 6.5% of all cancer-related deaths in that year [49]. The death rate has been decreasing since the last few decades. The decrease is due to the advancement in early diagnosis, treatment, and awareness about the symptoms [37]. However, in the past years women death rate was still high due to the diagnosis is frequently still too late. Early diagnosis prevents the patient from invasive tumor and it also increases the survival rate by five to ten years. Mammography is a reliable and initial diagnostic method for early diagnosis of breast cancer. Mammograms are low energy X-rays of the breast and radiologist use it to identify the abnormalities in the breast. Breast screening has been performed on two views: Cranio-Caudal (CC) and Medio-lateral Oblique (MLO) of the left and right breast. CC view is top-down screening and MLO view is taken under 45 degrees [19, 46].

Breast cancer includes calcifications and masses. Calcifications are the deposits of calcium in woman's breast and can be shown clearly as white dots in the screening process. There are further two types of calcification: macrocalcifications and microcalcifications [29]. Macrocalcifications are large white spots that are considered as the non-cancerous and are dispersed randomly in the breast. Microcalcifications are the small white deposits of calcium and are mostly considered as non-cancerous. Although, if these deposits are clustered together then this may be alarming as early breast cancer [47]. Masses are the lesions in woman's breast that can be cancerous or non-cancerous. The benign masses, that is, the non-cancerous ones are smooth or oval in shape with circumscribed boundary. The masses that are known as cancerous, that is, malignant, spread into their neighborhood by forming spicules. Diagnosis of masses is a challenging task due to the variations in their shape, appearance and size [29]. However, manual detection of the symptoms of cancer using mammograms is susceptible to human errors and laborious due to variability. In the current technical era, Computer-Aided Diagnosis (CADx) systems are used for reliable and fast diagnosis of disease. CADx systems have potential to reduce the heavy workload of the radiologist. These systems served as a second reader to improve the accuracy of the final decision.

In the last few years, deep learning has become one of the most successful methods in computer vision tasks [25]. Especially, Convolution Neural Networks (CNNs) have been proved as the reason for the boom of deep learning. Deep learning-based CADx systems [11, 13, 36] have attained the level appropriate for producing more realistic solutions in tumor diagnosis. The four major steps are involved in CNN-based CADx systems to assist the radiologist in making the final decision [50]. Firstly, the preprocessing step is performed to remove the noise from images. In the second step the region of the tumor is segmented out from the image. The feature extraction task is carried out for the region of the tumor in the third step. In the last step, the tumor classification task is performed. Traditional CADx systems were based on manual handcrafted features, which have shown the limited accuracy for complex problems. Several studies have been performed to build a CADx system for breast lesion classification and detection. In 2013, Kozegar et al. [27] used the traditional feature selection and machine learning techniques for iterative breast segmentation. Their proposed system had the ability to classify the segmented region of the lesion. Other results and the literature on the segmentation-based mammography analysis systems can be found for example in [7].

A number of recent studies have been published on fully automated CNN-based

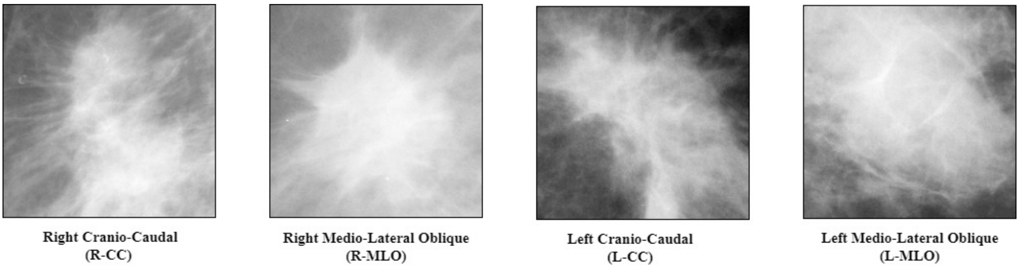

Fig. 1. Examples of ROIs of four mammographic views in the CBIS-DDSM dataset.

CADx system for tumor detection and classification tasks [8, 11, 13, 22]. The deep learning-based CADx systems have been introduced for different medical domains, for example brain tumor detection, lung disease diagnosis, lymph node, breast cancer diagnosis, and many others. We mainly focused on breast lesion classification [4, 5, 8, 10, 18, 32, 33, 34]. CNN is an end-to-end supervised learning process without any descriptor on the whole raw image. CNN learns the discriminant features automatically and its most surprising characteristic is that it achieves good generalization for vision tasks with the 2D input images [29].

Deep CNNs are more complex architectures than CNNs and require a large amount of data to train a model. Due to high computation complexity, training the model on a small amount of data leads to overfit. To overcome this problem, the transfer learning is used. Transfer learning is a technique of transferring the knowledge from one domain to another domain. In medical imaging, where small datasets are available, transferring of knowledge from another domain has been very effective. The knowledge transfer consists in using a network which is pre-trained on images coming from some domain. There are two modes of transferring the knowledge: first, transferring the knowledge from the medical domain, and second, transferring the knowledge from some other domain, for example, the domain of natural images. The current evidences show the high performance of using pre-trained models to achieve better accuracy [12, 22, 29, 36]. In recent years, the authors achieved reasonable accuracies for breast cancer detection and classification task using the transfer learning techniques [2, 12, 29].

Information extracted from multi-view images is more significant for decision making than that extracted from a single view. Multi-view mammograms are used by the radiologist to make a final decision. We will overcome the problem of not gaining profit from the multi-view nature of mammograms in CC and MLO views. In the previous studies, most of the research has been based on single-view images in the development of a CADx system. Breast screening provides the four views: Right MLO (R-MLO), Left MLO (L-MLO), Right CC (R-CC) and Left CC (L-CC) of mammograms as shown in Fig.1. Radiologists always start from the CC view, and when they find any abnormalities in this view they check the information from all views for making a final decision. Most of the studies focused on the CADx systems based on just two views (CC and MLO mammograms) [8,9]. Recent studies focused on the four-views information-based CADx systems which achieved the best accuracy for breast lesion classification. Multi-view information fusion mainly focuses on the analysis of mammograms using CC and MLO views of the left and right breast. Information fusion is based on two strategies: early fusion and late fusion. Early fusion is used to fuse the extracted features of different models and late fusion is based on combining the results of classification of the multiple models. The results of two-views CC and MLO models are fused to classify the breast lesion into malignant and benign [17] and produce significant results in terms of accuracy for the classification task. In the recent study, Khan et al. proposed a Multi-View Feature Fusion (MVFF) based CADx system that includes three stages [26].

Nevertheless, the multi-view information fusion has gained more success in recent years in context of breast cancer. According to the previous studies on the mammographic views, the breast screening is performed on bilateral view, CC and MLO, of right and left breast. Bassett et al. [6] believed that the CC view, with particular emphasis on the medial view imaging, conveys the most significant information. The CC view is the medial view in screening and has a great aspect of deep tissues to be visualized. Normally, these deep tissues in medial aspect of breast are not possible to capture in the MLO lateral view [6, 19, 45]. However, both projections are complementary to capture the most accurate information. In current era, one of the key challenges is to overcome the high False Positive Rate (FPR) that existed in the previous CADx systems. The four-view fusion systems reduce the high FPR [24]. Wei et al., in 2011, proposed a computer-aided detection system of four view information fusion for mass detection [48]. In comparison with single-view their system performed better in terms of accuracy and FPR. In 2015, Yanfeng Li et al. [30] proposed a bilateral image analysis scheme for mass detection to reduce the FPR. The results show the significance of proposed system in which the approach of bilateral analysis for mass detection reduce the FPR. Among the methods of breast mass detection [30, 31, 48] few of the research works on the multi-view information fusion for classification task [41, 51] use the multi-agent and feature fusion approach, respectively. The results show that the decision fusion mechanism reduces the problem for the classification task. Since the many masses are difficult to identify in one view and give more information in the other view, the late fusion approach reduces the FPR [51]. The four-view information fusion-based CADx systems can be considered as the simulation of radiologist's interpretation and are able to serve as a second reader.

The main focus of this research is to utilize the effectiveness of attention-based weighted late fusion in CADx systems to reduce the false positive rate for mammogram classification. In the late fusion, separate deep CNN models are trained for each view, i.e., L-CC, R-CC, L-MLO, and R-MLO of mammograms. The pre-trained CNN architectures are used to fine-tune on mammograms to classify the breast lesions. The obtained results of trained models are fused to achieve the best performance in terms of classification of breast masses. The proposed Multi-View Attention-based Late Fusion (MVALF) model outperforms the multi-view model and provides the state-of-the-art technique for mass classification tasks. Our proposed system is evaluated on benchmark dataset CBIS-DDSM (references will be given in Subsection 3.1). The main contributions of this research are as follows.

- A novel attention-based weighting algorithm is proposed to increase the effectiveness of our multi-view late fusion-based CADx system. Each model has its own predictive ability, therefore assigning the equal weights to all the models is not a good approach. In this regard, attention-based weighting algorithm assigns the higher weights to those models which have higher sensitivity.
- A Multi-View Attention-based Late Fusion (MVALF) system is proposed for the diagnosis of breast cancer. The main contribution of this work is to efficiently take the advantage of the four mammographic views of each patient because conventionally developed CADx systems have used two views information and ignored the importance of late fusion of separately trained multi-view models. The proposed MVALF approach yields good performance measures and shows the effectiveness of late fusion for four-view models to reduce the false positive rate.
- The end-to-end system is proposed, which is not limited to just classify the mammogram into cancerous or non-cancerous. The proposed MVALF-based CADx has the ability to classify the mammogram at different levels. The first level is about the classification into normal and abnormal. At the second level, the mammograms are classified on the basis of their abnormality. Finally, at the last level the mammograms are classified according to their level of pathology.

This paper proceeds as follows: Section 2 presents the literature review, Section 3 describes the methodology, Section 4 gives the details of experimentations and the results are discussed in it, and finally Section 5 concludes the paper.

2. Literature review

Many studies have been published on CNN-based CADx systems for breast cancer classification. Chakraborty et al. [10] proposed a novel method that was used to detect non-palpable breast cancer. The automatic diagnosis is difficult due to variability in size, irregularities in shape and occlusions in breast tissue. The proposed method classifies the masses along with characterized oriented tissue and multi-resolution features using Gray-Level Co-Occurrence Matrix (GLCM) and Angle Co-Occurrence Matrix (ACM). Recently Ribli et al. used fast Region-based CNN (R-CNN) for mass detection and classification into malignant and benign [34]. They achieved state-of-the-art performance on the INBreast dataset and their system reached high sensitivity with few false negatives, and with AUC of 0.85. Al-masni et al. [4] in 2018 proposed a YOLO-based CADx system for breast cancer detection. Their CADx system detects the location and diagnoses the masses and classifies them into benign and malignant class using CNN. The last fully connected layer of architecture is trained on ROI-based mammograms. In 2017, Lotter et al. [32] proposed a methodology for breast cancer mass detection and segmentation. The author proposed a patch-based CNN classifier for lesion classification and achieved 0.92 AUC. In another study, Akselrod-Ballin et al. [3] used fast R-CNN to detect the breast abnormalities on the INBreast dataset and achieved TPR 0.93 and FPI 0.56 for mass mammograms.

Chougrad et al. [12] explored the importance of a pre-trained model and determined the best strategy to train CNNs architectures. They focused on the use of the pre-trained model for classification of breast lesions. The pre-trained models VGG16, ResNet50 and InceptionV3were used instead of random initialization. The proposed full framework for breast cancer screening achieved AUC of 0.9 for masses classification into benign and malignant. Recently, in 2019 Hua Li et al. [29] proposed an improved DenseNet for mammogram classification into benign and malignant class based on a deep learning pre-trained model. The proposed model, DenseNet II, performs the classification task accurately and effectively. AlexNet, VGGNet, GoogleNet, DenseNet and the proposed DenseNet II were trained on processed data. The authors claimed that the system was robust and good at generalization. In the same year, Agarwal et al. [2] proposed a patch-based CNN for automated mass detection. The transfer learning models (ResNet50, VGG16, Inception) were used to train on the CBIS-DDSM dataset and the evaluation revealed that InceptionV3 performed the best on automatic mass detection. The evaluation results demonstrated that patch-based transfer learning CNNs performed substantially well for mass detection on CBIS-DDSM.

While the previous networks were trained on a single view and two views of mammograms, recent years witnessed great advancement in multi-view information-based CADx systems and information fusion of different models attained the state-of-the-art performance [1]. Carneiro et al. proposed a multi-view based CADx system for breast cancer risk prediction using two views of mammograms [8, 9]. Tan et al. proposed a four-view based feature fusion model for near term breast cancer risk prediction [43]. Jiao et al. [23] created and trained a CNN-based CADx system by combining the results of two classifiers and classified the mass mammograms into malignant and benign. They concluded that the results obtained from multi-view model fusion achieved higher classification performance than that using a single view. A similar work has been proposed in 2019, Khan et al. used the early fusion strategy to diagnose the tumor in breast. They utilized the extracted mammographic information of four views. The system had the capability to classify the tumor into malignant and benign. They achieved the classification accuracy of 77% and AUC of 0.84 [26].

In our work, we focus on the attention-based weighted late fusion technique by utilizing the four views of mammogram.

3. Materials and methods

In this section, we first describe the publicly available datasets, data pre-processing, data augmentation, CNN architectures used for our proposed system, evaluation metrics for testing the performance of CADx system, and the overall methodology with attention-based weighting algorithm.

3.1. Dataset

In this study, the dataset that we used to perform the experiments on our proposed MVALF based CADx system were CBIS-DDSM and mini-MIAS. DDSM [20,21] was the first version of CBIS-DDSM. It contains the digital images of mammographic screening of 2 620 patients. It contains the verified pathology information (benign and malignant) of each case. The four view information for each case is available with MLO and CC views of the left and right breasts. CBIS-DDSM [39, 44] is a subset of images selected from the original dataset and curated by expert radiologists [15,28]. It has been used for the training and also for performance evaluation of the proposed MVALF system. The images are compressed and converted into DICOM format. The Mammographic Image Analysis Society (MIAS) is another curated digital mammographic dataset of breast lesions [40] with images of resolution 1024×1024 pixels. The analysis is performed on extracted ROI images of 224×224 pixels of mini-MIAS [14] for normal class. Table 1 shows the detailed description of the train and test split of mammographic dataset using four views.

3.2. Data pre-processing

In order to enhance the performance of the CADx system, we need to perform some mandatory task to make the data clarity better for training a model. We used the ROI-based mammograms from the publicly available dataset. We also performed image

Tab. 1. Dataset description of mammograms in CBIS-DDSM and mini-MIAS.

Abnormality Type	**Training**	**Testing**	**Total**
Normal	3008	512	3520
Abnormal	2864	12	3376
Calcification	1546	256	1802
Mass	1318	256	1574

pre-processing such as contrast and brightness enhancement, resizing and image normalization on the selected datasets. The pre-processing helps to achieve better classification accuracy.

3.3. Data augmentation

Deep learning models perform better when we have a large amount of data. The data in medical imaging domain are very limited in size. The scarcity of the dataset in training the deep learning models leads them to overfit. Data enhancement or data augmentation is an approach to help increase dataset size. It also leads to better robustness and helps to prevent overfitting when training is done on a smaller dataset. We performed data enhancement on our dataset to improve the performance of the system. The images were augmented by rotating by a 0-45 degree angle, the shearing in the range of 0.2, zooming in the range of 0.2, horizontal shifting in the range of 0.2 of the image width, and vertical shifting in the range of 0.2 of the image height. The horizontal flip and vertical flip were performed, and to fill newly created pixels the fill mode strategy was applied. The augmented images were different from each other and there was no exact copy of any of the original images.

3.4. CNN architectures

CNNs are trained on images to recognize the visual pattern with minimal preprocessing. We analyzed the well-known transfer learning models on ImageNet (natural images) [16] along with fine-tuned layers on mammograms. The ImageNet is a dataset containing millions of natural images. ImageNet Large Scale Visual Recognition Challenge (ILSVRC) is a competition for classification and object detection held every year [1, 16]. We have evaluated the performance in the classification of mammograms of the three well known CNN architectures that have been the winners of ILSVRC.

3.4.1. VGGNet

Simonyan et al. in Visual Geometry Group (VGG) from University of Oxford proposed VGGNet [38]. It was much deeper than the previous networks. They used the filter size of 3×3 instead of 5×5, 7×7 or 11×11, as in AlexNet [35]. The network was runner-up of ILSVRC 2015 challenge for image classification with top five error rate of 7.3% and it also performed best in the image localization task. There are many versions of VGGNet; however, VGG16 and VGG19 are the most popular. VGG19 performed better than VGG16 although it is computationally more expensive.

3.4.2. InceptionV3

GoogLeNet was the winner of ILSVRC in 2014 for image classification with top five error rate of 6.7%. Szegedy et al. [42] from Google designed a much deeper network with

22 layers. A novel element known as the inception module was introduced to reduce the computational complexity of the network. In this network the number of parameters was reduced from 60 million (AlexNet) to 4 million.

3.4.3. ResNet50

Residual block network won the ILSVRC 2015 with 3.6% error rate [35]. It is a much deeper network than others with 152 layers. It consists of a residual block where each block contains two 3×3 convolution layers. Skip connections are used in ResNet to remove the vanishing gradient problem [25]. ResNet50 achieved good performance in all tasks such as localization, classification and object detection in ILSVRC.

3.5. Performance Evaluation

The CADx system is evaluated for the correct classification of mammograms. The model is evaluated using sensitivity, specificity, and accuracy as the measures of classification quality. Sensitivity is the True Positive Rate (TPR) and specificity is the True Negative Rate (TNR). Accuracy is measured by the performance of the model in terms of general correctness. We also evaluated the model using the ROC curve and the Area Under the ROC Curve (AUC). ROC curve is a 2-axis presentation with sensitivity on the y-axis and False Positive Rate (FPR) on the x-axis that is calculated as $1 -$ specificity. In the following Equations (1) to (3), sensitivity, specificity and accuracy are calculated in terms of the numbers of True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) classifications.

$$\text{Sensitivity} = \text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}} \,, \tag{1}$$

$$\text{Specificity} = \text{TNR} = \frac{\text{TN}}{\text{TN} + \text{FP}} \,, \tag{2}$$

$$\text{Accuracy} = \text{ACC} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{FP} + \text{TN}} \,. \tag{3}$$

3.6. Proposed four-view model fusion

A fully automated deep CNN-based framework is proposed for mammogram classification using Regions of Interest (ROI's) as input images. Firstly, the dataset is divided into four views: L-CC, R-CC, L-MLO, and R-MLO. Afterwards, the four models are trained on each view separately for all patients. The obtained results from four models of all views are combined to generate the final prediction for mass classification. The prediction fusion of multiple models is known as late fusion [24].

We applied the late fusion strategy on the trained model of each view to generate the final decision. The radiologists also examined the mammograms in the same manner to make the final decision about the abnormalities. We performed many experiments with variations in hyperparameters. The experiments were made with four view-based CADx systems with various pre-trained models along with the fine-tuning strategy. Fig. 2 shows the proposed MVALF based CADx system for the breast mammogram classification.

3.6.1. Network training

The first stage of the proposed system is related to the model training. At this stage, we fine-tuned the deep CNN models for each view, i.e. L-CC, R-CC, L-MLO, and R-MLO, separately. The best fitted fine-tuned layers have been selected after performing various experiments using different numbers of freezing layers. We also performed experiments for two-view and multi-view cases using pre-trained models. Finally, we concluded from the results that the pre-trained models performed better on multi-view information while the number of datasets was limited. It can be observed that the transferring of knowledge from one domain to another domain helps to achieve better accuracy.

3.6.2. Multiview late fusion strategy

The last level of our system represents the fusion of four view results, which were obtained from the model training phase of each view separately. In breast cancer the screening mammograms are taken from two angles: MLO and CC of left and right breasts. The radiologist makes a final decision after viewing the information from four views. Our proposed CADx system is capable of classifying the mammograms using the four views. Afterwards, the results of all models are fused using the attention-based weighted late fusion strategy and the final decision of the diagnostic task is achieved. The details of the personalized weighting algorithm to prioritize the models are discussed in the next paragraphs.

Attention-based weighting algorithm After training the M models (where $M = 4$) on the four views of mammogram, they have the ability to classify the unseen data into the respective binary classes. Their output is fused to make the final decision. Rather than considering the information of all views equally, the Attention based Weighting Algorithm (AWA) has been adopted. It calculates the weights of predictive score for each view of the models based on their sensitivity to increase the TPR and decrease the FPR.

Let $\mathrm{model}_1, \mathrm{model}_2, \mathrm{model}_3, \ldots, \mathrm{model}_M$ be the M models and $R_1, R_2, R_3, \ldots, R_n$ be the classification results, each of the specific model. Suppose that C is the number of classes of the given dataset labelled as $\mathrm{class}_1, \mathrm{class}_2, \mathrm{class}_3, \ldots, \mathrm{class}_C$. The matrix $W = (w_m)$, $1 \leq m \leq M$ is the weight matrix of M models. The testing image is classified by assigning the label of the model according to the highest score.

In our proposed framework, W is calculated based on TPR. According to the previous studies on the mammogram views, the breast screening is performed on bilateral view

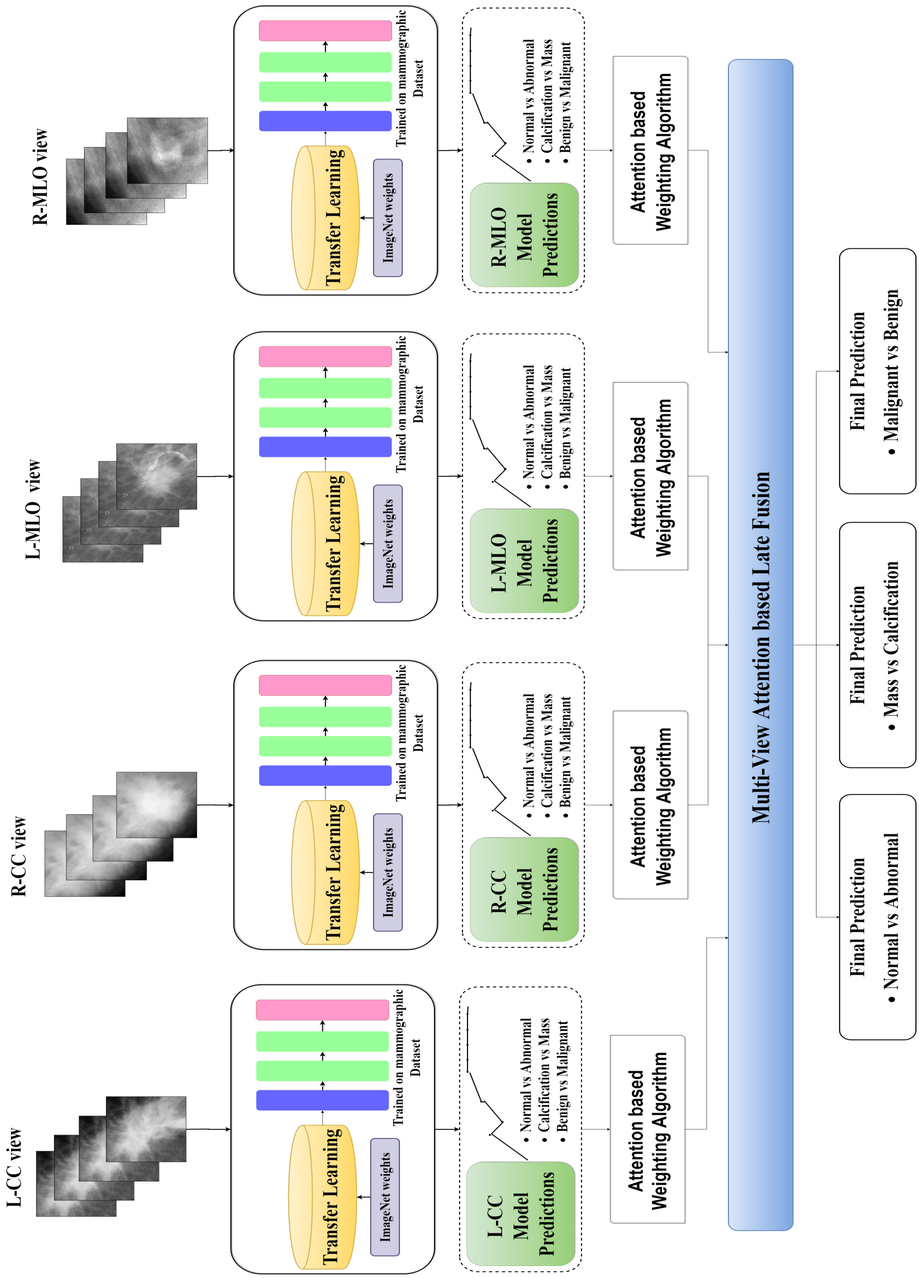

Fig. 2. Proposed Multi-View Attention based Late Fusion (MVALF) CADx system using pre-trained model.

CC and MLO of right and left breast. However, both projections are important to capture the more accurate information. The highest weight is assigned to the view with the highest sensitivity. In our case, the total number of classes is $C = 2$ and the number of models is $M = 4$. The pseudo code for our AWA is presented in Algorithm 3.1.

Algorithm 3.1 Attention-based Weighting Algorithm

for $\text{model}_m \leftarrow \text{model}_1$ **to** model_M **do**
 $\text{TruePos}(m) \leftarrow \langle\text{number of true positive instances in}\rangle(\text{model}_m)$
 $\text{FalseNeg}(m) \leftarrow \langle\text{number of false negative instances in}\rangle(\text{model}_m)$
 $\text{SensitivityM}(m) \leftarrow \text{TruePos}(m)/(\text{TruePos}(m) + \text{FalseNeg}(m))$
 $W(m)_{\text{sen}} \leftarrow \text{SensitivityM}(m)$
end for

4. Results and discussion

In this study, we used the attention-based late fusion strategy and evaluated the different CNN architectures for the classification of mammograms into three levels: mammogram classification, abnormality classification, and pathology classification. Furthermore, we performed experiments on a single view, two views and four views with the early fusion strategy for the comparative study with our proposed CADx system.

4.1. Experimental setup

In the experimental environment, the input size of the ROI image was 224×224. The ROI-based images were pre-processed before training on the CNN architectures. We used the stochastic gradient descent optimization algorithm with 0.0001 learning rate with a momentum of 0.9. The categorical-cross entropy was used as the loss function and the batch size was set between 20 to 50 for training. The dataset had a split of 0.2 for the validation set to evaluate the performance of the correct classification of mammograms. We used the experimental setup for training our models with the specification of NVIDIA Tesla P100, 16 gigabytes of memory, CUDA 10.1 version, Keras 2.2.5 version with TensorFlow 1.15.0 at the backend. The stopping criteria for training the model was set to 200 epochs with the patience level of 15.

4.2. Transfer learning and fine tuning

The transfer learning technique is used in our proposed methodology with fine-tuning strategy. The state-of-the-art pre-trained models (i.e. VGGNet, GoogleNet, ResNet) were trained on the public dataset of ImageNet that contains the natural images of 1000 classes. We removed the last fully connected classification layer of the pre-trained

Tab. 2. The total number of parameters that need to be trained on mammograms using CNN models.

CNN Models	Total Layers	Freezing Layer	Trainable Parameters	Batch Size
VGG19	22	14	14 158 848	50
InceptionV3	311	170	16 338 816	50
ResNet50	175	100	19 452 928	50

models and added two fully connected layers. The first layer has 300 connections and the second layer is used for final classification with two neurons. The approach of freezing layers in the pre-trained model reduces the number of trainable parameters. This helps overcome the problem of computational complexity in deep CNN models. The last, fully connected layers that are fine-tuned on mammograms surpass the overfitting which occurs due to random initialization in deep CNN networks.

The Table 2 shows the total number of layers, freezing layers of pre-trained models, total number of trainable parameters and batch size which was used in our experiments.

4.3. Monitoring the performance of our model

The basic structure of our proposed model is shown in Fig. 2. Our proposed MVALF based CADx system classifies the mammograms at three levels. The first level presents the classification of normal and abnormal mammograms. The second level describes the classification of abnormality into calcification and mass classes. The last level is about the classification of pathology into malignant and benign classes.

4.3.1. Classification into Normal and Abnormal

In the first level, classification of *Normal* and *Abnormal* classes is performed using the proposed MVALF based CADx system. The MVALF based CADx system outperformed the single view, two views and four views-based early fusion. Table 3 shows the performance of the proposed model. The model achieved a good balance between TPR and FPR. The use of transfer learning improves the performance of the proposed system. The four-view models use the weighted information fusion strategy on the basis of TPR, that helps to achieve the AUC of 0.996 shown in Fig. 3. Our proposed MVALF performed better on all the pre-trained models. InceptionV3 and ResNet50 performs slightly better with respect to VGG19. The achievements of the proposed model in comparison to previous studies are shown in Table 6. The proposed MVALF based CADx system performs 7% better than multi-view, two-view and single-view feature fusion.

4.3.2. Classification into Mass and Calcification

Secondly, experiments were performed to classify the abnormality into *Calcifications* and *Masses*. The experimental results in Table 4 show the preformance of the proposed

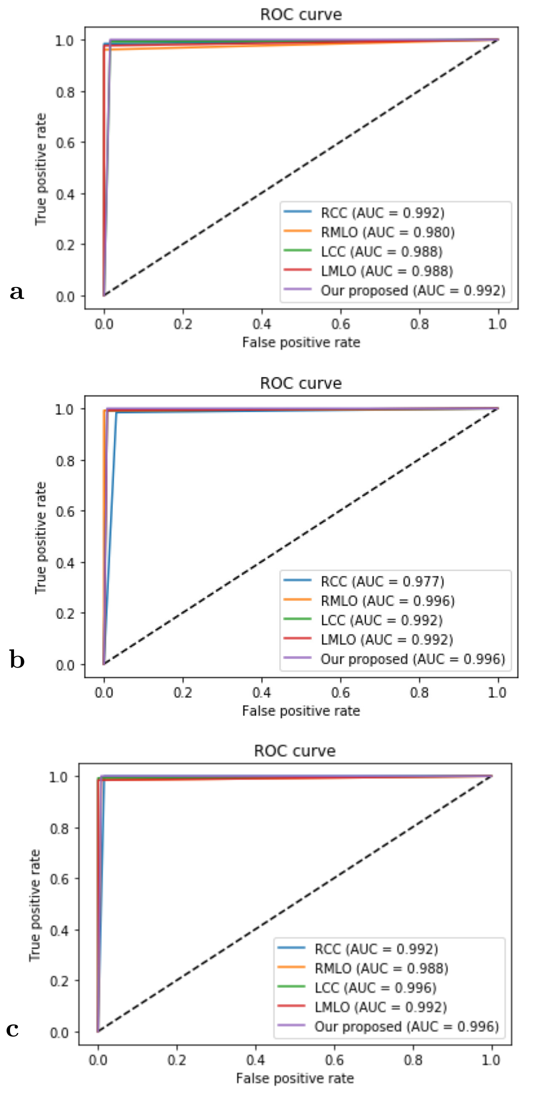

Fig. 3. ROC plotting for *Normal* and *Abnormal* classification. The testing performance of (**a**) VGG19, (**b**) InceptionV3, and (**c**) ResNet50 is presented, using the proposed MVALF-based CADx system.

Tab. 3. Performance measures of proposed MVALF for the classification of *Normal* vs. *Abnormal* mammograms.

Models	Views	Training Accuracy	Testing Accuracy	Sensitivity	Specificity	AUC
VGG19	R-CC	96.5%±0.88%	99.22%±0.68%	98.46%	100%	0.992
	L-CC	99.33%±0.57%	98.83%±0.22%	99.21%	98.45%	0.988
	L-MLO	99.00%±0.98%	98.83%±0.90%	97.71%	100%	0.989
	R-MLO	99.54%±0.22%	98.05%±0.90%	96.24%	100.00%	0.981
	Proposed Multiview (Late Fusion)	–	99.22%±0.78%	100%	98.44%	0.992
InceptionV3	R-CC	98.01%±1.2%	97.66%±1.50%	98.41%	96.92%	0.977
	L-CC	99.26%±0.53%	99.22%±0.71%	99.22%	99.22%	0.992
	L-MLO	99.93%±0.07%	99.22%±0.41%	99.22%	99.22%	0.992
	R-MLO	99.99%±0.10%	99.61%±0.59%	99.22%	99.00%	0.996
	Proposed Multiview (Late Fusion)	–	99.61%±0.29%	100%	99.22%	0.996
ResNet50	R-CC	98.45%±1.50%	99.22%±0.11%	100%	98.46%	0.992
	L-CC	97.44%±2.10%	99.61%±0.30%	99.22%	100%	0.996
	L-MLO	99.56%±0.15%	99.22%±0.13%	98.46%	100%	0.992
	R-MLO	98.28%±1.17%	98.83%±1.23%	98.45%	99.21%	0.988
	Proposed Multiview (Late Fusion)	–	**99.61%±100%**	**100%**	**99.22%**	**0.996**

MVALF-based CADx system. The late fusion of four-view models with their attentional mechanism VGG19 performs better with our proposed late fusion strategy in terms of AUC. However, the MVALF model achieved higher specificity with InceptionV3 in contrast with low sensitivity as compared to VGG19. The main reason behind the best performance of VGG19 for abnormality classification is the good quality of models for each view, i.e. R-CC, L-CC, L-MLO and R-MLO. The weights are assigned on the basis of sensitivity, as each separate model in VGG19 has high sensitivity, so that the model with higher weights improves the overall performance of the system. The model achieves the AUC of 0.922, testing accuracy of 92.12%, sensitivity of 93.55%, and specificity of 90.91%. Fig. 4 shows the ROC curvec of VGG19, InceptionV3 and ResNet50, and as it is clearly shown in the figure, this ensemble of the weighted information of all the views leads to achieving good performance in terms of AUC. The comparison study of the proposed model and the previous approach is shown in Table 6. This study shows the clear difference between the impact of different transfer learning models. The depth of each model has a different impact on the results of the classification task. The VGG19 with very few trainable parameters has achieved good accuracy and AUC for the abnormality classification.

4.3.3. Classification into Malignant and Benign

We performed different experiments for the two-class classification into *Benign* masses and *Malignant* masses. Table 5 shows the different experimental results of each view separately and for our proposed MVALF-based CADx system. The proposed system performed best for the classification task and achieved AUC of 0.896, testing accuracy of

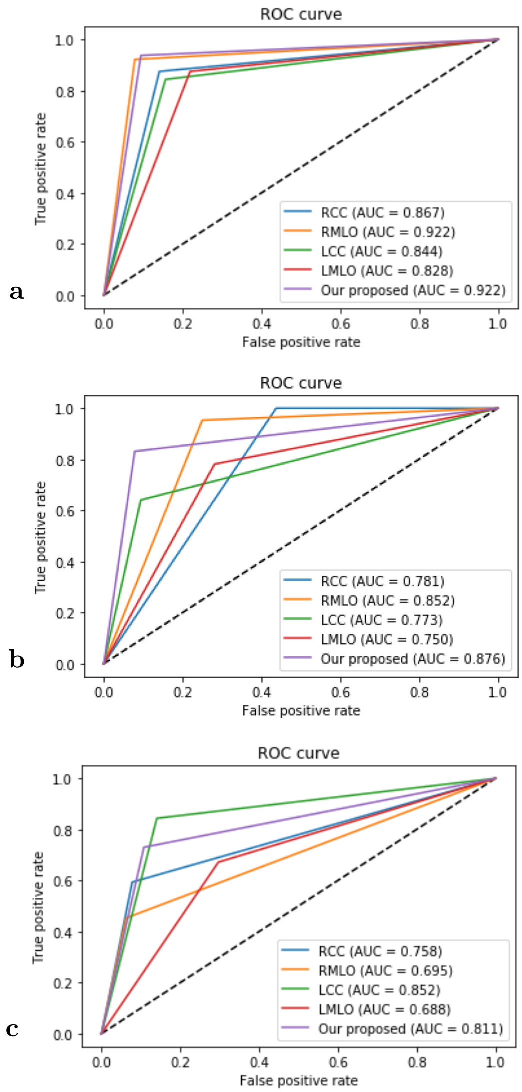

Fig. 4. ROC plotting for *Calcification* and *Mass* classification. The testing performance of (**a**) VGG19, (**b**) InceptionV3, and (**c**) ResNet50 is presented, using the proposed MVALF based CADx system.

Tab. 4. Performance measures of proposed MVALF for the classification of *Mass* vs. *Calcification* mammograms.

Models	Views	Training Accuracy	Testing Accuracy	Sensitivity	Specificity	AUC
VGG19	R-CC	95.09%±1.53%	86.72%±.23%	87.30%	86.15%	0.867
	L-CC	87.79%±1.98%	84.38%±1.57%	84.38%	84.38%	0.844
	L-MLO	88.67%±1.53%	82.81%±1.98%	83.21%	80.00%	0.828
	R-MLO	97.89%±0.98%	92.19%±0.54%	82.19%	82.19%	0.922
	Proposed Multiview (Late Fusion)	–	92.19%±1.56%	93.55%	90.91%	0.922
InceptionV3	R-CC	89.26%±1.98%	78.13%±2.14%	100%	69.57%	0.781
	L-CC	79.10%±2.34%	77.34%±2.19%	79.60%	87.23%	0.773
	L-MLO	79.93%±2.19%	75.00%±2.78%	76.67%	73.53%	0.750
	R-MLO	89.02%±1.78%	85.61%±1.78%	84.12%	79.22%	0.852
	Proposed Multiview (Late Fusion)	–	86.72%±1.57%	78.33%	94.12%	0.876
ResNet50	R-CC	86.45%±0.98%	75.78%±0.98%	69.41%	88.37%	0.758
	L-CC	87.44%±0.97%	85.16%±1.65%	84.62%	85.71%	0.852
	L-MLO	77.09%±2.19%	68.75%±2.45%	68.18%	69.35%	0.688
	R-MLO	78.21%±1.57%	69.53%±2.98%	63.16%	87.88%	0.695
	Proposed Multiview (Late Fusion)	–	**81.25%±1.54%**	**77.33%**	**86.79%**	**0.811**

89.91%, the sensitivity of 86.71%, and the specificity of 94.39%. The performance of our system in term of the ROC curve is shown in Fig. 5. Furthermore, for the comparative study we also performed experiments with single view, two views and four views feature fusion for the mass classification. The results presented in the Table 6 show that our proposed MVALF-based system outperformed and was able to surpass the state-of-art multi-view models.

The comparison between three different state-of-the-art pre-trained models are shown in Fig. 5. The pre-trained model VGG19 outperforms InceptionV3 and ResNet50 for the mass classification in MVALF system. However, our proposed system achieved best results with AUC of 0.896 in contrast with single view, two views and four view early fusion based system which have obtained AUC of 0.737, 0.842 and 0.769, respectively. The proposed MVALF model performs 5% better than the multi-view feature fusion model, 5–10% better than the single and two-views models. The MVALF based CADx system provides a benchmark approach of information fusion for classification tasks into the medical field, especially for breast cancer where four-view information of the patient is available. Table 5 depicts the performance measures of our proposed classifier into *Benign* and *Malignant* cases.

4.3.4. Comparison summary of our work with others

The comparison study was performed to evaluate the performance of our proposed MVALF-based CADx system in comparison to previous studies that use the deep CNN models for mammogram classification tasks. For instance, we compared between single view and two views. Furthermore, we compared our proposed system with the recent

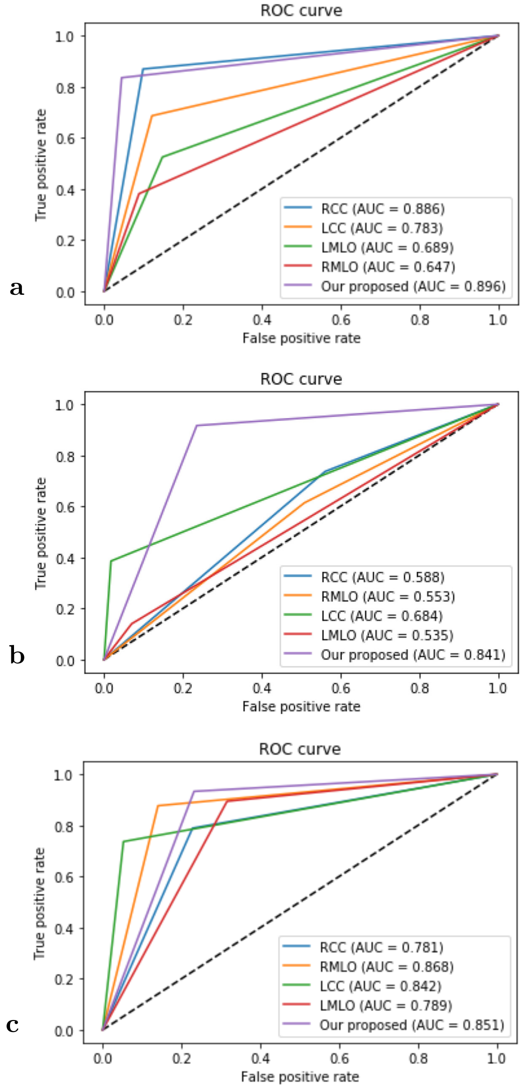

Fig. 5. ROC plotting for *Benign* and *Malignant* classification. The testing performance of (**a**) VGG19, (**b**) InceptionV3, and (**c**) ResNet50 is presented, using the proposed MVALF based CADx system.

Tab. 5. Performance measures of proposed MVALF for the classification of *Malignant mass* vs *Benign mass* mammograms

Models	Views	Training Accuracy	Testing Accuracy	Sensitivity	Specificity	AUC
VGG19	R-CC	96.57%±1.67%	88.64%±1.57%	88.59%	88.71%	0.886
	L-CC	81.33%±0.98%	78.82%±2.18%	75.90%	83.45%	0.783
	L-MLO	88.67%±0.45%	69.81%±2.14%	66.83%	75.89%	0.689
	R-MLO	75.54%±1.56%	66.20%±2.91%	62.35%	79.35%	0.647
	Proposed Multiview (Late Fusion)	–	89.91%±1.57%	86.71%	94.39%	0.896
InceptionV3	R-CC	89.26%±1.98%	77.77%±2.13%	70.50%	79.76%	0.811
	L-CC	73.10%±2.41%	67.40%±1.54%	75.44%	89.00%	0.851
	L-MLO	79.93%±1.58%	70.51%±3.20%	72.96%	77.67%	0.791
	R-MLO	84.02%±1.11%	75.26%±2.91%	73.00%	70.69%	0.785
	Proposed Multiview (Late Fusion)	–	80.07%±2.01%	98.73%	78.43%	0.860
ResNet50	R-CC	86.45%±1.45%	78.07%±2.10%	78.57%	77.59%	0.781
	L-CC	87.44%±1.98%	84.21%±1.98%	78.26%	93.33%	0.868
	L-MLO	77.09%±2.19%	78.95%±2.78%	76.67%	73.91%	0.842
	R-MLO	68.21%±2.98%	76.84%±1.98%	77.50%	86.21%	0.789
	Proposed Multiview (Late Fusion)	–	**83.33%±1.57%**	**94.64%**	**72.41%**	**0.851**

Tab. 6. Comparison with different mammography classification techniques using state-of-the-art pre-trained models on the CBIS-DDSM dataset.

Views	Models	Normal or Abnormal	Mass or Calcification	Malignant or Benign
Single View	VGG19	0.940	0.877	0.737
	InceptionV3	0.907	0.875	0.692
	ResNet50	0.914	0.862	0.644
Two View	VGG19	0.998	0.844	0.843
	InceptionV3	0.938	0.842	0.821
	ResNet50	0.971	0.883	0.811
Four Views (Early Fusion)	Small VGGNet [26]	0.934	0.923	0.769
Proposed Multiview (Late Fusion)	VGG19	0.992	0.922	0.896
	InceptionV3	0.996	0.876	0.860
	ResNet50	0.996	0.811	0.851

study performed on the four-view analysis using feature fusion strategy. Khan et al. in 2019 proposed a small VGGNet with the feature fusion strategy [26]. The system had the capability to classify the breast tumor using mammograms with four views. The results in Table 6 reveal that our proposed MVALF-based CADx system outperform the previous studies. We achieved the AUC of 0.996 for normal and abnormal mammogram classification, AUC of 0.922 for abnormality classification, and AUC of 0.896 for pathology classification.

5. Conclusion

In this work, we proposed a novel multi-view attention-based late fusion CADx system for mammogram classification using the transfer learning approach. We performed experiments using four views information and the results provide the evidence of achieving the best testing accuracy rate due to late information fusion. We observed that in the late fusion technique for mammogram classification, the overfitting problem occurs due to the unbalance and the limited size of the dataset. According to our assessment, data enhancement plays an important role in reducing the over-fitting problem. Furthermore, the comparison study shows that the proposed model achieves good classification performance and also reduces the computational complexity of the system with the help of the pre-trained model. We conclude that VGGNet pre-trained on ImageNet models with fine-tuning performs the best among all the pre-trained models for our proposed attention-based weighted late fusion approach. Table 6 demonstrates the comparative overview of the previous studies with the proposed MVALF-based CADx system. The results clearly show the effectiveness of the proposed technique. Our system provides a baseline for the new approach to attention-based weighted late fusion using the CBIS-DDSM for abnormality and pathology classification.

In the future work, we will experiment to analyze the impact of different sources for the improvement of the proposed CADx system.

Acknowledgement

This work has been supported by Higher Education Commission under Grant # 2 (1064), and is carried out at the Medical Imaging and Diagnostics (MID) Lab at COMSATS University Islamabad, under the umbrella of the National Center of Artificial Intelligence (NCAI), Pakistan.

References

[1] Large Scale Visual Recognition Challenge 2015 (ILSVRC2015), 2015. `http://www.image-net.org/challenges/LSVRC/2015/results`. [Aceessed Jun 2020].

[2] R. Agarwal, O. Diaz, X. Lladó, et al. Automatic mass detection in mammograms using deep convolutional neural networks. *Journal of Medical Imaging*, 6(3):31409, 2019. doi:10.1117/1.JMI.6.3.031409.

[3] A. Akselrod-Ballin, L. Karlinsky, A. Hazan, et al. Deep learning for automatic detection of abnormal findings in breast mammography. In M. J. Cardoso, T. Arbel, G. Carneiro, et al., editors, *Proc. Int. Workshops on Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support DLMIA, ML-CDS, in conjunction with MICCAI 2017*, volume 10553 of *Lecture Notes in Computer Science*, pages 321–329. Springer, Québec City, QC, Canada, 14 Sep 2017. doi:10.1007/978-3-319-67558-9_37.

[4] M. A. Al-masni, M. A. Al-antari, J.-M. Park, et al. Simultaneous detection and classification of breast masses in digital mammograms via a deep learning YOLO-based CAD system. *Computer methods programs in biomedicine*, 157:85–94, 2018. doi:10.1016/j.cmpb.2018.01.017.

[5] J. Arevalo, F. A. González, Ramos-Pollán, et al. Representation learning for mammography mass lesion classification with convolutional neural networks. *Computer methods programs in biomedicine*, 127:248–257, 2016. doi:10.1016/j.cmpb.2015.12.014.

[6] L. W. Bassett, I. A. Hirbawi, N. DeBruhl, and M. K. Hayes. Mammographic positioning: evaluation from the view box. *Radiology*, 188(3):803–806, 1993. doi:10.1148/radiology.188.3.8351351.

[7] M. Bator and M. Nieniewski. Detection of cancerous masses in mammograms by template matching: Optimization of template brightness distribution by means of evolutionary algorithm. *Journal of Digital Imaging*, 25(1):162–172, 2012. doi:10.1007/s10278-011-9402-1.

[8] G. Carneiro, J. Nascimento, and A. P. Bradley. Unregistered multiview mammogram analysis with pre-trained deep learning models. In N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, editors, *Proc. Int. Conf. Medical Image Computing and Computer-Assisted Intervention MICCAI 2015*, volume 9351 of *Lecture Notes in Computer Science*, pages 652–660, Munich, Germany, 5-9 Oct 2015. Springer. doi:10.1007/978-3-319-24574-4_78.

[9] G. Carneiro, J. Nascimento, and A. P. Bradley. Automated analysis of unregistered multi-view mammograms with deep learning. *IEEE Transactions on Medical Imaging*, 36(11):2355–2365, 2017. doi:10.1109/TMI.2017.2751523.

[10] J. Chakraborty, A. Midya, and R. Rabidas. Computer-aided detection and diagnosis of mammographic masses using multi-resolution analysis of oriented tissue patterns. *Expert Systems with Applications*, 99:168–179, 2018. doi:10.1016/j.eswa.2018.01.010.

[11] H. Chen, D. Ni, J. Qin, et al. Standard plane localization in fetal ultrasound via domain transferred deep neural networks. *IEEE Journal of Biomedical Health Informatics*, 19(5):1627–1636, 2015. doi:10.1109/JBHI.2015.2425041.

[12] H. Chougrad, H. Zouaki, and O. Alheyane. Deep convolutional neural networks for breast cancer screening. *Computer Methods Programs in Biomedicine*, 157:19–30, 2018. doi:10.1016/j.cmpb.2018.01.011.

[13] F. Ciompi, B. de Hoop, S. J. van Riel, et al. Automatic classification of pulmonary peri-fissural nodules in computed tomography using an ensemble of 2D views and a convolutional neural network out-of-the-box. *Medical Image Analysis*, 26(1):195–202, 2015. doi:10.1016/j.media.2015.08.001.

[14] A. F. Clark. The mini-MIAS database of mammograms, 2012. `http://peipa.essex.ac.uk/info/mias.html` [Accessed Jun 2020].

[15] K. Clark, B. Vendt, K. Smith, et al. The cancer imaging archive (TCIA): Maintaining and operating a public information repository. *Journal of Digital Imaging*, 26(6):1045–1057, 2013. doi:10.1007/s10278-013-9622-7.

[16] J. Deng, W. Dong, R. Socher, et al. ImageNet: A large-scale hierarchical image database. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition CVPR 2009*, pages 248–255, Miami, FL, USA, 20-25 Jun 2009. IEEE. doi:10.1109/CVPR.2009.5206848.

[17] S. Dhahbi, W. Barhoumi, and E. Zagrouba. Multi-view score fusion for content-based mammogram retrieval. In A. Verikas, P. Radeva, and D. Nikolaev, editors, *Proc. 8th Int. Conf. Machine Vision ICMV 2015*, volume 9875 of *Proc. SPIE*, page 987515, Barcelona, Spain, 8 Dec 2015. doi:10.1117/12.2228614.

[18] N. Dhungel, G. Carneiro, and A. P. Bradley. A deep learning approach for the analysis of masses in mammograms with minimal user intervention. *Medical Image Analysis*, 37:114–128, 2017. doi:10.1016/j.media.2017.01.009.

[19] G. W. Eklund. The art of mammographic positioning. In M. Friedrich and E. A. Sickles, editors, *Radiological Diagnosis of Breast Diseases*, pages 75–88. Springer, 2000. doi:10.1007/978-3-642-60919-0_6.

[20] M. Heath, D. Bowyer, R. Kopans, et al. The digital data base for screening Mammography. In M. J. Yaffe, editor, *Proc. 5th Int. Workshop on Digital Mammography*, pages 212–218, Toronto, Canada, 11-14 Jun 2000. Medical Physics Publishing, Madison, WI, USA. `http://www.eng.usf.edu/cvprg/Mammography/Database.html`.

[21] M. Heath, K. Bowyer, D. Kopans, et al. Current status of the digital database for screening mammography. In N. Karssemeijer, M. Thijssen, J. Hendriks, and L. van Erning, editors, *Digital Mammography*, pages 457–460. Springer Netherlands, Dordrecht, 1998. doi:10.1007/978-94-011-5318-8_75.

[22] B. Q. Huynh, H. Li, and M. L. Giger. Digital mammographic tumor classification using transfer learning from deep convolutional neural networks. *Journal of Medical Imaging*, 3(3):34501, 2016. doi:10.1117/1.JMI.3.3.034501.

[23] Z. Jiao, X. Gao, Y. Wang, and J. Li. A deep feature based framework for breast masses classification. *Neurocomputing*, 197:221–231, 2016. doi:10.1016/j.neucom.2016.02.060.

[24] A. Jouirou, A. Baâzaoui, and W. Barhoumi. Multi-view information fusion in mammograms: A comprehensive overview. *Information Fusion*, 52:308–321, 2019. doi:10.1016/j.inffus.2019.05.001.

[25] A. Khan, A. Sohail, U. Zahoora, and A. S. Qureshi. A survey of the recent architectures of deep convolutional neural networks. *Artificial Intelligence Review*, 53(8):5455–5516, 2020. doi:10.1007/s10462-020-09825-6.

[26] H. N. Khan, A. R. Shahid, B. Raza, A. H. Dar, and H. Alquhayz. Multi-view feature fusion based four views model for mammogram classification using convolutional neural network. *IEEE Access*, 7:165724–165733, 2019. doi:10.1109/ACCESS.2019.2953318.

[27] E. Kozegar, M. Soryani, B. Minaei, and I. Domingues. Assessment of a novel mass detection algorithm in mammograms. *Journal of Cancer Research and Therapeutics*, 9(4):592, 2013. doi:10.4103/0973-1482.126453.

[28] R. S. Lee, F. Gimenez, A. Hoogi, et al. A curated mammography data set for use in computer-aided detection and diagnosis research. *Scientific Data*, 4:170177, 2017. doi:10.1038/sdata.2017.177.

[29] H. Li, S. Zhuang, D.-A. Li, J. Zhao, and Y. Ma. Benign and malignant classification of mammogram images based on deep learning. *Biomedical Signal Processing Control*, 51:347–354, 2019. doi:10.1016/j.bspc.2019.02.017.

[30] Y. Li, H. Chen, Y. Yang, et al. A bilateral analysis scheme for false positive reduction in mammogram mass detection. *Computers in Biology and Medicine*, 57:84–95, 2015. doi:10.1016/j.compbiomed.2014.12.007.

[31] X. Liu, T. Zhu, L. Zhai, and J. Liu. Improvement of mass detection in mammogram using multi-view information. In C. M. Falco and X. Jiang, editors, *Proc. 8th Int. Conf. Digital Image Processing ICDIP 2016*, volume 10033 of *Proc. SPIE*, page 100334M, Chengdu, China, 29 Aug 2016. doi:10.1117/12.2244627.

[32] W. Lotter, G. Sorensen, and D. Cox. A multi-scale CNN and curriculum learning strategy for mammogram classification. In M. J. Cardoso, T. Arbel, G. Carneiro, et al., editors, *Proc. Int. Workshops on Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support DLMIA, ML-CDS, in conjunction with MICCAI 2017*, volume 10553 of *Lecture Notes in Computer Science*, pages 169–177. Springer, Québec City, QC, Canada, 14 Sep 2017. doi:10.1007/978-3-319-67558-9_20.

[33] W. Peng, R. V. Mayorga, and E. M. A. Hussein. An automated confirmatory system for analysis of mammograms. *Computer Methods Programs in Biomedicine*, 125:134–144, 2016. doi:10.1016/j.cmpb.2015.09.019.

[34] D. Ribli, A. Horváth, Z. Unger, et al. Detecting and classifying lesions in mammograms with deep learning. *Scientific Reports*, 8(1):4165, 2018. doi:10.1038/s41598-018-22437-z.

[35] O. Russakovsky, J. Deng, H. Su, et al. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015. doi:10.1007/s11263-015-0816-y.

[36] H.-C. Shin, H. R. Roth, M. Gao, et al. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. *IEEE Transactions on Medical Imaging*, 35(5):1285–1298, 2016. doi:10.1109/TMI.2016.2528162.

[37] R. L. Siegel, K. D. Miller, and A. Jemal. Cancer statistics, 2019. *CA: A Cancer Journal for Clinicians*, 69(1):7–34, 2019. doi:10.3322/caac.21551.

[38] K. Simonyan and A Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint*, 2014. arXiv:1409.1556v6.

[39] K. Smith, J. Kirby, D. Runbin, et al. CBIS-DDSM – Curated Breast Imaging Subset of DDSM. In TCIA Team [44]. [Accessed Jun 2020]. `https://wiki.cancerimagingarchive.net/display/Public/CBIS-DDSM`.

[40] J. Suckling, J. Parker, D. Dance, et al. The Mammographic Images Analysis Society digital mammogram database. In A. G. Gale, S. M. Astley, D. R. Dance, and A. Y. Cairns, editors, *Digital Mammography*, volume 1069 of *Exerpta Medica International Congress Series*, pages 375–378. Elsevier, 1994. `http://www.wiau.man.ac.uk/services/MIAS/` [Inoperative].

[41] L. Sun, L. Li, W. Xu, et al. A novel classification scheme for breast masses based on multi-view information fusion. In *Proc. 4th Int. Conf. Bioinformatics and Biomedical Engineering iCBBE 2010*, pages 1–4, Chengdu, China, 18-20 Jun 2010. IEEE. doi:10.1109/iCBBE.2010.5517742.

[42] C. Szegedy, W. Liu, Y. Jia, et al. Going deeper with convolutions. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition CVPR 2015*, pages 1–9, Boston, MA, USA, 7-12 Jun 2015. doi:10.1109/CVPR.2015.7298594. `https://www.cv-foundation.org/openaccess/content_cvpr_2015/html/Szegedy_Going_Deeper_With_2015_CVPR_paper.html`.

[43] M. Tan, J. Pu, S. Cheng, et al. Assessment of a four-view mammographic image feature based fusion model to predict near-term breast cancer risk. *Annals of Biomedical Engineering*, 43(10):2416–2428, 2015. doi:10.1007/s10439-015-1316-5.

[44] TCIA Team, editors. *The Cancer Imaging Archive*, 2021. [Accessed Jun 2020]. `https://www.cancerimagingarchive.net/`.

[45] P. D. Trieu, P. C. Brennan, W. Lee, E. Ryan, et al. The value of the craniocaudal mammographic view in breast cancer detection: a preliminary study. In C. K. Abbey and C. R. Mello-Thoms, editors, *Proc. Conf. SPIE Medical Imaging 2013: Image Perception, Observer Performance, and Technology Assessment*, volume 8673 of *Proc. SPIE*, page 86731J, Lake Buena Vista, FL, United States, 28 Mar 2013. doi:10.1117/12.2006821.

[46] C. J. Vyborny and R. A. Schmidt. Mammography as a radiographic examination: an overview. *RadioGraphics*, 9(4):723–764, 1989. doi:10.1148/radiographics.9.4.2667052.

[47] WebMD. Breast Calcifications, 2020. `https://www.webmd.com/women/guide/breast-calcification-symptoms-causes-treatments`. [Accessed Jun 2020].

[48] J. Wei, H.-P. Chan, C. Zhou, et al. Computer-aided detection of breast masses: Four-view strategy for screening mammography. *Medical Physics*, 38(4):1867–1876, 2011. doi:10.1118/1.3560462.

[49] World Health Organization. Cancer. Facts sheet, 2018. `https://www.who.int/news-room/fact-sheets/detail/cancer`. [Aceessed Jun 2020].

[50] N. I. R. Yassin, S. Omran, E. M. F. El Houby, and H. Allam. Machine learning techniques for breast cancer computer aided diagnosis using different image modalities: A systematic review. *Computer Methods and Programs in Biomedicine*, 156:25–45, 2018. doi:10.1016/j.cmpb.2017.12.012.

[51] H. Zhao, W. Xu, L. Li, and J. Zhang. Classification of breast masses based on multi-view information fusion using multi-agent method. In *Proc. 5th Int. Conf. Bioinformatics and Biomedical Engineering iCBBE 2011*, pages 1–4, Wuhan, China, 10-12 May 2011. IEEE. doi:10.1109/iCBBE.2011.5780304.

Hina Iftikhar completed the M.Sc. Computer Science degree from COMSATS University Islamabad (CUI), Islamabad Pakistan in the area of Medical Image Analysis. Her research interest includes the development of CADx system for early detection of breast cancer and computer vision task using machine learning and deep learning approaches. She has a number of conference papers in international conferences.

Ahmad Raza Shahid is currently working as Assistant Professor at COMSATS University Islamabad (CUI), Islamabad, Pakistan. He did his Ph.D. in Computer Science in York, UK in 2012. During his PhD he worked on automatically building a WordNet for four languages, namely, English, German, French and Greek. After his Ph.D. he has been working in the areas of Computer Vision and Pattern Recognition, Machine Learning, and Natural Language Processing. A few of the problems that he has worked on include cancer detection, pedestrian detection, driver fatigue detection, and data mining.

Basit Raza received his master's degree in computer science from the University of Central Punjab, Lahore, Pakistan. He received his Ph.D. in computer science from International Islamic University Islamabad and University Technology Malaysia in 2014. Currently, he is an Assistant Professor in the Department of Computer Science, COMSATS University Islamabad (CUI), Islamabad, Pakistan. He is member of Medical Imaging and Diagnostics Lab, National Center of Artificial Intelligence (NCAI) since 2018. His research interests are data science, medical imaging, database management systems, data mining, data warehousing, machine learning, deep learning and artificial intelligence. Dr. Raza has authored several papers in refereed journals and has been serving as a reviewer for prestigious journals, such as Applied Soft Computing, Swarm and Evolutionary Computation, Swarm Intelligence, Applied Intelligence, IEEE Access and Future Generation Computer Systems.

Hasan Nasir Khan received the B.Sc. degree in computer science from COMSATS University Islamabad, Sahiwal, Pakistan, in 2016. He pursued the M.Sc. degree in computer science at COMSATS University Islamabad, Islamabad, Pakistan in 2019. He is working as a Research Assistant with the Medical Imaging and Diagnostics Lab at COMSATS University Islamabad, under the umbrella of National Center of Artificial Intelligence, Pakistan. His research interest includes the development of computer-aided diagnosis systems for early diagnosis of breast cancer using artificial intelligence and computer vision techniques. Hasan Nasir Khan was a recipient of the Prime Minister of Pakistan's National ICT Scholarship Award in 2012. He has published 5 international conference proceedings and a journal paper.