International Journal of Computational Intelligence Systems (2024) 17:191
https://doi.org/10.1007/s44196-024-00593-7

RESEARCH ARTICLE

Multi-class Breast Cancer Classification Using CNN Features Hybridization

Sannasi Chakravarthy[1] · N. Bharanidharan[2] · Surbhi Bhatia Khan[3,4] · V. Vinoth Kumar[2] · T. R. Mahesh[5] · Ahlam Almusharraf[6] · Eid Albalawi[7]

Received: 28 November 2023 / Accepted: 5 July 2024

Abstract

Breast cancer has become the leading cause of cancer mortality among women worldwide. The timely diagnosis of such cancer is always in demand among researchers. This research pours light on improving the design of computer-aided detection (CAD) for earlier breast cancer classification. Meanwhile, the design of CAD tools using deep learning is becoming popular and robust in biomedical classification systems. However, deep learning gives inadequate performance when used for multilabel classification problems, especially if the dataset has an uneven distribution of output targets. And this problem is prevalent in publicly available breast cancer datasets. To overcome this, the paper integrates the learning and discrimination ability of multiple convolution neural networks such as VGG16, VGG19, ResNet50, and DenseNet121 architectures for breast cancer classification. Accordingly, the approach of fusion of hybrid deep features (FHDF) is proposed to capture more potential information and attain improved classification performance. This way, the research utilizes digital mammogram images for earlier breast tumor detection. The proposed approach is evaluated on three public breast cancer datasets: mammographic image analysis society (MIAS), curated breast imaging subset of digital database for screening mammography (CBIS-DDSM), and INbreast databases. The attained results are then compared with base convolutional neural networks (CNN) architectures and the late fusion approach. For MIAS, CBIS-DDSM, and INbreast datasets, the proposed FHDF approach provides maximum performance of 98.706%, 97.734%, and 98.834% of accuracy in classifying three classes of breast cancer severities.

Keywords Breast cancer · Deep neural networks · Mammogram images · Feature fusion · Late fusion · Transfer learning

✉ T. R. Mahesh
trmahesh.1978@gmail.com

Sannasi Chakravarthy
sannasi@bitsathy.ac.in

N. Bharanidharan
bharanidharan.n@vit.ac.in

Surbhi Bhatia Khan
s.khan138@salford.ac.uk

V. Vinoth Kumar
drvinothkumar03@gmail.com

Ahlam Almusharraf
Aialmusharraf@pnu.edu.sa

Eid Albalawi
ealbalawi@kfu.edu.sa

[1] Department of ECE, Bannari Amman Institute of Technology, Sathyamangalam 638401, India

[2] School of Computer Science Engineering and Information Systems (SCORE), Vellore Institute of Technology, Vellore 632014, India

[3] School of Science, Engineering and Environment, University of Salford, Manchester, UK

[4] Department of Electrical and Computer Engineering, Lebanese American University Byblos, Byblos, Lebanon

[5] Department of Computer Science and Engineering, Faculty of Engineering and Technology, JAIN (Deemed-to-Be University), Bangalore 562112, India

[6] Department of Business Administration, College of Business and Administration, Princess Nourah Bint Abdulrahman University, P.O. Box 84428, 11671 Riyadh, Saudi Arabia

[7] Department of Computer Science, School of Computer Sciences and Information Technology, King Faisal University, Al Hofuf 400-31982, Al Ahsa, Saudi Arabia

Published online: 22 July 2024

1 Introduction

Breast cancer, noticed in the breast part of humans, is the second most affecting cancer worldwide. According to the statistics released by World Health Organization (WHO), approximately 2 million women were diagnosed with this cancer type, resulting in the global deaths of almost 685,000 in 2020 [1]. At the end of the same year, it was around 7.6 million women survived who were under-diagnosis in the past five years. This makes breast cancer the earth's most predominant cancer type. And it is revealed that this type of cancer affects women much more as compared with men. Additionally, this cancer type affects women of any age group in every nation after puberty [2]. And the impact of this cancer type is more in women's later life. Furthermore, around half of tumors developed in the breast part have no recognizable and no common risk factors other than gender (women) and ageing. This makes breast cancer deadly for women since those who have it do not experience any symptoms. [3]

The improvements in survival of people with breast cancer began in the late 1980s due to earlier detection of the disease combined with advanced diagnosis modes. Thus early identification and appropriate diagnosis of breast cancer play a significant part in minimizing mortality and ensuring improved survival rates. This could be possible because of computer-aided diagnosis (CAD) tools. Nowadays, researchers are continuously working towards designing robust CAD tools for efficient and adequate diagnosis of this cancer type. Herein, medical imaging and analysis play a vital role in efficient diagnosis. For this, several imaging modalities are employed using X-rays (digital mammograms), imaging with magnetic resonance approach (MRI), imaging with acoustic waves (ultrasound), and imaging with infrared rays (thermography) [4]. In this, mammograms provide breast images with better visualization of the anatomy. This makes breast imaging using mammography the most often approach for early diagnosis [5]. Digital mammogram images can provide a higher quality representation of masses, microcalcification, and structural deformities. Among these, microcalcifications and masses are important in tumor detection at an earlier stage of breast cancer, whereas the last indicator is the least significant in tumor detection. Clinical people always encounter some difficulty in providing accurate evaluation during the diagnosis. This is especially because of the different shapes and sizes of the breast and the larger amount of mammograms implicated in the screening for breast cancer [6]. Consequently, there is always a demand for a robust CAD system for detecting and classifying breast cancer severities. The proposed CAD system for breast cancer problem is given in Fig. 1. From this, it is inferred that:

Step 1: The digital mammograms are first retrieved from publicly available datasets [mammographic image analysis society (MIAS), CBIS-DDSM (curated breast imaging subset of digital database for screening mammography), and INbreast] and taken as training inputs. The details of the databases will be illustrated in Sect. 4.1.

Step 2: The preprocessing of mammograms and image augmentation are done to attain the best classification performance. The preprocessing and data augmentation details will be presented in Fig. 2 and Sects. 3.1 and 3.2.

Step 3: Next to preprocessing, the mammograms are applied to four convolutional neural networks (CNN) architectures, namely VGG16. VGG19, ResNet50, and DenseNet121, where the deep features are extracted by fine-tuning the models. For example, as given in Step 3 of Fig. 1, the last convolution block (layer-13 to 18) is fine-tuned, whereas freezing the first fewer convolution blocks (layer-0 to 12), together with the top-level classifier. The details of transfer learning-based feature extrication will be discussed in Sect. 3.3.

Step 4: Here, the features are extracted from the last layer before the softmax layers of every CNN architecture, and then the feature arrays of 1024 feature vectors are created. Afterwards, these arrays are applied further into the sequential model for performing classification. The details about the feature fusion network will be presented in Sect. 3.3.

Step 5: The step involves the classification of unseen test mammograms.

Step 6: The classification performance of the proposed approach is finally evaluated using the standard performance measures, and a comparative analysis is done.

In CAD frameworks, feature-based approaches are commonly adopted for medical classification problems by feature extraction; these feature vectors will be used to train the classifiers. The outcomes of classification models are highly dependent on the extracted feature vectors, so several research works primarily emphasize two things: (i) developing qualitative feature descriptors and (ii) extracting better feature vectors. As compared with conventional handcrafted feature-based approaches, CNNs can automatically extricate more abstract features. Furthermore, deep feature fusion approaches are emerging in order to enhance the better feature representation of applied images. In this way, the paper proposed a fusion of hybrid deep features (FHDF) approach for capturing more potential information and thereby improving the classification performance of breast cancer.

2 Related Works

The section gives an overview of deep learning techniques used for the problem of breast cancer classification and different feature fusion approaches used for computer vision tasks.

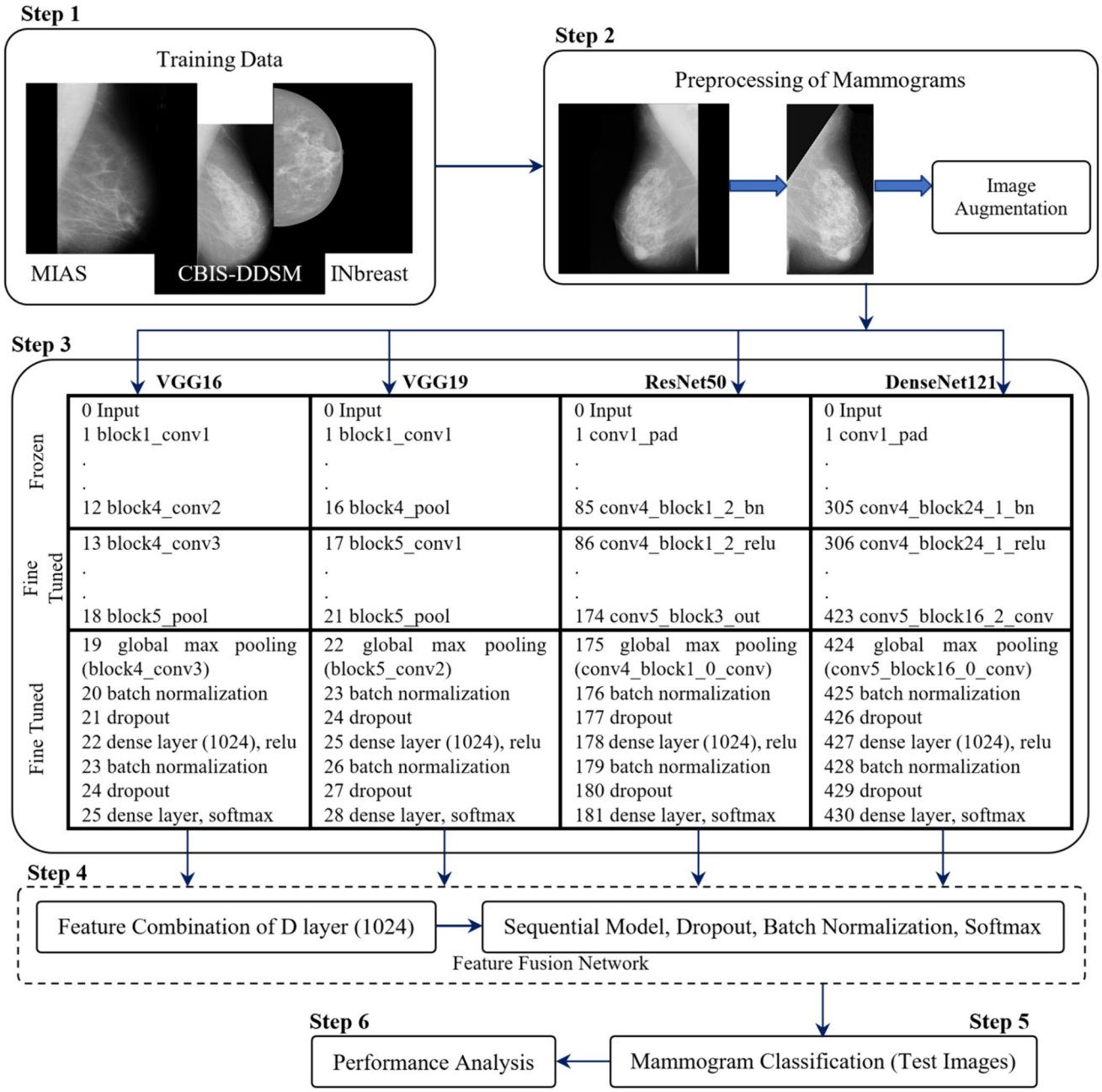

Fig. 1 Proposed workflow for the three-class classification of breast cancer

2.1 Deep Learning for Breast Cancer Classification

In recent days, several machine learning (ML) and deep learning (DL) techniques have emerged for classifying breast tumors using different input datasets. In the year of 2017, Neeraj et al. [7] developed a CAD system for breast mass detection and classification of mammograms. For the classification part, they used a DL architecture which was pretrained with hand-crafted features. And they used mammograms from the INbreast database for implementation. The result of the work revealed that the model provided a classification performance of 90% accuracy. Here, the study investigates on providing solutions for a binary classification task (benign vs malignant). In the same year, Thijs et al. [8] presented the design of large-scale DL for the classification problem of breast cancer.

The authors presented a detailed comparison between the recently evolved mammographic CAD tool, which relies on manually extracted features and a CNN. The authors trained the above systems using privately obtained mammogram

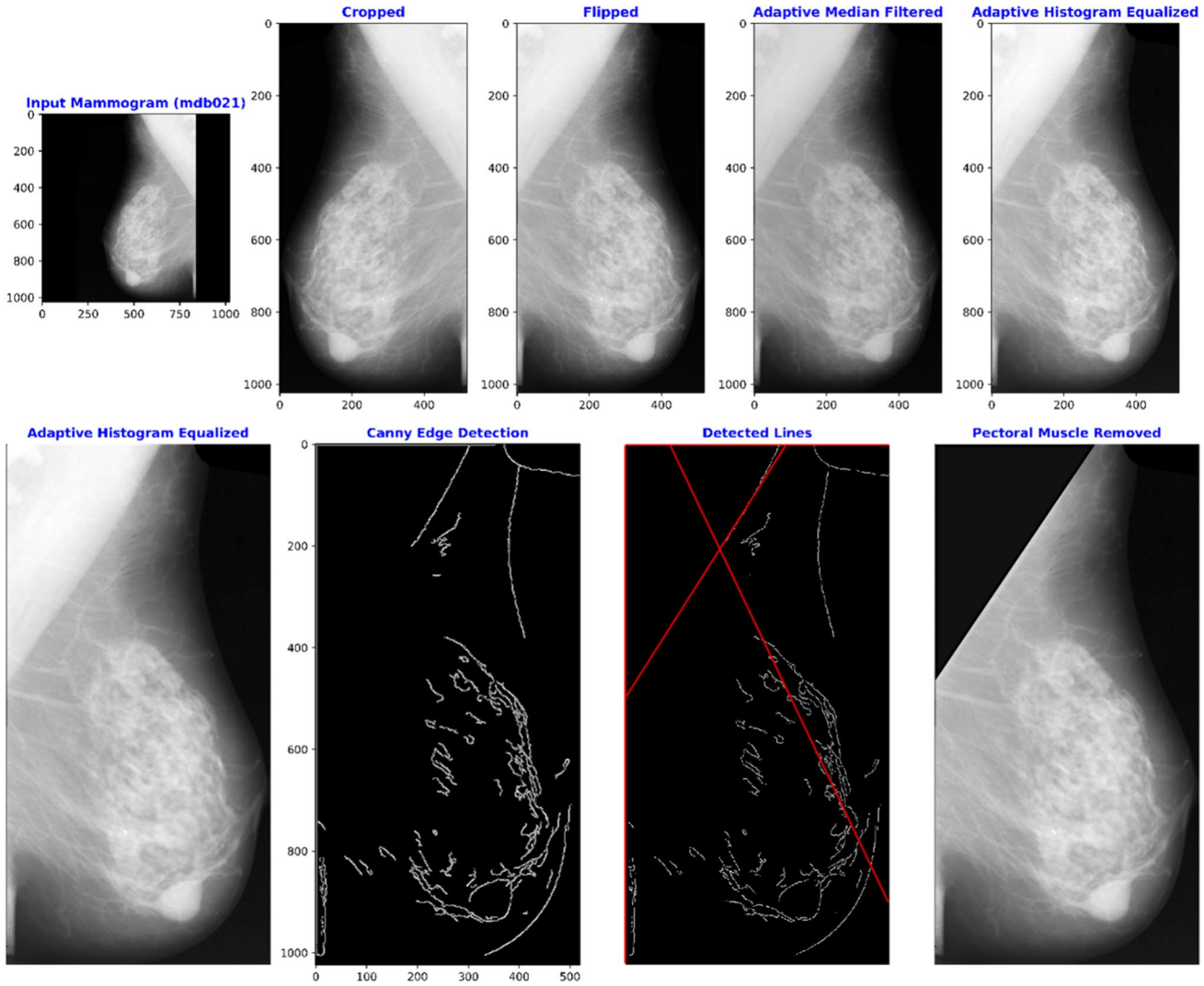

Fig. 2 Sample Pre-processing of a digital mammogram in MIAS database (mdb021)

data of around 45,000 images. And concluded that the deep CNN architectures performed well in classification and obtained results of 85.2% accuracy. Here, the study involves the investigation of a binary classification task. In the year of 2018, Xiaofei et al. [9] evaluated ten distinct deep CNN models and revealed that integrating both image augmentation and CNN-based transfer learning techniques is the most efficient way to improve classification performance in breast cancer problems. For this, the authors utilized privately obtained mammogram datasets. Here, the study analyses the binary classification problem. In the same year, Yemini [10] developed a CAD tool using CNN based transfer learning approach with Google Inception-V3 as a base model. They evaluated this using the digital mammograms taken from the INbreast dataset and obtained a result of 0.78 AUC. Here, the study investigates on providing solutions for a binary classification task (normal vs abnormal).

Chougrad et al. [11], in the year 2020, proposed a CAD system that intends to portray spontaneous label correlation relationships for mammogram classification. For this, they utilized the pretrained CNN models for the attractive nature of transfer learning. The authors used a different approach for fine-tuning the models by utilizing an optimization technique that uses Stochastic gradient descent (SGD) adopted with a decaying learning rate. The work resulted in 0.687 and 0.617 of F1 score performance as obtained for the classification problem using INbreast and MIAS databases. Here, the study investigates on providing solutions for a multi-label classification task using transfer learning. In the same year, Shu et al. [12] presented a CAD system using CNNs for breast cancer classification using two pooling structures that are different from the conventional one. Here, the extraction of features is done, and the pooling structures are used in dividing the mammogram input regions with higher

malignant probabilities in accordance with the extricated features. The researchers used the DenseNet169 architecture for feature learning. In addition, they modified the architecture's last layer in accordance with the pooling structure for classifying the input feature vectors. The work was tested using the INbreast dataset and attained a classification result of 92.2% accuracy. Here, the study investigates on providing solutions for a multi-label classification task.

In addition to the above works, we the authors did some experimentations using transfer learning approaches for binary and multi-class classification problems. In the year 2021 [13], the deep features from mammograms are extracted using AlexNet, DarkNet19, GoogleNet, VGG16, and ResNet CNN models where classification is done typical ML algorithms such as K-nearest neighbour (KNN), Naïve Bayes (NB), Ensemble, and support vector machines (SVM) algorithms. Here, the hyperparameters are tuned automatically using the Bayes optimization techniques. In the year 2022 [14], ResNet18-based deep feature extraction is done, and the classification is further proceeded using extreme learning machine (ELM) model optimized with an enhanced crow-search algorithm. In the same year of 2022 [15], experimentations using transfer learning approaches are done with different strategies used for deep feature extraction, feature selection, feature fusion, and feature classification. And all these works were carried out using MIAS, CBIS-DDSM, and INbreast datasets with a maximum performance of 95% classification accuracy. In the same way, a new approach of feature fusion (FHDF) is proposed in this paper to enhance the performance of multi-class classification problems further.

From the literature [7–14], it is inferred that most researchers have focused on the problem of binary classification; however, multi-class classification is significant in real-time scenarios. And noted that some researchers employed pre-segmented image inputs for performing classification tasks. Furthermore, the transferred architecture is often incapable of capturing the representations of image inputs, and conventional feature vectors cannot provide the optimality of CAD systems in a promising manner. Thus, this work examines the hybrid fusion approach to address the above-said problems.

2.2 Related Works in Computer Vision Tasks Using Feature Fusion Approach

Several research works employ the fusion of extracted features, some of which are given below. In the work of [16], the researchers developed a hybrid fusion CAD model based on the integration of early and late fusion for the problem of glaucoma classification. Here, the central and Hu moments, and gray level co-occurrence matrix (GLCM) features are fused with CNN features, whereas classification is performed using the SVM algorithm. Another work of [17] portrays that the authors employed multi-structure-based fusing of CNN features used for the classification of satellite remote sensing scenes. Here, GoogLeNet, VGG-16, and CaffeNet are adopted for extracting the feature vectors and are fused using the fusion network. In the research of [18], an ensemble of multiple deep architectures is fused for classifying the medical image inputs. The results of this work revealed that the ensemble technique provides better classification when combined with fused features. And in the research work of [19], they developed a CAD model for the classification of skin lesions using the fused features from VGG16 and AlexNet models. The researchers found that the classification performance of fused features provides better accuracy than the individual feature vectors. The research works of [16–19] reveal that deep learning using CNNs emerges as one of the most substantial machine learning tools in medical classification problems. It has outpaced the classification performance of conventional classification models and human recognition. The convolution operation in CNNs simplifies an input image from several thousands of pixels to smaller feature maps. This makes the input dimension as a reduced one with significant representations. Here, it is also noted that the employment of the transfer learning concept is much more helpful in extracting deep features. It is one of the machine learning approaches where a CNN architecture trained on solving one problem is re-used on an another related problem. Moreover, the mentioned research works utilized the concept of feature fusion for improved and better feature representation of applied images. As a result, these deep feature fusion-based approaches provide supreme classification results as compared with the conventional handcrafted and individual deep features.

The summary of significant contributions of the proposed work are:

1. To the extent of our knowledge, this paper is the first one to use the FHDF approach for the three-class classification of breast cancer.
2. A better preprocessing approach is employed for pectoral muscle removal in mammograms.
3. The deep learning models with improved architecture, namely VGG16, VGG19, ResNet50, and DenseNet121, are presented for extricating the complementary feature vectors pertaining to the different depths of the CNN models.
4. An enhanced FHDF approach is proposed to adaptively fuse the CNN features through dense layer combined with softmax, batch normalization, and dropout layers.

3 Proposed Framework

This section presents information on how the mammogram inputs are preprocessed for further stages. How are the resultant mammograms augmented? How deep features are extracted from these augmented data? How is the proposed fusion of hybrid deep features network constituted?

3.1 Preprocessing of Mammograms

In MIAS and CBIS-DDSM databases, the dark and thickened borders on either side of mammogram images are cropped manually. In this dataset, the mammograms are obtained with medio-lateral oblique (MLO) viewpoint. Herein, the significant part of preprocessing lies in the removal of pectoral muscles (PM). These PMs are the regions located on either the top right or left side of the breast and quite the opposite to the direction of nipple location. For successful PM removal, the left-view mammogram images are flipped in an uniform manner. This flipping of images is done to make all the inputs as right-MLO view mammogram images so that the PM is located uniformly at the upper-left-side portion. A quite rudimentary idea for automatic mammogram flipping is detecting the image orientation. This could be easier since the background pixel areas of inputs are totally black and consequently provides us with the breast orientation on either half of the mammograms. Before proceeding further, the impulse noise present in the images is filtered with an adaptive median filter [23] approach without disturbing the non-affected pixels. In addition, the contrast of the mammograms is adaptively enhanced using the adaptive histogram equalization (AHE) [24] technique. After noise removal and appropriate contrast enhancement, Sobel [25] filter with canny edge detection [26] is employed with a threshold value of 1.8 for better detection of edges. Then Hough transform [27] technique is applied to obtain a list of output lines. Here, every detected line is characterized by an object using three parameters: the first one is the distance (dist) i.e. calculating the perpendicular distance of lines from the origin, the second one is the angle (degrees) i.e. calculating the degree made by the perpendicular from the x-axis on the positive side (nearer to the origin), and the third one is calculating the two points (point1 and point2) on the detected line. Now possible lines for PM segmentation can be shortlisted by examining a simple condition of whether the values of two parameters (dist and angle) of each line lie inside the below-given intervals;

$$\text{MIN\_ANGLE} <= \text{angle} <= \text{MAX\_ANGLE and}$$
$$\text{MIN\_DIST} <= \text{dist} <= \text{MAX\_DIST}.$$

If more than one line is obtained by using the above procedure, then the line which provides the least loss of information will be selected. Finally, the values of pixels that are covered by the shortlisted lines will be set as zero (black), and thus the PM is removed. A sample illustration of PM removal in mdb021 mammogram of MIAS dataset is presented in Fig. 2. Furthermore, the mammogram images of INbreast database are FFDM, so every finding and its details are substantial for further classification stage; however, the above-used approach of adaptive median filtering [23] is adopted for impulse noise removal in the mammograms of INbreast dataset.

3.2 Data Augmentation

The architecture of deep learning models works well if the models are trained using a larger sample of input images [28]. However, the adopted mammogram databases are composed of fewer hundred samples due to limited patient availability. Moreover, the overfitting problem of the employed classification problem needs to be addressed. And the above issues are taken care of using the process of image augmentation that intends to focus on increasing the amount of mammograms using existing samples. These newly generated mammograms are actually distinct variants of the original mammograms. The proposed work employs augmentation utilising the rotation of mammograms by the degrees of 45, 90, 135, 180, 235, and 270 and through horizontal and vertical flipping of inputs. In this way, each input sample for every class is augmented eight times which can be illustrated graphically in Fig. 3.

3.3 Feature Extraction

3.3.1 Transfer Learning Approach

In recent days, DL has been the emerging approach for solving several real-time classification and recognition problems. Here, CNNs are vital in providing real-time solutions for biomedical allied fields [29]. CNNs are the key network of deep learning and are prevalent for research in wider areas. Compared to conventional machine learning (ML) algorithms, CNNs are much more robust to noise and uneven transformation. And this makes it more popular in solving problems of biomedical image analysis [30]. The CNNs are composed of tens or hundreds of layers in which each layer can learn to detect distinct features of an input image. Here, the filters play a major role in applying them to every training image with a distinct resolution, and the obtained output is applied to further layers [31]. In this way, the architecture of CNN is composed of convolution layers (learning low and high-level features), pooling layers (for reducing the size of the convoluted feature vectors either through average or

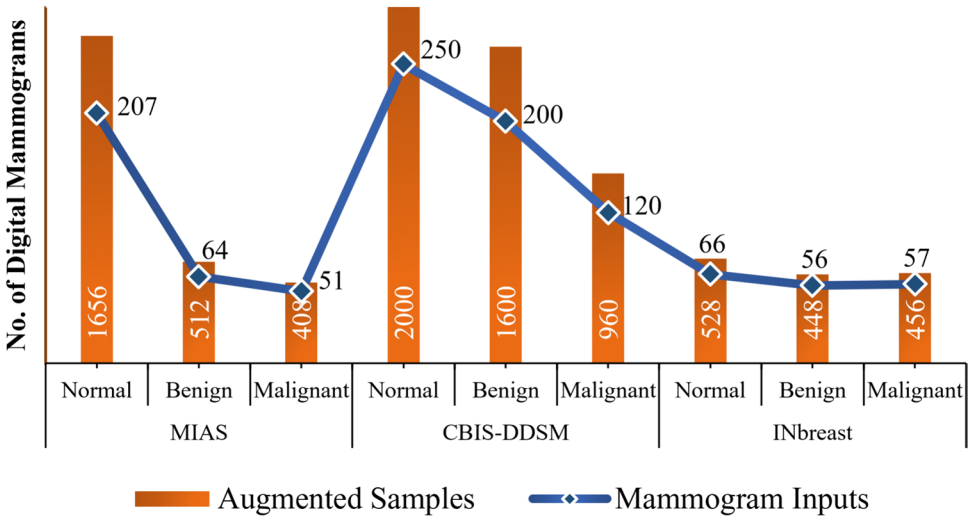

Fig. 3 Mammogram inputs before and after augmentation

max-pooling), and a fully connected (fc) layer that connects each neuron of every layer to its succeeding ones for image analysis based on the multilayer perceptron [32].

For training a CNN from scratch, it always claims more time with higher computing power and data. In the biomedical field, the imaging databases are generally in the order of 10^2 to 10^4 since sorting a larger annotated database is quite impractical. In addition, the quality of the image will also become substandard. For this, the solution uses an interesting part of DL, the transfer learning (TL) approach, which intends to utilise knowledge gained while generating a solution for one task and employing it on another but related task [34]. In place of learning from scratch, TL uses patterns already trained on the related task. Herein, the approach has two phases: the first one involves the selection of a pretrained network trained on a larger volume of a standard database, which is necessarily related to the task that we need to solve; and the second one is fine-tuning of the selected model in accordance with the size and similarity of the considered problem (image inputs) [35]. Since the input datasets differ from the input of the pretrained model, the work fine-tuned and freeze some layers in the employed deep CNN models as given in Fig. 1.

The work involves the training and testing of some advanced pretrained DL architectures, namely VGGNet, InceptionNet, ResNet, ResNet-V2, Inception-ResNet-V2, NasNet, XceptionNet, and DenseNet models and noted that the combination of VGG16, VGG19, ResNet50, and DenseNet121 gives the superior performance for this breast cancer classification problem in ablation analysis presented in Sect. 4.2. Here, the principle of VGG models is the use of smaller-sized convolutional filter kernels, which allow the networks to possess a larger amount of weight layers [30]. This means that more layers will result in enhanced performance. The concept of both VGG16 and VGG19 models is the same, except that VGG16 has three fewer convolution layers than VGG19. For reducing the errors, ResNet models use shortcut or skip connections that merely perform identity mapping [31]. ResNet50 is one variant with 48 convolutional, one maxpooling and one average pooling layer. The skip connections in ResNet50 bypass some layers and send the output as an input to the subsequent layers. Thus, providing an alternate path for the gradient with backpropagation. Rather than deriving representational power from highly wider or deeper models, DenseNet architectures utilize the potential of the network through the concept of feature reuse [33]. The layers in DenseNet121 model spread their weights across several inputs and thus make use of deep layers to reuse features that are extricated earlier. The degradation problem [30] encountered in deep learning is alleviated by using skip connections in ResNet50 and feature reusability in DenseNet121 models. The structure of the four transfer learning models is illustrated in Fig. 1.

The work employed the VGG series, ResNet50, and DenseNet121 models in the transfer learning approach where the weights were pre-trained originally in the database of ImageNet [33]. This database comprises a training set of about 1.2 million images, a validation set of about 50,000 images, and a testing set of about 1,00,000, and all these inputs correspond to 1000 class labels. As illustrated in step 3 of Fig. 1, the early layers of each DL architecture are frozen where more generic features are captured. And the successive layers of the architectures are retrained using fine-tuning by training on digital mammogram inputs to further acquire more database-specific features. In the end, the work fine-tuned the own FC classifier as shown in step 3 of Fig. 1. For example, as illustrated in Fig. 4, the VGG16 model is considered where the first fewer convolutional blocks utilize the parameters $(W_1, W_2, \ldots W_k)$ that are already trained (pre-trained) on the ImageNet database.

The size of the preprocessed mammograms for all four TL models is $(224 \times 224 \times 3)$ as shown in Fig. 4b. The

learning rate is tuned as 10^{-3} for the first fifty epochs and further, the training is continued for another fifty epochs with 10^{-5} as the learning rate. The batch size for the training data is kept as 32 whereas for testing data, the batch size is considered as 1, and adaptive moment estimation, Adam approach [36] is used for optimization. Figure 4a illustrates the entire transfer learning approach using the VGG TL model where the first fewer layers are frozen i.e., pre-trained on the ImageNet database, and later convolutional block layers followed by FC layers are fine-tuned in the proposed work. Figure 4b shows a sample feature map visualization of the VGG16 model where the output of the first convolutional layer ($224 \times 224 \times 64$) is visualized. The 64 feature maps are plotted as an 8×8 square of images. These feature maps illustrate how deep the mammogram's interior parts, edges, and other fine details are learned for further classification. Herein, for better visualization of feature maps, the cmap of 'hot' is used in matplot library as given in Fig. 4b.

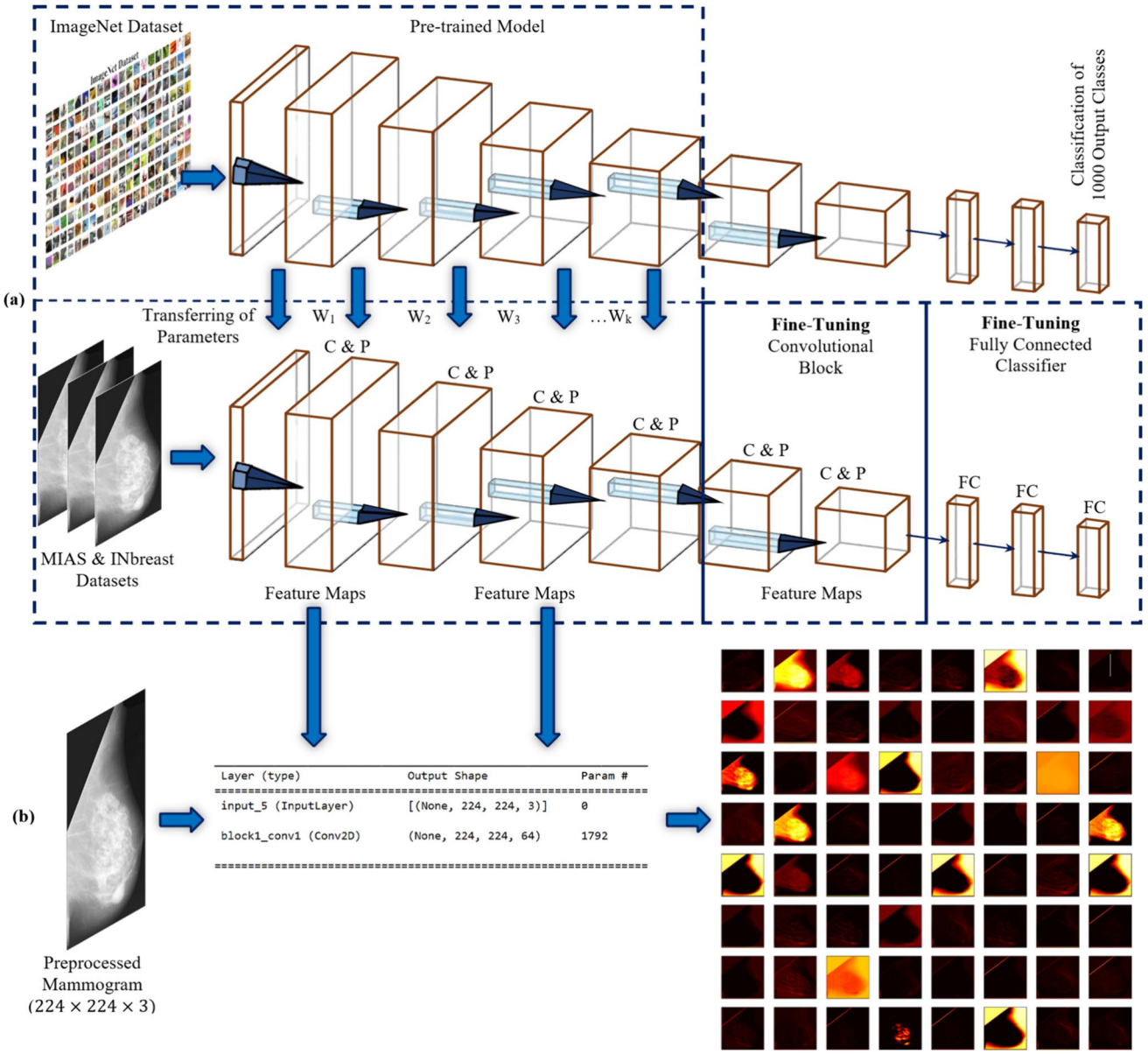

Fig. 4 **a** Visualization of transfer learning approach where parameters are transferred from pre-trained CNN and fine-tuned on digital mammogram databases [fully connected layer (FC), pooling (P), convolution (C)], **b** visualization of the feature maps of first convolutional layer ($224 \times 224 \times 64$) of VGG16 as an 8×8 image matrix

3.3.2 Late Fusion (LF) Approach

The approach of the late fusion technique is one of the ensemble methods of classification where the final output is based on the maximum number of decision by individual classifiers and weights. This approach is generally used in ML problems to improve classification performance. In the proposed work, the final classification result obtained using the four distinct TL networks (VGG16, VGG19, ResNet50, and DenseNet121) is integrated by adopting a majority voting approach. Here, each output class is calculated according to the majority of votes obtained for that particular class target. If $m = 1,2,3,\ldots X$ and $n = 1,2,3,\ldots Y$, then the decision of ith classifier can be given as $E(m,n) \in (0,1)$. Thus, the LF approach for majority voting is illustrated as

$$\sum_{m=1}^{x} E(m,n) = \max_{n=1}^{y} \sum_{m=1}^{x} E(m,n), \tag{1}$$

where m and n represent the number of classifiers used and output classes; X and Y represent the maximum available classifiers and output classes.

3.3.3 Proposed Fusion of Hybrid Deep Features (FHDF) Network

In the problems of image analysis and classification, the role of feature representation is significant in improving classification performance. As from the literature [16–19], the approach of feature fusion (FF) is found to be a noteworthy and efficient one in biomedical image classification. This approach integrates multiple related feature vectors into a single one, which includes rich information and provides more contribution (representation) as compared with the initial feature inputs. In the literature, there are two techniques followed for feature fusion namely serial and parallel approaches [18]. In the first approach, the idea is to concatenate two feature sets into a union vector. For example, for an image with a dimension of (x, y), if F_1 and F_2 are the two feature sets extricated, then the serially fused one can be represented as $F_S = (x + y)$. In the latter approach, the idea is to concatenation of feature sets using a complex vector. For the above example, the parallel feature fusion with an imaginary component (i) can be represented as $F_P = F_1 + iF_2$.

The above two feature fusion approaches have the limitation of being unable to utilize the original feature inputs because the two methods are aimed at creating a new feature set, either F_S or F_P. And the above approaches suffer from the idea of concatenating multiple feature vectors. In the proposed work, an idea of the fusion of hybrid deep features (FHDF) is employed by combining feature inputs extricated from multiple deep-TL models. Figure 5 illustrates the outline of the proposed FHDF network. In this figure, $F_{V16}, F_{V19}, F_{Res}$, and F_{Den} represent the normalized features extricated from the dense layer (FCL) with 1024 neurons of the four employed TL models: VGG16, VGG19, ResNet50, and DenseNet121. The proposed network is composed of a concatenation layer and a fully connected layer with an activation function as softmax for integrating distinct features. Furthermore, batch normalization and dropout layers are utilized between the above two layers for avoiding overfitting and to optimize the performance during training of data. Herein, the concatenation layer provides the fused feature vectors with a size of 4096. This way of effective feature fusion can be represented as

$$F(i) = \bigcup_{i=1}^{4} F^n(i), \tag{2}$$

where $\bigcup$ indicates the concatenation operation, $F^n(i)$ represents the nth feature vector, and $F(i)$ denotes the output vector of i^{th} fused features.

4 Experiments and Analysis

4.1 Preparation of Input Data for Evaluation

The research evaluation considers three different mammogram datasets, namely mammographic image analysis (MIAS) [20], curated breast imaging subset of digital database for screening mammography (CBIS-DDSM) [21] and INbreast [22] databases. Here, the MIAS database is constituted by a UK research crew. The digital mammograms available in this dataset are publicly accessible in peipa archive of Essex University [20] and downloaded in.pgm format. Here, during the acquisition, the digitization of films is done with a fifty-micro-meter pixel edge, creating the mammogram output of 1024×1024. The image corpus consists of a sum of 322 digital mammogram images corresponding to both side breast parts. The dataset is

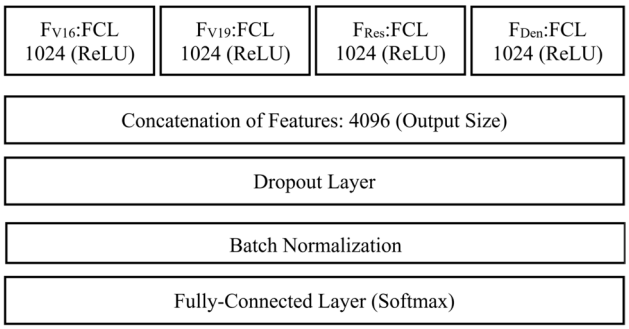

Fig. 5 Proposed framework of fusion of hybrid deep features (FHDF) network

composed of a Mediolateral oblique (MLO) view of acquisition. These mammograms are separated in the dataset as either normal or abnormal samples. In this, well-defined, spiculated, ill-defined, architectural distortion, calcification masses, and asymmetry are characterized as abnormal lesions. In addition, the benign and malignant severities are characterized as abnormal samples.

The second dataset taken for evaluation is the DDSM database which is constituted by the University of South Florida [21]. The database is acquired using 2500 approximate cases with forty-three volumes. In addition, the dataset is constituted using 2 basis views of angles: craniocaudal (CC) and MLO for every patient. The work adopts an MLO view of acquired images as found in the MIAS dataset. Moreover, the research employs the mammogram images from the updated DDSM i.e., the CBIS-DDSM database. The last one is the INbreast dataset [22] where the acquisition device used is Mammo-Novation Siemens which employed amorphous selenium-based solid-state detectors for supporting the resolution of 14-bit with 70-mm pixel sizes. Here, the breast images are available in DICOM format and obtained at an imaging center in association with the National Committee of Data Protection from 2008 to 2010.

In the above-said datasets, standard and good-quality digital mammograms are available. However, the INbreast dataset contains breast images in the form of full-field digital mammograms—FFDM images, which provide better recognition of microcalcification than digital mammograms [22]. Herein, the MIAS and CBIS-DDSM datasets are commonly used benchmark databases that can be useful for evaluating many research methods. In this work, we chose to use the INbreast database because it contains high-quality FFDM images. Furthermore, this dataset is the only available public dataset that comprises FFDM images that give precise and accurate information about every detail. By using these three datasets, the paper aims in classifying the mammogram inputs as either normal or benign or as malignant tumors. The number of mammogram inputs taken for evaluating the proposed CAD system is given in Table 1.

After preprocessing and augmentation, the MIAS, CBIS-DDSM, and INbreast databases comprise of a total of 2576, 4560, and 1432 digital mammograms. The proposed work involves the stratified fashion of data preparation where training and testing sets take 70% and 30% of inputs from both datasets. Herein, the testing set is further subdivided for validation of the work. In addition, the work employed a fivefold cross-validation strategy which makes use of stratified partitioning for its split. This claims that the proposed work confirms that every mammogram input is being tested in an equal manner and thus avoiding any bias error.

Table 1 Digital mammograms for evaluating the proposed work

Database	Output class	Mammogram inputs
MIAS	Normal	207
	Benign	64
	Malignant	51
CBIS-DDSM	Normal	250
	Benign	200
	Malignant	120
INbreast	Normal	66
	Benign	56
	Malignant	57

Table 2 Ablation experimentation on fusion of different features (MIAS dataset)

VGG16	✓	✓	✓	✓
VGG19	✓	✓	✓	✓
ResNet50		✓		✓
DenseNet121			✓	✓
Overall accuracy (%)	92.755	96.507	95.213	98.706
Kappa (κ)	0.861	0.933	0.909	0.975

4.2 Experimental Setup and Ablation Analysis

The proposed work is carried out in a computer system having 16 GB RAM, 1 TB Hard-disk, and an Intel Core i7 processor running on Windows 10 operating system. In addition, the employed system was equipped with a 2 GB configuration of NVIDIA GPU. Moreover, the work utilized Jupyter-based python IDE for implementation and evaluation. The IDE is configured with many machine learning libraries such as Pandas, OpenCV, Sklearn, MatplotLib, Keras, TensorFlow, and PyTorch. For the evaluation of the work, the research adopted the standard overall accuracy and total misclassification cost as metrics for performance analysis. Further, the results are validated using Cohen's kappa (κ) measurement [37]. The above metrics are calculated from the elements of the confusion matrix: TP, FP (true and false positives), TN, and FN (true and false negatives).

With the above experimental setup, an ablation study is carried out to further demonstrate the effectiveness of selecting the best combination of deep features. This is done by considering the fusion of different features, as illustrated in Step 4 of Fig. 1. The ablation experimentation results on the MIAS dataset are summarized in Table 2. Here, the results reveal that every deep feature we consider plays a key role in the classification performance, especially the fusion of all four features. Also reveals clearly that even if only one combination is used, the proposed approach can

be very competitive compared to others. In this way, the work utilizes the fusion of appropriate deep features for the remaining two datasets, which has brought the classification performance supreme.

4.3 Results and Analysis

4.3.1 Overall Performance Analysis

The overall performance of classifiers for the three datasets along with the existing ones is listed in Table 3. This performance is calculated for three-class breast cancer problems with normal, benign, and malignant targets. Moreover, the performance analysis is graphically illustrated in Fig. 6. Here, the total misclassification represents how often the classification model is incorrect in predicting the actual negative and positive output targets, i.e., it can be otherwise termed as classification error. This metric is calculated as a concatenated result of normal vs benign and malignant, benign vs normal, and malignant vs normal cases. The overall classification accuracy is calculated in percentage (%), which gives us the amount of correct outcomes in predicting the actual negative and positive targets. As from the literature [7–14], the overall classification accuracy could be very misleading since the metric does not consider the class-imbalance of input datasets. To overcome this, a robust statistic metric, Cohen's kappa (κ) parameter ($0 \rightarrow 1$) is considered in this work. Here, the metric assesses the degree of agreement among the employed classification models by calculating the inter-rater reliability. In Fig. 6, the overall accuracy (%) is plotted in the primary axis under the total misclassification which is plotted in the secondary axis. And the obtained range ($0 \rightarrow 1$) of the kappa statistic measure is augmented into the range of ($0 \rightarrow 100$) for better visualization of result comparison. As from Figure, VGG16 performs well as compared with the VGG19 model for all three datasets. That is, VGG16 provides better results of accuracy of 92.367% (MIAS), 89.839% (CBIS-DDSM), and 92.308% (INbreast) as compared to the performance of VGG19. The skip connections used in ResNet50 make it to provide a better classification accuracy of 94.049% (MIAS), 93.202% (CBIS-DDSM), and 94.172% (INbreast) when compared with the above two models. Due to improved feature propagation and reduced vanishing-gradient ability, the DenseNet121 model provides a better classification accuracy of 94.825% (MIAS), 94.363% (CBIS-DDSM), and 95.338% (INbreast) as compared to the performance of the above three models. In addition, the ensemble-based LF-approach provides a higher classification accuracy of 96.378% (MIAS), 96.199% (CBIS-DDSM), and 97.203% (INbreast) over the above-discussed models. Consequently, the proposed FHDF technique yields a supreme classification accuracy of 98.706% (MIAS), 97.734% (CBIS-DDSM), and 98.834% (INbreast) over others. The above-attained results are validated further using the kappa coefficient where the highest value of the agreement is obtained for the proposed FHDF method, i.e., 0.975 (MIAS), 0.965 (CBIS-DDSM), and 0.982% (INbreast). In addition, the graph in Fig. 6 shows that whenever the accuracy values are found to be higher, the misclassification rate will become lower.

Table 3 Performance analysis of the proposed work

Database	Classification models	Total misclassification	Overall classification accuracy (%)	Kappa (κ)
MIAS	VGG16	38	92.367	0.854
	VGG19	54	89.651	0.804
	ResNet50	29	94.049	0.886
	DenseNet121	25	94.825	0.901
	LF	19	96.378	0.931
	FHDF	6	98.706	0.975
CBIS-DDSM	VGG16	94	89.839	0.842
	VGG19	117	87.833	0.809
	ResNet50	65	93.202	0.894
	DenseNet121	52	94.363	0.912
	LF	36	96.199	0.941
	FHDF	22	97.734	0.965
INbreast	VGG16	22	92.308	0.884
	VGG19	32	90.21	0.853
	ResNet50	18	94.172	0.912
	DenseNet121	13	95.338	0.930
	LF	7	97.203	0.958
	FHDF	3	98.834	0.982

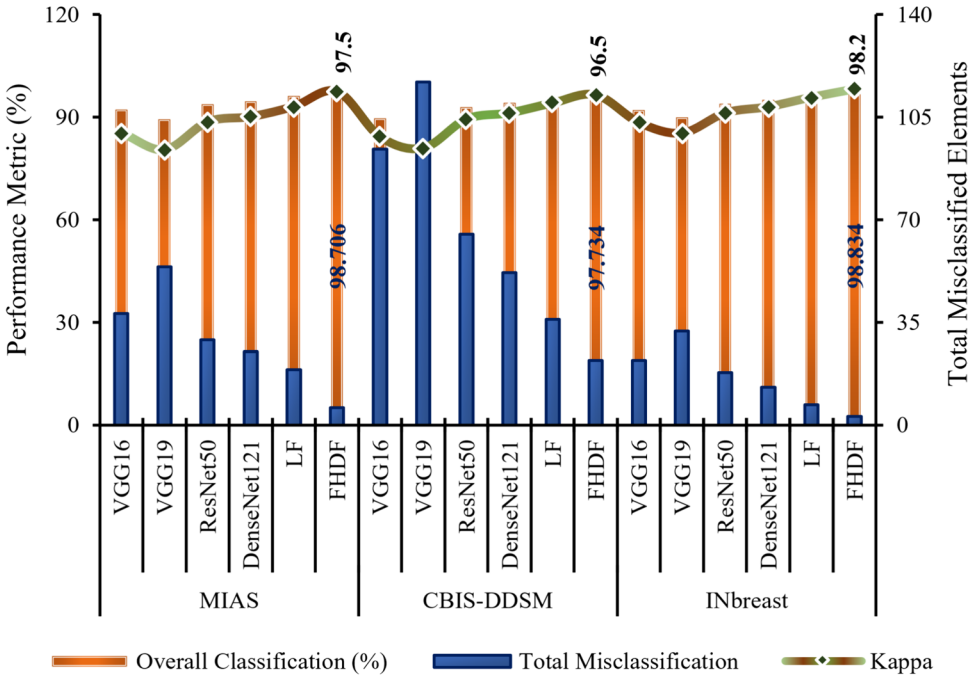

Fig. 6 Graphical performance analysis of the proposed method

Accordingly, the proposed method has the least misclassification rate, corresponding to all three datasets.

4.3.2 Insight Performance Analysis

The above discussion based on Table 3 and Fig. 6 focussed on the overall performance analysis. However, the research focuses on both detection and classification of severities. That is, detecting the disease as either normal or abnormal, and further classifying the abnormal severities as either benign cases or malignant. This formulates the solution to a three-class classification problem where the mammogram inputs need to be classified into three output targets namely normal, benign, and malignant. Hence, the individual or insight performance analysis of all classification models should need to be done for each output target, respectively. Furthermore, the insightful analysis of the classifier's performance is significant because of the unavoidable class imbalance problem in the employed input datasets.

Accuracy metric highlights how well the model correctly discriminates normal, benign, and malignant cases with respect to the total inputs [13]. Precision metric concentrates on providing how much fraction of predicted positive cases is actually positives [13]. Recall metric calculates how well the model predicts the positive cases correctly with respect to total actual positives [13]. F1 score is calculated as a harmonic mean of two metrics: recall and precision [14]. Here, accuracy metric is greater only if the input dataset is symmetric, i.e., the values of false negatives and false positives are almost the same [14]. But the research employed three different asymmetric datasets. When the amount of false negatives and false positives are not same, then precision and recall measures can be used. As from the precision and recall measures definition, both cannot be higher. For a model, if recall is increased, then precision will be lower and vice-versa. Thus, F1 score is a metric which gives a harmonic mean of the above two measures. Here, the harmonic mean is more suitable for calculating ratios between recall and precision. So, F1 score will be higher only if both recall and precision are higher. Thus the research work utilizes the above-discussed metrics for assessing the employed models.

Table 4 illustrates the confusion matrix obtained for the test data of MIAS, CBIS-DDSM, and INbreast datasets using the proposed FHDF technique. In this way, the individual performance analysis of classification models for the three-class classification is tabulated in Tables 5, 6, and 7, respectively. The third column (no. of classified outputs) represents the overall classified samples for each output class, as shown in Table 4. And Fig. 7 illustrates a plot that shows the performance analysis of LF and the proposed FHDF approach for each class of the MIAS, CBIS-DDSM, and INbreast databases.

From Tables 5, 6, and 7 for the employed mammogram databases, the VGG16, VGG19, ResNet50, DenseNet121, and LF models give their maximum performance of classification while classifying the normal cases. Hence, the substantial difficulty lies in discriminating the abnormal severities (benign/malignant), which is why these models provide overall poor performance as portrayed before. Accordingly, the VGG16 model yields the highest classification performance of accuracy (95.08%), precision (96.36%), recall (95.97%), and F1 score (96.12%) in

Table 4 Test data confusion matrix for MIAS, CBIS-DDSM, and INbreast datasets using FHDF approach

Truth Data

Classifier Results	Class Targets	Normal	Benign	Malignant	Classification Overall	User's Accuracy (Precision %)
	Normal	**494**	1	2	497	99.39
	Benign	2	**151**	2	155	97.41
	Malignant	1	2	**118**	121	97.52
	Truth Overall	497	154	122	773	
	Producer's Accuracy (Recall %)	99.39	98.05	96.72		

(a) Test Data – MIAS Dataset

Truth Data

Classifier Results	Class Targets	Normal	Benign	Malignant	Classification Overall	User's Accuracy (Precision %)
	Normal	**587**	7	2	596	98.49
	Benign	9	**469**	5	483	97.10
	Malignant	4	4	**281**	289	97.23
	Truth Overall	600	480	288	1368	
	Producer's Accuracy (Recall %)	97.83	97.70	97.56		

(b) Test Data – CBIS-DDSM Dataset

Truth Data

Classifier Results	Class Targets	Normal	Benign	Malignant	Classification Overall	User's Accuracy (Precision %)
	Normal	**156**	0	1	157	99.36
	Benign	1	**133**	1	135	98.51
	Malignant	1	1	**135**	137	98.54
	Truth Overall	158	134	137	429	
	Producer's Accuracy (Recall %)	98.73	99.25	98.54		

(c) Test Data – INbreast Dataset

discriminating the normal cases for the mammograms of the MIAS dataset. The VGG19 model yields the highest classification performance of accuracy (93.01%), precision (95.48%), recall (93.56%), and F1 score (95.44%) in discriminating the normal cases for the mammograms of the MIAS dataset. The ResNet50 model yields the highest classification performance of accuracy (96.25%), precision (97.36%), recall (96.78%), and F1 score (97.49%) in discriminating the normal cases for the mammograms of the MIAS dataset. The DenseNet121 model yields the highest classification performance of accuracy (96.77%), precision (97.77%), recall (97.18%), and F1 score (97.57%) in discriminating the normal cases for the mammograms of the MIAS dataset. But in the case of LF and FHDF models, their obtained classification result is good irrespective of the database type, i.e., in specific, the proposed FHDF approach of classification provides superior classification accuracy in the range of 98.17–99.3%. Furthermore, as compared with the four transfer learning models, Fig. 7

Table 5 Individual performance analysis of the classification models for MIAS dataset

Target label	Mammogram inputs	No. of classified outputs	Acc (%)	Pre (%)	Recall (%)	F1 score (%)
VGG16 model						
Normal	497	495	95.08	96.36	95.97	96.12
Benign	154	159	94.70	85.53	88.31	87.43
Malignant	122	119	94.95	84.87	82.78	84.59
VGG19 model						
Normal	497	487	93.01	95.48	93.56	95.44
Benign	154	161	92.88	80.74	84.41	83.16
Malignant	122	125	93.4	78.40	80.32	79.28
ResNet50 model						
Normal	497	494	96.25	97.36	96.78	97.49
Benign	154	158	95.60	87.97	90.26	88.96
Malignant	122	121	96.25	88.43	87.71	88.17
DenseNet121 model						
Normal	497	494	96.77	97.77	97.18	97.57
Benign	154	155	96.25	90.32	90.90	91.42
Malignant	122	124	96.64	88.71	90.16	89.76
LF technique						
Normal	497	498	97.54	97.99	98.18	98.31
Benign	154	155	97.54	93.54	94.15	94.29
Malignant	122	120	97.67	93.33	91.80	93.68
FHDF technique						
Normal	497	497	99.22	99.39	99.39	99.56
Benign	154	155	99.09	97.41	98.05	98.49
Malignant	122	121	99.09	97.52	96.72	97.34

reveals that the proposed approach performs better for the input of FFDM images taken from INbreast data. In addition, the proposed approach provides superior results in discriminating both normal and abnormal severity cases for all data inputs. This only makes the proposed FHDF classification approach to yield supreme overall-classification performance as illustrated in Fig. 6 and Table 3. Hence, from Tables 5, 6, 7 and Figs. 6, 7, the proposed methodology has outperformed in discriminating whether the mammogram is normal or abnormal and if there is any abnormality, then it is fine enough to discriminate further the severities as either benign case or malignant class. The above results are attained not only due to the use of the FHDF model but also because of the suitable preprocessing approach (Fig. 2) applied with the appropriate fusion of deep features. In addition to the above performance and comparative analysis, the ANOVA test is performed for the employed classification models for further statistical validation. Table 8 illustrates the analysis of variance (ANOVA) results and its statistical examination for the employed problem. As listed, the higher F value (42.06386) and the very small P value (3.38E−07) illustrate the significance of the proposed methodology for multiclass breast cancer classification.

4.4 Performance Comparison of Proposed CAD Model with the Existing Research Models

While comparing the research on breast cancer classification problems with other biomedical research works, the researchers are actively endeavoring to give new solutions for early breast cancer diagnosis. However, the comparison among the research works is implicitly difficult due to several factors such as employed mammograms with distinct databases, the amount of data inputs, input samples chosen for assessment, the approach of extricating and selecting feature vectors, parameter tuning, classification strategy, and the way of evaluating the performance. The performance comparison of the proposed approach is listed and summarized from several findings as given in Table 9.

Table 6 Individual performance analysis of the classification models for CBIS-DDSM dataset

Target label	Mammogram inputs	No. of classified outputs	Acc (%)	Pre (%)	Recall (%)	F1 score (%)
VGG16 model						
Normal	600	578	93.13	93.77	90.33	91.96
Benign	480	484	93.27	90.08	90.83	90.44
Malignant	288	306	93.27	82.02	87.15	85.28
VGG19 model						
Normal	600	571	91.45	92.29	87.83	90.39
Benign	480	477	91.74	88.47	87.91	88.27
Malignant	288	320	92.11	78.12	86.80	82.43
ResNet50 model						
Normal	600	593	95.25	95.11	94.00	95.24
Benign	480	481	95.39	93.34	93.54	93.44
Malignant	288	294	95.76	89.11	90.97	90.57
DenseNet121 model						
Normal	600	594	96.19	95.96	95.31	95.87
Benign	480	483	95.97	96.99	94.58	94.51
Malignant	288	289	96.56	91.69	92.01	92.39
LF technique						
Normal	600	596	97.37	97.31	96.67	97.49
Benign	480	478	97.08	96.02	95.62	96.34
Malignant	288	294	97.95	94.21	96.18	95.41
FHDF technique						
Normal	600	596	98.39	98.49	97.83	98.19
Benign	480	483	98.17	97.10	97.70	97.46
Malignant	288	289	98.90	97.23	97.56	97.33

5 Discussion on the Findings

In recent years, the evolution of DL algorithms has helped more in solving real-time problems in the bio-medical field. Breast cancer classification using digital mammograms can support physicians in identifying the tumors in earlier stages, which is crucial to preventing cancer deaths.

The proposed work of three-class classification is evaluated using three different mammogram datasets: MIAS, CBIS-DDSM, and INbreast. And these databases are publicly available one for research purposes. In the preprocessing stage, the unwanted noise is removed using a simple adaptive median filter. But in the literature [38–44], a few works have not employed any filtering techniques for noise removal, whereas some works employed filters such as simple median filters. But the thing to be noted is the noise has to be removed without disturbing the unaffected pixels. So, the work utilized an adaptive median filtering approach. In the next step, mammograms are enhanced using an adaptive histogram technique to improve the contrast of microcalcification and pectoral muscle regions without overexposure. As a result, the Hough transform and canny edge detection provides clear pectoral removed mammograms with a better enhancement of microcalcification as shown in Fig. 2.

Then, the challenge is to detect whether the input is normal or abnormal; if found to be abnormal, it needs to be classified as benign or malignant. For this, the research proposes the FHDF approach using transfer learning to detect and classify breast cancer. Here, the important thing is the selection of deep CNNs used for feature extrication. The research performed a lot of ablation experiments and found that the fusion of VGG16, VGG19, ResNet50, and DenseNet121 gives a very competitive classification performance, as illustrated in Tables 2 and 3. While assessing the overall performance, the results need to be validated through any consistent validation metric. The works of [38–44] note that the research findings should be properly validated. After validating using Cohen's kappa (κ), the attained results were validated. The value for the proposed approach is highly closer to 1, representing that the proposed approach provides supreme classification performance for breast cancer problems. Since it is a multiclass classification, the insight performance analysis is presented in Tables 5, 6, and 7. The findings of these tables illustrate that the utilized classification architectures are well at discriminating the normal and abnormal mammograms. And they lag in further classifying benign and malignant samples. However, the proposed research of the FHDF approach provides superior

Table 7 Individual performance analysis of the classification models for INbreast dataset

Target label	Mammogram inputs	No. of classified outputs	Acc (%)	Pre (%)	Recall (%)	F1 score (%)
VGG16 model						
Normal	158	156	94.87	93.59	92.40	93.38
Benign	134	134	95.34	92.53	92.53	93.29
Malignant	137	139	94.41	90.64	91.97	91.63
VGG19 model						
Normal	158	158	92.54	89.87	89.87	90.22
Benign	134	134	93.47	89.55	89.55	90.47
Malignant	137	137	94.41	91.24	91.24	91.50
ResNet50 model						
Normal	158	158	95.80	94.30	94.30	94.19
Benign	134	134	96.74	94.77	94.77	95.36
Malignant	137	137	95.80	93.43	93.43	93.58
DenseNet121 model						
Normal	158	157	96.97	96.17	95.57	96.46
Benign	134	136	97.20	94.85	96.26	96.39
Malignant	137	136	96.50	94.85	94.16	95.27
LF technique						
Normal	158	157	98.37	98.08	97.46	98.39
Benign	134	135	98.37	97.03	97.76	97.18
Malignant	137	137	97.67	96.35	96.35	96.44
FHDF technique						
Normal	158	157	99.30	99.36	98.73	98.96
Benign	134	135	99.30	98.51	99.25	99.47
Malignant	137	137	99.07	98.54	98.54	98.83

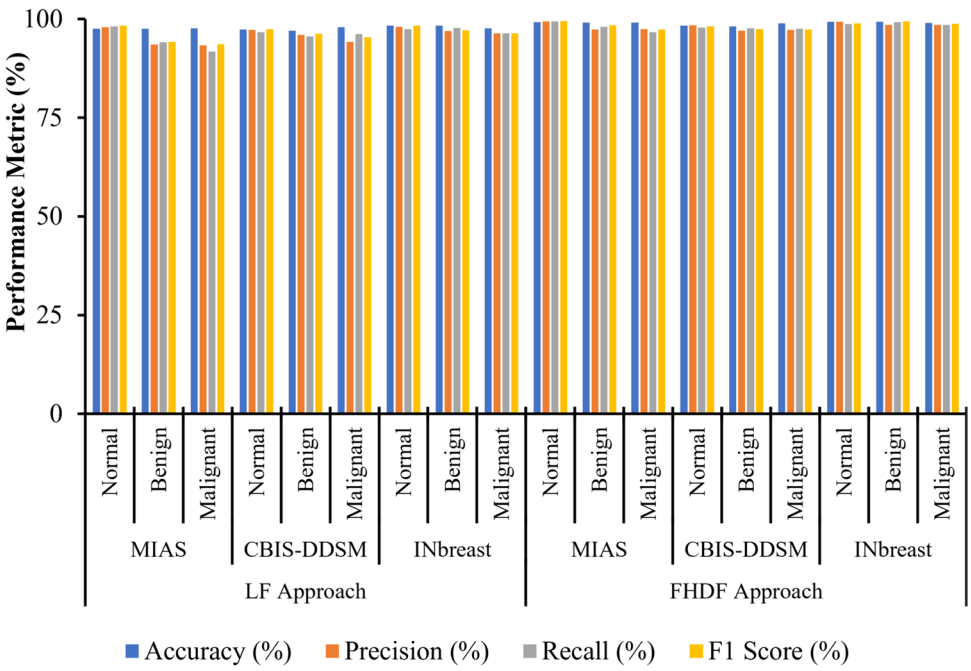

Fig. 7 Individual performance analysis of LF and FHDF approaches for each class of DDSM, MIAS, and INbreast datasets

Table 8 ANOVA statistical analysis of the proposed methodology for multi-class breast cancer classification

Source of variation	SS	*df*	MS	*F*	*P* value	*F* crit
Between classifier models	167.9662	5	33.59323	42.06386	3.38E−07	3.105875
Within classifier models	9.583495	12	0.798625			
Total	177.5497	17				

Table 9 Performance comparison of the proposed CAD model with the existing research models for breast cancer classification

Reference works	Techniques used	Target problem	Accuracy (%)
Prathibha and Mohan (2018) [38]	Multi-resolution transform with a CNN model	Multiclass classification	85.4 (DDSM)
Safdarian and Hedyezadeh (2019) [39]	Support vector machine (SVM) with boundary descriptor feature inputs	Multiclass classification	97 (DDSM)
Akila et al. (2019) [40]	Multiscale all CNN model	Multiclass Classification	96 (MIAS)
Figlu et al. (2020) [41]	Optimized kernel ELM architecture	Multiclass classification	97.4 (MIAS) 92.6 (DDSM)
Abeer et al. (2021) [42]	Transfer learning using VGG16 model	Multiclass classification	96.8 (MIAS)
Karthiga et al. (2022) [43]	Deep CNN with pretrained models	Binary classification	95.9 (MIAS) 96.5 (INbreast)
Khaoula et al. 2022 [44]	Apriori dynamic selection with SVM	Multiclass classification	96.4 (DDSM) 75.8 (INbreast)
Proposed Work	Fusion of hybrid deep features (FHDF) approach	Multiclass classification	98.7 (MIAS) 97.7 (CBIS-DDSM) 98.8 (INbreast)

classification performance as compared with the existing works. Finally, the paper compared the classification performance of the proposed method with standard pretrained models, late fusion technique, and other existing approaches. It revealed that the proposed FHDF approach outperforms them, thus establishing the novelty of the framework. The potential limitations of the proposed work involve the computational complexity involved during the fusion of deep features obtained from distinct models. In addition, as from Tables 5, 6, and 7, it is noted that the proposed approach modestly struggled to recognize the malignant mammograms as compared with the other cases. The above limitations will be looked out in our future proposals.

6 Conclusion and Future Work

The proposed study discusses the design of a robust CAD model for enhancing the multiclass classification of breast cancer data. For this, the work employed the recent emerging deep learning strategy, i.e., four distinct pre-trained convolutional neural networks are employed. After freezing and fine-tuning the pretrained models, each model's deep features are extricated. Before this task, the mammogram images are appropriately pre-processed for their removal of noise, pectoral muscle, and unwanted regions. In addition, pre-processed mammograms are augmented enough and partitioned in a stratified manner to overcome the problem of overfitting and bias errors. In this way, the above work is evaluated using the digital mammograms of MIAS, CBIS-DDSM, and INbreast databases with VGG16, VGG19, ResNet50, DenseNet121, Late Fusion, and Fusion of Hybrid Deep Features models. For evaluation, the overall and insight performance analysis is done for better analysis of classification models. Accordingly, the proposed FHDF approach provides a supreme result of 98.70% (MIAS), 97.73% (CBIS-DDSM), and 98.83% (INbreast) classification accuracy as compared with the standalone and existing classification models. Moreover, the above results are validated properly through kappa analysis: 0.975 (MIAS), 0.965 (CBIS-DDSM), and 0.982 (INbreast). The future direction will involve extending the FHDF approach for clinical mammograms with different preprocessing methods. The proposed approach involves an effective way of fusing deep features extracted from different mammogram datasets. Furthermore, the effectiveness of the proposed approach will be applied to the same breast cancer problem but for multimodal datasets.

Acknowledgements This work was supported by Princess Nourah bint Abdulrahman University Researchers Supporting Project number(PNURSP2024R432), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia. This work was supported by the Deanship of Scientific Research, Vice President for Graduate Studies and Scientific Research, King Faisal University, Saudi Arabia [KFU241404].

Author Contributions S.C and B.N took care of the review of literature and methodology. S.B.K and M.T.R have done the formal analysis,

data collection and investigation. A.A has done the initial drafting and statistical analysis. V.K.V and E.A have supervised the overall project. All the authors of the article have read and approved the final article.

Funding Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2024R432), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

Data Availability Statement Data used for the findings will be shared by the corresponding author upon request.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Ethics approval and consent to participate Not applicable.

Consent for publication Not applicable as the work is carried out on publicly available dataset.

References

1. Wilkinson, L., Gathani, T.: Understanding breast cancer as a global health concern. Br. J. Radiol. **95**(1130), 20211033 (2022)
2. Zemni, I., Kacem, M., Dhouib, W., Bennasrallah, C., Hadhri, R., Abroug, H., Ben Fredj, M., Mokni, M., Bouanene, I., Belguith, A.S.: Breast cancer incidence and predictions (Monastir, Tunisia: 2002–2030): a registry-based study. PLoS ONE **17**(5), e0268035 (2022)
3. Yari, Y., Nguyen, T.V., Nguyen, H.T.: Deep learning applied for histological diagnosis of breast cancer. IEEE Access **8**, 162432–162448 (2020)
4. Abirami, C., Harikumar, R., Chakravarthy, S.S.: Performance analysis and detection of micro calcification in digital mammograms using wavelet features. In: 2016 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET), pp. 2327–2331. IEEE (2016)
5. Yu, Z., Song, M., Chouchane, L., Ma, X.: Functional genomic analysis of breast cancer metastasis: implications for diagnosis and therapy. Cancers **13**(13), 3276 (2021)
6. SR, S.C., Rajaguru, H.: A systematic review on screening, examining and classification of breast cancer. In: 2021 Smart Technologies, Communication and Robotics (STCR), pp. 1–4 (2021)
7. Dhungel, N., Carneiro, G., Bradley, A.P.: A deep learning approach for the analysis of masses in mammograms with minimal user intervention. Med. Image Anal. **37**, 114–128 (2017). https://doi.org/10.1016/J.MEDIA.2017.01.009
8. Thijs, K., et al.: Large scale deep learning for computer aided detection of mammographic lesions. Med. Image Anal. **35**, 303–312 (2017)
9. Xiaofei, Z., et al.: Classification of whole mammogram and tomosynthesis images using deep convolutional neural networks. IEEE Trans. Nanobiosci. **17**(3), 237–242 (2018)
10. Yemini, M., Zigel, Y., Lederman, D.: Detecting masses in mammograms using convolutional neural networks and transfer learning. In: 2018 IEEE International Conference on the Science of Electrical Engineering in Israel (ICSEE), pp. 1–4. IEEE (2018)
11. Chougrad, H., Zouaki, H., Alheyane, O.: Multi-label transfer learning for the early diagnosis of breast cancer. Neurocomputing **392**, 168–180 (2020)
12. Shu, X., Zhang, L., Wang, Z., Lv, Q., Yi, Z.: Deep neural networks with region-based pooling structures for mammographic image classification. IEEE Trans. Med. Imaging **39**, 2246–2255 (2020)
13. Sannasi Chakravarthy, S.R., Rajaguru, H.: Deep-features with Bayesian optimized classifiers for the breast cancer diagnosis. Int. J. Imaging Syst. Technol. **31**(4), 1861–1881 (2021)
14. Chakravarthy, S.S., Rajaguru, H.: Automatic detection and classification of mammograms using improved extreme learning machine with deep learning. IRBM **43**(1), 49–61 (2022)
15. Sannasi Chakravarthy, S.R., Bharanidharan, N., Rajaguru, H.: Multi-deep CNN based experimentations for early diagnosis of breast cancer. IETE J. Res. **69**(10), 7326–7341 (2022)
16. Benzebouchi, N.E., Azizi, N., Ashour, A.S., Dey, N., Sherratt, R.S.: Multi-modal classifier fusion with feature cooperation for glaucoma diagnosis. J. Exp. Theor. Artif. Intell. **31**(6), 841–874 (2019). https://doi.org/10.1080/0952813X.2019.1653383
17. Xue, W., Dai, X., Liu, L.: Remote sensing scene classification based on multi-structure deep features fusion. IEEE Access **8**(1), 28746–28755 (2020). https://doi.org/10.1109/ACCESS.2020.2968771
18. Kumar, A., Kim, J., Lyndon, D., Fulham, M., Feng, D.: An ensemble of fine-tuned convolutional neural networks for medical image classification. IEEE J. Biomed. Health. Inf. **21**(1), 31–40 (2016). https://doi.org/10.1109/JBHI.2016.2635663
19. Amin, J., Sharif, A., Gul, N., Anjum, M.A., Nisar, M.W., Azam, F., Bukhari, S.A.C.: Integrated design of deep features fusion for localization and classification of skin cancer. Pattern Recogn. Lett. **131**, 63–70 (2020). https://doi.org/10.1016/j.patrec.2019.11.042
20. Suckling, J., Parker, J., Dance, D., Astley, S., Hutt, I.: Mammographic image analysis society (mias) database v1. 21 (2015). https://www.repository.cam.ac.uk/handle/1810/250394. Accessed 28 Mar 2021
21. Heath, M., Bowyer, K., Kopans, D., Moore, R., Kegelmeyer, P.: In: Yaffe, M.J. (ed.), Proceedings of the fifth international workshop on digital mammography, pp. 212–218. Medical Physics Publishing (2001)
22. Moreira, I.C., Amaral, I., Domingues, I., Cardoso, A., João Cardoso, M., Cardoso, J.S.: INbreast: toward a full-field digital mammographic database. Acad. Radiol. **19**, 236–248 (2012). https://doi.org/10.1016/j.acra.2011.09.014
23. Sannasi Chakravarthy, S.R., Rajaguru, H.: Detection and classification of microcalcification from digital mammograms with firefly algorithm, extreme learningmachine and non-linear regression models: a comparison. Int. J. Imaging Syst. Technol. **30**(1), 126–146 (2020). https://doi.org/10.1002/ima.22364
24. Rao, B.S.: Dynamic histogram equalization for contrast enhancement for digital images. Appl. Soft Comput. **89**, 106114 (2020)
25. Yaman, S., Karakaya, B., Erol, Y.: Real time edge detection via IP-core based sobel filter on FPGA. In: 2019 International Conference on Applied Automation and Industrial Diagnostics (ICAAID), vol. 1, pp. 1–4. IEEE (2019)
26. Gong, L.H., Tian, C., Zou, W.P., Zhou, N.R.: Robust and imperceptible watermarking scheme based on Canny edge detection

and SVD in the contourlet domain. Multimed. Tools Appl. **80**(1), 439–461 (2021)
27. Iqbal, B., Iqbal, W., Khan, N., Mahmood, A., Erradi, A.: Canny edge detection and Hough transform for high resolution video streams using Hadoop and Spark. Clust. Comput. **23**(1), 397–408 (2020)
28. Zubair Rahman, A.M.J., Gupta, M., Aarathi, S., et al.: Advanced AI-driven approach for enhanced brain tumor detection from MRI images utilizing EfficientNetB2 with equalization and homomorphic filtering. BMC Med. Inform. Decis. Mak. **24**, 113 (2024). https://doi.org/10.1186/s12911-024-02519-x
29. Satheesh Kumar, J., Vinoth Kumar, V., Mahesh, T.R., et al.: Detection of Marchiafava Bignami disease using distinct deep learning techniques in medical diagnostics. BMC Med. Imaging **24**, 100 (2024). https://doi.org/10.1186/s12880-024-01283-8
30. Ahmed, S.T. et al.: PrEGAN: privacy enhanced clinical EMR generation: leveraging GAN model for customer de-identification. In: IEEE Transactions on Consumer Electronics. https://doi.org/10.1109/TCE.2024.3386222
31. Fourcade, A., Khonsari, R.H.: Deep learning in medical image analysis: a third eye for doctors. J. Stomatol. Oral Maxillofac. Surg. **120**(4), 279–288 (2019)
32. Chlap, P., Min, H., Vandenberg, N., Dowling, J., Holloway, L., Haworth, A.: A review of medical image data augmentation techniques for deep learning applications. J. Med. Imaging Radiat. Oncol. **65**(5), 545–563 (2021)
33. Morid, M.A., Borjali, A., Del Fiol, G.: A scoping review of transfer learning research on medical image analysis using ImageNet. Comput. Biol. Med. **128**, 104115 (2021)
34. Wan, Z., Yang, R., Huang, M., Zeng, N., Liu, X.: A review on transfer learning in EEG signal analysis. Neurocomputing **421**, 1–14 (2021)
35. Li, W., Huang, R., Li, J., Liao, Y., Chen, Z., He, G., Yan, R., Gryllias, K.: A perspective survey on deep transfer learning for fault diagnosis in industrial scenarios: theories, applications and challenges. Mech. Syst. Signal Process. **167**, 108487 (2022)
36. Mehta, S., Paunwala, C., Vaidya, B.: CNN based traffic sign classification using Adam optimizer. In: 2019 International Conference on Intelligent Computing and Control Systems (ICCS), pp. 1293–1298. IEEE (2019)
37. Bing, P., Liu, Y., Liu, W., Zhou, J., Zhu, L.: Electrocardiogram classification using TSST-based spectrogram and ConViT. Front. Cardiovasc. Med. (2022). https://doi.org/10.3389/fcvm.2022.983543
38. Xue, X., Zhao, S., Xu, M., Li, Y., Liu, W., Qin, H.: Circular RNA_0000326 accelerates breast cancer development via modulation of the miR-9-3p/YAP1 axis. Neoplasma **70**(3), 430–442 (2023). https://doi.org/10.4149/neo_2023_220904N894
39. Jiang, Z., Yang, L., Jin, L., Yi, L., Bing, P., Zhou, J., Yang, J.: Identification of novel cuproptosis-related lncRNA signatures to predict the prognosis and immune microenvironment of breast cancer patients. Front. Oncol. (2022). https://doi.org/10.3389/fonc.2022.988680
40. Yang, C., Sheng, D., Yang, B., Zheng, W., Liu, C.: A dual-domain diffusion model for sparse-view CT reconstruction. IEEE Signal Process. Lett. (2024). https://doi.org/10.1109/LSP.2024.3392690
41. Zheng, W., Lu, S., Yang, Y., Yin, Z., Yin, L., Ali, H.: Lightweight transformer image feature extraction network. PeerJ Comput. Sci. **10**, e1755 (2024). https://doi.org/10.7717/peerj-cs.1755
42. Saber, A., Sakr, M., Abou-Seida, O., Keshk, A.: A novel transfer-learning model for automatic detection and classification of breast cancer based deep CNN. Kafrelsheikh J. Inf. Sci. **2**(1), 1–9 (2021)
43. Karthiga, R., Narasimhan, K., Amirtharajan, R.: Diagnosis of breast cancer for modern mammography using artificial intelligence. Math. Comput. Simul **202**, 316–330 (2022)
44. Soulami, K.B., Kaabouch, N., Saidi, M.N.: Breast cancer: three-class masses classification in mammograms using apriori dynamic selection. Concurr. Comput. Pract. Exp. **34**(24), e7233 (2022)