ORIGINAL ARTICLE

An efficient deep convolutional network model using mask images for multiclass classification of breast cancer ultrasound images

Kadir Can Burçak[1]

Received: 11 January 2025 / Accepted: 3 September 2025 / Published online: 16 September 2025

Abstract
Breast cancer begins in the breast tissue when mutated cells grow out of control and eventually form a tumor. One of the most common causes of death among women worldwide is breast cancer. Early diagnosis and treatment can increase the likelihood of cancer prevention and recovery. Breast ultrasound analysis performed by medical professionals requires high competence in interpreting images, is time-consuming, and creates a negative situation in terms of the treatment process. Artificial intelligence methods have shown great success in the development of medical diagnosis and diagnostic models. When combined with artificial intelligence techniques, breast ultrasound images can produce good results in the detection and classification of breast cancer. This study focused on multiclass classification of breast cancer ultrasound images collected via ultrasound scanning via deep learning methods. In the first stage of the model, image preprocessing was performed to reduce the negative effects of poor contrast, unclear target areas, image transients, and unbalanced image classes. In the second stage, a new dataset was created by using mask images. The third stage consists of training two different datasets on a block basis with the fine-tuned MobiLenetV3Large, MobiLenetV3Small, MobiLenetV2, ResNet101 and EfficientNetV2M deep learning models and, finally, the classification stage. As a result of the experiments conducted on the breast ultrasound image datasets prepared under different conditions, the proposed model achieved 100% accuracy and F1 score for the training data, a 99.15% accuracy and a 99.21% F1 score for the test data, and better performance efficiency than the other models did.

Keywords Deep learning · Medical ultrasound image classification · Breast cancer · Transfer learning · Depthwise separable convolutions

1 Introduction

According to research conducted in 2018, cancer caused the deaths of ten million people worldwide and is the second leading cause of death. The most common types of cancer in men are lung, liver, stomach, and prostate cancers, whereas the most common cancers in women are breast, lung, and cervical cancers [1, 2]. The World Health Organization reported that breast cancer was the most common type of cancer worldwide as of 2021. The most commonly diagnosed cancer in women is breast cancer and skin cancer. In addition, the report states that breast cancer accounts for 12% of new cancer cases annually. Approximately 30% of newly diagnosed cancers in women in the coming years are estimated to be breast cancer [3]. Effective diagnosis of breast cancer in the early stages can increase a patient's chances of recovery, prevent the cancer from spreading to healthy cells, and reduce

mortality. The radiologists' correct classification of breast cancer is highly important in the course of the disease. However, visual analysis of ultrasound images is a process that requires high competence. There may be differences in the diagnosis of the same case among medical experts. Recently, researchers have focused on computer-aided systems (CAD) to overcome current problems. Cancer, which causes the death of millions of people worldwide, highlights the importance of the use of artificial intelligence in medical imaging [4, 5]. Machine learning and deep learning, which are advanced image recognition methods, have achieved promising results for researchers. In particular, deep learning can automatically extract hidden shapes and features in medical images, making better use of manual methods based on expert experience [6–8]. Deep learning is an area where significant investment and research have been made, especially to create systems that can perform feature detection from large amounts of unlabeled training data. Deep learning, a subdiscipline of artificial intelligence, has significantly increased the use of CAD systems in the healthcare sector owing to the use of digital medical image formats [5]. The main purpose of using CAD in medical imaging is to transform the time and costly process of manual processing of medical images into a process that benefits patients and institutions conducting research. The most successful model is created by reflecting the experiences of medical experts, such as radiologists, to the CAD systems to be developed [6]. In addition, ultrasound imaging stands out as an appropriate image capture method. Although the quality of the images is low, this problem can be eliminated for successful detection with appropriate image preprocessing algorithms. A hybrid technique compatible with image enhancement and deep learning can be developed [9]. Deep learning algorithms need large datasets to achieve high performance. However, the insufficiency of annotated data in medical images and the low sharing of data due to ethical rules make it difficult to obtain large datasets [10, 11]. On the basis of studies conducted in recent years, there have been significant developments in the field of medical image analysis with deep learning, and it is progressing rapidly. The inadequacy of the dataset in medical image sets, the difficulty of the study, and the permeability and complexity of ultrasound images also reveal limitations in breast cancer diagnosis. The main purpose of our study is to propose a new breast cancer diagnosis mechanism with high reliability and performance, which uses depthwise separable convolutions in multiclass classification of breast cancer ultrasound images by providing solutions to the above limitations. It also provides efficient deployment of these models on mobile devices and embedded systems. The main contribution of this article is the presentation of a new deep learning architecture for breast cancer detection, which is influenced by the well-known Mobile LeNet (MobiLeNet) network, uses depthwise separable convolutions for information transfer between layers, and it outperforms other methods in terms of usability of mask images, usage efficiency and superior performance. In this study, two approaches are adopted in the multiclass classification of medical images of patients with breast cancer collected via ultrasound (US) with deep learning methods. The proposed model minimizes the negative effects of uncertain target regions, and image permeability in the first stage of ultrasound images and image preprocessing are performed for unbalanced image classes. In the second stage, new datasets were created on the basis of the idea that mask images could be used. In the third stage, different datasets were trained with block-based fine-tuned deep learning models, namely, MobiLeNetV2, MobiLeNetV3Small, MobiLeNetV3Large, ResNet-RS101 and EfficientNetV2M, and finally, the classification stages. As a result of the experiments, the proposed model achieved a 100% F1score and accuracy for the training data, 99.15% accuracy and 99.57% F1 score for the test data, and outperformed the other models. In addition, detailed information about the structure of our proposed framework, a block-based comparison of MobiLeNet architectures and the performances of other deep learning methodologies on the datasets are reported.

The rest of this paper is organized as follows. Section 2 provides an overview of the background concepts used in this paper, along with an in-depth review of related work. Section 3 presents the architecture, methodology, and operation of the proposed approach in detail. Section 4 reports the experimental setup, performance index, findings of the experiments, and comparison with state-of-the-art technologies. Finally, Sect. 5 concludes the paper by highlighting the open research implications.

2 Related works

In this section, the methods in the literature related to the classification of ultrasound images with artificial intelligence methods and image enhancement in the medical decision-making process are discussed.

Chiao et al. developed a deep learning-based model using mask regions for benign and malignant classification of ultrasound images in breast lesion detection. It achieved 85% accuracy in binary classification of ultrasound images [12]. Moon et al. performed binary classification via VGG, DenseNet-121 and DenseNet-40 ensemble learning methods on a specially collected dataset and the BUSI breast cancer ultrasound dataset. The diagnostic performance for the BUSI dataset was 90.77% [13]. Shinohara et al. trained a network on three pretrained models via a transfer learning technique with ultrasound images to diagnose carpal tunnel syndrome. The network achieved 96%accuracy and 94% recall performance. The proposed technique provides accurate estimation of the median nerve cross-sectional area without measuring it [14]. Podda et al. evaluated and compared the superiority of ensemble methods over individual networks using the publicly available dataset BUSI with a recurrent multilayer approach using convolutional neural networks. In the BUSI dataset, 91% classification accuracy was achieved [15]. Xu et al. proposed a deep neural network consisting of a restricted Boltzmann machine for the determination of tumor regions with multiple ROI regions and the comprehensive selection of features to diagnose breast cancer. In the binary classification of breast cancer images with a support vector machine (SVM), 87.1% AUC and 84.9% accuracy were achieved [16]. Wu et al. proposed the MCRUNet model, which is based on a multistage U-Net architecture, for the evaluation of infiltrating lymphocytes in breast cancer tumor ultrasound images. By using the gray feature enhancement module for image reconstruction and normalization, it achieved an accuracy of 89.31%, an accuracy of 85.71%, an accuracy of 83.3% and an F1 score of 86.5% in the test group [17]. Alotaibi et al. developed a three-stage image processing scheme using the VGG-19 deep learning model with a transfer learning method and RGB fusion. The developed model achieved 85.6% accuracy, 87.4% recall, 84.8% specificity, 75.4% sensitivity and 87.4% F1score on the BUSI dataset with denoising, RGB fusion and ROI highlighting preprocessing procedures [18]. Gonk et al. defined hard sample mining rules to solve the complex features of ultrasound images. They detected hard samples for benign and malignant images with overlapping features via specialized shape descriptors such as concavity, circle variance, compactness, ellipticity variance and growth direction of ultrasound images. The network proposed a one-stage PGAN model to increase the number of hard examples in the dataset and achieved a classification accuracy of 97.7% in breast ultrasound image classification with the general dataset [11]. Yi et al. proposed a CAS (breast cancer diagnosis framework) based on the lesion area via deep learning. Compared with the raw classification, the two-stage model improved the accuracy by approximately 4%, the precision by 3.98% and the F1 score by 4% [19]. Harikumar et al. classified fetal ultrasound images with a convolution neural network (CNN) for prenatal diagnosis. They applied local interpretable model-agnostic explanations (LIME) to interpret the CNN classification and identified the critical regions that contributed positively and negatively to the classification decision. The model provides a reliable solution for prenatal diagnosis [20]. Tian et al. proposed an effective binary classification model for breast tumor ultrasound images via radiomics and transfer learning features on a dataset consisting of single-lesion two-dimensional ultrasound images, including five hundred and ninety-three benign and three hundred and fifty-seven malignant tumor cases obtained from patients at a local clinic [21]. Islam et al. proposed an EDCNN (deep convolutional neural network model) with high accuracy by comparing different transfer learning models (MobilLeNet and Xception) with the Vision Transformer model. They created two different datasets by using various preprocessing techniques (image resizing, data normalization, etc.) to prepare the data for training. The EDCNN model achieved 87.82% accuracy on the first dataset and 85.69% accuracy on the second dataset. In addition, the model achieved an AUC value of 0.91 on the first dataset [22]. Dar and Ganivada used the iterative feature of a genetic algorithm to extract the salient features of the BUSI dataset. They reported that this technique used to improve the categorization of the BUSI dataset resulted in a 4% increase in accuracy with the MobiLeNet network with minimum parameters and high validation accuracy [23]. Sahu et al. proposed a new model based on deep learning. In the data preprocessing stage, a Gaussian-based Laplacian was used. With the transfer learning

technique, the AlexNet, ResNet and MobileLeNetV2 networks achieved 96.92% abnormality detection accuracy and 94.62% malignant detection accuracy on the BUSI ultrasound dataset [24]. Uzen developed an encoder-decoder network architecture using the BUSI dataset. The created network architecture focused on pixel-level classification via classification success at the image level. In addition, the features obtained by the ConvMixer and DenseNet121 networks in the designed model were integrated to extract local details and semantic information. The model achieved a 69.23% Jaccard score and 80.23% Dice score on the BUSI dataset [25]. Atrey et al. proposed a hybrid machine learning framework including the ResNet-SVM architecture. The model presents a combined classification using ultrasound and mammography images. Data augmentation was performed on the images. In this model, where an SVM is used as a classifier, the classification accuracy is 99.22% [26]. Lanjewar et al. proposed a model based on LSTM and transfer learning using BUSI ultrasound images. They combined three models, MobiLeNetV2, VGG-16 and ResNet50, with LSTM to extract features from ultrasound images. The data balancing method was used to improve the performance. The LSTM model achieved an F1 score of 99%, with an AUC of 1 in the VGG-16 model used as a transfer learning network [27].

3 Materials and methods

This section details the methodologies, architecture, datasets, block diagram, and pseudocode used in developing the proposed breast cancer diagnosis model. In this study, EfficiNetV2M, ResNet101, MobiLeNetV3Large, MobiLeNetV3Small, MobiLeNetV2 deep learning models, which have proven their classification success in medical tasks, were used. They were chosen to cover a wide range of architectures, from highly representative, high resource consuming architectures to optimal resource consuming architectures with different residual blocks. In addition, residual block and inverted residual block layers provide a solution to the problem of network performance degradation and improve the generalization ability in deep neural networks, which has brought the use of these five different network architectures to the fore. The proposed model is inspired by the use of fine-tuned deep learning architectures based on transfer learning and the availability of mask images in this field. Mask images are commonly used for segmentation tasks. In our study, a new dataset was created by overlaying real images and mask images. The impact of the new dataset on the model's performance was investigated. Thus, it provides detailed information examining the results of using five different transfer learning architectures under classification tasks with varying difficulty levels. Figure 1 shows the block diagram of the proposed breast cancer diagnostic model for classifying breast cancer ultrasound images. The pseudocode of the proposed model during the training process is given below.

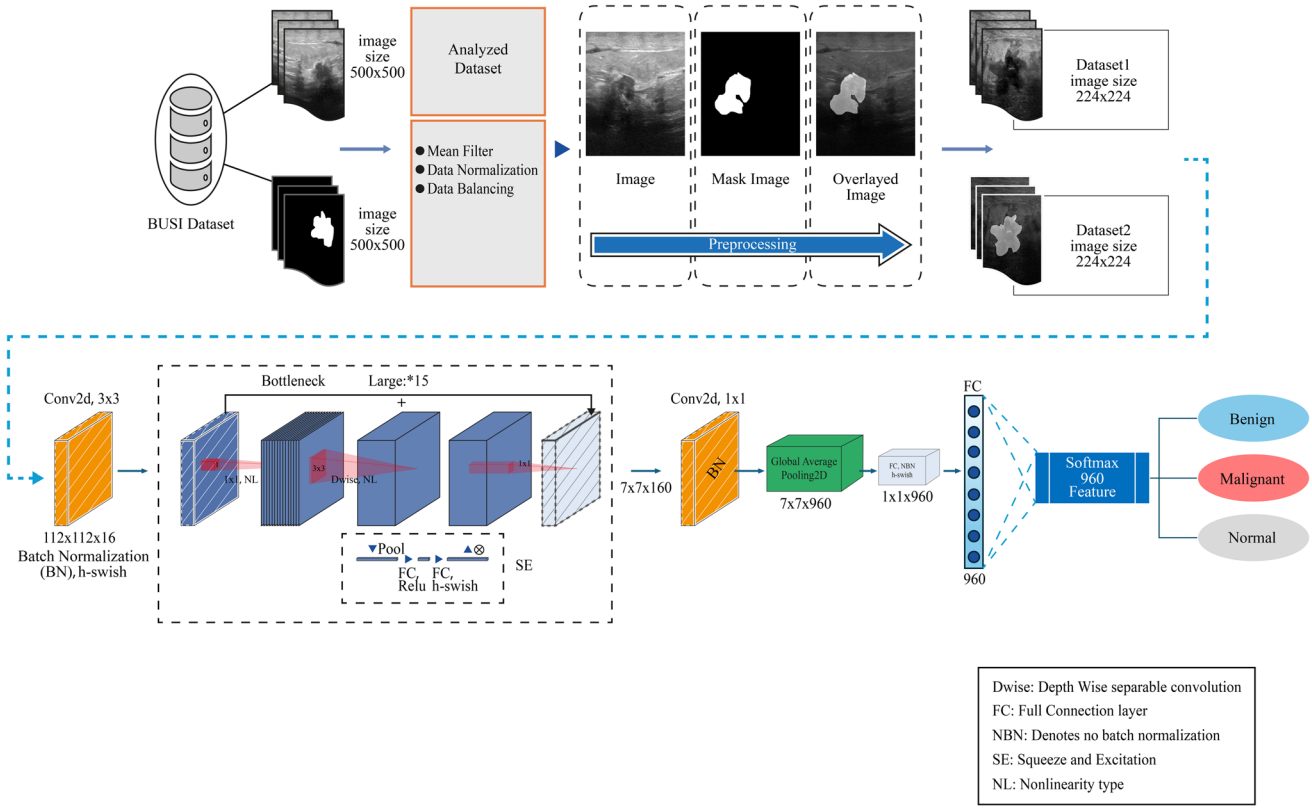

Fig. 1 Block diagram of the proposed diagnostic model for multiclass classification of breast cancer ultrasound images

Algorithm 1: Proposed model

Require: $X:$ *data*, $y:$ *labels*, $f:$ *number of images*, $l, b:$ *image dimensions*,
$b:$ *Base Model*, $h:$ *Head Model*
*Step*1: *Loading this Busi dataset*
*Step*2: *Resize images and masks* $\Leftarrow (224x224)$
*Step*3: *Analyzed dataset and mean filter*
*Step*4: *Dataset data balancing and normalization*
*Step*5: *Overlaying images and masks*
*Step*6: ***for*** $i = 0$ *to* $f - 1$ ***do***
*Step*7: $y \Leftarrow$ *imageLabel*
*Step*8: $image \Leftarrow imageresize(l, b)$
*Step*9: $X \Leftarrow image$
*Step*10: $i++$
*Step*11: ***end for***
Model():
*Step*12: $b \Leftarrow MobilNetV3Large - MobilNetV3Small - MobilNetV2 - ResNet101 - EfficientNetV2M$
*Step*13: $h \Leftarrow Sequential()$
*Step*14: $h.add(f)$
*Step*15: $h.add(Dropout)$
*Step*16: $h.add\left(global_{average_{poolingd}}\right)$
*Step*17: $h.add\left(Dense(activation = 'softmax')\right)$
*Step*18: $metric\ (\ ('accuracy', 'AUROC', 'AUPRC', 'F1 - Score', 'Precision', 'Recall')$
*Step*19: ***return*** *Model*

3.1 Datasets

The breast ultrasound dataset contains breast ultrasound images collected from six hundred female patients between the ages of twenty-five and seventy-five in 2018. The dataset contains a total of 780 images. The images

are approximately 500 × 500 pixels in.png format. Example breast ultrasound images of benign, malignant, and normal classes from the BUSI dataset are shown in Fig. 2.

The real images are presented together with the original images. The images are divided into three classes: normal, benign and malignant [28]. A detailed list of the number of images for each class is shown in Table 1.

3.2 Data preprocessing

Under this heading, the image preprocessing step for the dataset used in the study is presented. One of the important steps in the preparation phase of image processing is the application of different preprocessing steps [29, 30]. Ultrasound imaging, which is generally used in breast cancer diagnosis despite its low resolution and noise problems, is safer and more economical than magnetic resonance imaging and biopsies [31]. To eliminate these disadvantages and increase the quality of ultrasound images, they can be processed via methods such as denoising and enhancement. Enhanced ultrasound images stand out as reasonable alternatives to MR images since they can achieve greater accuracy [31]. Traditional data augmentations such as rotation and translation in medical images are frequently used methods in the context of training deep learning models. However, these

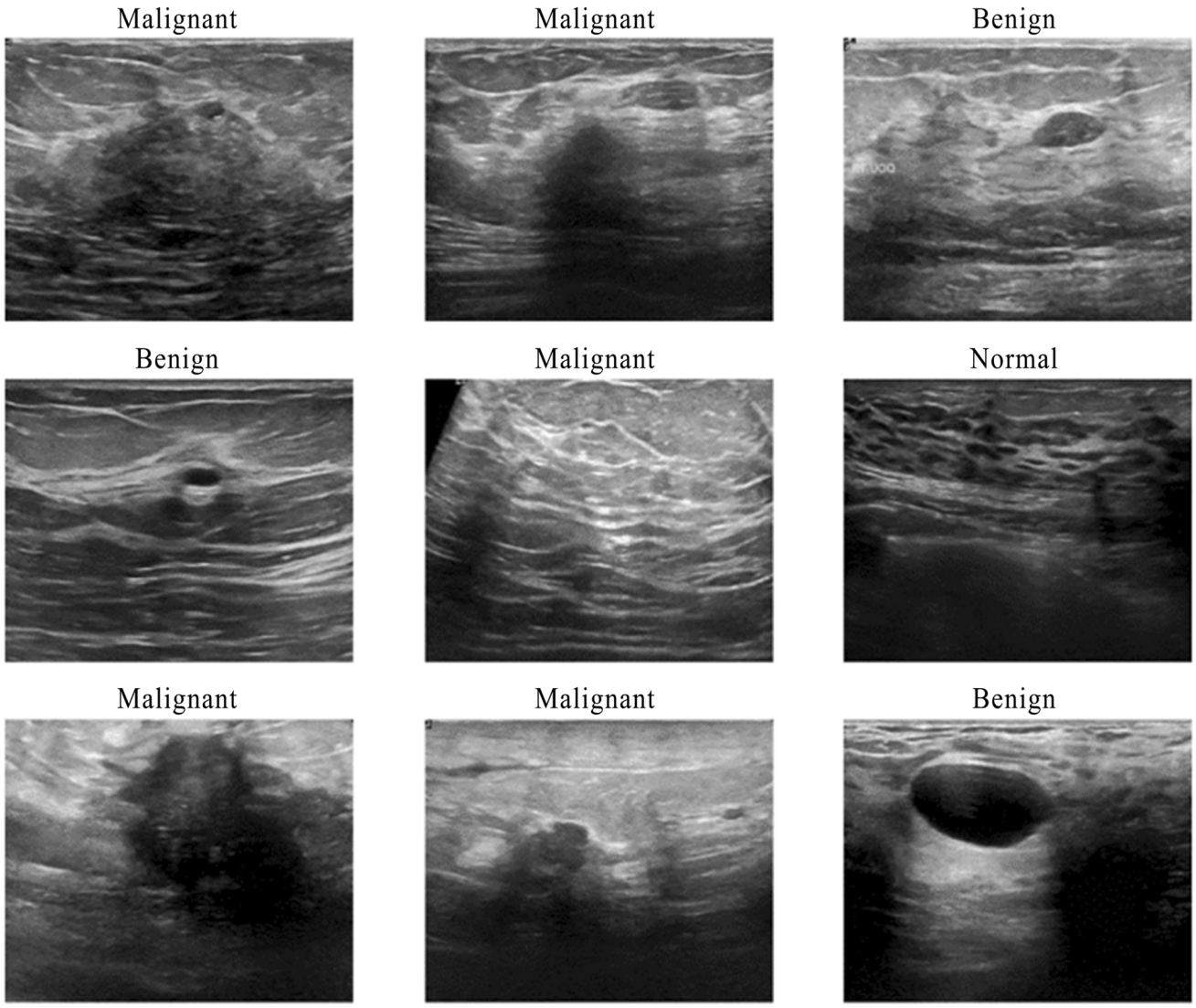

Fig. 2 Example breast ultrasound images of the benign, malignant and normal classes in the BUSI dataset

Table 1 Image distributions for the normal, benign and malignant classes in the BUSI dataset

Types	Number of images
Benign	437
Malignant	210
Normal	133
#Total	780

methods can increase the performance of the models, but they can also distort the semantic content of the image and lead to the learning of incorrect attributes in the network [11]. Alnowami et al. investigated the effects of various data preprocessing steps on model accuracy on the raw dataset they used to detect brain tumors in magnetic resonance imaging scans. It was reported that data preprocessing steps improved learning convergence in the training of deep learning networks [32]. Training deep learning networks without compromising the semantic integrity of raw images labeled by medical experts will produce more realistic results in terms of reliability in critical tasks such as medical diagnosis. For this purpose, our work is inspired by the usability of raw mask images together with real images. In Fig. 3, the flow diagram of the superimposition of benign, malignant and normal images in the BUSI dataset and the labeled mask images of these classes is given in detail. In Fig. 4, sample ultrasound images belonging to three classes from the new dataset formed by superimposing the real images of the BUSI dataset and the mask images are given.

The normalization transform, one of the preprocessing transforms, extracts the channel mean for each value in the image and divides this mean value by the channel's standard deviation. The mathematical formula for the normalization transform used in our dataset is shown in Eq. (1).

$$\text{output}[:, \text{None}, \text{None}] = \frac{(\text{input}[:, \text{None}, \text{None}] - \text{mean}[;, \text{None}, \text{None}])}{\text{std}[:, \text{None}, \text{None}]} \tag{1}$$

In our study, the normalization transformation was performed in four steps. In the first stage, the input image is cropped to 224 $\times$ 224 by preserving its aspect ratio, in the second stage, the channel size of the image is moved to the beginning, in the third stage, the mean value used in Imagenet is entered [0.485, 0.456, 0.406] and the standard deviation [0.229, 0.224, 0.225], and in the last stage, the mean is subtracted from the input channel value and divided by the standard deviation of the floating point values in the range [0,1].

3.3 Depthwise separable convolutions

While standard convolutional networks perform channel and spatial computations in a single step, depth-separable networks divide the computations into two steps. In the first step, depth convolution applies a single convolution filter per input channel, and in the last step, pointwise convolution uses pointwise convolution to create a linear combination of each input channel. MobiLeNet networks are also called depth-separable convolutional layers. The MobiLeNet network was developed by Google engineers in 2017. It was developed to achieve better performance in neural networks because of the power, memory and computational constraints of mobile devices and embedded systems. It is also designed for efficient deployment on mobile devices and embedded systems and has laid a solid foundation in mobile neural network applications [33]. The MobiLeNet network has achieved high success in image net competition. Depth-separable layers are used as building blocks in MobiLeNet architectures to reduce model size and complexity. Depth-separable methods make these mobile networks very fast. One of the networks used in our study is the MobiLeNetV3 network. MobiLeNetV3 was designed for efficient deployment on mobile devices and embedded systems in 2019 [34]. Figure 5 shows the MobiLeNetV3 block structure [34]. In the first convolution block of the network, if the input channels and expanded channels are equal, the expansion step is not applied, but there is definitely a projection step even if the

https://doi.org/10.1007/s00521-025-11653-0

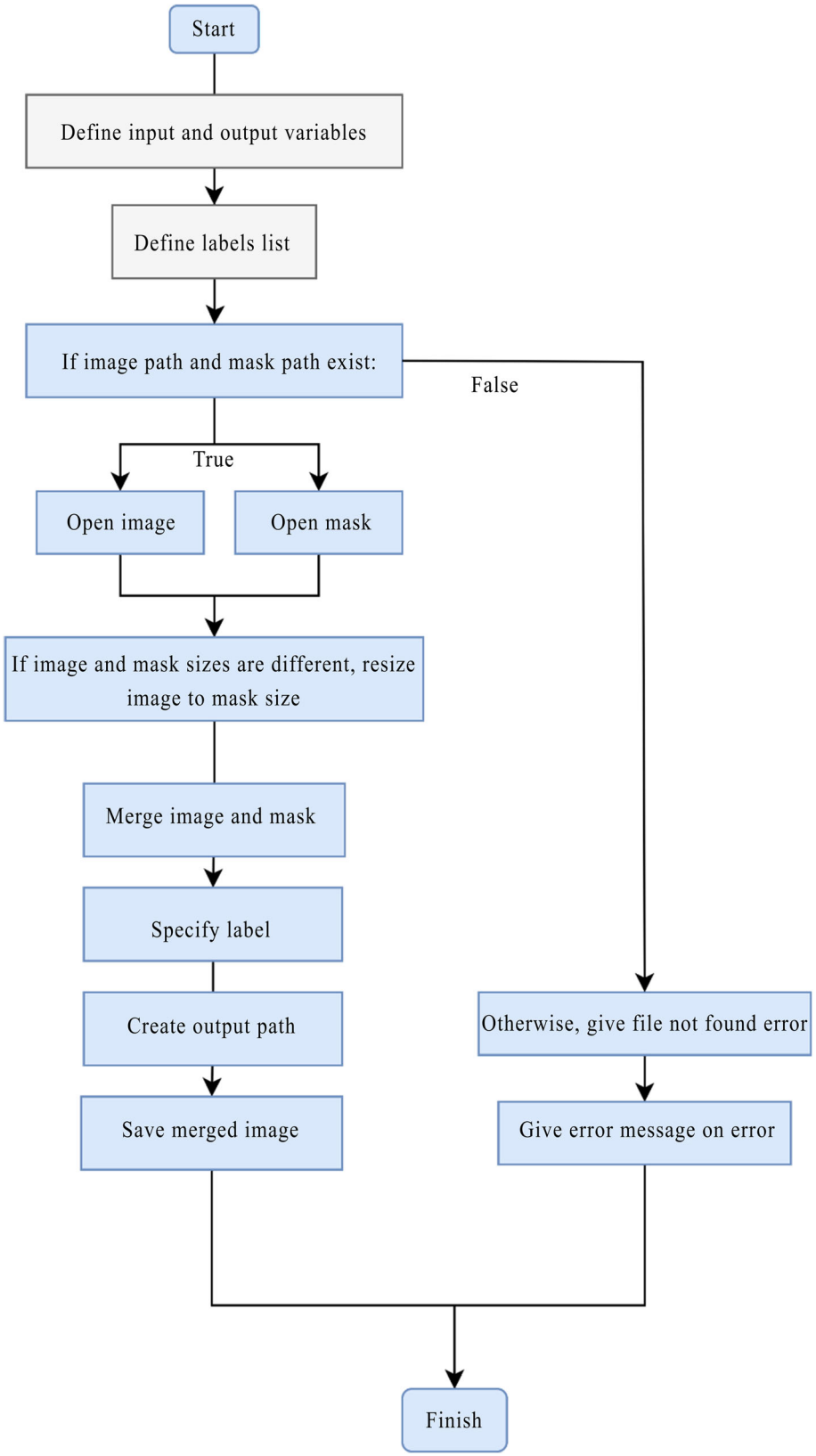

Fig. 3 Flowchart for the steps of overlaying images in the dataset

expanded channels are the same as the output channels. In addition, the activation method of the depth block is implemented before the squeeze-and-excitation layer to increase the accuracy [35].

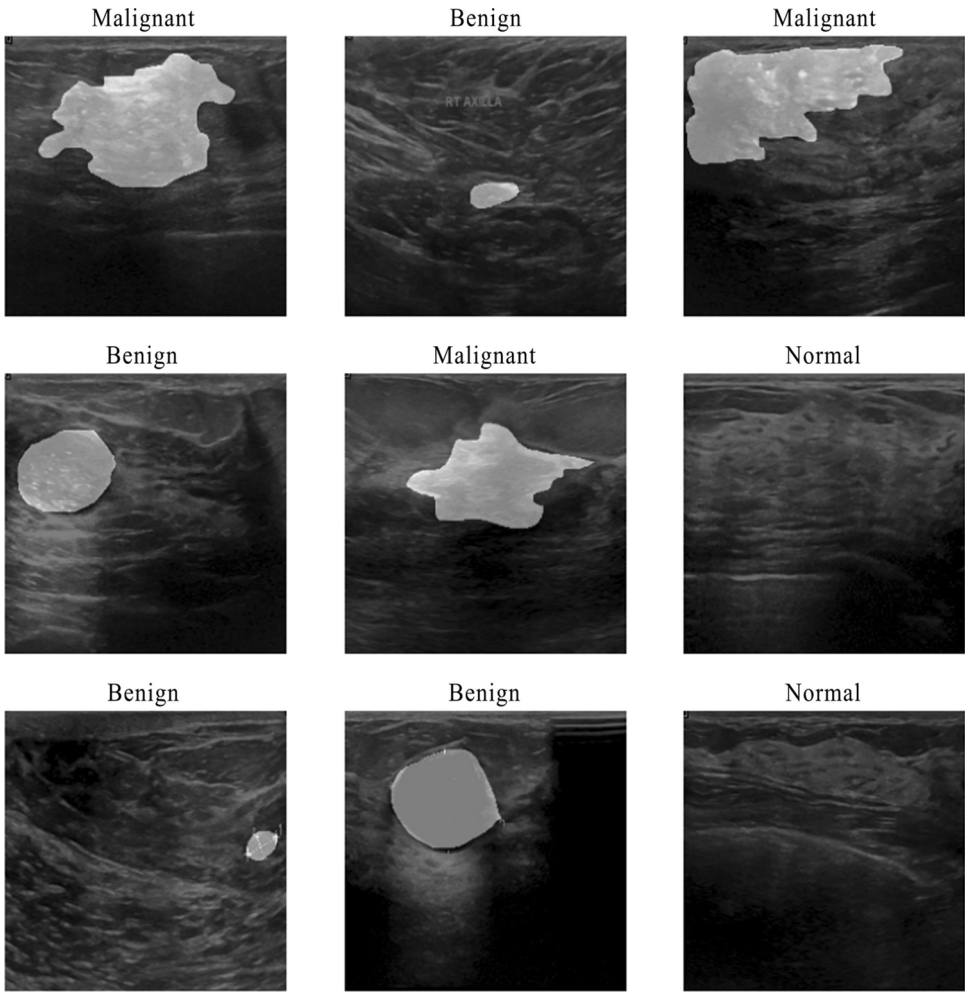

Fig. 4 Sample ultrasound images from three classes from the new dataset

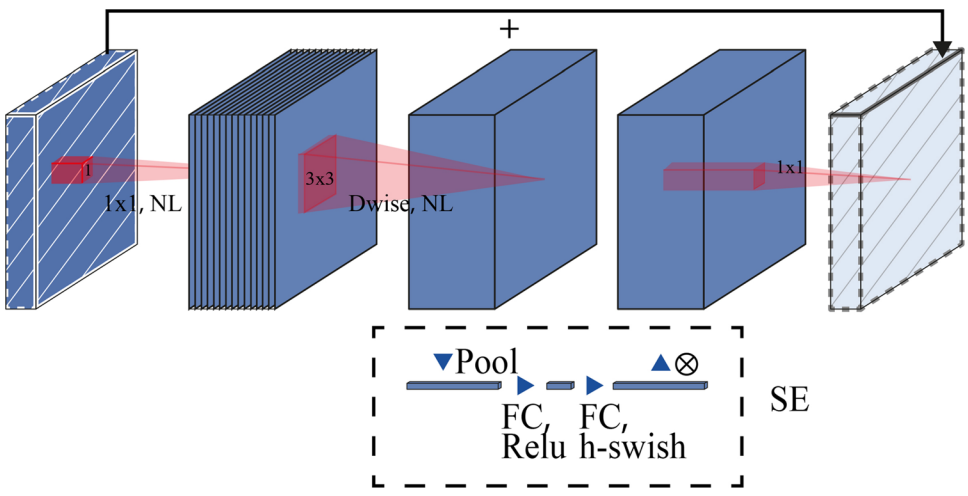

Fig. 5 MobiLeNetV3 block structure

The MobiLeNetV3 architecture, which is formed by adding the compression and excitation module to the MobiLeNetV2 network, supports the network by giving equal weights to different channels from the input while creating the output feature maps.

The MobiLeNetV3 network was developed with a complementary approach that improves the general situation with the NetAdapt algorithm [36, 37]. MobiLeNetV3-Large achieved 3.2% greater accuracy in ImageNet

classification [34]. Unlike the MobiLeNetV2 model, it uses hard-swish nonlinearity. When a nonlinearity called hard-swish is used as a backup for ReLU, it significantly increases the accuracy of neural networks. Hard-swish, a piecewise linear function, is computationally efficient because it does not require the storage of the gradients of the neural networks during backpropagation and saves memory. Therefore, it is generally preferred in deep neural networks when resources are limited. The nonlinearity is shown in Eq. (2).

$$\text{hard} - \text{swish}(x) = x\frac{\text{ReLU6}(x+3)}{6} \tag{2}$$

In Formula 1, x is the input value, and ReLU6 is the rectified linear unit (ReLU6) function that outputs 0 when $x < 0$ is a value, outputs below $x > 6$ and outputs x when $0 < x < 6$ is a value. Here, the hard-swish(x) function is a linear function of the output value x for $x < -3$, the output x for $x > 3$ and the output value x for $-3 \leq x \leq 3$. Like other activation functions, the hard-swish(x) function enables neural networks to learn complex patterns and relationships in images by providing nonlinearity. As a result, the reasons why separable networks are widely used in terms of depth are as follows: first, they have fewer parameters than normal CNNs do, reducing overfitting in network training; second, they are less expensive and lower in cost because they require fewer computations. Table 2 summarizes the input and output values and parameters of the MobiLeNetV3Large architecture according to the layer order. The number of steps in each block is denoted by S. SE is used as an abbreviation for squeeze-and-excite and indicates whether it is in that block or not. NL indicates the type of nonlinearity used. Here, HS represents h-swish, and RE represents ReLU. NBN indicates that there is no batch normalization.

3.4 Proposed network model

Transfer learning is an artificial intelligence technique where a model developed for a particular task is reused as the starting point for a model on a different task. It allows the model to apply knowledge gained in one area to

Table 2 Detailed parameters of each layer of the MobiLeNetV3 large architecture

Input	Layer (type)	Expsize	Out	SE	NL	S
224 × 224 × 3	conv2d	–	16	–	HS	2
112 × 112x16	bnec,3 × 3	16	16	–	RE	1
112 × 112x16	bnec,3 × 3	64	24	–	RE	2
56 × 56x24	bnec,3 × 3	72	24	–	RE	1
56 × 56x24	bnec,5 × 5	72	40	✔	RE	2
28 × 28x40	bnec,5 × 5	120	40	✔	RE	1
28 × 28x40	bnec,5 × 5	120	40	✔	RE	1
28 × 28x40	bnec,3 × 3	240	80	–	HS	2
14 × 14x80	bnec,3 × 3	200	80	–	HS	1
14 × 14x80	bnec,3 × 3	184	80	–	HS	1
14 × 14x80	bnec,3 × 3	184	80	–	HS	1
14 × 14x80	bnec,3 × 3	480	112	✔	HS	1
14 × 14x112	bnec,3 × 3	672	112	✔	HS	1
14 × 14x112	bnec,5 × 5	672	160	✔	HS	1
14 × 14x112	bnec,5 × 5	672	160	✔	HS	2
7 × 7x160	bnec,5 × 5	960	160	✔	HS	1
7 × 7x160	conv2d,1 × 1	–	960	–	HS	1
7 × 7x960	Pool,7 × 7	–	–	–	HS	–
1 × 1x960	conv2d,1 × 1,NBN	–	1280	–	HS	1
1 × 1x1280	conv2d,1 × 1,NBN	–	k	–		–

different but related problems, thereby improving efficiency and performance, especially when data for the new task are limited [38]. Transfer learning allows us to produce high-performance models and models that learn faster with less training data. Our approach is based on the use of fine-tuning, which is a type of transfer learning technique that performs better.

The first layer of the proposed deep learning model consists of a convolutional layer (conv2d) using 16 filters with a kernel size of 3 × 3. Images with an input size of 224 × 224x3 are passed through the first conv2d layer, and batch normalization and h-swish activation are applied to improve model efficiency. After this layer, depth-based separable convolutions consisting of two layers are defined. In the first layer, a depth-separable convolutional operation with a 3 × 3 filter is applied for spatial filtering. The second layer is used for feature generation. In these layers, the channel size is increased, followed by a 1 × 1 bottleneck expansion convolution, followed by depth-based convolutions and a 1 × 1 projection layer. If the input and output have the same number of channels, they are connected by residual connections. This is to increase the representativeness of the transformations per channel. The squeeze and excitation attention modules are added after the depth filters in the expansion. The extracted features are passed through the conv2d layer, which uses 960 filters with a 1 × 1 kernel size to increase the number of channels, thus generating 960 new channels for each spatial location. Thus, more detailed feature vectors are transferred to the classification layer. The global average pooling2d layer, which comes after the conv2d layer, averages the transferred attribute values. Then, the h-swish activation function is applied to these vectors and transferred to the fully connected layer without batch normalization. Finally, the softmax function compatible with the multiclass classification task is used for activation. Figure 6 shows the network architecture of the proposed model in detail.

3.5 Classifiers

Softmax is an important component of CNN architectures and has achieved good performance in data classification tasks, especially image recognition [39]. In our study, categorical cross-entropy is used as an additional loss function to Softmax in multiclass classification of benign, malignant and normal breast cancer ultrasound images. It produces a probability output on three classes for each image. The softmax and categorical cross-entropy calculation formulas are shown in Eqs. (3)-(4), respectively.

where f is the Softmax unit function and where C is the class number:

$\Re^C \rightarrow (0,1)^C$, S is a vector, $S = \{(S_1, \ldots, S_C)\} \in \Re^C, f_S \in (0,1)^C$

$$f_{(S)_i} = \frac{e^{S_i}}{\sum_{j=1}^{C} e^{S_j}} \tag{3}$$

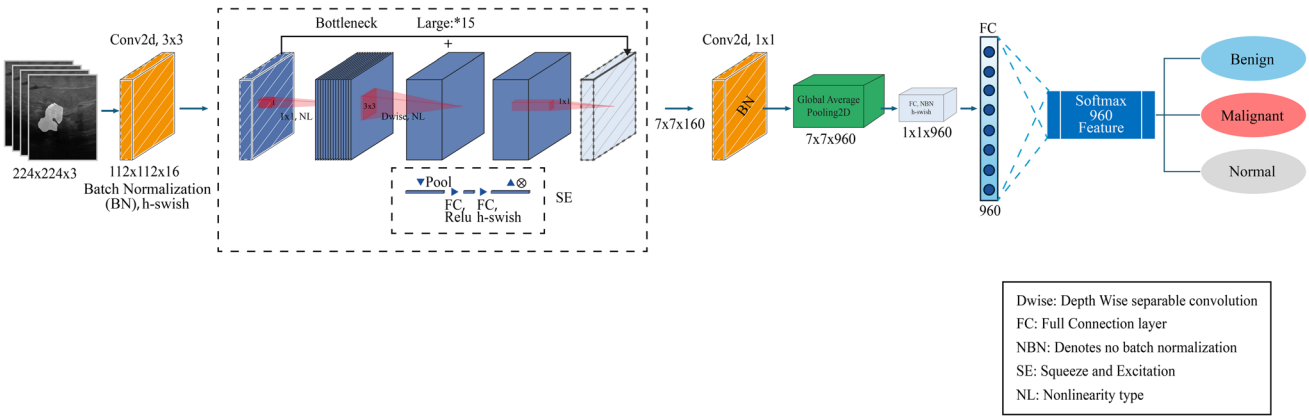

Fig. 6 Network architecture of the proposed model

The sum of the losses calculated for each time t is L:

$$L = -\sum_{i}^{C} t_i \log\left(f(S)_i\right) \tag{4}$$

4 Experimental results and analysis

This section presents the experimental results, comparative analysis and superiority of the proposed model over existing methods. The model explained in Sect. 3 is implemented via the Python library. The results are processed in parallel on a Tesla T4 GPU in the Google Compute Engine backend. The proposed model and other fine-tuned deep learning models are evaluated on the BUSI breast cancer ultrasound image dataset. Training, validation and test sets were created to test the performance of the models. In addition, the datasets were trained by dividing them into two categories. The first training set was trained on the network, with 80% training data points and 20% test data points randomly selected from the ultrasound image dataset groups. The second training set is the training set created by performing data balancing to optimize the unbalanced distribution of the dataset images. The first of these two different training sets has a validation rate of 30%, and the second has a validation rate of 40%. Table 3 shows the categorization of the dataset used in the training process in detail.

4.1 Evaluation metrics

In our study, fine-tuned deep learning models (MobiLeNetV2, MobiLeNetV3Small, MobiLeNetV3Large, ResNet101 and EfficientNetV2M) were used in multiclass classification of the BUSI dataset to compare the predicted labels with the real labels of the test set; these models are expressed as positive and model-generated positive values (TPs), negative and model-generated negative values (TNs), negative and model-generated positive values (FPs), and positive and model-generated negative values (FNs). The FPR represents the false positive rate, and the TPR represents the true positive rate. Different evaluation metrics used in comparing the performances are given in Eqs. (5)-(9). Accuracy is calculated by the ratio of correctly predicted images to the total number of images in the dataset. Precision calculates how many of the images predicted to be positive by the classifier are actually positive. Recall is the metric that calculates how many of the correctly predicted true positive images we predicted as positive. The F1 score is the harmonic mean of precision and recall. Measuring model performance with only the accuracy value may not be sufficient. The AUROC is an important performance metric for measuring the performance of unbalanced datasets [40]. AUROC is the area under the curve where the FPR value is on the x-axis and the TPR value is on the y-axis. The AUPRC represents the region between the recall value on the x-axis and the precision value on the y-axis.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FN} + \text{FP}} \tag{5}$$

Table 3 Dataset categorization for training

Datasets	Dataset Balancing	Training	Validation	Test
Dataset1	–	546	106	128
Dataset2	✔	563	100	117

$$\text{Recall(TPR)} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{6}$$

$$\text{FPR} = (1 - \text{Specificity}) = \frac{\text{FP}}{\text{FP} + \text{TN}} \tag{7}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{8}$$

$$F1 - \text{score} = 2 * \frac{\text{Recall} * \text{Precision}}{\text{Recall} + \text{Precision}} \tag{9}$$

4.2 Network configuration

This subsection discusses the general network configuration and parameter settings. Before starting the experiments, various tuning steps were performed on the proposed approach and other deep neural networks used in our study. An image size of 224 × 224 pixels, a batch size of 32, a learning rate set to 1e-3, and the Adam optimizer were used for training the classification networks. Tables 4 and 5 list the training parameters, layer types, and output shapes of the proposed network in detail.

4.3 Evaluations and analysis

We divided our method into two approaches. In the first approach, we used only the real images of the BUSI dataset (Dataset1) in the classification process. In the second approach, we performed the classification process with the second dataset (Dataset2), which we prepared as a combination of the real images of the BUSI dataset and the mask images. We tested and reported comparative experiments of the fine-tuned deep learning methodologies we used in our study with balanced and unbalanced datasets by performing data testing separately for these datasets. The experimental results clearly demonstrate the usability of mask images in medical image analysis. As shown in Table 6, the classification performance of the mask images improved by 3.73–4.33% in terms of the accuracy and F1 score values, respectively.

The graphs of the accuracy, precision, recall, AUROC, AUPRC and loss metric values of the proposed approach are shown in Fig. 7. In Fig. 7, the high-value AUROC curve is very close to the upper left corner; in other words, the area under the curve is large, which is evidence that the proposed model performs well. The best working point is the minimum FPR value, and the maximum TPR value is the desired situation in classification performance. The model reaches high efficiency, with an AUROC value of 1 for the training data and 0.9991 for the test data. Table 7 shows the performance of the proposed model on different datasets for the test data in detail.

In Table 7, our model achieved a 99.15% accuracy, 99.21% F1 score, 99.14% precision, and 99.29% recall for the test data in the balanced Dataset2 dataset. In both datasets, the balanced datasets contributed to the performance of the network between 1.91 and 1%. In Table 6, when the classification performance with only real images is examined, the balanced dataset achieves high performance. However, when we look at the performance of the network according to the test data, a difference of 3.57% occurs between the AUROC and AUPRC performance values. This difference is due to overfitting of the network. In general, the balanced Dataset2 dataset

Table 4 Training parameters of the proposed network

Software used	CNN architecture	Image size	Optimization	Batch Size	Learning rate
Python	Our model	224 × 224	Adam	32	1e-3

Table 5 Layer type, output shape and parameters of the proposed network

Layer (type)	Output shape	Param #
input_2 (InputLayer)	[(None, 224, 224, 3)]	0
MobiLeNetV3large (Functional)	(None, 7, 7, 960)	2,996,352
global_average_pooling2d (GlobalAveragePooling2D)	(None, 960)	0
dense (Dense)	(None, 3)	2883
Total params: 2,999,235 (11.44 MB)		
Trainable params: 2883 (11.26 KB)		
Non trainable params: 2,996,352 (11.43 MB)		

Table 6 The performance of the proposed model on different datasets for training data

Datasets	Preprocessing	Recall(%)	Precision (%)	F1-score (%)	Acc (%)	AUROC (%)	AUPRC (%)
Dataset1	Unbalanced	93,59	96,78	95,15	95,42	99,55	99,17
	Balanced	94,32	97,07	95,67	96,27	99,68	99,4
Dataset2	Unbalanced	99,82	99,82	99,82	99,82	100	99,99
	Balanced	100	100	100	100	100	100

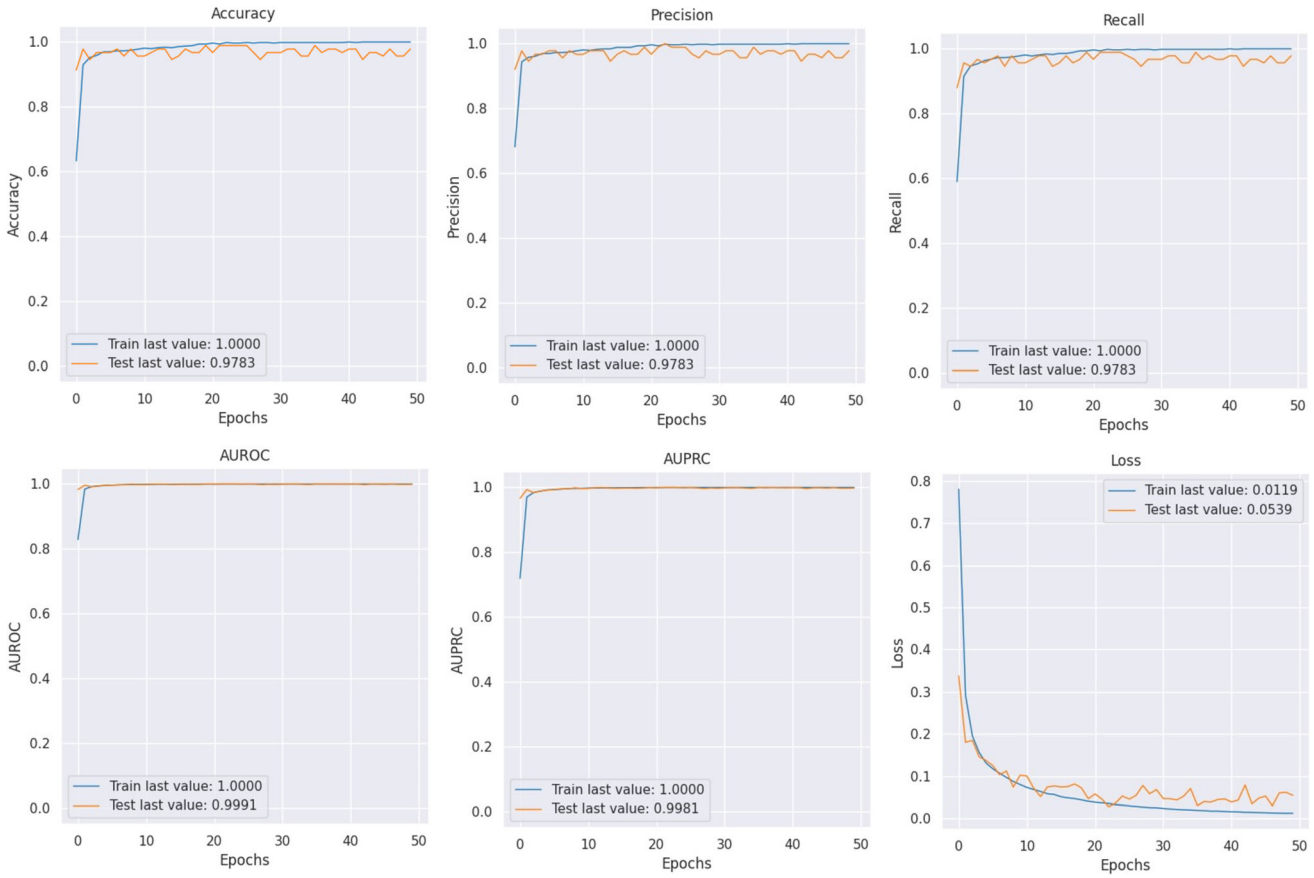

Fig. 7 The performance graph of the proposed model for training data

will be used as the dataset in other fine-tuned network training since it provides high performance in our training. The MobiLeNetV3 large network differs from the MobiLeNetV3 small network in that it has 4 convolution

Table 7 The performance of the proposed model on different datasets for testing data

Datasets	Preprocessing	Recall(%)	Precision (%)	F1-score (%)	Acc (%)	AUROC (%)	AUPRC (%)
Dataset1	Unbalanced	77,78	85,85	81,61	83,76	95,41	91,84
	Balanced	79,69	84,3	81,93	82,03	95,95	92,5
Dataset2	Unbalanced	98,44	98,44	98,44	98,44	99,94	99,88
	Balanced	99,29	99,14	99,21	99,15	99,96	99,93

blocks. The MobiLeNetV3 large network consists of twenty layers, whereas the MobiLeNetV3 small network consists of sixteen layers. These differences allowed us to compare the trained models on a block basis. Therefore, the MobiLeNetV3 Large and MobiLeNetV3 Small networks trained with a fine-tuned transfer learning strategy were tested. Table 8 shows that the classification accuracy, F1 score and AUROC performance of the MobiLeNetV3 large model on the balanced Dataset2 dataset reached 100%. Table 8 shows the blockwise comparative results of the MobiLeNet versions. The network training was performed via the balanced Dataset2 dataset for training all three models. The use of mask images and data balancing increased the classification accuracy of all three models. The MobiLeNetV3 Large model outperforms the other two methods. This shows that the image-level features extracted by the proposed model are robust and most suitable for describing ultrasound images. Therefore, our final proposed model is the fine-tuned MobileLeNetV3 network. Adding the network lookup component to different versions of the MobileLeNetV3 model, reorganizing some expensive layers of the network, and adding h-swish had an impact of 8.8% on the network performance. Figure 8 shows a comparison of the block-based performance of the MobiLeNet networks.

Other fine-tuned models used in our study are EfficientNetV2 and ResNet101. The EfficientNetV2 training-aware neural architecture is designed to jointly optimize training speed and parameter efficiency by using a combination of search and scaling. In the experiments, EfficientNetV2 models are reported to train up to seven times faster than the state-of-the-art models. EfficientNetV2 outperforms previous models on the ImageNet and CIFAR/Flowers datasets. EfficientNetV2 achieved 87.3% accuracy on ImageNet ILSVRC2012 by pretraining on the same ImageNet21k and training five to eleven times faster using the same computing resources [41]. ResNet networks reformulate stacked nonlinear layers to fit a residual mapping instead of a deep residual frame to fit a base mapping. These residual networks make the model easier to optimize and achieve high accuracy with significantly increased depth. Compression and excitation channels are usually added separately to ResNet blocks. It has lower complexity despite being eight times deeper than VGG networks [42]. ResNet finished the competition in the first place in the ILSVRC 2015 classification by achieving 3.57% error on the ImageNet test set. It achieved a 28% improvement on the COCO object detection dataset [43]. Table 9 shows the accuracy, F1 score, AUROC and AUPRC performances of the proposed model and other fine-tuned deep learning architectures for the test values on the balanced Dataset2 dataset. Figure 9 shows a comparison of the performances of the proposed model and other models used in our study. In the experiments conducted for multiclass classification of breast cancer ultrasound images, the ResNet101 model achieved a 98.44% accuracy and F1 score for the test data of the balanced Dataset2 dataset. EfficientNetV2M achieved a 95.31% F1 score and accuracy performance. Our proposed model outperformed all the other models.

Table 8 Performance of the balanced Dataset2 dataset on MobiLeNet models

MobiLeNetV3 Large			MobiLeNetV3 Small			MobiLeNetV2		
Acc(%)	F1 Score(%)	AUROC(%)	Acc(%)	F1- Score(%)	AUROC(%)	Acc(%)	F1-Score(%)	AUROC(%)
100	100	100	98,58	98,39	99,95	91,12	83,62	98,32

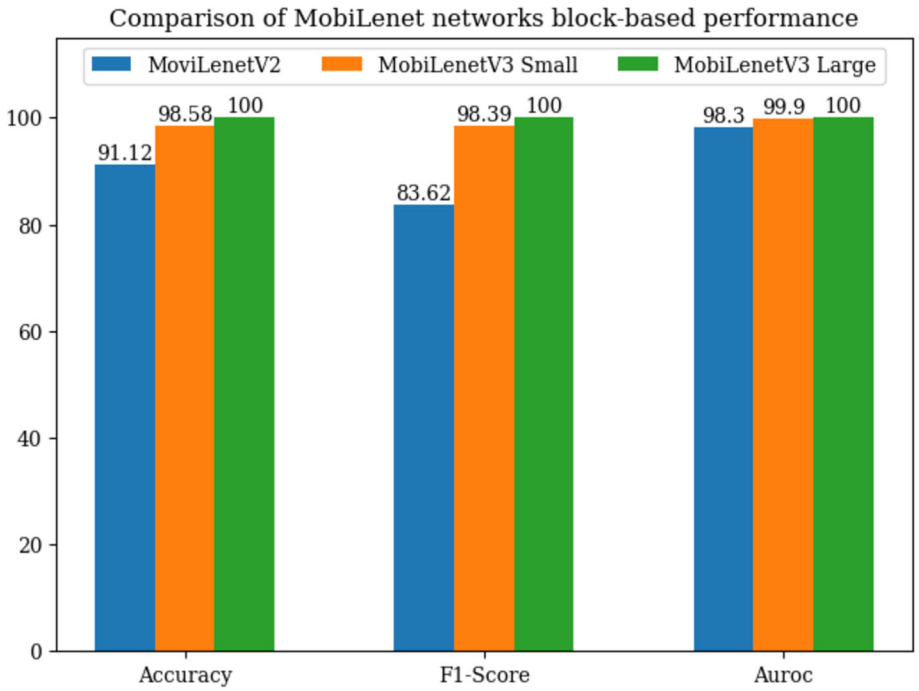

Fig. 8 Comparison of block-based performance of MobiLeNet networks

Table 9 Accuracy, F1 score, AUROC and AUPRC performance of the proposed model and other fine-tuned deep learning architectures for test values on the balanced Dataset2 dataset

	Multiclass classification			
CNN models	Test accuracy(%)	Test F1-score(%)	Test AUROC(%)	Test AUPRC(%)
ResNet-101	98,44	98,44	99,96	99,93
EfficientNetV2M	95,31	95,31	99,78	99,58
MobiLeNetV3Small	96,58	96,58	99,74	99,49
MobiLeNetV2	82,91	83,62	93,32	86,94
Proposed Model	99,15	99,21	99,96	99,93

In this study, classification was performed in three groups—benign cancer cells, malignant cancer cells and normal cells—using breast cancer ultrasound images. In the first approach of our study, we fine-tuned the high-performance networks trained on ImageNet (MobiLeNetV3Large, MobiLeNetV3Small, MobiLeNetV2, ResNet101 and EfficientNetV2M) and trained them on raw datasets and compared them with alternative networks. In the second approach, we retrained the new dataset that we prepared using mask images with fine-tuned networks. With respect to multiclass classification with the data consisting of original ultrasound images and mask images, we achieved 100% accuracy in the training data and 99.15% accuracy in the test data. In published studies on the BUSI dataset, Moon et al. achieved 90.77% diagnostic performance in binary classification via VGG, DenseNet-121 and DenseNet-40 ensemble learning methods on the BUSI breast cancer ultrasound dataset [13]. Podda et al. reported the superiority of ensemble learning methods over individual networks using the publicly available dataset BUSI via a recursive multilayer approach with CNNs. They reported 91% classification accuracy on the BUSI dataset [15]. Alotaibi et al. developed a three-stage image processing scheme using RGB fusion with a transfer learning method on the VGG-19 deep learning model. The developed model achieved 85.6% accuracy on the BUSI dataset with denoising, RGB fusion and ROI highlighting preprocessing procedures [18]. Gonk et al. defined hard sample mining rules to decipher the complex features of ultrasound images. The network proposed a one-stage PGAN model to increase the number of hard examples in the dataset and achieved a classification accuracy of 97.7% in breast ultrasound image classification with a general dataset [11]. Dar and Ganivada used the iterative feature of a genetic algorithm to extract the salient features of the BUSI dataset. They reported that this technique used to improve the categorization of the BUSI dataset achieved an accuracy of

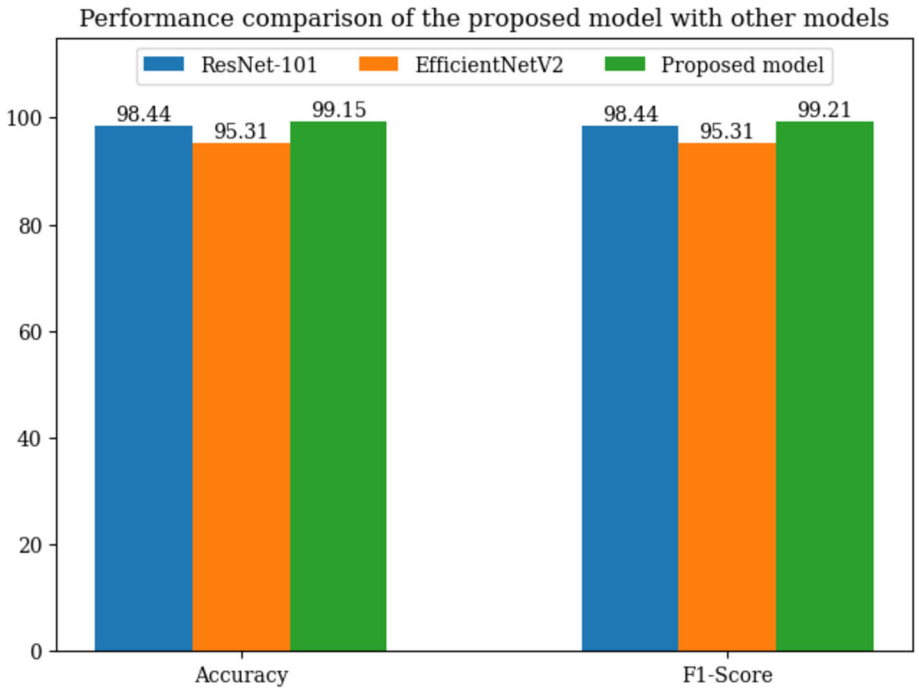

Fig. 9 Comparison of the performance of the proposed model and other models used in our study

97.51% and an F1 score of 95.87%, indicating that it has minimum parameters and high validation accuracy with the MobileLeNet network [23]. Üzen developed an encoder-decoder network architecture using the BUSI dataset. The created network architecture focused on pixel-level classification via image-level classification success. The model he designed was integrated with the features obtained by the ConvMixer and DenseNet121 networks to extract local details and semantic information. The model achieved an accuracy of 91.89% on the BUSI dataset [25]. Sahu et al. proposed a new deep learning-based model. In the data preprocessing stage, a Gaussian-based Laplacian was used on the BUSI ultrasound data. It was reported that the AlexNet, ResNet and MobileLeNetV2 networks achieved 96.92% abnormality and 94.62% malignant detection accuracy on the BUSI ultrasound dataset with the transfer learning technique [24]. Lanjewar et al. proposed a model based on LSTM and transfer learning using BUSI ultrasound images. To extract features from ultrasound images, three models, namely, MobiLeNetV2, VGG-16 and ResNet50, were combined with long short-term memory (LSTM). The data balancing method was used to improve the performance. LSTM was reported to achieve a 99% F1 score in the VGG-16 model used as a transfer learning network [27]. Table 10 shows the comparison of the proposed model with the state-of-the-art methods. As shown in Table 10, our model has proven to be highly accurate and reliable compared with other related studies because of its metric values.

Table 10 Comparative analysis of the results of related studies in the literature

Category	Approaches to breast cancer detection	Dataset	Accuracy	F1-Score
Deep Learning	Moon et al. [13]	BUSI	%90,77	%82,86
Deep Learning	Alotaibi et al. [18]	BUSI	%85,6	%87,4
Deep Learning	Podda et al. [15]	BUSI	%91	–
Encoder-Decoder and DenseNet121	Üzen [25]	BUSI	%91,89	–
Transfer Learning	Sahu et al. [24]	BUSI	%96,92	–
Deep Learning	Dar and Ganivada [44]	BUSI	%97,51	%95,87
GAN(Generative Adversarial Networks)	Gonk et al. [11]	BUSI	%97,7	
LSTM and Transfer Learning	Lanjewar et al. [27]	BUSI	%99	%99

5 Conclusion

In this study, we propose a new deep learning model for multiclass classification of breast cancer ultrasound images. The proposed model was trained and tested on the BUSI dataset with different datasets and different data volumes and subclasses. Two different datasets were created from the raw data. The first dataset is the BUSI dataset, and the second dataset is the dataset created by merging images and masks. We examined the performances of the models by performing data balancing in the training of the networks. We found that the parameters of the EfficientNetV2M, ResNet101 and MobiLeNetV3L networks performed well in terms of recognition performance and computational amount on the BUSI dataset. To ensure reliability, the models were evaluated with accuracy, sensitivity, F1 score, AUROC and AUPRC metrics. The model achieves an accuracy of 99.15% for the test data and has an F1 score of 99.21%, which outperforms related studies. The experimental study using this wide range of evaluation metrics indicates that the proposed model is robust and has high performance. We also reported the results of training the network with other architectures in which mask images may have the potential to be generalized. We observed that mask images play an important role in training deep learning methodologies for classification tasks. The integration of MobiLeNet architectures, which are deepwise convolution networks, into our work allowed us to use smaller input shapes and lower batch sizes. We plan to use the current model features and approach in segmentation tasks in the future and to extend it to other datasets.

Declarations

Conflict of interest The author declares that she has no financial or personal relationships with other persons or organizations that could inappropriately influence the work reported in this article.

References

1. World Health Organization, 2022, February 3
2. de Martel C, Georges D, Bray F, Ferlay J, Clifford GM (2020) Global burden of cancer attributable to infections in 2018: a worldwide incidence analysis. Lancet Glob Health 8:e180–e190
3. breastcancer.org, U.S. Breast Cancer Statistics,, 2022, January 13 at 5:10 AM
4. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A (2018) Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 68:394–424
5. Kiani A, Uyumazturk B, Rajpurkar P, Wang A, Gao R, Jones E, Yu Y, Langlotz CP, Ball RL, Montine TJ (2020) Impact of a deep learning assistant on the histopathologic classification of liver cancer. NPJ Digit Med 3:23
6. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, van der Laak J, van Ginneken B, Sanchez CI (2017) A survey on deep learning in medical image analysis. Med Image Anal 42:60–88
7. Ren S, He K, Girshick R, Zhang X, Sun J (2016) Object detection networks on convolutional feature maps. IEEE Trans Pattern Anal Mach Intell 39:1476–1481
8. Tian Y, Fu S (2020) A descriptive framework for the field of deep learning applications in medical images. Knowl-Based Syst 210:106445
9. Sawant A, Kulkarni S (2022) Ultrasound image enhancement using super resolution. Biomed Eng Adv 3:100039
10. George K, Sankaran P, K PJ (2020) Computer assisted recognition of breast cancer in biopsy images via fusion of nucleus-guided deep convolutional features. Comput Methods Programs Biomed 194:105531
11. Gong X, Zhou H, Gu Y, Guo Y (2023) Breast ultrasound image classification with hard sample generation and semisupervised learning. Biomed Signal Process Control 86:105196
12. Chiao J-Y, Chen K-Y, Liao KY-K, Hsieh P-H, Zhang G, Huang T-C (2019) Detection and classification the breast tumors using mask R-CNN on sonograms. Medicine (Baltimore) 98:e15200
13. Moon WK, Lee Y-W, Ke H-H, Lee SH, Huang C-S, Chang R-F (2020) Computer-aided diagnosis of breast ultrasound images using ensemble learning from convolutional neural networks. Comput Methods Programs Biomed 190:105361
14. Shinohara I, Inui A, Mifune Y, Nishimoto H, Yamaura K, Mukohara S, Yoshikawa T, Kato T, Furukawa T, Hoshino Y (2022) Using deep learning for ultrasound images to diagnose carpal tunnel syndrome with high accuracy. Ultrasound Med Biol 48:2052–2059

15. Podda AS, Balia R, Barra S, Carta S, Fenu G, Piano L (2022) Fully automated deep learning pipeline for segmentation and classification of breast ultrasound images. J Comput Sci 63:101816
16. Xu Z, Wang Y, Chen M, Zhang Q (2022) Multiregion radiomics for artificially intelligent diagnosis of breast cancer using multimodal ultrasound. Comput Biol Med 149:105920
17. Wu R, Jia Y, Li N, Lu X, Yao Z, Ma Y, Nie F (2023) Evaluation of breast cancer tumor-infiltrating lymphocytes on ultrasound images based on a novel multi-cascade residual U-shaped network. Ultrasound Med Biol 49:2398–2406
18. Alotaibi M, Aljouie A, Alluhaidan N, Qureshi W, Almatar H, Alduhayan R, Alsomaie B, Almazroa A (2023) Breast cancer classification based on convolutional neural network and image fusion approaches using ultrasound images. Heliyon. https://doi.org/10.1016/j.heliyon.2023.e22406
19. Yi S, Chen Z, Yi L, She F (2023) CAS: Breast cancer diagnosis framework based on lesion region recognition in ultrasound images. J King Saud Univ-Comput Inf Sci 35:101707
20. Harikumar A, Surendran S, Gargi S (2024) Explainable AI in deep learning based classification of fetal ultrasound image planes. Procedia Comput Sci 233:1023–1033
21. Tian R, Lu G, Tang S, Sang L, Ma H, Qian W, Yang W (2024) Benign and malignant classification of breast tumor ultrasound images using conventional radiomics and transfer learning features: a multicenter retrospective study. Med Eng Phys 125:104117
22. Islam MR, Rahman MM, Ali MS, Nafi AAN, Alam MS, Godder TK, Miah MS, Islam MK (2024) Enhancing breast cancer segmentation and classification: an ensemble deep convolutional neural network and U-net approach on ultrasound images. Mach Learn Appl 16:100555
23. Dar MF, Ganivada A (2024) Deep learning and genetic algorithm-based ensemble model for feature selection and classification of breast ultrasound images. Image Vis Comput 146:105018
24. Sahu A, Das PK, Meher S (2024) An efficient deep learning scheme to detect breast cancer using mammogram and ultrasound breast images. Biomed Signal Process Control 87:105377
25. Üzen H (2024) Convmixer-based encoder and classification-based decoder architecture for breast lesion segmentation in ultrasound images. Biomed Signal Process Control 89:105707
26. Atrey K, Singh BK, Bodhey NK (2024) Integration of ultrasound and mammogram for multimodal classification of breast cancer using hybrid residual neural network and machine learning. Image Vis Comput 145:104987
27. Lanjewar MG, Panchbhai KG, Patle LB (2024) Fusion of transfer learning models with LSTM for detection of breast cancer using ultrasound images. Comput Biol Med 169:107914
28. Al-Dhabyani W, Gomaa M, Khaled H, Fahmy A (2020) Dataset of breast ultrasound images. Data Brief 28:104863
29. Zerouaoui H, Idri A (2022) Deep hybrid architectures for binary classification of medical breast cancer images. Biomed Signal Process Control 71:103226
30. Saikia AR, Bora K, Mahanta LB, Das AK (2019) Comparative assessment of CNN architectures for classification of breast FNAC images. Tissue Cell 57:8–14
31. Sawant A, Kulkarni S, Sawant M (2024) Ultrasound super resolution imaging for accurate uterus tumor detection and malignancy prediction. J Pharm Biomed Anal Open 3:100029
32. Alnowami M, Taha E, Alsebaeai S, Anwar SM, Alhawsawi A (2022) Mr image normalization dilemma and the accuracy of brain tumor classification model. J Radiat Res Appl Sci 15:33–39
33. Chu X, Zhang B, Xu R (2020) Moga: searching beyond mobilenetv3, ICASSP 2020–2020 IEEE international conference on acoustics, speech and signal processing (ICASSP), IEEE, pp 4042–4046
34. Howard A, Sandler M, Chu G, Chen L-C, Chen B, Tan M, Wang W, Zhu Y, Pang R, Vasudevan V (2019) Searching for mobilenetv3, Proceedings of the IEEE/CVF international conference on computer vision, pp 1314–1324
35. Hu J, Shen L, Sun G (2018) Squeeze-and-excitation networks, Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7132–7141
36. Yang T-J, Howard A, Chen B, Zhang X, Go A, Sandler M, Sze V, Adam H (2018) Netadapt: platform-aware neural network adaptation for mobile applications, Proceedings of the European conference on computer vision (ECCV), pp 285–300
37. Zhang X, Zhao H (2021) Hyperspectral-cube-based mobile face recognition: a comprehensive review. Inf Fusion 74:132–150
38. Pan SJ, Yang Q (2009) A survey on transfer learning. IEEE Trans Knowl Data Eng 22:1345–1359
39. Shao H, Wang S (2023) Deep classification with linearity-enhanced logits to softmax function. Entropy 25:727
40. Liu Y, Han L, Wang H, Yin B Classification of papillary thyroid carcinoma histological images based on deep learning, J Intell Fuzzy Syst 1–11
41. Tan M, Le Q (2021) Efficientnetv2: smaller models and faster training. International conference on machine learning, PMLR, pp 10096–10106
42. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556
43. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition, Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778

44. Dar MF, Ganivada A (2024) Deep learning and genetic algorithm-based ensemble model for feature selection and classification of breast ultrasound images. Image Vis Comput. https://doi.org/10.1016/j.imavis.2024.105018

Authors and Affiliations

Kadir Can Burçak[1]

✉ Kadir Can Burçak
kcburcak@ahievran.edu.tr

[1] Department of Information Systems and Technologies, Kırşehir Ahi Evran University, Kırşehir, Turkey