Article

MFAN: Multi-Feature Attention Network for Breast Cancer Classification

Inzamam Mashood Nasir [1,*], **Masad A. Alrasheedi [2]** and **Nasser Aedh Alreshidi [3]**

1 Faculty of Informatics, Kaunas University of Technology, 51368 Kaunas, Lithuania
2 Department of Management Information Systems, College of Business Administration, Taibah University, Al-Madinah Al-Munawara 42353, Saudi Arabia; mrshedi@taibahu.edu.sa
3 Department of Mathematics, College of Science, Northern Border University, Arar 73213, Saudi Arabia; nasser.alreshidi@nbu.edu.sa
* Correspondence: inzamam.nasir@ktu.edu; Tel.: +92-333-810-6946

Abstract: Cancer-related diseases are some of the major health hazards affecting individuals globally, especially breast cancer. Cases of breast cancer among women persist, and the early indicators of the diseases go unnoticed in many cases. Breast cancer can therefore be treated effectively if the detection is correctly conducted, and the cancer is classified at the preliminary stages. Yet, direct mammogram and ultrasound image diagnosis is a very intricate, time-consuming process, which can be best accomplished with the help of a professional. Manual diagnosis based on mammogram images can be cumbersome, and this often requires the input of professionals. Despite various AI-based strategies in the literature, similarity in cancer and non-cancer regions, irrelevant feature extraction, and poorly trained models are persistent problems. This paper presents a new Multi-Feature Attention Network (MFAN) for breast cancer classification that works well for small lesions and similar contexts. MFAN has two important modules: the McSCAM and the GLAM for Feature Fusion. During channel fusion, McSCAM can preserve the spatial characteristics and extract high-order statistical information, while the GLAM helps reduce the scale differences among the fused features. The global and local attention branches also help the network to effectively identify small lesion regions by obtaining global and local information. Based on the experimental results, the proposed MFAN is a powerful classification model that can classify breast cancer subtypes while providing a solution to the current problems in breast cancer diagnosis on two public datasets.

Keywords: breast cancer; deep learning; attention module; computer-aided diagnosis; Wiener Filter; CBIS-DDSM; ultrasound dataset

MSC: 68T07

1. Introduction

Cancer is a disease characterized by rapid and unregulated cell development and is also a primary cause of death in society. The IARC estimates that one in five people will be diagnosed with cancer, and one in eight men and one in eleven women will die from it. Women are most commonly affected by breast cancer, which is a primary cause of death [1]. Most importantly, emerging countries have more female breast cancer cases than industrialized ones. Pakistan has 1.38 million cases every year, and one third of them die. Cancer accounts for 9.6 million annual deaths globally [2], with a survival rate of 1.7 million [3]. Statistics indicate that breast cancer accounts for 24.5% of cancer-related fatalities in women (GLOBOCAN, 2020). Early-stage cancer must be discovered and treated soon to prevent breast or other metastases. Correct diagnosis and therapy can boost breast cancer survival by up to 80% [4]. New diagnostic and therapeutic options for breast cancer patients are emerging from imaging techniques and biochemical biomarkers targeting micro RNAs, proteins, DNAs, and mRNAs [5].

Deep Learning (DL) is a branch of Machine Learning (ML) based on brain neurons. DL networks are made of artificial neural layers and neurons. Every neuron is coupled with every neuron in the following layer through weighting factors known as weights. The recent increase in DL methods has caused several fields to employ this technology to solve different problems or enhance the outcomes of existing research. DL models using SVM have recently achieved tremendous success [6–8]. Recently, researchers have employed DL models in human action recognition [9–11] and statistical models [12,13]. An increase in the utilization of DL approaches in medical fields has yielded optimistic results, particularly since DL offers more rapid and precise analysis than physicians. We use two public datasets in our paper. The CBIS-DDSM dataset is a new, standardized version of the Digital Database for Screening Mammography. Starting with 2620 scanned film mammography examinations, the DDSM covers a wide spectrum of cases, from benign to malignant, with well-reviewed pathological data. The CBIS-DDSM dataset, a current breast imaging standard, updates and refines this repository. The second collection, the Breast Cancer Ultrasound dataset, includes baseline breast ultrasound images from 25–75-year-old women. The data were obtained in 2018. There are 600 female patients. The collection includes 780 photos, averaging 500 by 500 pixels. The images are PNG. Images are classified as normal, benign, or malignant.

Computer-Aided Diagnostic (CAD) tools are widely used to identify and characterize breast USI lesions. Radiotherapists choose these systems to diagnose breast cancer and determine prognoses. Statistical approaches [14] are commonly used to examine extracted features such as posterior acoustic attenuation, lesion shape, margins, and homogeneity, as previously noted in literature surveys. However, assessing the margins and morphology of lesions in USIs remains challenging [15]. Furthermore, ML methods have been commonly used for investigating and categorizing tumors according to lesion-specific handcrafted texture and morphologic features [16]. However, this source of features is still very much dependent on medical knowledge for the extraction of the features. Due to the issues related to handcrafted features, new algorithms have been adopted, including DL algorithms, which can learn features from data in a more efficient way, especially when it comes to learning high-order and non-linear features. DL models are equally efficient in the classification of USIs, especially where feature engineering may be challenging [17]. Other research works have applied DL approaches of different types, including pretrained CNNs, in classifying tumors in breast USI [18]. The following is a summary of this work's main contributions:

- Our Multi-Feature Attention Classification Network (MFAN) makes use of two ground-breaking modules: the Multi-Scale Spatial Channel Attention Module (Mc-SCAM) and the Global–Local Attention Module for Feature Fusion. Both of these modules have unprecedented capabilities. The MFAN method solves the low classification accuracy of the NFSC method, which is caused by confined lesion zones and backdrops that are comparable.
- Concerning information loss in some dimensions during feature extraction, the Mc-SCAM can perform comprehensive feature extraction. Thus, it learns about the dependency of the channel and allows for the expanding of the perception of depth-wise convolution to gain a better understanding of the model.
- The GLAM aggregates various scale features in the network and simultaneously adopts both global and local information by employing the dual attention branches. This module integrates characteristics from different scales and optimizes the information extraction step, improving the capability of the model to focus on subtle lesion regions, thus enhancing classification capacity.

This article is divided as follows: Section 2 summarizes the past literature; The proposed model is fully detailed in Section 3; Section 4 covers the experimental results and discussions; and Section 5 concludes the paper.

2. Literature Review

Previous research has extensively examined the analysis phenomena as an effective way to improve Convolutional Neural Networks' (CNNs) representation capacities across different domains [19–23]. In addition to their application in object recognition [24], CNNs are also employed for the detection of human movement [25,26], the detection of targets [27], and the identification of individuals [28]. Advanced computer vision-based ultrasound breast cancer classification methods have been developed. The new CNN-based DL model BreastNet classifies breast cancer from histopathology images. Using a hypercolumn attention mechanism and residual blocks, this model achieved 98.80% accuracy on the BreakHis dataset [29]. Using transfer learning, a deep learning model improved breast cancer detection and prediction in cytology pictures, outperforming current algorithms [30]. For feature selection, the classification technique uses an information gain-directed simulated annealing genetic algorithm wrapper. This method uses information gain to prioritize features and a cost-sensitive SVM to identify optimal characteristics [31].

An architecture for breast cancer classifier was suggested by incorporating the attention mechanism into the VGG16 model. These improvements have helped this enhanced model to effectively distinguish between background features and lesion-specific features in ultrasound images. A complex ensemble loss function with binary cross-entropy loss and the logarithm of hyperbolic cosine loss was also introduced. This combination aimed to improve the distinction between labels and lesion classifications, resulting in a more detailed classification process [32].

Breast mass classification from ultrasound images utilizes transfer learning (TL) and deep representation scaling (DRS) layers. The DRS layer settings of this pre-trained CNN were optimized for breast mass classification, outperforming state-of-the-art approaches such as Byra [33] during training. Similar to LUPI, a deep doubly supervised transfer learning network classifies breast cancer using MMD criteria. Stiff improved classification performance using a unique doubly supervised transfer learning network (DDSTN) using these strategies [34]. Using ultrasound pictures, image fusion with content representation, and CNN models, a computerized diagnostic system was created to discriminate benign from malignant breast cancer. Using BUSI and other datasets, the authors achieved significant classification accuracy gains [35].

Using a complex image-processing structure meant for multi-view examination [36] improved diagnostic outcomes. The first-order local entropy was applied to threshold the tumor regions from the texture features, and the derived measures were used to decide the radius and area of possible malignancies. This method provided the detection accuracies of 88.0% and 80.5% of images obtained with CC and MLO in breast cancer screening. In [37], the authors proposed a framework based on transfer learning and used several augmentation techniques to increase the size of the mammogram dataset and reduce overfitting, which would lead to more accurate outcomes. Almalki et al. [38] put forward a multi-stage technique and applied it to a large mammography image base. First, the classification step was conducted, and afterwards, the pectoral muscle extraction took place at the second stage. The last approach entailed identifying abnormal regions in the enhanced images through a segmentation module that yielded a classification accuracy of 92% when compared with the BI-RADS dataset that has five categories. Further advancement and optimization are being conducted to make the DL approach the key technology for medical diagnostics, and deep learning methods have indicated high efficacy in the diverse field of medical applications. Despite their astounding capabilities, DL algorithms are not perfect. The quality and diversity of training data determines AI breast cancer detection accuracy. Diagnostic disparities can result from data biases or training dataset restrictions. AI models must be validated and refined to reduce errors and ensure consistent outcomes across demographics and clinical contexts. These invasion approaches have shown potential progress; however, there is still plenty of room for improvement and development in this area [39,40]. In the next section, the proposed model will try to overcome these identified gaps.

3. Materials and Methods

The proposed model shown in Figure 1 explains the overall architecture. It starts with an image fed into the system for processing. Initial feature extraction uses attention modules. These modules allow the model to focus on the most important parts of the image, resulting in high-quality, contextually relevant features. The GLAM (Global–Local Attention Module) modules fuse the extracted features. Fusion creates a more complete view and improves the model's capacity to recognize complex nuances. Finally, fused features provide output labels for input image categorization or recognition. This structured technique uses attention-based feature extraction and enhanced feature fusion to increase model accuracy and performance.
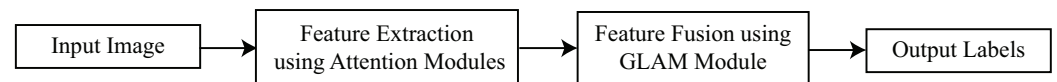
```
┌─────────────┐      ┌──────────────────┐      ┌──────────────────┐      ┌───────────────┐
│ Input Image │ ───> │ Feature Extraction│ ───> │ Feature Fusion using│ ───> │ Output Labels │
│             │      │ using Attention Modules│  │   GLAM Module     │      │               │
└─────────────┘      └──────────────────┘      └──────────────────┘      └───────────────┘
```

Figure 1. Simplified architecture of the proposed model for breast cancer classification.

3.1. Pre-Processing

In the first step, the images are pre-processed to remove noise via the Wiener Filter (WF) method. In general, noisy images are pre-processed using the noise removal method that helps to increase the image features which have been affected by the noise. An adaptive filter is a more specific approach, within which the denoising process solely relies on the presence of noise at the local level within that image. Let the noisy image be assessed as $\hat{I}(x,y)$, and the variance of noise throughout the area as σ_y^2. The mean of the local window is represented by the symbol $\hat{\mu}_L$, and the variance of the local window is represented by the symbol $\hat{\sigma}_y^2$.

$$\hat{I} = I'(x,y) - \frac{\sigma_y^2}{\hat{\sigma}_y^2}\left(I'(x,y) - \hat{\mu}_L\right) \tag{1}$$

At this point, the image has no noise fluctuations; hence, $\sigma_y^2 = 0 \geq \hat{I} = I'(x,y)$. This ratio approaches one when global noise variance diminishes and local variation exceeds global variance. If $\hat{\sigma}_y^2 \gg \sigma_y^2$, then $\hat{I} = I'(x,y)$. High local variance indicates an edge in the image window. For equal local and global variance, the formula is adjusted to $\hat{I} = \hat{\mu}_L$, with $\hat{\sigma}_y^2 \approx \sigma_y^2$.

This may indicate average intensity from the average area. Contrast enhancement is achieved using contrast-limited adaptive histogram equalization (CLAHE) [41]. CLAHE is a sophisticated form of adaptive histogram equalization that controls contrast amplification to reduce noise enhancement. The CLAHE increases pixel contrast proportionally to transfer function slope. Unlike the conventional image, CLAHE works on "tiles". Using bilinear interpolation, nearby tiles are merged to eliminate spurious split lines. This approach boosts visual contrast, making it helpful.

3.2. Proposed Architecture

3.2.1. Network Overview

The proposed network model is thus an improved version of the EfficientNetB2 [42] architecture, incorporated with the discussed network optimization techniques. In parsing the EfficientNet models from EfficientNetB0 to EfficientNetB7, the parameters of depth, width, and resolution are optimal within this sequence. The proposed MFAN, illustrated in Figure 2, is designed to support multi-feature and multi-scale classification. The major components are seven attention modules (AM) for characteristics extraction, two additional modules, and a classification convolutional layer. The AM block extracts extensive information with interesting properties. The AM highlights the data that improve model performance by combining scale features. The SE approach [43] picks features using global average pooling but overlooks crucial spatial information in feature maps. McSCAM is used to retain and extract spatial information from feature maps. Initially, the model's

discriminability can be improved through the use of multi-scale feature extraction in the channel attention technique without sacrificing spatial feature details.

In Figure 2, the AM6 block is responsible for the initial expansion of the feature map channels by a factor of six thousand. Following this, depth-wise convolution (DWC) is performed in order to keep the channel number consistently the same. Last but not least, the feature map is subjected to McSCAM in order to capture both channel and spatial detail. While AM1 is the activation map that has a constant channel number, the size of the convolution kernel k that is used by DWC in the AM block is the size of the convolution kernel. The ability to concentrate more intently on minute problematic regions within an image is made possible through the process of extracting features from photos of varying sizes.

The features generated from seven AM blocks are given as *F*1 to *F*7. The obtained features are then used to embed the multi-global–local attention module for feature fusion (GLAM) to achieve global and detailed multi-scale local semantic feature map information. A dedicated McSCAM is used to enhance these features and determine the final subtype for breast cancer.
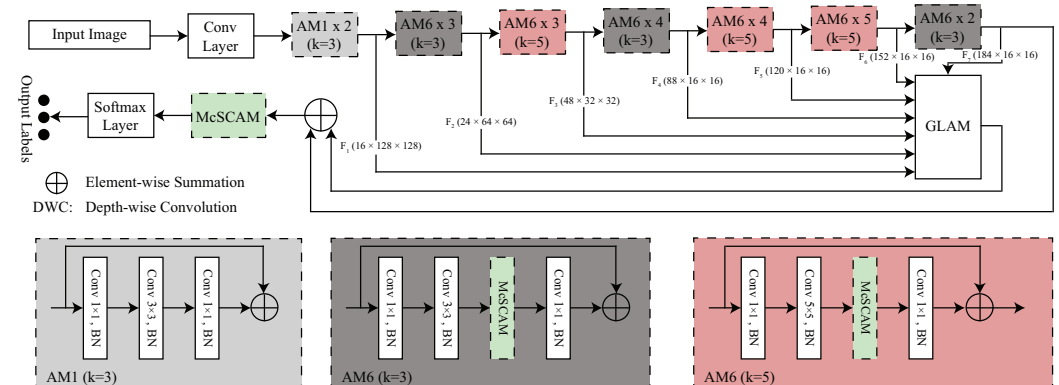

Figure 2. Architecture of proposed MFAN model for multi-feature and multi-scale classification.

3.2.2. Multi-Scale Spatial Channel Attention Module

A mechanism known as channel attention concentrates on critical channels while simultaneously reducing the impact of noise that is unrelated to those channels. Focusing attention on specific channels improves the model's performance in detecting and differentiating small subtypes of breast cancer. However, the fusion process of features basically considers the connection between channels, without considering the spatial data, and the latter is critical when aiming at localized breast cancer classification. Employing a multi-layer perceptron while learning features through subtraction hinders the channel's independence and has a detrimental effect on learning since it necessitates more parameters. The spatial pyramid pooling employed in visual task detection networks [32] can also be used for census map features from differently sized maps to sharpen tasks and keep precise spatial information. In achieving this goal, we design the attention mechanism and propose the McSCAM framework that considers multi-scale ideas. Thus, we maintain spatial feature details while learning channels by substituting the global average pooling layer with multi-scale pooling techniques. Also, the utilization of DWC in feature extraction tends to maintain distinct and independent channel interactions.

To obtain spatial characteristics, input features are pooled at three scales, as shown in Figure 3. The DWC approach is used to transform multi-scale spatial input into a consistent spatial dimension, hence aiding the process of feature learning. After applying the Sigmoid activation function, the input activation map is multiplied along the channel dimension. This procedure enables the network to provide priority to the vital channel by assigning it a greater weight value. Ultimately, the desired outcome is achieved by concentrating on the output of stored spatial data. The channel attention mechanism is formulated by incorporating multi-scale spatial information in the following manner:
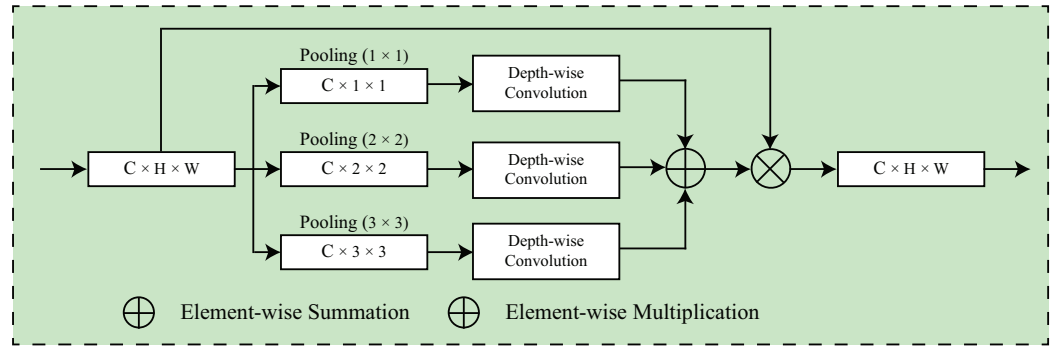
Figure 3. Architecture of the proposed McSCAM module.

$$g_{\text{out}} = \Sigma \left(\sum_{S \in \text{scales}} \frac{1}{S \times S} \text{DWC}(\text{Pool}(g_{\text{in}}, S)) \right) \otimes g_{\text{in}} \tag{2}$$

Assume that the input feature is denoted by g_{in}, the scale pooling is denoted by $Pool(., S)$, and the scales are denoted by S, with the purpose of preventing larger-scale pooling from dominating a larger channel weight while neglecting the benefits of small-scale pooling. Through the utilization of the coefficient as the square of the scale size, the reciprocal multiplication process allows for the equalization of the unique information that is obtained by pooling each scale. Both $DWC(.)$ and $Sig(.)$ are activation functions that are used to perform depth wise convolution and sigmoid activation, respectively. The symbol $\otimes$ is used to indicate element-wise multiplication, which results in the output feature g_{out} while maintaining multi-scale spatial information.

3.2.3. Multi-Global–Local Attention Module for Feature Fusion

The multi-scale feature fusion module (MSFFM) and global and local attention module make up GLAM blocks. The MSFFM combines and increases features from different scales, while the attention module uses two different mechanisms to obtain both global and local information appropriately. The technique that combines features from a specific scale to create a new feature space is called multi-scale feature fusion. This module builds on the popular Fully Convolutional Network (FCN) [44] and PANet [35] used in medical segmentation. Medical image segmentation demands acquiring additional feature information from images at different sizes to obtain complicated and diversified high-dimensional information for pixel-by-pixel classification.

This strategy also helps medical imaging subtype classification. The network's seven AM blocks' features are input modules. The output of layers with uneven spatial sizes is down-sampled to match high-dimensional characteristics. The fused features are obtained using the convolution in which channel expansion ensures that features are fused and concatenated along the channel dimension. Thus, after passing through a convolutional layer, the amount of information in the channel dimension is reduced and is beneficial for decreasing computational costs to calculate fused features. Down-sampled features are shown in Figure 3, which illustrates the fusion process for the down-sampled features. The multi-scale feature fusion process can be mathematically defined by the following equation. Here, $g_i \in R^{C_i \times H_i \times W_i}$, where i ranges from 1 to 7 to symbolize the features obtained from seven AM blocks. g_7 specifically refers to the most profound feature.

$$g_{di} = W_{di}^T g_i, \quad i = 1, 2, \ldots, 7, \tag{3}$$

$$g_{fi} = W_{e(i-1)}^T g_{d(i-1)} \oplus g_{di}, \quad i = 1, 2, \ldots, 7, \tag{4}$$

$$g_{f1} = W_{e1}^T g_{d1}, \tag{5}$$

where the symbol $W_{di}^T g_i$ represents the process of down sampling the feature g_i to obtain g_{di}. The expression $W_{ei}^T g_{di}$ reflects the result of performing channel expansion on g_{di} to

match the size of the next level $g_{d(i-1)}$ and then adding it to g_{di} to obtain g_{fi}. The symbol $\otimes$ denotes the operation of element-wise summing between two features. The computation of the output of the MSFFM is given as

$$g_{\text{out}} = W_{oi}^{T} C\left([g_{f1}, g_{f2}, \ldots, g_{f7}]\right). \tag{6}$$

The function $C(.)$ combines features, while W_{oi}^{T} convolutions channel compression to produce the aggregated feature g_{out}.

Integration of the multi-feature attention mechanism is now being implemented on both the global and the local levels. An individual attention mechanism is capable of concentrating on certain information; however, it typically only examines information that is either global or local in scope. To emphasize both global and local features in formation, it is important to take into consideration a structure that consists of two branches and includes both global and local attention modules, as depicted in Figure 4. The GLAM is composed of two branches: the global attention branch involves the utilization of global average pooling in order to obtain information from the aggregated feature g_{out} that is provided by MSFFM; a branch of convolution, known as the local attention branch, is responsible for extracting local information. Both types of data are integrated immediately. The convolution produces the feature GLA$_{\text{out}}$, emphasizing global and local information. The computations are detailed below.
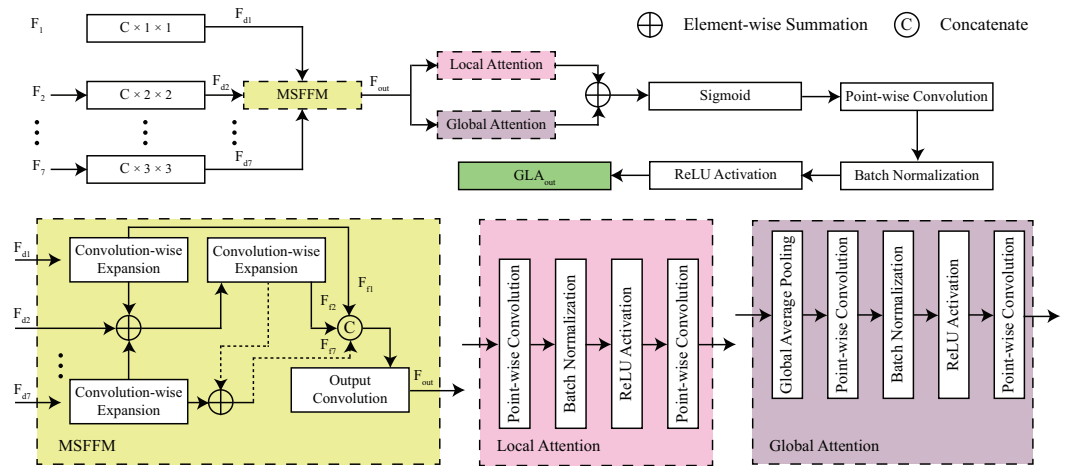

Figure 4. Architecture of the GLAM with global and local attention branches.

$$\text{GA}_f = \text{PWC}(\text{Re}(\text{BN}(\text{PWC}(\text{GAP}(g_{\text{out}}))))), \tag{7}$$

$$\text{LA}_f = \text{PWC}(\text{Re}(\text{BN}(\text{PWC}(g_{\text{out}})))), \tag{8}$$

$$\text{GLA}(g_{\text{out}}) = \text{Re}\left(\text{BN}\left(\text{PWC}\left(\text{Sig}\left(\text{GA}_f \oplus \text{LA}_f\right)\right)\right)\right), \tag{9}$$

where function $GA(.)$ represents a global attention branch, whereas $LA(.)$ represents another local attention branch. The notation $GAP(.)$ represents global average pooling, whereas $PWC(.)$ represents point-wise convolution with a kernel size of 1×1. These operations are specifically intended to decrease the number of parameters. Batch normalization, denoted as $BA(.)$, is a technique that is used after the convolution procedure and before the activation function. The functions ReLU and Sigmoid are denoted by $Re(.)$ and $Sig(.)$, respectively. The symbol $\oplus$ is used to denote element-wise summing operations.

3.3. Loss Function

To prevent network overfitting, we employ the cross-entropy loss function for label smoothing [45]. This is crucial when dealing with limited data as the model cannot adequately represent all sample properties determined by

$$\text{LS}_{\text{CELossi}} = \begin{cases} -(1-\varepsilon)\sum_{i=1}^{k} p_i \log(q_i) & \text{if } (i=y) \\ -\varepsilon \sum_{i=1}^{k} p_i \log(q_i) & \text{if } (i \neq y). \end{cases} \tag{10}$$

k is the number of categories, while $\in$ is a smoothing hyperparameter. Actual sample and model-predicted distribution are p_i and q_i, respectively. Label smoothing improves model generalization through regularization. Combining the p_i value of one-hot encoding with the noise softens the version, reducing the relevance of the category corresponding to the sample label in the loss function:

$$p_i = \begin{cases} 1-\varepsilon & \text{if } (i=y) \\ \frac{\varepsilon}{k-1} & \text{if } (i \neq y). \end{cases} \tag{11}$$

In the next section, the experimental results are presented, which proves the authenticity of the proposed model.

4. Experiments

4.1. Dataset Description and Experimental Setup

For this purpose, we employed the CBIS-DDSM dataset (D1) [46], which is an enhanced version of the DDSM [47], including 2620 mammography cases with associated pathology. This short subset is managed by an expert mammographer and consists of 6671 images of mammograms in PNG format, with image sizes of 255×255 pixels. The data are split into 80% training and 20% testing, with new ROI segmentation and pathologic diagnoses for better CAD systems assessment. The Breast Cancer Ultrasound dataset (D2) [48] includes a set of images obtained through ultrasound scans of females varying within the age bracket of 25 to 75 years from the sample data collection year of 2018. The collection includes 780 PNG photos (about 500×500 pixels) from 600 female patients. Photographs are classified as superficial, benign, or malignant. The dataset includes 133 normal, 437 benign, and 210 cancerous images.

The experiment is based on the Intel Core i7-11700K CPU, 32 GB DDR4 RAM, and NVIDIA GeForce RTX 3080 GPU enabler for effective computing and high-performance GPU acceleration. Internal storage uses 1 TB NVMe SSD, and the main OS is Windows 11 Pro. The Python 3.12 environment is fine-tuned with the main libraries and frameworks, including TensorFlow. Hyperparameters are trained using a learning rate of 0.001 with 32 batches and 50 epochs using Adam optimizer (Adam Technologies, Islamabad, Pakistan).

4.2. Evaluation Metrics

The confusion matrix proves to be one of the most important metrics for assessing the performance of a trained classifier since it compares its predictions to the actual results in the test dataset. This matrix comprises four key metrics. These are referred to as the true positives (TP), the true negatives (TN), the false positives (FP), and the false negatives (FN). From these basic measures mentioned above, it is possible to derive other indicators such as the true positive rate (TPR), false positive rate (FPR), and true negative rate (TNR), which give a clear picture of how accurate and adhesive the classifier is. In the confusion matrix, TP and TN are represented by the elements lying on the diagonal, which defines the correct classifications of the model. On the other hand, FP and FN represent incorrect classifications; for instance, FN means that a malignant tumor has been classified as benign. These basic parameters allow for the determination of more specific characteristics of the model, including Accuracy (Acc), recall, Precision (Pre), and F1-score (F1). A fair basis for comparison in the current work was ensured by the selection of these evaluation metrics,

which were chosen to coincide with those utilized by methodologies that are considered to be state of the art.

4.3. Classification Results

4.3.1. Comparison with Pretrained Models

To compare with the current advanced DL models, Table 1 shows the performance of the proposed model following five pretrained models, which are EfficientNetB2, MobileNetv2, DenseNet201, ShuffleNetv3, and NasNetLarge, on two public datasets: D1 and D2. The findings reveal that the proposed model yields an outstanding performance, better than all other models for both datasets. In the case of D1, the proposed architecture of the DL model obtains an exceptional level of accuracy of 98.67%, which is significantly higher than the second-best model, namely, EfficientNetB2, which has scored 84.54%. In the same regard, with respect to recall, the proposed model scores 98.16%, surpassing EfficientNetB2's 83.13%. Precision and F1-score also behave similarly, and for the proposed model, it achieved a precision of 99.42% and 97.79%, respectively, compared to EfficientNetB2, which achieved and accuracy of 86.12% and 83.84%, respectively. Thus, on D2, the proposed model sustains it superior performance with an accuracy of 98.21%, which is higher than that of the MobileNetv2, which scored 83.97%, thereby ranking second on this dataset. The present model also obtained a recall of 98.06% and a precision of 99.36%, far exceeding those of NasNetLarge, which scored 76.23% and 77.63% in the same metrics. This quantitative comparison highlights the proposed model's exceptional ability to deliver high accuracy, precision, and recall across different datasets, establishing its effectiveness in breast cancer image classification tasks. Comparative analysis of evaluation metrics for the selected pretrained model and the proposed model on D1 is shown in Figure 5, while same data for D2 are shown in Figure 6.
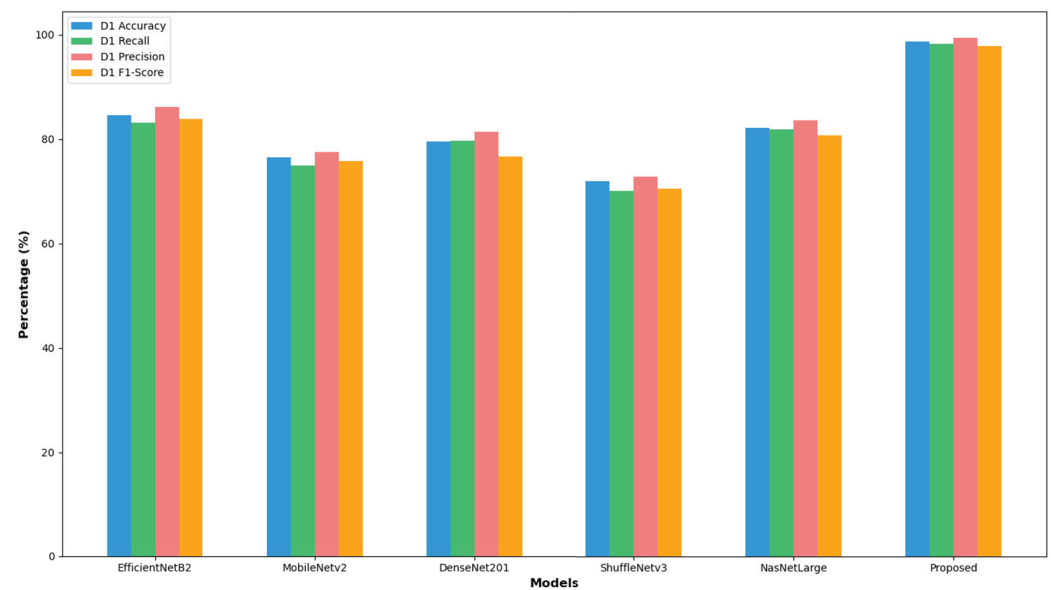

Figure 5. Comparative analysis of evaluation metrics for selected pretrained model and proposed model on D1.
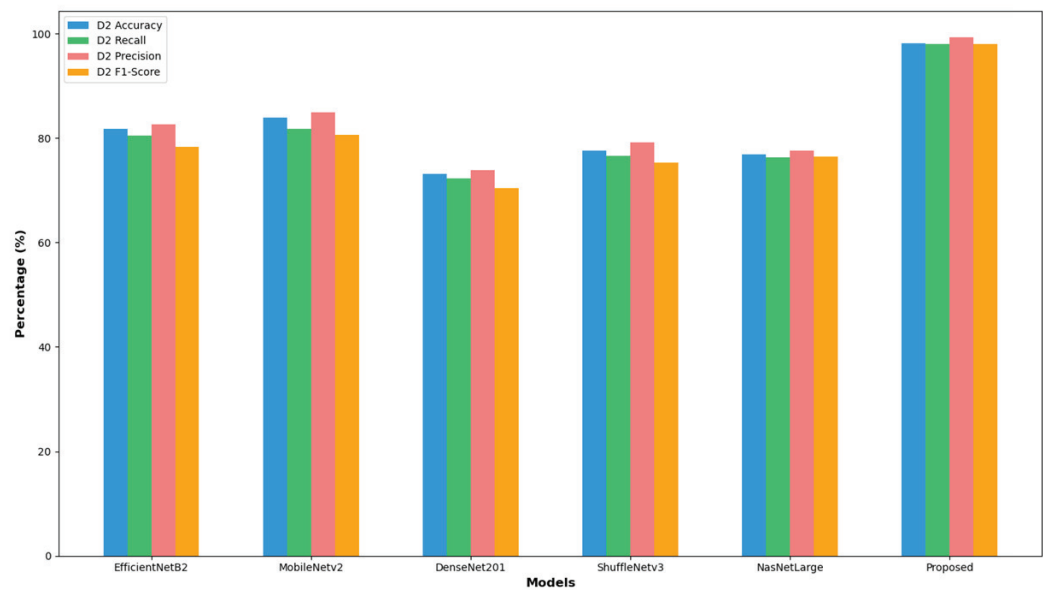
Figure 6. Comparative analysis of evaluation metrics for selected pretrained model and proposed model on D2.

Table 1. Comparison of pretained models with proposed model on D1 and D2.

Models	D1 Acc (%)	Re (%)	Pre (%)	F1 (%)	D2 Acc (%)	Re (%)	Pre (%)	F1 (%)
EfficientNetB2	84.54	83.13	86.12	83.84	81.72	80.45	82.67	78.32
MobileNetv2	76.49	74.89	77.56	75.77	83.97	81.77	84.99	80.54
DenseNet201	79.51	79.68	81.39	76.65	73.18	72.34	73.89	70.41
ShuffleNetv3	71.88	70.07	72.77	70.44	77.58	76.58	79.11	75.22
NasNetLarge	82.16	81.84	83.51	80.61	76.88	76.23	77.63	76.39
Proposed	98.67	98.16	99.42	97.79	98.21	98.06	99.36	98.01
Improvement	14.13	15.03	13.30	13.95	14.24	16.29	16.29	17.47

4.3.2. Comparison with Loss Function

Based on the results presented in Table 2, the performances of three loss functions, including Mean Square Error (MSE), Mean Absolute Error (MAE), and Categorical Cross Entropy (CCE), have been compared. The performance of CCE is significantly higher than all the other losses. For CCE on dataset D1, accuracy was found to be 98.67%; on D2, accuracy reaches 98.21%. However, MSE and MAE indicate low accuracy rates. Specifically, MSE achieves 97.15% on D1 and 95.89% on D2, while MAE achieves 96.44% on D1 and 97.64% on D2.

Table 2. Comparison of loss function for proposed model on D1 and D2.

Loss Function	D1 Acc (%)	F1 (%)	D2 Acc (%)	F1 (%)
Mean Square Error	97.15	95.76	95.89	94.72
Mean Absolute Error	96.44	95.21	97.64	95.99
Categorial Cross Entropy	98.67	97.79	98.21	98.01

4.3.3. Comparison with Different Scale Size of McSCAM

Table 3 shows variations of scale size for the multi-scale module between two datasets: D1 and D2. Among all the combinations, the 1,2,3 combination outperforms other combinations on both datasets, with an accuracy of 98.67% for D1 and 98.21% for D2. For other scale sizes, 1,3,5 and 1,3,5,7 reveal drops of 1.55–2.91% on D1 and 1.47–2.38% on D2. The 1,2,3,4,5 combination has a fairly good performance; however, the difference made is still insufficient by 2.09% on D1 and 0.93% on D2.

Table 3. Comparison of different scale sizes for McSCAM on D1 and D2.

Scale	D1		D2	
	Acc (%)	F1 (%)	Acc (%)	F1 (%)
1,2,3	98.67	97.79	98.21	98.01
1,3,5	97.12	96.33	96.74	94.72
1,3,5,7	95.76	93.64	95.83	94.18
1,2,3,4,5	96.58	95.21	97.28	95.54

4.3.4. Analysis of Proposed Modules

The results in Table 4 show the performance of the baseline, McSCAM, and GLAM models for various permutations on two datasets. Even based on the baseline alone, there is a decrease in performance in both datasets. Therefore, incorporating MsSCAM increases the outcomes by around 5.66% on D1 and 10.45% on D2. Even incorporating the GLAM alone also results in improvements of about 4.41% on D1 and 10.69% on D2. The combination of all three modules (baseline, MsSCAM, and GLAM) achieves the highest performance, outperforming the baseline by over 11% on D1 and by nearly 14% on D2. This highlights the substantial impact of combining the MsSCAM and GLAM for superior results.

Table 4. Comparison of performance of baseline, McSCAM, and GLAM models on D1 and D2.

Baseline	McSCAM	GLAM	D1		D2	
			Acc (%)	F1 (%)	Acc (%)	F1 (%)
✓			87.46	86.21	84.33	82.73
✓	✓		93.12	92.07	94.78	93.11
✓		✓	91.87	89.99	95.02	93.46
✓	✓	✓	98.67	97.79	98.21	98.01

4.3.5. Comparison with State-of-the-Art Methods

Table 5 compares different methods for breast cancer classification based on the CBIS-DDSM and Breast Cancer Ultrasound datasets, focusing on their accuracies. For the CBIS-DDSM dataset, existing methods demonstrate accuracy ranging from 77.60% to 91.50%, with Method [49] being the highest at 91.50%. However, the proposed method outperforms others, and it has an accuracy of 98.67%. This significant improvement demonstrates its enhanced performance in classifying malignant cases from benign ones in mammographic images. Likewise, current methods for the Breast Cancer Ultrasound dataset range from 85.83% to 96.69% accuracy, with Method [50] reaching the upper limit at 96.69%. The proposed method once more establishes a new state of the art, obtaining 98.21% accuracy. The improved overall performance of the system is demonstrated by the increase in the level of accuracy from 14% to 21%. These results underscore the proposed method's strength and innovation, positioning it as a highly reliable and advanced solution for breast cancer classification across different imaging modalities, surpassing the performance of existing state-of-the-art techniques.The next section concludes this article with future directions.

Table 5. Comparison with state-of-the-art methods on D1 and D2.

Method	Dataset	Accuracy (%)
DCNN-based multi-class classification [51]	D1	88.00
Integrating Segmentation with CNN [52]	D1	77.60
Deep features fused with local features [49]	D1	91.50
Cross-model semantic mining [53]	D1	87.05
Shallow CNN [54]	D1	89.20
Proposed	D1	98.67
Deep learning-based model [55]	D2	85.83
Image enhancement techniques for detection [50]	D2	96.69
Proposed	D2	98.21

5. Conclusions

This work presents an enhanced approach for breast cancer categorization using both mammography and ultrasound images, which involves a series of steps, including image preprocessing and categorization. Preprocessing is utilized for contrast enhancement, and it makes a dramatic difference in the image quality. The images used for enhancing are employed in training the proposed Multi-Feature Attention Network (MFAN) and the results obtained from these enhanced images. Evaluation of the proposed MFAN model with respect to mammography and ultrasound data proves that it has high accuracy and a high generalization ability in different types of imaging. However, one limitation with of method is that the use of strictly labeled data are limited, thus reducing the availability of a large volume of mostly unlabeled data that could be used to improve the model even further. Multiple attention modules and significant memory utilization from feature maps make the proposed model computationally demanding and unsuitable for real-time and resource-constrained applications. Perfect performance may require huge labeled datasets, risking overfitting on small data and lowering generalizability. The intricate structure makes hyperparameter adjustment and interpretability difficult, and scaling to higher-resolution inputs may require large processing resources. For future work, it is suggested that researchers incorporate MFAN with unsupervised or weakly supervised learning approaches to utilize more unlabeled data for frequently overseeing the mentioned constraints and advancing the technology of breast cancer classification. Techniques such as autoencoders for feature extraction, contrastive learning for improved representation learning, or clustering algorithms (such as K-means or hierarchical clustering) could be incorporated into the model in order to enhance its capacity to recognize subtle patterns within the data. These techniques could provide valuable insights.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Sung, H.; Ferlay, J.; Siegel, R.L.; Laversanne, M.; Soerjomataram, I.; Jemal, A.; Bray, F. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* **2021**, *71*, 209–249. [CrossRef] [PubMed]
2. Ahmad, S.; Ur Rehman, S.; Iqbal, A.; Farooq, R.K.; Shahid, A.; Ullah, M.I. Breast cancer research in Pakistan: A bibliometric analysis. *Sage Open* **2021**, *11*, 21582440211046934. [CrossRef]
3. Miller, K.D.; Nogueira, L.; Devasia, T.; Mariotto, A.B.; Yabroff, K.R.; Jemal, A.; Kramer, J.; Siegel, R.L. Cancer treatment and survivorship statistics, 2022. *CA Cancer J. Clin.* **2022**, *72*, 409–436. [CrossRef]
4. World Health Organization. *WHO Position Paper on Mammography Screening*; World Health Organization: Geneva, Switzerland, 2014.
5. Widiana, I.K.; Irawan, H. Clinical and subtypes of breast cancer in Indonesia. *Asian Pac. J. Cancer Care* **2020**, *5*, 281–285. [CrossRef]
6. Nasir, I.M.; Raza, M.; Shah, J.H.; Wang, S.H.; Tariq, U.; Khan, M.A. HAREDNet: A deep learning based architecture for autonomous video surveillance by recognizing human actions. *Comput. Electr. Eng.* **2022**, *99*, 107805. [CrossRef]
7. Nasir, I.M.; Rashid, M.; Shah, J.H.; Sharif, M.; Awan, M.Y.; Alkinani, M.H. An optimized approach for breast cancer classification for histopathological images based on hybrid feature set. *Curr. Med. Imaging* **2021**, *17*, 136–147. [CrossRef]
8. Nasir, I.M.; Raza, M.; Shah, J.H.; Khan, M.A.; Rehman, A. Human action recognition using machine learning in uncontrolled environment. In Proceedings of the 2021 1st International Conference on Artificial Intelligence and Data Analytics (CAIDA), Riyadh, Saudi Arabia, 6–7 April 2021; IEEE: New York, NY, USA, 2021.
9. Nasir, I.M.; Raza, M.; Shah, J.H.; Khan, M.A.; Nam, Y.C.; Nam, Y. Improved Shark Smell Optimization Algorithm for Human Action Recognition. *Comput. Mater. Contin.* **2023**, *76*, 2667–2684.
10. Nasir, I.M.; Raza, M.; Ulyah, S.M.; Shah, J.H.; Fitriyani, N.L.; Syafrudin, M. ENGA: Elastic Net-Based Genetic Algorithm for human action recognition. *Expert Syst. Appl.* **2023**, *227*, 120311. [CrossRef]
11. Tehsin, S.; Nasir, I.M.; Damaševičius, R.; Maskeliūnas, R. DaSAM: Disease and Spatial Attention Module-Based Explainable Model for Brain Tumor Detection. *Big Data Cogn. Comput.* **2024**, *8*, 97. [CrossRef]
12. Tehsin, S.; Hassan, A.; Riaz, F.; Nasir, I.M.; Fitriyani, N.L.; Syafrudin, M. Enhancing Signature Verification Using Triplet Siamese Similarity Networks in Digital Documents. *Mathematics* **2024**, *12*, 2757. [CrossRef]
13. Malik, D.S.; Shah, T.; Tehsin, S.; Nasir, I.M.; Fitriyani, N.L.; Syafrudin, M. Block Cipher Nonlinear Component Generation via Hybrid Pseudo-Random Binary Sequence for Image Encryption. *Mathematics* **2024**, *12*, 2302. [CrossRef]
14. Zhang, X.; Lin, X.; Tan, Y.; Zhu, Y.; Wang, H.; Feng, R.; Tang, G.; Zhou, X.; Li, A.; Qiao, Y. A multicenter hospital-based diagnosis study of automated breast ultrasound system in detecting breast cancer among Chinese women. *Chin. J. Cancer Res.* **2018**, *30*, 231. [CrossRef] [PubMed]
15. Mohammed, M.A.; Al-Khateeb, B.; Rashid, A.N.; Ibrahim, D.A.; Abd, Ghani, M.K.; Mostafa, S.A. Neural network and multi-fractal dimension features for breast cancer classification from ultrasound images. *Comput. Electr. Eng.* **2018**, *70*, 871–882. [CrossRef]
16. Wang, Y.; Wang, N.; Xu, M.; Yu, J.; Qin, C.; Luo, X.; Yang, X.; Wang, T.; Li, A.; Ni, D. Deeply-supervised networks with threshold loss for cancer detection in automated breast ultrasound. *IEEE Trans. Med. Imaging* **2019**, *39*, 866–876. [CrossRef]
17. Shen, L.; Margolies, L.R.; Rothstein, J.H.; Fluder, E.; McBride, R.; Sieh, W. Deep learning to improve breast cancer detection on screening mammography. *Sci. Rep.* **2019**, *9*, 12495. [CrossRef]
18. Mambou, S.J.; Maresova, P.; Krejcar, O.; Selamat, A.; Kuca, K. Breast cancer detection using infrared thermal imaging and a deep learning model. *Sensors* **2018**, *18*, 2799. [CrossRef]
19. Tehsin, S.; Rehman, S.; Saeed, M.O.B.; Riaz, F.; Hassan, A.; Abbas, M.; Young, R.; Alam, M.S. Self-organizing hierarchical particle swarm optimization of correlation filters for object recognition. *IEEE Access* **2017**, *5*, 24495–24502. [CrossRef]
20. Tehsin, S.; Rehman, S.; Bilal, A.; Chaudry, Q.; Saeed, O.; Abbas, M.; Young, R. Comparative analysis of zero aliasing logarithmic mapped optimal trade-off correlation filter. In Proceedings of the Pattern Recognition and Tracking XXVIII, Anaheim, CA, USA, 9–13 April 2017; SPIE: Bellingham, WA, USA, 2017.
21. Tehsin, S.; Rehman, S.; Riaz, F.; Saeed, O.; Hassan, A.; Khan, M.; Alam, M.S. Fully invariant wavelet enhanced minimum average correlation energy filter for object recognition in cluttered and occluded environments. In Proceedings of the Pattern Recognition and Tracking XXVIII, Anaheim, CA, USA, 9–13 April 2017; SPIE: Bellingham, WA, USA, 2017.
22. Tehsin, S.; Asfia, Y.; Akbar, N.; Riaz, F.; Rehman, S.; Young, R. Selection of CPU scheduling dynamically through machine learning. In Proceedings of the Pattern Recognition and Tracking XXXI, Online, 27 April–9 May 2020; SPIE: Bellingham, WA, USA, 2020.
23. Saad, S.M.; Bilal, A.; Tehsin, S.; Rehman, S. Spoof detection for fake biometric images using feature-based techniques. In Proceedings of the SPIE Future Sensing Technologies, Online, 9–13 November 2020; SPIE: Bellingham, WA, USA, 2020.
24. Tehsin, S.; Rehman, S.; Awan, A.B.; Chaudry, Q.; Abbas, M.; Young, R.; Asif, A. Improved maximum average correlation height filter with adaptive log base selection for object recognition. In Proceedings of the Optical Pattern Recognition XXVII, Baltimore, MD, USA, 17–21 April 2016; SPIE: Bellingham, WA, USA, 2016.
25. Akbar, N.; Tehsin, S.; Bilal, A.; Rubab, S.; Rehman, S.; Young, R. Detection of moving human using optimized correlation filters in homogeneous environments. In Proceedings of the Pattern Recognition and Tracking XXXI, Online, 27 April–9 May 2020; SPIE: Bellingham, WA, USA, 2020.

26. Yousafzai, S.N.; Shahbaz, H.; Ali, A.; Qamar, A.; Nasir, I.M.; Tehsin, S.; Damaševičius, R. X-News dataset for online news categorization. *Int. J. Intell. Comput. Cybern.* **2024**, *17*, 737–758. [CrossRef]

27. Akbar, N.; Tehsin, S.; ur Rehman, H.; Rehman, S.; Young, R. Hardware design of correlation filters for target detection. In Proceedings of the Pattern Recognition and Tracking XXX, Baltimore, MD, USA, 14–18 April 2019; SPIE: Bellingham, WA, USA, 2019.

28. Asfia, Y.; Tehsin, S.; Shahzeen, A.; Khan, U.S. Visual person identification device using raspberry Pi. In Proceedings of the 25th Conference of FRUCT Association, Helsinki, Finland, 5–8 November 2019.

29. Toğaçar, M.; Özkurt, K.B.; Ergen, B.; Cömert, Z. BreastNet: A novel convolutional neural network model through histopathological images for the diagnosis of breast cancer. *Phys. A Stat. Mech. Its Appl.* **2020**, *545*, 123592. [CrossRef]

30. Khan, S.; Islam, N.; Jan, Z.; Din, I.U.; Rodrigues, J.J.C. A novel deep learning based framework for the detection and classification of breast cancer using transfer learning. *Pattern Recognit. Lett.* **2019**, *125*, 1–6. [CrossRef]

31. Liu, N.; Qi, E.S.; Xu, M.; Gao, B.; Liu, G.Q. A novel intelligent classification model for breast cancer diagnosis. *Inf. Process. Manag.* **2019**, *56*, 609–623. [CrossRef]

32. Kalafi, E.Y.; Jodeiri, A.; Setarehdan, S.K.; Lin, N.W.; Rahmat, K.; Taib, N.A.; Ganggayah, M.D.; Dhillon, S.K. Classification of breast cancer lesions in ultrasound images by using attention layer and loss ensemble in deep convolutional neural networks. *Diagnostics* **2021**, *11*, 1859. [CrossRef] [PubMed]

33. Byra, M. Breast mass classification with transfer learning based on scaling of deep representations. *Biomed. Signal Process. Control.* **2021**, *69*, 102828. [CrossRef]

34. Han, X.; Wang, J.; Zhou, W.; Chang, C.; Ying, S.; Shi, J. Deep Doubly Supervised Transfer Network for Diagnosis of Breast Cancer with Imbalanced Ultrasound Imaging Modalities. *arXiv* **2020**, arXiv:2007.06634.

35. Moon, W.K.; Lee, Y.W.; Ke, H.H.; Lee, S.H.; Huang, C.S.; Chang, R.F. Computer-aided diagnosis of breast ultrasound images using ensemble learning from convolutional neural networks. *Comput. Methods Programs Biomed.* **2020**, *190*, 105361. [CrossRef]

36. Hikmah, N.F.; Sardjono, T.A.; Mertiana, W.D.; Firdi, N.P.; Purwitasari, D. An image processing framework for breast cancer detection using multi-view mammographic images. EMITTER *Int. J. Eng. Technol.* **2022**, *10*, 136–152. [CrossRef]

37. Alruwaili, M.; Gouda, W. Automated breast cancer detection models based on transfer learning. *Sensors* **2022**, *22*, 876. [CrossRef]

38. Almalki, Y.E.; Soomro, T.A.; Irfan, M.; Alduraibi, S.K.; Ali, A. Computerized analysis of mammogram images for early detection of breast cancer. *Healthcare* **2022**, *10*, 801. [CrossRef]

39. Dhar, T.; Dey, N.; Borra, S.; Sherratt, R.S. Challenges of deep learning in medical image analysis—Improving explainability and trust. *IEEE Trans. Technol. Soc.* **2023**, *4*, 68–75. [CrossRef]

40. Wetstein, S.C.; de Jong, V.M.; Stathonikos, N.; Opdam, M.; Dackus, G.M.; Pluim, J.P.; van Diest, P.J.; Veta, M. Deep learning-based breast cancer grading and survival analysis on whole-slide histopathology images. *Sci. Rep.* **2022**, *12*, 15102. [CrossRef]

41. Zuiderveld, K. Contrast limited adaptive histogram equalization. In *Graphics Gems IV*; Academic Press Professional, Inc.: San Diego, CA, USA, 1994; pp. 474–485.

42. Tan, M.; Le, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In Proceedings of the International Conference on Machine Learning, PMLR, Long Beach, CA, USA, 10–15 June 2019.

43. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018.

44. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015.

45. Müller, R.; Kornblith, S.; Hinton, G.E. When does label smoothing help? *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 1–17.

46. Lee, R.S.; Gimenez, F.; Hoogi, A.; Miyake, K.K.; Gorovoy, M.; Rubin, D.L. A curated mammography data set for use in computer-aided detection and diagnosis research. *Sci. Data* **2017**, *4*, 170177. [CrossRef] [PubMed]

47. Heath, M.; Bowyer, K.; Kopans, D.; Kegelmeyer, P., Jr.; Moore, R.; Chang, K.; Munishkumaran, S. Current status of the digital database for screening mammography. In *Digital Mammography: Nijmegen*; Springer: Dordrecht, The Netherlands, 1998; pp. 457–460.

48. Al-Dhabyani, W.; Gomaa, M.; Khaled, H.; Fahmy, A. Dataset of breast ultrasound images. *Data Brief* **2020**, *28*, 104863. [CrossRef]

49. Sajid, U.; Khan, R.A.; Shah, S.M.; Arif, S. Breast cancer classification using deep learned features boosted with handcrafted features. *Biomed. Signal Process. Control.* **2023**, *86*, 105353. [CrossRef]

50. Da Silva, D.S.; Nascimento, C.S.; Jagatheesaperumal, S.K.; Albuquerque, V.H.C.D. Mammogram image enhancement techniques for online breast cancer detection and diagnosis. *Sensors* **2022**, *22*, 8818. [CrossRef]

51. Heenaye-Mamode Khan, M.; Boodoo-Jahangeer, N.; Dullull, W.; Nathire, S.; Gao, X.; Sinha, G.R.; Nagwanshi, K.K. Multi-class classification of breast cancer abnormalities using Deep Convolutional Neural Network (CNN). *PLoS ONE* **2021**, *16*, e0256500. [CrossRef] [PubMed]

52. Tsochatzidis, L.; Koutla, P.; Costaridou, L.; Pratikakis, I. Integrating segmentation information into CNN for breast cancer diagnosis of mammographic masses. *Comput. Methods Programs Biomed.* **2021**, *200*, 105913. [CrossRef]

53. Zhang, H.; Wu, R.; Yuan, T.; Jiang, Z.; Huang, S.; Wu, J.; Hua, J.; Niu, Z.; Ji, D. DE-Ada*: A novel model for breast mass classification using cross-modal pathological semantic mining and organic integration of multi-feature fusions. *Inf. Sci.* **2020**, *539*, 461–486.

54. Das, H.S.; Das, A.; Neog, A.; Mallik, S.; Bora, K.; Zhao, Z. Breast cancer detection: Shallow convolutional neural network against deep convolutional neural networks based approach. *Front. Genet.* **2023**, *13*, 1097207. [CrossRef]

55. Uysal, F.; Köse, M.M. Classification of breast cancer ultrasound images with deep learning-based models. *Eng. Proc.* **2022**, *31*, 8. [CrossRef]