mathematics

Article

An interpretable Breast Ultrasound Image Classification Algorithm Based on Convolutional Neural Network and Transformer

Xiangjia Meng [1,2,*], Jun Ma [3], Feng Liu [1,2], Zhihua Chen [1,2] and Tingting Zhang [1,2]

[1] School of Information Engineering, Shandong Youth University of Political Science, Jinan 250103, China
[2] New Technology Research and Development Center of Intelligent Information Controlling in Universities of Shandong, Shandong Youth University of Political Science, Jinan 250103, China
[3] School of Cyber Science and Engineering, Southeast University, Nanjing 219302, China
* Correspondence: mxj@sdyu.edu.cn

Abstract: Breast cancer is one of the most common causes of death in women. Early signs of breast cancer can be an abnormality depicted on breast images like breast ultrasonography. Unfortunately, ultrasound images contain a lot of noise, which greatly increases the difficulty for doctors to interpret them. In recent years, computer-aided diagnosis (CAD) has been widely used in medical images, reducing the workload of doctors and the probability of misdiagnosis. However, it still faces the following challenges in clinical practice: one is the lack of interpretability, and another is that the accuracy is not high enough. In this paper, we propose a classification model of breast ultrasound images that leverages tumor boundaries as prior knowledge and strengthens the model to guide classification. Furthermore, we employ the advantages of convolutional neural network (CNN) to extract local features and Transformer to extract global features to achieve information balance and complementarity between the two neural network models which increase the recognition performance of the model. Additionally, an explanation method is used to generate visual results, thereby improving the poor interpretability of deep learning models. Finally, we evaluate the model on the BUSI dataset and compare it with other CNN and Transformer models. Experimental results show that the proposed model obtains an accuracy of 0.9870 and an F1 score of 0.9872, achieving state-of-the-art performance.

Keywords: breast cancer; ultrasound imaging classification; artificial intelligence; ensemble learning

MSC: 68T05

check for updates

Citation: Meng, X.; Ma, J.; Liu, F.; Chen, Z.; Zhang, T. An interpretable Breast Ultrasound Image Classification Algorithm Based on Convolutional Neural Network and Transformer. *Mathematics* **2024**, *12*, 2354. https://doi.org/10.3390/math12152354

Academic Editors: Vladimir Balan and Jakub Nalepa

Received: 12 April 2024
Revised: 8 July 2024
Accepted: 26 July 2024
Published: 28 July 2024

1. Introduction

Breast cancer is one of the major cancers that have attracted attention in the world, and it has become first in the incidence of cancer in women. According to "Cancer Statistics, 2022" [1], breast cancer has accounted for 31% of new cancer cases in women, greatly outpacing the percentage of other malignancies. Due to the etiology of breast cancer being unclear, it is difficult to prevent it. Physical examination is able to find breast suspicious lesions and tumors as early as possible. The effective way to increase the breast cancer survival rate is early identification and diagnosis. With the right therapy, early-stage breast cancer can be permanently healed.

The methods for breast disease diagnosis mainly include clinician touch examination, radiological technology and cell histopathological biopsy. Benefiting from advances in imaging techniques, the accuracy of diagnosing breast lesions has improved significantly over the past few years. Among the advantages of non-invasive, radiological techniques, such as mammography, ultrasound, and computer tomography, have become essential and important procedures in the diagnosis of breast cancer. By using these technologies, abnormal signs in the breast can be effectively detected and located. Due to the characteristics of its practical inspection approach, cheap cost, no radiation, high real-time performance, and robust equipment mobility, ultrasound imaging has emerged as a key tool for early breast diagnosis and as an aid in the localization of minimally invasive breast rotational surgery [2,3]. Figure 1 presents normal, benign and malignant cases in the breast ultrasound dataset BUSI.

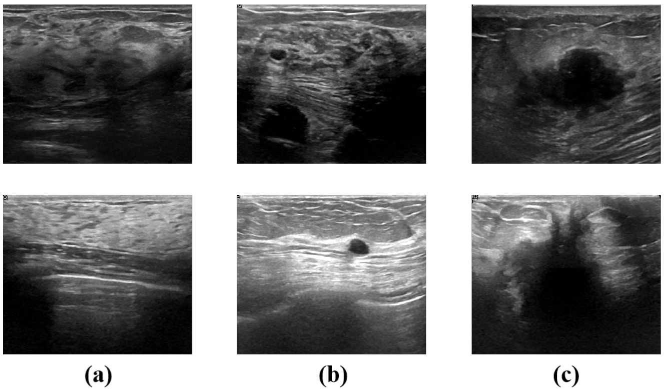

Figure 1. Examples of images in dataset BUSI (**first row**) and dataset (**second row**). (**a**) shows an example of normal case images, (**b**) images with benign lesion and (**c**) examples with malignant lesion.

Although breast ultrasound has multiple advantages, it still faces some challenges in practical application. On the one hand, due to the high acoustic impedance of various tissues and organs of the human body, it is easy to produce noise and artifacts in ultrasound images, resulting in blurred images. On the other hand, reading ultrasound images requires professional doctors with extensive clinical experience, which undoubtedly increases the time and labor costs. These challenges are not conducive to the identification of benign and malignant tumors and restrict the use of ultrasound for breast cancer diagnosis. Therefore, it is critical to interpret breast ultrasound images accurately and objectively. CAD is a popular way of interpreting medical images and shows a brilliant prospect, which improves the efficiency of diagnosis and increases the survival rate of patients. Many researchers have successfully applied CAD to breast cancer diagnosis, such as classifying breast imaging [4], detection of suspicious lesions [5], and segmenting tumor regions [6].

Most of the previous deep learning-based CAD are all convolutional neural networks (CNNs) with local filters, which neglect to take into account global features. However, local features are merely a portion of an image's features, and stronger global features with great recognition abilities are disregarded. Recently, the Vision Transformer (ViT) performed a global self-attention computation of the relationship between tokens [7,8], achieving comparable performance to CNN. Different from CNNs, induction biases and translational equivalence, are lost in ViT [9]. Scholars have demonstrated that the limitation of lack of inductive biases on ViT can break when the amount of data is large enough. Due to privacy and ethical requirements, it is difficult to obtain millions of labeled medical images. Therefore, we propose to integrate learning CNN and Transformer models to achieve better performance.

In addition, current deep-learning CAD algorithms are still black-box models, making them difficult to interpret. In order to build trust in intelligent systems and apply them to real breast cancer diagnosis scenarios, it is clear that we must build a transparent model to explain why the breast image is predicted to be normal, benign, or malignant. Some researchers attempt to open the deep neural network black box model through feature visualization, such as saliency map, class activation mapping (CAM) [10] and gradient weighted class activation mapping (Grad-CAM) [11]. These approaches resize the gradient or weight maps of the active feature maps to the size of the input image and overlap them on the original images to highlight the area of focus of the model. In contrast to these methods, we adopt a gradient-free interpretation method to evaluate the confidence increase to show the importance of each feature.

In this work, we propose an interpretable ensemble model to classify breast ultrasound images into normal, benign, and malignant. Specifically, tumor boundaries are considered as prior knowledge to improve the perception of tumor boundaries and reduce the incorrect recognition caused by blurred tumor boundaries. Furthermore, we train a CNN-based model and a multi-scale hierarchy Transformer, respectively, and then optimize the classification results using ensemble learning to improve the performance. In addition, we introduce a gradient-free interpretation method to improve model interpretability, in which the importance of feature activation is measured by calculating the change in confidence scores rather than measuring local sensitivities. The proposed model was validated on the breast ultrasound dataset BUSI.

In summary, our contributions are as follows:

- We propose an interpreted ensemble model for breast ultrasound image classification. During the training phase, the tumor boundaries mask is employed as prior knowledge to assist the model in identifying the region of interest.
- We integrate a CNN-based model and a multi-scale Transformer model to optimize the predictions and improve the average accuracy of the model.
- Moreover, we visualize the confidence increase map according to the prediction results to improve the interpretability of the model.
- Finally, we evaluate the model on the BUSI dataset and compare it with CNN and Transformer models. Experimental results show that the proposed model achieves state-of-the-art performance with 0.9870 in accuracy and 0.9872 in F1 score.

2. Related Work

2.1. Computer-Aided Diagnosis of Breast Cancer

CNNs have been widely employed in the analysis of breast images over the past few years. Compared with the traditional machine learning methods, CNN can automatically extract high-level nonlinear features from images, eliminating the need for a feature engineering stage and saving a significant amount of time [12–15]. Reference [2] proposed the abnormal breast ultrasound system for cancer detection to accelerate reviewing while keeping high detection sensitivity with low false positives. Reference [16] designed a novel method to segment the breast tumor via semantic classification and merging patches. Reference [17] introduced attention blocks into the U-Net and learned the feature representation of the spatial region with high salience. Most of these CNN-based methods were trained on a fully convolutional network to semantically segment lesion or tumor regions on ultrasound images and adopt the post-processing step to obtain more accurate segmentation results. In addition to breast lesion segmentation, scholars also adopted the CNN for classifying breast images into benign and malignant [4,18], judging the BI-RADS grading [19] and predicting the subclass of the tumors like fibroadenoma and lobular carcinoma [20,21]. Moreover, several recent works were focused on localizing target objects of interest and classifying given ROIs into benign or malignant [22–25]. The above breast CAD systems are expected to assist doctors by improving the diagnosis of breast cancer in clinical practice [26].

2.2. Transformer of Computer Version

Transformer was originally proposed for natural language processing (NLP) and has achieved great success in several fields, such as text classification, sentiment analysis, translation and so on [27–29]. Recently, inspired by the functionality of the Transformer in NLP, researchers have extended the Transformer to computer vision tasks and shown the ability to replace CNN. For example, to solve the difficulty of image identification, Dosovitskiy et al. proposed the visual transformer ViT that divided the whole image into several image patches and embedded them in a sequence of tokens [7]. Then, multiple Transformer layers were applied to tokens to build global attention. Experiments on

ViT demonstrated that it outperformed CNNs on several image recognition benchmarks. Li et al. [30] designed a tokens-to-token Vision Transformer (T2T-ViT), which recursively aggregated neighboring tokens into a single token and progressively structuring images to tokens so that local structures represented by surrounding tokens can be modeled. Liu et al. [9] introduced a multi-scale hierarchical Transformer, Swin-Transformer, which calculated the representation using shifting Windows. The window-shifting scheme improved efficiency by limiting self-attention calculations to non-overlapping local windows while allowing cross-window connections. This hierarchical structure allowed modeling feature mapping at different scales with linear computational complexity associated with image size. Therefore, in this paper, Swin-Transformer is used as a backbone network to solve the classification problem of breast ultrasound images.

2.3. Attribution Methods in Deep Learning

Although CNN has achieved great success in several areas [31–34], interpreting its results still faces challenges and limitations in medical scenarios. In recent years, several studies have provided attribution methods, which can visualize the class activation and class discriminative regions (i.e., locate the category in the image) to help the researchers understand the predictions. Several previous works, such as Guided Backpropagation [35] and Deconvolution [36], visualized CNN predictions by highlighting important pixels (i.e., the value variation of these pixels has the greatest impact on the prediction score). Despite producing fine-grained visualizations, these methods are not class-discriminative [11]. Gradient-based interpretation methods reflect the sensitivity of each feature to the predicted category by calculating the forward and backward gradient scores of all input features when they pass through the network, such as the typical saliency maps. Nevertheless, saliency maps frequently contain noise [37], and using their absolute value can prevent the detection of positive and negative evidence that might be present in the input. In this paper, we employ a gradient-free method, Score-CAM, to generate the visual interpretation result, which evaluates the increase in confidence score and thus provides reliable results.

3. Method

To solve the problem of identifying benign and malignant breast cancers, we construct a breast cancer identification method based on the improved Swin Transformer network. Figure 2 depicts the framework of the proposed method, which consists of three modules: image embedding, multi-scale feature extraction, and classifier. Specifically, the tumor mask is employed as prior knowledge to improve the model's perception of the tumor location. Moreover, we calculate the class contribution score map by measuring the importance of each feature element for understanding the model's decision-making.

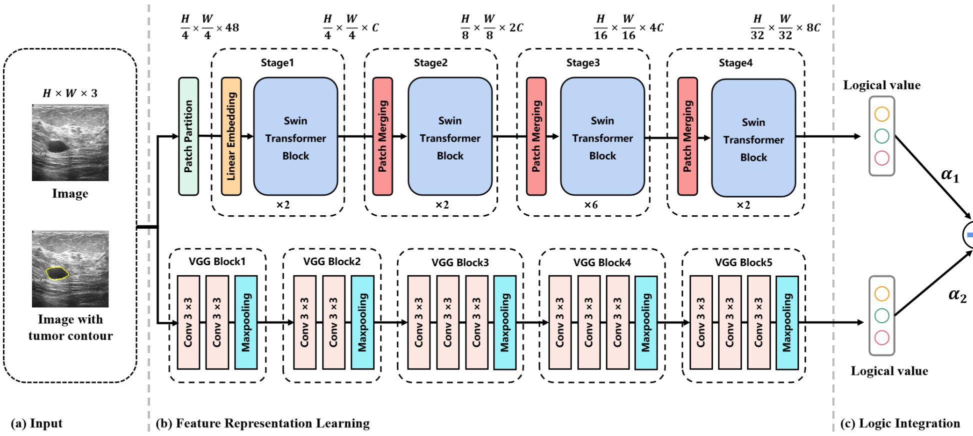

Figure 2. The overall framework of the proposed method. It consists of three modules: image partition module, feature representation learning module and tumor classifier.

3.1. Improved Swin-Transformer

3.1.1. Patches Partition

Firstly, in contrast to the earlier feature extraction technique which only uses images, we concatenate the tumor mask with the original images as input and send them to the network. This design gives the model a head start in its performance in identifying benign and malignant breast tumors by providing the model with prior knowledge about the tumor's shape, size, and location. Then, the patch partition will generate several patches with patch resolution from breast ultrasound images and matching masks. Each patch is considered a "token" characterized by the channel concatenation of image gray values and corresponding tumor mask. In this module, the patch partition divides the input image with a size of 224 × 224 and the tumor mask into non-overlapping patches with a size of 4 × 4. Each patch has a feature dimension of 4 × 4 × 3. After that, a linear embedding layer is used to map patch tokens to C dimensions, where $C = 96$.

3.1.2. Multi-Scale Feature Extraction

In the feature extraction module, we divide the module into four stages, from coarse to fine, inspired by the structure of Resnet-50, to construct multi-scale features. Specifically, each stage consists of an even number of Swin Transformer blocks to control the number of tokens. The Swin Transformer blocks alternately contain a window multi-head self-attention (W-MSA) layer and a shifted window multi-head self-attention (SW-MSA) layer. Figure 3 illustrates the structure of two successive Swin Transformer blocks. Notably, the WS-MSA layer in the second Swin Transformer block is designed to introduce a cross-window connection while maintaining efficient computation of non-overlapping windows.

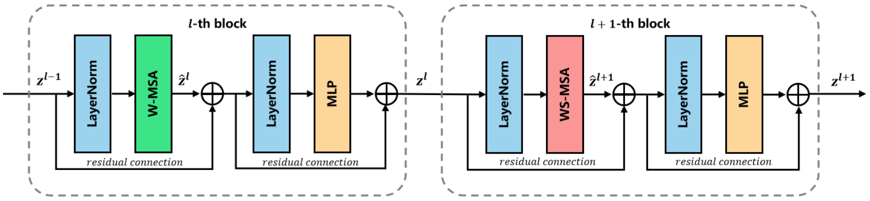

Figure 3. An illustration of two successive Swin Transformer blocks. The l-th block has a common window multi-head self attention layer (green rectangle) and the $l+1$-th block has a window-shifted multi-head self attention layer (red rectangle).

Swin Transformer Block. We first use a normalization layer on the output feature maps of $l-1$-th block to scale the feature distribution. Then, the feature maps will be divided into multiple patches through a window partitioning strategy and fed into the W-MSA or WS-MSA layer for calculating local attention. In this way, the computation of attention is restricted in each window, which reduces the computational cost. After that, each patch is sequentially restored to the shape of z_{l+1} by patch merging to obtain the hidden feature $\hat{z}_l$. Furthermore, the classic residual skip connection is used to concatenate the features $\hat{z}_l$ with z_{l+1}. Finally, the output of block l will be obtained by a similar structure which consists of a normalization layer, multiple layer perception (MLP), and residual concatenation. The following formula shows the calculation process of two successive Swin Transformer:

$$\begin{aligned}
\hat{z}^l &= \text{W-MSA}(LN(z^{l-1})) + z^{l-1} \\
z^l &= MLP(LN(\hat{z}^l)) + \hat{z}^l, \\
\hat{z}^{l+1} &= \text{SW-MSA}(LN(z^l)) + z^l, \\
z^{l+1} &= MLP(LN(\hat{z}^{l+1})) + \hat{z}^{l+1}
\end{aligned} \tag{1}$$

where z^l represents the output features of l-th block, $\hat{z}^l$ and $\hat{z}^{l+1}$ are the output of W-MSA module and WS-MSA module, respectively. The WS-MSA layers introduce adjacent non-overlapping windows in the previous layer, which plays an important role in the connection of windows, thus helping capture global information and local window information.

Window Shifted Multi-head Self Attention. As demonstrated in Figure 4, in block l, we employ the common window partition to compute self-attention in each window. The feature map with 8×8 is uniformly divided into four windows of size 4×4 ($M = 4$). In block $l + 1$, a shifted window partition is adopted, which cyclic-shifts the window toward the top-left direction by 2 pixels. If we only simply refine the window partition, this will lead to the different sizes of newly generated windows, which means some windows will be smaller than 4×4. Unfortunately, in practice, only windows of the same size are suitable for batch processing. One common way for calculating self-attention is to fill all windows into 4×4. However, as can be seen in Figure 4, the number of windows increases from 2×2 to 3×3, resulting in rapid growth of the calculation cost. Therefore, to improve the efficiency of batch computation, we shift new windows cyclically and merge them to reduce the number of windows used for computing self-attention. The state of windows after shifting used to calculate the self-attention is shown in Figure 5. It is worth noting that the merged window may consist of several sub-windows that are not adjacent to each other in the feature map. Therefore, a masking method is applied to limit the calculation of self-attention to each sub-window.

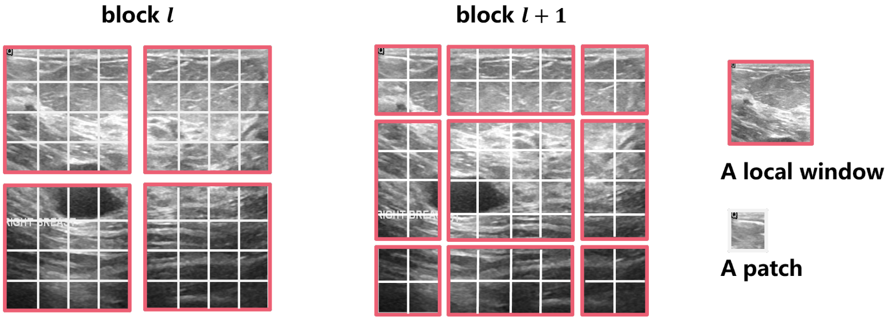

Figure 4. An illustration of two window partitioning strategies. In block l, we employ a common window partition to compute self-attention in each window. In block $l + 1$, a shifted window partition is adopted, which produces new Windows that cross the boundaries of previous windows l and thus provides the windows' connection.

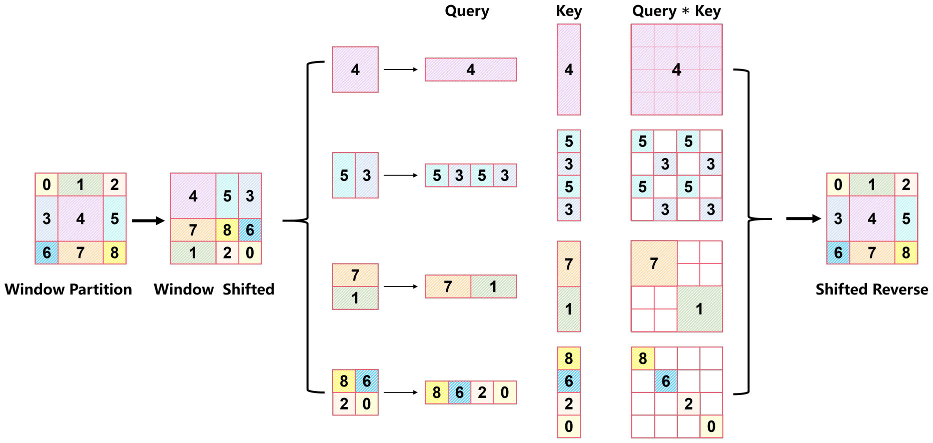

Figure 5. The masking method is used to limit the computation of self-attention to sub-windows. The colored region of the "Query ∗ Key" matrix takes part in the computation of self-attention, whereas the white region does not.

Figure 5 shows the process of the masking method in block $l + 1$, where "Query" and "Key" are the Query vector and Key vector in self-attention. To balance the computational overhead, the sub-windows are reconstructed and merged into 2×2 windows according to the standard window partitioning method of block l. Considering the window calculated in a batch may consist of several windows that are not adjacent to each other in the feature map. Therefore, we generate an attention mask that allows the Query and Key vectors with the same index to perform attention calculation while ignoring the results of attention evaluations with different indexes. In Figure 5, the white squares do not participate in the calculation of self-attention. Thanks to the use of circular shifting, the number of batch windows is the same as the number of regular window partitions, which improves computational efficiency.

Computational overhead. In ViT, the standard MSA is used for global attention to compute the dependencies between patches. The computational complexity of an image with $h \times w$ patches is as follows:

$$\Omega(MSA) = 4hwC^2 + 2(hw)^2C \tag{2}$$

where h and w are the height and width of patches. C represents the feature dimension and is a constant. Its computational complexity is the quadratic of patch size, making it unsuitable for high-resolution images.

While for W-MSA, self-attention is computed in the local window. Here, the window refers to a set of patches that uniformly and non-overlappingly segment the entire image. The following is the computational complexity of the W-MSA:

$$\Omega(W\text{-}MSA) = 4hwC^2 + 2M^2hwC \tag{3}$$

where M^2 represents the number of patches in each window, when the window size ($M \times M$) is fixed, the complexity of the W-MSA is linear. It can be easily seen from Equations (2) and (3) that global self-attention computation is typically prohibitively expensive for big $h \times w$, while window-based self-attention is scalable.

Multi-head Self-Attention in Vision Transformer. This work introduces the multi-head self-attention structure in Transformer into ViT. In particular, we add relative position bias into the similarity calculation of each head, which can be expressed as follows:

$$Attention(Q, K, V) = Softmax(\frac{QK^T}{\sqrt{d_k}} + B)V, \tag{4}$$

where $Q, K, V \in R^{n \times d}$ are the *query*, *key* and *value* metrics, and $B \in R^{n \times n}$ is the relative position parameter introducing the position embedding. n is the number of patches in a window and d_k is the dimension of query or key.

3.2. Interpretation Method

The deep learning model provides a large number of parameters to fit the distribution of samples and achieves great detection accuracy. However, deep learning is mostly a black-box model, and it is hard to understand the reasons for model decisions. Therefore, in this paper, we employ Score-CAM, an improved CAM-based approach, to provide a visual interpretation of the predictions. Score-CAM evaluates the contribution of each feature by measuring the degree of increase in confidence, which gets rid of the gradient dependency in the standard CAM method. We will describe this algorithm below.

Score CAM. Generally, the mapping between input and output of the trained Swin-Transformer can be defined as a function $Y = F(x)$, which takes an input vector $X = [x_0, x_1, \cdots, x_n]^\top$ and outputs the prediction scalar Y. For a given input baseline X_b, the contribution score C_i of x_i for the label category is the change in confidence score by replacing x_i with the i-th output in X_b. Therefore, the definition of the contribution score is formulated as:

$$C_i(X_i) = f(X_b \circ H_i) - f(X_b). \tag{5}$$

where H_i is a vector, which has the same shape as X_b. For each entry h_j in H_i, $h_j = \mathbb{I}[i = j]$, i.e., $h_j = 1$ when $i = j$, and $h_j = 0$ otherwise. $\circ$ represents Hadamard Product.

Similarly, specific to a certain layer of the network, we define the trained Swin Transformer as $Y = f(X)$ and outputs a class probability scalar Y. We pick the second normalization layer in the last Swin Transformer block and corresponding activation as A and denote the k-th channel as A_k. Therefore, the contribution score A_k towards Y can be defined as:

$$C(A_k) = f(X \circ H_k) - f(X_b) \tag{6}$$

where

$$H_k = s(Up(A_k)) \tag{7}$$

$Up(\cdot)$ represents the upsample operation which expands A_k into the input size. In this way, each upsampled activation map not only provides the spatial locations that are most relevant to an internal activation map but also can directly work as a mask to perturb the input image. $s(\cdot)$ is a normalization function that maps each element in the activation map matrix into $[0,1]$ to a generate smoother mask H_k. The normalization function $s(\cdot)$ is represented as follows:

$$s(A_k) = \frac{A_k - minA_k}{maxA_k - minA_k}, \tag{8}$$

then, the final visualization is obtained by a linear combination of weights and activation mappings. In addition, ReLU is also applied to the linear combination of mappings, since we are only interested in those features that have a positive impact on the category. Finally, we show the visualization in the form of a heatmap and apply it to the input image to explain the decision process.

3.3. Loss Function

The loss function of the breast ultrasound images classification is shown in Equation (9). The corresponding formula for this loss is shown as follows:

$$CE\left(Y, \widehat{Y}\right) = \sum_{i=1}^{n} y_i log \widehat{y_i}, \tag{9}$$

where Y is the ground truth, and $\widehat{Y}$ is the prediction.

4. Experiments

4.1. Dataset

In this paper, we evaluate the proposed method using the breast dataset BUSI, which collects breast ultrasound images from 600 patients between 25 and 75 years old, covering 780 images with an average resolution of 500 × 500 pixels. According to the relevant domain knowledge, BUSI datasets are divided into three categories: normal, benign and malignant. Furthermore, the dataset also provides expert manual annotation of tumor regions in benign and malignant images. All images and ground truth are saved in PNG format. During the experiment, we used 80% of the dataset as a training set, 10% as a validation set, and the rest as a test set. Table 1 shows the specific information of this dataset.

Table 1. The number of samples of raw and augmented data in training, validation and test sets.

Data State	Classes	Train	Validation	Test
Raw	Bengin	351	168	107
	Malignant	43	21	13
	Normal	43	21	13
Augmentated	Bengin	500	500	500
	Malignant	100	100	100
	Normal	43	21	13

4.2. Data Preprocessing and Data Enhancement

For deep learning models, the generalizability of the model is closely related to the number of training samples. However, medical images need to be interpreted by professional doctors, which are difficult to obtain. In this paper, we employ data augmentation for expanding the dataset before training to reduce overfitting and improve the performance of generalizability. At the beginning, we applied median filtering to reduce speckle noise. This step was followed by contrast enhancement using histogram equalization to improve the visibility of structures within the images. Then, the data augmentation methods adopted in the training and validation sets include random rotation, horizontal flipping, image scaling and so on. After that, the training set is increased to 1500 images while the validation set is enlarged to 300 images. Then, we reshape the enhanced image to 224 × 224 × 3 and normalize it to [0, 1]. Finally, all images will be converted to tensors and trained on the GPU.

4.3. Implement Details

In this work, we implement the proposed method in Python using Pytorch. To avoid overfitting, we use the ReLU function as an activation function between hidden layers, which can increase the nonlinearity between networks. To speed up the model convergence, we leverage the models that have been trained on ImageNet as pre-trained models. We set an initial learning rate of the VGG16 to 1×10^{-3} and the Swin-Transformer to 1×10^{-6}, as well as set the same batch size of 16. The parameters of the network model are optimized by Adam optimizer with the weight decay to 1×10^{-4}. We trained the model on the Ubuntu 20.04 system using the Nvidia 1080Ti GPU to speed up the training process and converged in 300 epochs. The model which achieves the best performance on the validation set is used for the final test.

4.4. Evaluation Metrics

In this paper, we adopt accuracy, recall, precision, F1-score and AUC (Area Under Curve) to comprehensively evaluate the performance of the proposed model. The calculation formula of the evaluation metrics is as follows:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{10}$$

$$Recall = \frac{TP}{TP + FN} \tag{11}$$

$$Precision = \frac{TP}{TP + FP} \tag{12}$$

$$F1\text{-}score = \frac{2Precision \times Recall}{Precision + Recall} \tag{13}$$

where TP (True positive) indicates the number of positive samples predicted as positive, FN (False negative) denotes the number of positive samples predicted as negative, FP (False positive) records the number of negative samples predicted as positive, and TN (True negative) represents the probability of negative samples predicted as negative. Accuracy reflects the ability of the classifier to predict all samples correctly. Recall reports the proportion of positive samples in the dataset that are predicted correctly. Precision is the proportion of samples predicted as positive by the classifier that are actually positive. When the classification confidence is high, precision is high, while the classification confidence is low, recall will be increased. Therefore, to comprehensively consider these metrics, F1-score is proposed. The F1 score is a summed average of the precision and recall, which maximizes precision and minimizes recall while minimizing the disparity between them. The greater the value for these evaluation metrics, the better the model's performance.

5. Results

5.1. Ablation Experiments

In this subsection, we perform ablation analysis on data augment, data balancing, introducing boundary contours of tumors, and integrating CNN and Swin-Transformer learning, respectively, to explore the impact of these methods on model classification performance. Table 2 shows the prediction results of the model under various conditions.

Specifically, experiments are first conducted on the BUSI dataset using VGG16, the model's weighted average accuracy, recall, precision, F1 score and AUC are 0.8803, 0.8808, 0.8803, 0.8805 and 0.9411, respectively. Afterward, we adopt data augment and balance on the data and use the tumor region as prior knowledge to facilitate the model's perception of this region. In this way, the model achieves an accuracy of 0.963, an F1 score of 0.9611, and an AUC of 0.9937, indicating that effective data augmentation and prior knowledge are helpful in improving model performance. Similarly, the data set after data augmentation is also used to train the Swin-Transformer model, and the obtained evaluation metrics are slightly higher than VGG16, with weighted average accuracy, precision, recall, F1 score, and AUC of 0.9740, 0.9755, 0.9740, 0.9741 and 0.9892, respectively. Meanwhile, without introducing prior knowledge of tumor margins, we further evaluate the performance of the model (Ensemble *) when adopting the ensemble learning method. After integrating CNN and Swin-Transformer, Ensemble * achieves comparable performance to Swin-T. Furthermore, based on the Ensemble *, we combine the tumor's prior knowledge to train the model, which improves the accuracy, precision, recall, and F1 score by nearly 1.3%, respectively, 0.9870, 0.9880, 0.9870, 0.9872, and achieve an AUC of 0.9982. The ablation experiments show that data augmentation, tumor prior knowledge, and integrated learning strategies all help to enhance the performance of the model.

Table 2. The ablation analysis results on breast ultrasound images classification, containing data augment and balance, prior knowledge of tumor boundaries and ensemble learning. $^+$ represents the data augmentation of the original dataset and the introduction of prior knowledge of tumor boundaries. The ensemble * represents the use of augmented data for ensemble learning, and no tumor margin prior is introduced. Our model uses data augmentation and ensemble learning for VGG16 and Swin-T combined with tumor margin priors.

Method	Class	Accuracy	Precision	Recall	F1-Score	AUC
VGG16	Bengin	0.9302	0.9070	0.8966	0.9017	0.9316
	Malignant	0.7619	0.8372	0.8571	0.8471	0.9447
	Normal	0.9231	0.8462	0.8462	0.8462	0.9675
	Average	0.8803	0.8808	0.8803	0.8805	0.9411
VGG16 $^+$	Bengin	0.9767	0.9546	0.9767	0.9655	0.9891
	Malignant	1.0000	1.0000	1.0000	0.9756	0.9966
	Normal	0.9231	0.9231	0.8462	0.9231	0.9964
	Average	0.9610	0.9616	0.9610	0.9611	0.9937
Swin-T $^+$	Bengin	1.0000	0.9555	1.0000	0.9773	0.9932
	Malignant	1.0000	1.0000	1.0000	1.0000	1.0000
	Normal	0.8462	1.0000	0.8462	0.9167	0.9964
	Average	0.9740	0.9783	0.9765	0.9923	0.9978
Ensemble *	Bengin	0.9767	1.0000	0.9535	0.9762	0.9850
	Malignant	1.0000	0.9546	1.0000	0.9767	0.9940
	Normal	0.9231	0.9286	1.0000	0.9630	0.9964
	Average	0.9740	0.9755	0.9740	0.9741	0.9892
Ours	Bengin	0.9767	0.9773	1.0000	0.9760	0.9973
	Malignant	1.0000	1.0000	1.0000	0.9950	1.0000
	Normal	1.0000	1.0000	0.9880	0.9990	0.9976
	Average	0.9870	0.9880	0.9870	0.9872	0.9982

5.2. Compare with Other Methods

In this subsection, we compare the performance of the proposed method with other models on the dataset BUSI, and the results are shown in Table 3. Specifically, we compare the proposed method with CNN models such as VGG16, ResNet50, ResNet101 and other advanced methods like [38,39]. Compared with CNN models, our proposed model has clear advantages in stability and robustness, with slight differences in multiple evaluation metrics. Furthermore, we also compare with Transformer-based models including Vit, Swin-T, [5,40]. Among them, [40] is a weakly supervised Transformer model. When compared to Transformer models, our model combines tumor boundary prior knowledge and ensemble learning, so it has a better recognition rate. By comparing multiple models of the two classes of deep learning methods (i.e., CNN and Transformer) in Table 2, our model achieves optimal results on various metrics, demonstrating the effectiveness of the proposed model.

Table 3. Comparison of breast ultrasound images classification performance of different models on dataset BUSI.

Model Type	Model	Accuracy	Precision	Recall	F1-Score	AUC
CNN	VGG16	0.9610	0.9616	0.9610	0.9611	0.9937
	ResNet50	0.9481	0.9525	0.9481	0.9475	0.9844
	ResNet101	0.9495	0.9429	0.9523	0.9476	0.9862
	[38]	0.9162	0.9318	0.9148	0.9666	0.9678
	[41]	0.9280	-	-	-	0.9869
	[42]	0.9412	0.9613	0.8993	92.93	-
	[43]	0.9000	-	-	0.9000	-
	[44]	0.8996	0.8933	0.9997	-	-
	[39]	0.9319	0.9318	0.8875	-	-
Transformer	ViT	0.9345	0.9350	0.9345	0.9243	0.9960
	Swin-T	0.9740	0.9783	0.9765	0.9923	0.9978
	[40]	0.9529	0.9629	0.9601	0.9615	-
	[5]	0.8670	-	-	-	0.9500
CNN + Transformer	Ours	0.9870	0.9880	0.9870	0.9872	0.9982

5.3. Visualization

Although deep learning-based algorithms have achieved great success in many fields, most of them are black-box models, which make it difficult to understand the decision-making of the model intuitively. Therefore, in this section, we further explore the internal decision-making of the model by adopting a visual explanation method named Score-CAM. Specifically, we show a heatmap on the image to emphasize the key attention regions of the model. The redder the color, the higher the contribution score assigned to the corresponding pixel, i.e., the point significantly increases the confidence score.

Figure 6 shows the visual interpretation results of using Swin-Transformer or VGG as feature extractors after ensemble learning. For each category, we give three examples. To facilitate comparison with the visual interpretation of the results, we show the original images and draw the tumor boundaries on the benign and malignant images. In the first three columns of the figure, we show the samples with high similarity in Score-CAM generated by Swin-transformer and VGG16. Columns 4–9 show the instances with differences in the Score-CAM generated by the two models. As can be seen from these images, it is effective to improve the accuracy of predictions by integrating two models. We discover that for both benign and malignant breast images, the region of the tumor plays a key role in increasing the prediction confidence. In other words, if the model reflects the tumor region accurately, it can differentiate between benign and malignant samples. In addition, we also find that both the Swin-Transformer and VGG16 models focus on regions of normal samples with rich textural features. These visualizations help researchers to visualize the characteristics of model selection and understand model decisions.

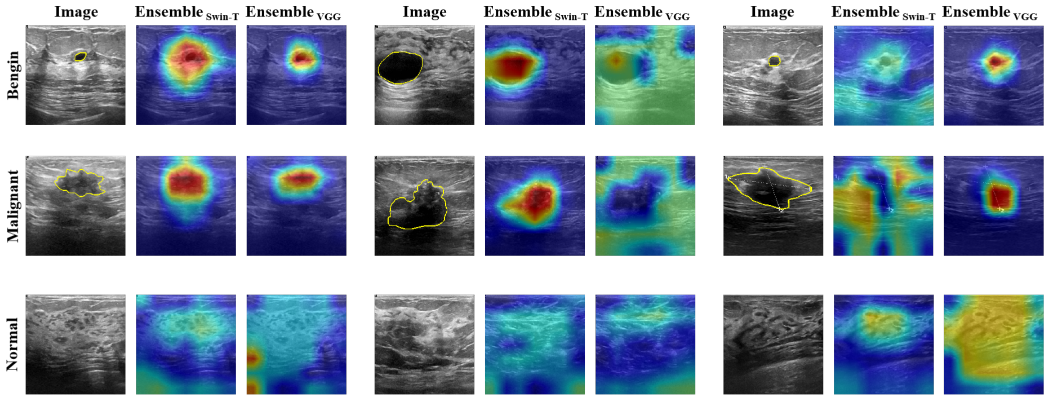

Figure 6. Visual interpretation results generated by Score-CAM. Ensemble$_{\text{Swin-T}}$ and Ensemble$_{\text{VGG}}$ represent the visual results generated by Swin-Transformer and VGG16 after training with logic integration, respectively.

6. Conclusions

Breast cancer is one of the life-threatening cancers for women. In clinical practice, breast cancer classification models face the challenges of poor interpretability and low accuracy. In this paper, we propose an interpretable breast image recognition model that classifies breast ultrasound images into normal, benign, and malignant. Firstly, the method introduces manually annotated tumor contours into the model as a priori knowledge, which strengthens the model's ability to determine the location and extent of the tumor, thereby guiding the classification. Furthermore, we integrate and train two deep learning models, i.e., the CNN-based model and the Transformer-based model, to achieve information balance and complementarity between the two neural network models, which solves the problems of few training samples and limits feature extraction capabilities. Finally, we validate the proposed method on the BUSI dataset. Experiments show that our model outperforms the current mainstream CNN model and Transformer model in several classification metrics, reaching the state of the art. In future work, we will consider the fusion of features extracted by CNN and Transformer, as well as combine clinical diagnosis reports to further improve the model's accuracy and robustness.

Author Contributions: Conceptualization, X.M. and J.M.; methodology, X.M.; software, F.L. and T.Z.; validation, X.M. and Z.C.; formal analysis, J.M.; writing—original draft preparation, J.M.; writing—review and editing, X.M.; funding acquisition, X.M. All authors have read and agreed to the published version of the manuscript.

Funding: X.M. acknowledges support from the Youth Innovation Science and Technology Support Program of Universities in Shandong Province (Grant No.2021KJ082), the National Natural Science Foundation of China (Grant No.11647013) and the Shandong Provincial Natural Science Foundation of China (Grant No.ZR2014AL015).

Data Availability Statement: The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding author.

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

Abbreviations

The following abbreviations are used in this manuscript:

CAD	Computer-aided diagnosis
CNN	Convolutional neural network
ViT	Vision Transformer
CAM	Class activation mapping
Grad-CAM	Gradient weighted class activation mapping
NLP	Natural language processing

W-MSA	Window multi-head self attention
SW-MSA	Shifted window multi-head self attention
MLP	Multiple layer perception
AUC	Area Under Curve

References

1. Giaquinto, A.N.; Miller, K.D.; Tossas, K.Y.; Winn, R.A.; Jemal, A.; Siegel, R.L. Cancer statistics for African American/Black People 2022. *CA A Cancer J. Clin.* **2022**, *72*, 202–229. [CrossRef] [PubMed]
2. Fujioka, T.; Kubota, K.; Mori, M.; Kikuchi, Y.; Katsuta, L.; Kimura, M.; Yamaga, E.; Adachi, M.; Oda, G.; Nakagawa, T.; et al. Efficient Anomaly Detection with Generative Adversarial Network for Breast Ultrasound Imaging. *Diagnostics* **2020**, *10*, 456. [CrossRef]
3. 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* **2020**, *39*, 866–876. [CrossRef] [PubMed]
4. Wang, Y.; Choi, E.J.; Choi, Y.; Zhang, H.; Jin, G.Y.; Ko, S.B. Breast Cancer Classification in Automated Breast Ultrasound Using Multiview Convolutional Neural Network with Transfer Learning. *Ultrasound Med. Biol.* **2020**, *46*, 1119–1132. [CrossRef] [PubMed]
5. Yap, M.H.; Pons, G.; Martí, J.; Ganau, S.; Sentís, M.; Zwiggelaar, R.; Davison, A.K.; Martí, R. Automated Breast Ultrasound Lesions Detection Using Convolutional Neural Networks. *IEEE J. Biomed. Health Inf.* **2018**, *22*, 1218–1226. [CrossRef]
6. Wang, K.; Liang, S.; Zhong, S.; Feng, Q.; Ning, Z.; Zhang, Y. Breast Ultrasound Image Segmentation: A Coarse-to-Fine Fusion Convolutional Neural Network. *Med. Phys.* **2021**, *48*, 4262–4278. [CrossRef] [PubMed]
7. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv* **2021**, arXiv:2010.11929.
8. Wu, J.; Luo, T.; Zeng, J.; Gou, F. Continuous Refinement-based Digital Pathology Image Assistance Scheme in Medical Decision-Making Systems. *IEEE J. Biomed. Health Inf.* **2024**, *28*, 2091–2102. [CrossRef]
9. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, BC, Canada, 11–17 October 2021; pp. 9992–10002.
10. Simonyan, K.; Vedaldi, A.; Zisserman, A. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv* **2013**, arXiv:1312.6034.
11. Selvaraju, R.R.; Das, A.; Vedantam, R.; Cogswell, M.; Parikh, D.; Batra, D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *Int. J. Comput. Vis.* **2019**, *128*, 336–359. [CrossRef]
12. Anwar, S.M.; Majid, M.; Qayyum, A.; Awais, M.; Alnowami, M.; Khan, M.K. Medical image analysis using convolutional neural networks: A review. *J. Med. Syst.* **2018**, *42*, 1–13. [CrossRef] [PubMed]
13. Hu, Z.; Tang, J.; Wang, Z.; Zhang, K.; Zhang, L.; Sun, Q. Deep learning for image-based cancer detection and diagnosis—A survey. *Pattern Recognit.* **2018**, *83*, 134–149. [CrossRef]
14. Huang, Z.; Ling, Z.; Gou, F.; Wu, J. Medical Assisted-segmentation System based on Global Feature and Stepwise Feature Integration for Feature Loss Problem. *Biomed. Signal Process. Control* **2024**, *89*, 105814. [CrossRef]
15. Zhou, Z.; Xie, P.; Dai, Z.; Wu, J. Self-supervised Tumor Segmentation and Prognosis Prediction in Osteosarcoma Using Multiparametric MRI and Clinical Characteristics. *Comput. Methods Programs Biomed.* **2024**, *244*, 107974. [CrossRef]
16. Huang, Q.; Huang, Y.; Luo, Y.; Yuan, F.; Li, X. Segmentation of breast ultrasound image with semantic classification of superpixels. *Med. Image Anal.* **2020**, *61*, 101657. [CrossRef] [PubMed]
17. Vakanski, A.; Xian, M.; Freer, P.E. Attention Enriched Deep Learning Model for Breast Tumor Segmentation in Ultrasound Images. *Ultrasound Med. Biol.* **2020**, *46*, 2819–2833. [CrossRef]
18. Cao, Z.; Yang, G.; Chen, Q.; Chen, X.; Lv, F. Breast tumor classification through learning from noisy labeled ultrasound images. *Med. Phys.* **2019**, *47*, 1048–1057. [CrossRef]
19. Huang, Y.; Han, L.; Dou, H.; Luo, H.; Yuan, Z.; Liu, Q.; Zhang, J.; Yin, G. Two-stage CNNs for computerized BI-RADS categorization in breast ultrasound images. *BioMed. Eng. OnLine* **2019**, *18*, 8. [CrossRef] [PubMed]
20. Nawaz, M.A.; Sewissy, A.A.; Soliman, T.H.A. Multi-Class Breast Cancer Classification using Deep Learning Convolutional Neural Network. *Int. J. Adv. Comput. Sci. Appl.* **2018**, *9*, 316–332. [CrossRef]
21. Chopra, J.; Kumar, A.; Aggarwal, A.K.; Marwaha, A. Biometric System Security Issues and Challenges. *Second Int. Conf. Innov. Trends Electron. Eng.* **2016**, *20*, 83–87.
22. Yap, M.H.; Goyal, M.; Osman, F.; Ahmad, E.; Martí, R.; Denton, E.R.E.; Juette, A.; Zwiggelaar, R. End-to-end breast ultrasound lesions recognition with a deep learning approach. In Proceedings of the Medical Imaging 2018: Biomedical Applications in Molecular, Structural, and Functional Imaging, Houston, TX, USA, 11–13 February 2018.
23. Shin, S.Y.; Lee, S.; Yun, I.D.; Kim, S.M.; Lee, K.M. Joint Weakly and Semi-Supervised Deep Learning for Localization and Classification of Masses in Breast Ultrasound Images. *IEEE Trans. Med. Imaging* **2019**, *38*, 762–774. [CrossRef] [PubMed]
24. Mo, W.; Zhu, Y.; Wang, C. A Method for Localization and Classification of Breast Ultrasound Tumors. *Adv. Swarm Intell.* **2020**, *12145*, 564–574.

25. Wu, J.; Dai, T.; Guan, P.; Liu, S.; Gou, F.; Taherkordi, A.; Li, Y.; Li, T. FedAPT: Joint Adaptive Parameter Freezing and Resource Allocation for Communication-Efficient. *IEEE Internet Things J.* **2024**, *11*, 1–12. [CrossRef]
26. Tanaka, H.; Chiu, S.W.; Watanabe, T.; Kaoku, S.; Yamaguchi, T. Computer-aided diagnosis system for breast ultrasound images using deep learning. *Ultrasound Med. Biol.* **2019**, *64*, 235013.
27. Han, K.; Wang, Y.; Chen, H.; Chen, X.; Guo, J.; Liu, Z.; Tang, Y.; Xiao, A.; Xu, C.; Xu, Y.; et al. A Survey on Vision Transformer. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 87–110. [CrossRef] [PubMed]
28. Li, L.; He, K.; Zhu, X.; Gou, F.; Wu, J. A pathology image segmentation framework based on deblurring and region proxy in medical decision-making system. *Biomed. Signal Process. Control* **2024**, *95*, 106439. [CrossRef]
29. Wu, J.; Yuan, T.; Zeng, J.; Gou, F. A Medically Assisted Model for Precise Segmentation of Osteosarcoma Nuclei on Pathological Images. *IEEE J. Biomed. Health Inf.* **2024**, *27*, 3982–3993. [CrossRef] [PubMed]
30. Yuan, L.; Chen, Y.; Wang, T.; Yu, W.; Shi, Y.; Tay, F.E.H.; Feng, J.; Yan, S. Tokens-to-Token ViT: Training Vision Transformers from Scratch on ImageNet. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, BC, Canada, 11–17 October 2021; pp. 538–547.
31. Jaeger, P.F.; Kohl, S.A.A.; Bickelhaupt, S.; Isensee, F.; Kuder, T.; Schlemmer, H.; Maier-Hein, K. Retina U-Net: Embarrassingly Simple Exploitation of Segmentation Supervision for Medical Object Detection. *arXiv* **2019**, arXiv:1811.08661.
32. Li, X.; Chen, H.; Qi, X.; Dou, Q.; Fu, C.W.; Heng, P.A. H-DenseUNet: Hybrid Densely Connected UNet for Liver and Tumor Segmentation From CT Volumes. *IEEE Trans. Med. Imaging* **2018**, *37*, 2663–2674. [CrossRef]
33. Coudray, N.; Ocampo, P.; Sakellaropoulos, T.; Narula, N.; Snuderl, M.; Fenyö, D.; Moreira, A.; Razavian, N.; Tsirigos, A. Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. *Nat. Med.* **2018**, *24*, 1559–1567. [CrossRef]
34. Aggarwal, A.K. A Review on Genomics Data Analysis using Machine Learning. *Wseas Trans. Biol. Biomed.* **2023**, *20*, 119–131. [CrossRef]
35. Springenberg, J.T.; Dosovitskiy, A.; Brox, T.; Riedmiller, M.A. Striving for Simplicity: The All Convolutional Net. *arXiv* **2015**, arXiv:1412.6806.
36. Zeiler, M.D.; Fergus, R. Visualizing and Understanding Convolutional Networks. In Proceedings of the Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, 6–12 September 2014.
37. Samek, W.; Binder, A.; Montavon, G.; Lapuschkin, S.; Müller, K.R. Evaluating the Visualization of What a Deep Neural Network Has Learned. *IEEE Trans. Neural Netw. Learn. Syst.* **2017**, *28*, 2660–2673. [CrossRef]
38. 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] [PubMed]
39. Roy, K.; Bhattacharjee, D.; Kollmann, C. BUS-Net: A Fusion-based Lesion Segmentation Model for Breast Ultrasound (BUS) Images. *Lect. Notes Netw. Syst.* **2023**, *404*, 313–321.
40. Wang, W.; Jiang, R.; Cui, N.; Li, Q.; Yuan, F.; Xiao, Z. Semi-supervised vision transformer with adaptive token sampling for breast cancer classification. *Front. Pharmacol.* **2022**, *13*, 929755. [CrossRef]
41. Lazo, J.F.; Moccia, S.; Frontoni, E.; De Momi, E. Comparison of different CNNs for breast tumor classification from ultrasound images. *arXiv* **2013**, arXiv:2012.1451.
42. Zhang, G.; Zhao, K.; Hong, Y.; Qiu, X.; Zhang, K.; Wei, B. SHA-MTL: Soft and hard attention multi-task learning for automated breast cancer ultrasound image segmentation and classification. *Int. J. Comput. Assist. Radiol. Surg.* **2021**, *16*, 1719–1725. [CrossRef] [PubMed]
43. Yang, T.; Yu, X.; Ma, N.; Zhang, Y.; Li, H. Deep representation-based transfer learning for deep neural networks. *Knowl.-Based Syst.* **2022**, *253*, 109526. [CrossRef]
44. Podda, A.S.; Balia, R.; Barra, S.; Carta, S.M.; Fenu, G.; Piano, L.C. Fully-Automated Deep Learning Pipeline for Segmentation and Classification of Breast Ultrasound Images. *J. Comput. Sci.* **2022**, *63*, 101816. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.