

Explainable Hybrid Multi-Feature Fusion Model with Transformer for Medical Image Classification

Mumtaz Qabulio (Corresponding Author)

Department of Software Engineering, FET, University of Sindh, Jamshoro

Muhammad Suleman Memon

Department of Information Technology, Dadu Campus, University of Sindh

Nimra memon

Department of computer science Government Girls Degree College Nawabshah

Shabana

Department of Computer Science Govt. Aisha Girls Degree College Nawabshah

ABSTRACT

Breast Cancer disease detection is one of the important areas. Breast classification is an essential part in computer-aided diagnosis, which allows detecting and treating different diseases at the earliest stage. Nonetheless, traditional deep learning architectures usually use convolutional neural networks (CNNs) only, which might not be able to extract complementary information in handcrafted features. To overcome this drawback, this paper presents a proposed explainable hybrid multi-feature fusion framework, which combines deep and handcrafted features with a Transformer-based fusion mechanism. Deep features are computed with the help of a pre-trained ResNet50 architecture, whereas handcrafted features, such as texture and color descriptors, are calculated with the help of Gray-Level Co-occurrence Matrix (GLCM) and histogram-based approaches. These heterogeneous features are then mapped into a shared feature space and combined using a Transformer encoder to effectively capture inter-feature relationships.

The model suggested is tested on a multi-class breast cancer dataset and shows a better performance as compared to baseline CNN models. The accuracy, precision, recall, F1-score, and area under the curve (AUC) are used to conduct comprehensive experiments. Moreover, explainability is added with the help of Grad-CAM to visualize attentions of the model, and t-SNE analysis is used to illustrate that the fusion has enhanced the separability of features. The findings indicate that the suggested hybrid system has a substantial positive impact on the classification and interpretability of the data, which is why it is a promising solution to the analysis of real-life medical images.

Keywords— Breast Cancer Classification, Medical Image Classification, Deep Learning Transformer, Hybrid System

Introduction

The breast cancer disease has increased in recent years. The women with benign breast disease are at a risk of getting breast cancer [1].

Medical imaging classification[2] has become one of the fundamental aspects of a modern healthcare system, as it allows diagnosing diseases automatically, and helps clinicians in decision-making. As the field of deep learning has progressed swiftly, convolutional neural networks (CNNs) have proven to be highly successful in a range of different medical imaging applications, such as breast disease classification, breast lesion detection, and disease segmentation. ResNet50[3] and EfficientNet are examples of architecture that have shown a high ability to extract high-level semantic features of breast cancer images. Although effective, such models tend to use learned features only, which may miss out on valuable domain-specific information contained in traditional handcrafted descriptors.

Medical image analysis Handcrafted features, including texture and color information, have long been employed to define very subtle variations in tissue structure and disease patterns. Methods such as Gray-Level Co-occurrence Matrix (GLCM)[4] give statistical data of texture, whereas methods based on histograms characterize color distributions. In medical imaging, these characteristics come in especially handy because in many instances, texture anomalies and color differences are indicators of pathological conditions. Nevertheless, more conventional machine learning methods that use only handcrafted features usually do not have the representational capability of deep learning models.

Hybrid feature fusion has become a promising direction to take advantage of the strengths of both approaches. Integration of deep features, handcrafted features can help achieve a more detailed medical image representation. However, most fusion techniques, which are currently in use, are based on either simple concatenation or fixed weighting approaches, which are not adequate to reflect complex interactions between features of different types. In the recent past, Transformer-based architecture has demonstrated tremendous success in the modeling of long-range dependencies and contextual relationships in several fields. Their self-attention mechanism allows them to dynamically weight features, which makes them suitable to feature fusion tasks.

We introduce an explainable hybrid multi-feature fusion model (a combination of CNN-based deep features and handcrafted features) in this paper, which employs a Transformer encoder. The proposed framework initially derives deep features of breast diseases images with the help of a pre-trained CNN backbone, and handcrafted features are calculated with the help of texture and color descriptors. These characteristics are then mapped into a single embedding space and merged by a Transformer-based attention module, enabling the model to learn inter-feature relationships that are complicated. Moreover, Grad-CAM is used to visualize significant parts to the decisions of classification to increase model interpretability.

The key contributions of this work are as follows:

- An integrated system that incorporates the deep learning characteristics with the handcrafted characteristics to enhance breast cancer image classification.
- A Transformer-based fusion model to be able to model heterogeneous features

Online ISSN

3007-3197

Print ISSN

3007-3189

<http://amresearchreview.com/index.php/Journal/about>

relationships.

- Grad-CAM explainability of model predictions in a visual way.
- Detailed experimental analysis and evaluation with various performance measures and visualizations, such as ROC curves, precision-recall curves, confusion matrices, and t-SNE analysis.

The rest of the paper is structured in the following way. Section II gives related work, Section III details the proposed methodology, Section IV gives the results of the experiment.

Related Work

The author in [5] proposed a hybrid deep learning model that combines convolutional neural networks with EfficientNetV2B3 to detect breast cancer with histopathology images. The model uses the concept of transfer learning and integrates various pre-trained architectures to enhance feature extraction and classification effectiveness. They were tested on large-scale whole slide image data, with a precision of 96.3% and high recall and AUC. The proposed hybrid model was also compared to other variants like MobileNet and DenseNet and shown to be of better performance. Also, data augmentation and preprocessing methods were used to overcome the issue of class imbalance and increase strength. Although effective, the approach is mainly based on CNN-based feature fusion, without sophisticated attention processes of dynamic interaction of features.

An Ensemble Deep Convolutional Neural Network (EDCNN) was proposed by [6]. to detect and classify breast cancer with the help of ultrasound images. The model integrates MobileNet and Xception models to improve feature extraction with a hybrid model. Also, the U-Net-based segmentation is used to isolate the regions of interest to enhance the accuracy of the diagnosis. Preprocessing techniques like normalization and data augmentation are used as a part of the study to increase the generalization of the models. The experimental findings indicate that the proposed model with an accuracy of 87.82 and an AUC of 0.91 are higher than a number of transfer learning and Vision Transformer models. Moreover, it uses Grad-CAM to give interpretability through highlighting the important areas that affect the predictions. In spite of its good performance, the method can be limited by a grayscale image dependency and possible overfitting because of the small datasets.

In the proposed work [7] the author proposed a hybrid deep learning (NSGA-II-DL) framework to classify HER2 in breast cancer combining transfer learning and metaheuristic feature selection. The model uses ResNet50 to extract features and then optimizes using NSGA-II to select few and best features and SVM to classify the final features. The method is highly effective in terms of dimensionality reduction of features and better performance with the highest accuracy being 94.4 percent. The paper points out the significance of multi-objective optimization in the management of complicated histopathological image structures.

The author in [8] suggested a computer-aided diagnosis system to classify breast tumor based on improved transfer learning and chaotic optimization. This paper uses EfficientNet-B4 to extract deep features, then serial mid-value-based feature fusion,

Online ISSN

3007-3197

Print ISSN

3007-3189

<http://amresearchreview.com/index.php/Journal/about>

and chaotic crow search optimization to choose the best features. The method is successful in enhancing the performance of the classification on the INbreast and CBIS-DDSM datasets with accuracies of up to 98.45%. The findings indicate the importance of integrating feature improvement and metaheuristic optimization in the development of robust breast cancer detection.

The author in [9] came up with an ensemble framework called attention-based ensemble framework (ECS-A-Net) to classify breast cancer based on mammographic images. The model incorporates adapted SE-ResNet50 and InceptionV3 to extract features, then channel and spatial attention modules to boost discriminative features learning. The algorithm makes use of data augmentation to overcome data constraints and enhance generalization. Experimental results on DDSM and MIAS datasets indicate better performance with accuracies of 96.50% and 95.33, respectively which is better than many of the state-of-the-art approaches.

The author in [10] suggested a completely automated CAD framework, which integrates YOLOv4 and Vision Transformers to detect and classify mass of breasts in CESM and FFDM images. The model uses YOLO to perform precise localization with transformers to gain access to global contextual features, which enhance classification performance. Their findings indicate high detection and classification rates, especially indicating the high sensitivity of CESM imaging in the diagnostic process. This paper demonstrates how the use of detection networks and transformer-based models can be combined to enhance breast cancer diagnosis.

The author in [11] proposed a deep learning-based computer-aided diagnosis (CAD) framework for breast cancer detection, integrating YOLO-based lesion detection, Associated-ResUNet for segmentation, and a customized BreastNet-SVM classifier. The model was evaluated on the CBIS-DDSM dataset and demonstrated outstanding performance, achieving 98.5% detection accuracy and 99.16% classification accuracy. The author in [12] proposed a hybrid deep learning framework which combines CNN, Bi-LSTM and EfficientNet-B0 for the detection and classification of breast cancer. The proposed study utilizes transfer learning for effective feature extraction. The study used temporal modeling using Bi-LSTM to identify complex patterns in mammographic images. The proposed methodology was evaluated on the CBIS-DDSM and MIAS datasets which achieved a high accuracy rate of 99.2% and exceeding various conventional architectures. The study demonstrates the effectiveness of integrating spatial and temporal features to improve diagnostic performance.

The author in [13] developed a method for classifying breast cancer in ultrasound images employing transfer learning and a CNN model. To enhance the performance of the model image preprocessing and dataset augmentation were used to addressing the issue of an unbalanced dataset. Their model achieved an accuracy of approximately 90% on the BUSI dataset which shows a good performance of transfer learning in medical image analysis. Nonetheless, the model encounters challenges related to its dependence on dataset size and quality, as well as its adaptability to various contexts.

The author in [14] used a new way to deal with class imbalance in breast cancer images. The DCGAN was used to create more images and then fine-tuning a model called ResNet50 to make it work better. This helped balance out the BreakHis dataset which is a collection of breast cancer images. The model made a good classification by

achieving an accuracy 97%. The DCGAN approached used in the study shows that it improved an accuracy and is helpful in medical images. At the same time proposed approach can be complicated and can take more computing power.

The author in [15] suggested a new way to classify breast cancer using a combination of advanced techniques. First the quality of the images was improved using special filters and made the differences between light and dark areas more clear. Then the clustering method was used to separate the images into different groups and picked the most important features using a tool called MCSO. To help make a better classification fuzzy logic was used. This approach worked really well on two big sets of data, called MIAS and DDSM and was very accurate. However it's a bit complicated and has many steps which could make it slower and more expensive to use.

Methodology

3.1 Overview of the Proposed Framework

In this study we used a Hybrid Multi-Feature Fusion Model using Transformer to classify medical images. The framework is a combination of deep features obtained with the help of convolutional neural network (CNN) and handcrafted features obtained with the help of texture and color descriptors. These diverse features are mapped to a single feature space and merged with a Transformer-based attention model and classified by a classification layer.

The overall pipeline consists of the following stages:

- i. Input preprocessing
- ii. Deep feature extraction
- iii. Handcrafted feature extraction
- iv. Feature projection
- v. Transformer-based feature fusion
- vi. Classification
- vii. Explainability and visualization

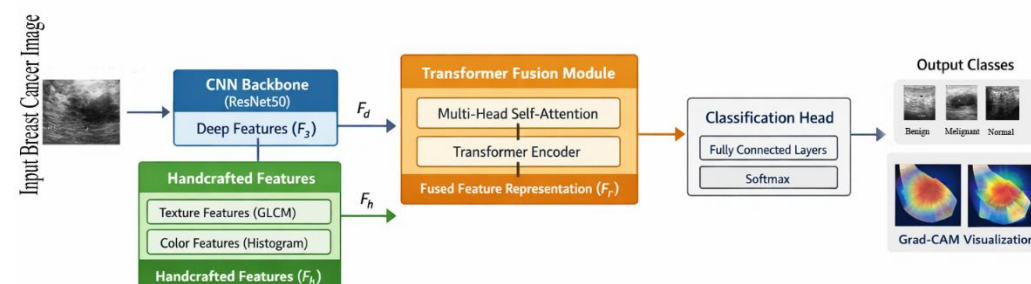


Figure. 1 Model Architecture

3.2 Input Preprocessing

Preprocessing was used to ensure consistent input size, improving generalization and reducing overfitting. Input preprocessing refers to the set of operations applied to raw images to standardize them before feature extraction. The Images are resized to 224×224 and Pixel values are normalized. The Data augmentation techniques (e.g.,

Online ISSN

3007-3197

Print ISSN

3007-3189

<http://amresearchreview.com/index.php/Journal/about>

rotation, flipping) are optionally applied.

3.3 Deep Feature Extraction

Deep feature extraction is the extraction of hierarchical feature representations of images with the help of a CNN model. A pre-trained ResNet50 is used as the backbone. The last classification layer is eliminated to get feature embeddings. This Captures high-level semantic information such as shapes, edges, and structures

3.4 Handcrafted Feature Extraction

Handcrafted feature extraction involves computing domain-specific descriptors using traditional image processing techniques.

3.4.1 Texture Features (GLCM)

GLCM (Gray-Level Co-occurrence Matrix) measures the spatial relationship between pixel intensities.

3.4.2 Color Features

Color features represent the distribution of pixel intensities across color channels.

3.4.3 Combined Handcrafted Features

Concatenation of texture and color features into a single vector.

3.5 Feature Projection

Feature projection maps heterogeneous feature vectors into a common embedding space.

3.6 Transformer-Based Feature Fusion

Feature fusion is the process of combining multiple feature representations into a unified representation. In this work, a Transformer encoder is used to model relationships between features.

3.6.1 Token Representation

Each feature vector is treated as a token in the Transformer.

3.6.2 Self-Attention Mechanism

Self-attention computes the importance of each feature relative to others.

3.6.3 Fusion Output

3.7 Classification Layer

The classification layer maps fused features to output class probabilities.

3.8 Loss Function

The loss function measures the difference between predicted and true labels.

3.9 Explainability (Grad-CAM)

Online ISSN

3007-3197

Print ISSN

3007-3189

<http://amresearchreview.com/index.php/Journal/about>

Grad-CAM is an explainability technique that highlights regions in the image contributing to the prediction.

3.10 Feature Visualization (t-SNE)

The t-SNE is a dimensionality reduction technique used to visualize high-dimensional features in 2D space.

3.11 Training Strategy

Training strategy defines how the model learns from data.

Table 1: Training Strategy of the Proposed Model

Parameter	Value / Description
Input Image Size	$224 \times 224 \times 3$
Batch Size	16
Number of Epochs	30
Optimizer	Adam
Learning Rate	0.0001
Loss Function	Categorical Cross-Entropy
Backbone Model	ResNet50 (Pre-trained)
Feature Fusion	Transformer Encoder
Activation Function	ReLU (hidden layers), Softmax (output layer)
Regularization	Dropout (0.5)
Data Augmentation	Rotation, Flipping, Zooming
Evaluation Metrics	Accuracy, Precision, Recall, F1-score, AUC
Hardware Used	NVIDIA RTX 3050 (8GB VRAM)
Framework	PyTorch

3.12 Evaluation Metrics

Evaluation metrics quantify model performance. In this study we used several metrics to evaluate the performance of the model. The metrics include accuracy, precision, recall, f1 score, AUC and confusion matrix.

Results & Discussion

Online ISSN

3007-3197

Print ISSN

3007-3189

<http://amresearchreview.com/index.php/Journal/about>

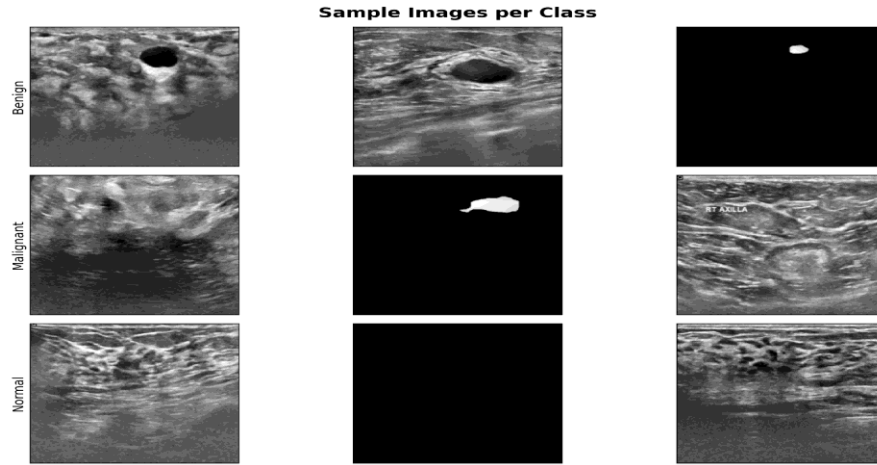


Figure. 2 Sample Images

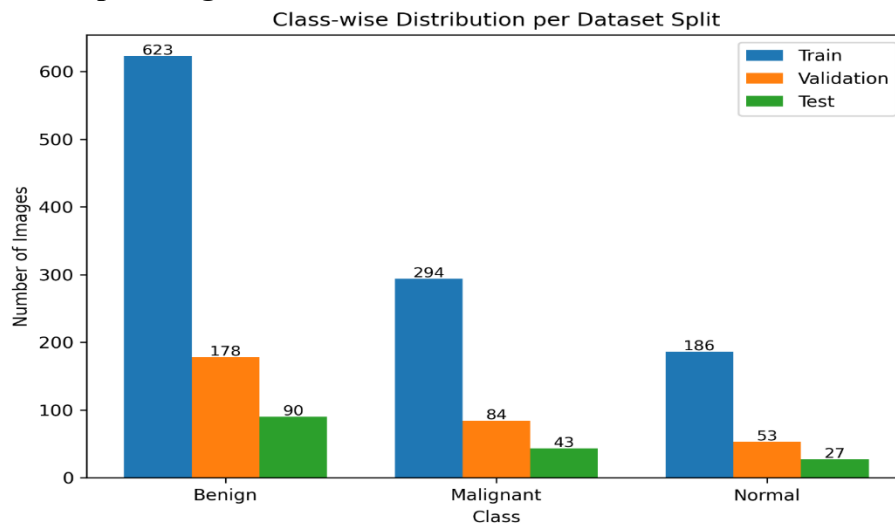


Figure. 3 Class-wise Distribution

Table 2: Hyperparameter Tuning Strategy

Hyperparameter	Tested Values	Optimal Value
Learning Rate	0.01, 0.001, 0.0001	0.0001
Batch Size	8, 16, 32	16
Number of Epochs	10, 20, 30	30
Optimizer	SGD, Adam, RMSprop	Adam
Dropout Rate	0.3, 0.4, 0.5	0.5
Weight Decay	0, 1e-4, 1e-5	1e-4
Transformer Heads	2, 4, 8	4
Transformer Layers	1, 2, 3	2
Embedding Dimension	128, 256, 512	256
Activation Function	ReLU, LeakyReLU	ReLU
Learning Rate	None, StepLR, ReduceLROnPlateau	

Online ISSN

3007-3197

Print ISSN

3007-3189

<http://amresearchreview.com/index.php/Journal/about>

Scheduler	ReduceLROnPlateau	
Data Augmentation	Basic, Moderate, Extensive	Moderate

Figure 4 shows the validation accuracy curve and train loss curve.

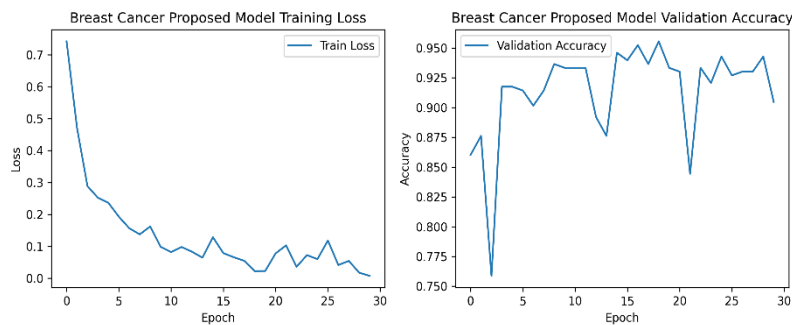


Figure. 4 Accuracy Curve

Figure 5 shows the performance of the proposed model using class-wise precision, recall and F1-score for each category (Benign, Malignant, and Normal). Precision measures the proportion of correctly predicted positive samples among all predicted positives for each class, indicating the model's ability to avoid false positives. Recall represents the proportion of correctly identified positive samples among all actual samples of a class, reflecting the model's sensitivity.

The F1-score is used to measure how well a model works. It is a balance between two things the precision and the recall. Precision shows how accurate the model is and recall shows how well it can find all the cases. This score is very useful when the classes are not equal in size. The results show that our model does a great job across all three measures for each type of case.

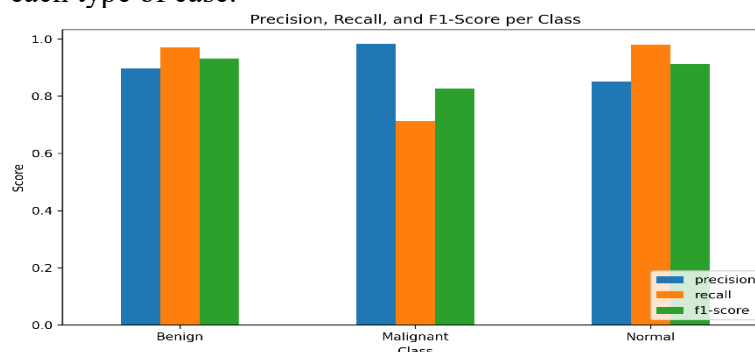


Figure. 5 Precision, Recall, and F1 Score per Class

Figure 6 shows the performance of the model which was evaluated using ROC, Precision-Recall Curve and Confusion Matrix. The ROC curve shows the trade-off between the true positive rate and false positive rate across different threshold values where a larger area under the curve (AUC) indicates better classification performance. The Precision-Recall curve shows the balance between precision and recall particularly useful for imbalanced datasets demonstrating the model's ability to maintain high precision while achieving strong recall. The confusion matrix shows the number of correct and incorrect predictions for each class (Benign, Malignant, and Normal). The proposed model shows a good performance.

Online ISSN

3007-3197

Print ISSN

3007-3189

<http://amresearchreview.com/index.php/Journal/about>

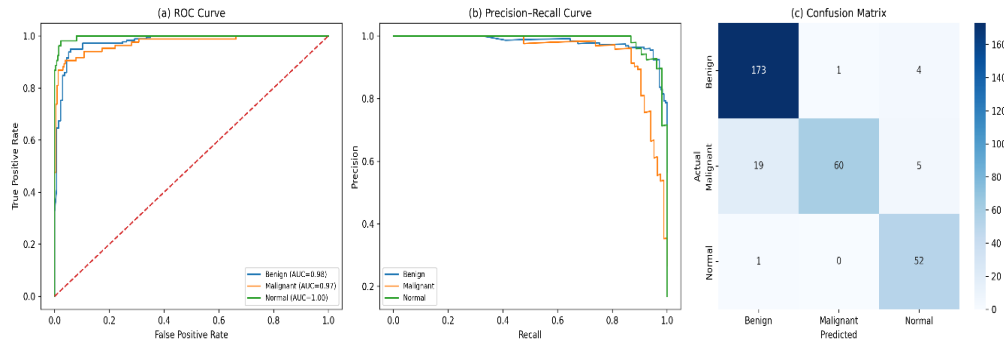


Figure. 6 ROC, Precision-Recall Curve, Confusion Matrix

The Figure 7 uses a technique called t-Distributed Stochastic Neighbor Embedding or t-SNE for short visualize the complex data learned by our model in a simpler two-dimensional space. Different types of samples Benign, Malignant, and Normal are spread out with each point on the figure representing a specific image feature embedding. This gives us a sense of how these samples relate to each other in terms of their features.

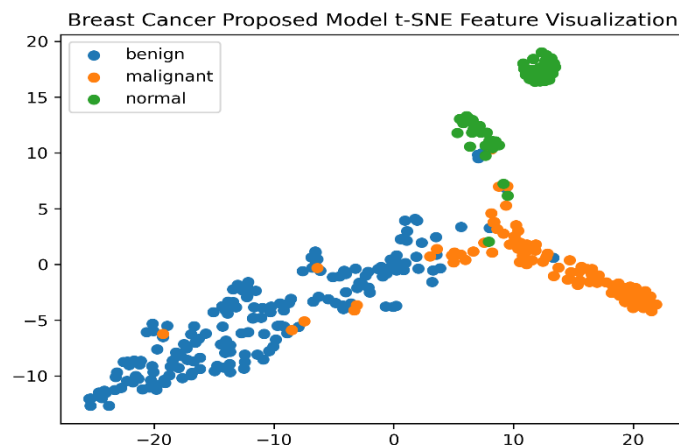


Figure. 7 t-SNE Feature Visualization

The gradient weighted class activation (Grad-CAM) helps understand the predictions. The Grad-CAM gives a visual explanation by indicating different parts of the image which are most important. The red and yellow colors are important for predictions. The result show that the proposed model focuses on meaningful cases. Additionally, Grad-CAM enhances the interpretability and transparency of the model, which is crucial for medical applications. Overall the results confirm that the model not only achieves high classification performance but also provides reliable visual justifications for its predictions.

Online ISSN

3007-3197

Print ISSN

3007-3189

<http://amresearchreview.com/index.php/Journal/about>

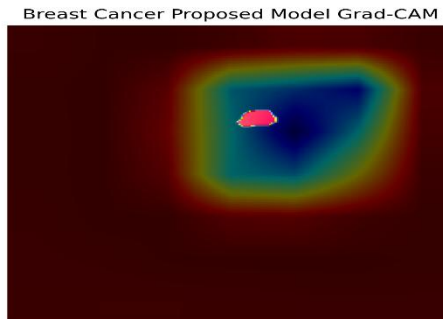


Figure. 8 Grad-CAM

Conclusion

In this study we combined multiple features using transformers to classify the breast cancer. The proposed study achieved a good accuracy of 90% which means it is a reliable way to classify breast cancer. This new method is really good at helping doctors figure out what kind of breast cancer someone has. This new approach will assist doctors in making better decisions about how to treat their patients and this will lead to better outcomes for people with breast cancer.

The class-wise analysis indicate that the model performed very well in detecting benign and normal cases by achieving the high recall values of 0.97 and 0.98 which highlights the effectiveness in minimizing false negatives for these classes. The malignant class achieves the highest precision (0.98) which indicates strong confidence in positive predictions. However lower recall (0.71) suggests that some malignant cases remain challenging to detect.

The results show that our model is really good at learning what makes things different and can be used for many types of classes. By combining feature fusion and a special kind of attention called Transformer-based attention we were able to make our model better at classifying the breast cancer images. Next, we want to make our model even better at finding bad cases, use more and different types of data, and try out new ways of paying attention to important details to make our model more reliable and useful in real-life medical situations.

References

- [1] V. Patil *et al.*, "Multiplicity of benign breast disease lesions and breast cancer risk in African American women," *Front. Oncol.*, vol. 14, 2024, doi: 10.3389/fonc.2024.1410819.
- [2] G. Murtaza *et al.*, "Deep learning-based breast cancer classification through medical imaging modalities: state of the art and research challenges," *Artif. Intell. Rev.*, vol. 53, no. 3, pp. 1655–1720, 2020, doi: 10.1007/s10462-019-09716-5.
- [3] N. Behar and M. Shrivastava, "ResNet50-Based Effective Model for Breast Cancer Classification Using Histopathology Images," *CMES - Computer Modeling in Engineering and Sciences*, vol. 130, no. 2, pp. 823–839, 2022, doi: 10.32604/cmcs.2022.017030.

Online ISSN

3007-3197

Print ISSN

3007-3189

<http://amresearchreview.com/index.php/Journal/about>

- [4] S. Gao, "Gray level co-occurrence matrix and extreme learning machine for Alzheimer's disease diagnosis," *International Journal of Cognitive Computing in Engineering*, vol. 2, no. July, pp. 116–129, 2021, doi: 10.1016/j.ijcce.2021.08.002.
- [5] S. Zarif, H. Abdulkader, I. Elaraby, A. Alharbi, W. S. Elkilani, and P. Pławiak, "Using hybrid pre-trained models for breast cancer detection," *PLoS One*, vol. 19, no. 1 January, Jan. 2024, doi: 10.1371/journal.pone.0296912.
- [6] M. R. Islam *et al.*, "Enhancing breast cancer segmentation and classification: An Ensemble Deep Convolutional Neural Network and U-net approach on ultrasound images," *Machine Learning with Applications*, vol. 16, p. 100555, Jun. 2024, doi: 10.1016/j.mlwa.2024.100555.
- [7] T. A. Rashid *et al.*, "NSGA-II-DL: Metaheuristic Optimal Feature Selection With Deep Learning Framework for HER2 Classification in Breast Cancer," *IEEE Access*, vol. 12, pp. 38885–38898, 2024, doi: 10.1109/ACCESS.2024.3374890.
- [8] S. Chakravarthy, B. Nagarajan, V. V. Kumar, T. R. Mahesh, R. Sivakami, and J. R. Annand, "Breast Tumor Classification with Enhanced Transfer Learning Features and Selection Using Chaotic Map-Based Optimization," *International Journal of Computational Intelligence Systems*, vol. 17, no. 1, Dec. 2024, doi: 10.1007/s44196-024-00409-8.
- [9] S. M. Thwin, S. J. Malebary, A. W. Abulfaraj, and H. S. Park, "Attention-Based Ensemble Network for Effective Breast Cancer Classification over Benchmarks," *Technologies (Basel)*, vol. 12, no. 2, Feb. 2024, doi: 10.3390/technologies12020016.
- [10] N. M. Hassan, S. Hamad, and K. Mahar, "YOLO-based CAD framework with ViT transformer for breast mass detection and classification in CESM and FFDM images," *Neural Comput. Appl.*, vol. 36, no. 12, pp. 6467–6496, Apr. 2024, doi: 10.1007/s00521-023-09364-5.
- [11] J. Ahmad *et al.*, "Deep learning empowered breast cancer diagnosis: Advancements in detection and classification," *PLoS One*, vol. 19, no. 7 July, Jul. 2024, doi: 10.1371/journal.pone.0304757.
- [12] U. K. Lilhore *et al.*, "Hybrid convolutional neural network and bi-LSTM model with EfficientNet-B0 for high-accuracy breast cancer detection and classification," *Sci. Rep.*, vol. 15, no. 1, Dec. 2025, doi: 10.1038/s41598-025-95311-4.
- [13] S. Armoogum, K. Motean, D. A. Dewi, T. B. Kurniawan, and J. Kijssomporn, "Breast Cancer Prediction Using Transfer Learning-Based Classification Model," *Emerging Science Journal*, vol. 8, no. 6, pp. 2373–2384, Dec. 2024, doi: 10.28991/ESJ-2024-08-06-014.
- [14] R. B. Eshun, M. Bikdash, and A. K. M. K. Islam, "A deep convolutional neural network for the classification of imbalanced breast cancer dataset," Jun. 01, 2024, *Elsevier Inc.* doi: 10.1016/j.health.2024.100330.
- [15] A. A. Hayum, J. Jaya, R. Sivakumar, and B. Paulchamy, "An efficient breast cancer classification model using bilateral filtering and fuzzy convolutional neural network," *Sci. Rep.*, vol. 14, no. 1, Dec. 2024, doi: 10.1038/s41598-024-56698-8.