Multimedia Tools and Applications
https://doi.org/10.1007/s11042-024-18947-w

Forward attention-based deep network for classification of breast histopathology image

Sudipta Roy[1] · **Pankaj Kumar Jain**[1,2] · **Kalyan Tadepalli**[3] · **Balakrishna Pailla Reddy**[4]

Received: 8 September 2023 / Revised: 1 January 2024 / Accepted: 13 March 2024

Abstract
Breast cancer diagnosis via histopathology is clinically important but challenges remain. We develop a Forward Attention-based deep network (FA-VGG16) for classifying breast histopathology images. For binary classification, FA-VGG16 achieves 90.4% accuracy, outperforming VGG16 (89.3%). Solving class imbalance boosts performance to 97.7% accuracy. For quaternary classification of benign subtypes, FA-VGG16 obtains individual accuracy between 77.1 and 88.5% and overall, 77.8%. For malignant subtypes, individual accuracy ranges from 77.2 to 98.3% and overall, 92.4%. Receiver operating characteristic analysis yields areas under the curve values exceeding 95.7% for all benign and malignant subtypes. Paired t-testing of variants indicates FA-VGG16 significantly outperforms others ($p < 0.0001$). The attention modules in FA-VGG16 enhance feature extraction, evident from its progressive ablation study performance drop. In summary, FA-VGG16 demonstrates 97.7% accuracy for binary and 92.4% for quaternary classification, numerically validating its potential to enhance breast cancer diagnosis through attention mechanisms.

Keywords Deep learning · Histopathology images · Breast cancer · Class imbalance · Classification

✉ Sudipta Roy
sudipta1.roy@jioinstitute.edu.in

Pankaj Kumar Jain
pankaj.jain@jioinstitute.edu.in

Kalyan Tadepalli
kalyan.tadepalli@rfhospital.org

Balakrishna Pailla Reddy
balakrishna.pailla@ril.com

[1] Artificial Intelligence & Data Science, Jio Institute, Navi Mumbai 410206, India

[2] Artificial Intelligence & Data Science, Jio Institute, Navi Mumbai, Maharashtra 410206, India

[3] Sir HN Reliance Foundation Hospital, Girgaon, Mumbai 400004, India

[4] Reliance Jio - Artificial Intelligence Centre of Excellence (AICoE), Hyderabad 500081, India

Published online: 22 March 2024

1 Introduction

Breast cancer is among the most common cancers worldwide in women. According to American Cancer Society's most recent update, there has been a 0.5% increase in incidence for 2010–2019. While early detection through expanded screening programs have helped to bring down the mortality in some countries, 5-year survival rate in countries like India (66%) and South Africa (40%) is still a major challenge [1]. The classification, diagnosis, and treatment of breast cancer has undergone significant changes in recent years, with more emphasis on the molecular and genetic profiling of tumors. While histopathology remains the gold standard for breast cancer diagnosis, it is a challenging and error-prone process that depends heavily on the expertise of pathologists. The diagnosis of breast cancer requires a thorough evaluation of various morphological features, including tumour stage, grade, histological type, proliferation status, and lymph vascular invasion [2, 3]. The difficulty in obtaining accurate and reproducible diagnoses is due to the heterogeneity of breast cancer, inter-observer variability among pathologists, and the inherent limitations of traditional microscopy-based techniques.

Proper classification of breast cancer helps determine the type and stage of the disease, which in turn informs the choice of treatment options. Accurate classification leads to more effective treatments, improved outcomes, and a better quality of life for breast cancer patients. It can also help in the development of new and improved treatment methods. Additionally, breast cancer classification is important for monitoring and tracking the progress of the disease, as well as for conducting research and advancing our understanding of breast cancer. In this article we addressed these issues in an analytical way.

Recent advances in computer vision and deep learning have made tremendous strides in improving the accuracy of diagnosis. All aspects of breast cancer diagnosis like classification of subtypes, detection of subtle histopathological differences with the highest diagnostic yield can be improved through the utilization of deep learning-based techniques [4, 5]. The development of such systems could have a significant impact on breast cancer diagnosis and management, improving patient outcomes and reducing the disease burden. Successful applications of AI in medical imaging include breast cancer and brain cancer in traditional way and well preprocessing [6–11]. But much faster progress is happening after using deep learning. In this work, an attention-based iterative approach has been proposed for classifying breast cancer histopathology images.

The mammography serves as the initial step in deciding whether a biopsy might be needed. Mammographic images are reviewed for indications of malignant lesions such as masses, microcalcifications, and architectural distortions. If seen, they undergo classification according to the Breast Imaging Reporting and Data System (BI-RADS), which allocates a numerical score based on the probability of malignancy. Lesions designated as BI-RADS category 4 or 5 are regarded as suspicious and necessitate biopsy. Nevertheless, determining which lesions warrant biopsy is a subjective undertaking that relies on the expertise of the radiologist or oncologist.

Therefore, our aim is to design a multiclass deep learning model where microscopic slides could be diagnosed. In the first instance the model should be able to classify the images into benign and malignant classes. Further, these classes can be subdivided into their subclasses to determine the type of abnormality in the images. We studied related literature in the same area and presented their work.

1.1 Problem identification

Previous studies have mainly focused on simple classification images into binary classes using either visual, machine learning or DL-based methods. While reviewing above literature we identified some gaps in previous studies such as:

- These methods lack the ability to extract features from a wide range of breast cancer images.
- These methods often use cross-entropy loss functions which are not very effective.
- Lack of high accuracy in binary and multiclass classification.
- The above methods also suffer from high class imbalance problem.
- Data augmentation is not performed in many studies.
- Additionally, most literatures do not perform ablation studies to show the robustness of their models or provide statistical parameters to validate their results.

1.2 Our contribution

We identified these gaps in previous studies and addressed them in our research article.

- We developed an attention-based deep network FA-VGG16 which enhances the feature extraction capability from a wide range of breast tumors.
- In our proposed model we used Focal loss function and Adam optimizer to optimize the loss between the predicted and ground truth images.
- We also applied data augmentation and class imbalance solution, to avoid the accuracy, recall and generalization bias in the model.
- We conducted ablation study and changing database experiment to check the robustness of our proposed model towards change in model layers, parameters, and database size.
- Further, we compared our proposed model with state-of-the-art models (SOTA) to benchmark our results against them.
- Lastly, we performed statistical tests which show a complete analysis of breast cancer classification and make our model acceptable in the clinical decision support system.

While prior works have advanced methodologies for breast histopathology classification, limitations remain that this study aims to address. Mainly, few investigations have conducted algorithmic evaluations with the level of rigor necessary to establish clinical validity and drive the field forward. Through systematic experimentation accounting for known challenges, we develop robust techniques to surpass discriminative performance alone.

Specifically, we introduce an attention-based deep learning framework to perform iterative benign-malignant subtype classification of breast tissue slides. Integrated attentional mechanisms enhance feature representation learning from histopathological imagery. Moreover, our tailored focal loss function and extensive augmentation/imbalance handling protocols achieve discriminability while addressing dataset biases. Thorough statistical validation and comparison to established models demonstrate technical merit. Collectively, these methodological contributions substantiate potential clinical applicability by establishing new benchmarks for analytical evaluations in computational pathology. Going

forward, our findings may help translate such techniques into pragmatic decision support systems to positively impact cancer diagnosis.

Inclusion of the above gaps in our proposed model not only improved the accuracy but the overall performance of the system, compared to other studies. A summary of our methodology is presented in Fig. 1.

To build upon previous research efforts in this area, we conducted an in-depth review of related work.

2 Related work

In our literature review, we found that significant progress has been made in analyzing the BreakHis histopathology image database [12]. This seminal work laid the foundation for subsequent research by providing a standardized dataset and focusing initial efforts on data representation. However, as the field advanced, new opportunities emerged to improve performance. A multiclass benign and malignant microscopic histopathological database was firstly provided by the authors in this study [12]. Their database consists of a total of 7909 images from 40X, 100X, 200X, and 400X magnifying factors. Their research was mainly focused on data extraction and representation therefore, they used only state of the art methods for feature extraction and classification. They used state of the art (SOTA) features extraction methods such as local binary pattern, gray level cooccurrence matrix (GLCM), parameter free threshold adjacency statistics (PFTAS), oriented FAST and Rotated BRIEF (ORB). For classification also they used SOTA quadratic linear analysis (QDA), support vector machine (SVM) and nearest neighbour and random forests methods. The methods used here are very common and their database has very high degree of similarity and class imbalance. By systematically reviewing related efforts (See Table 1), we aim to position our approach within the broader context of the field and avoid duplicating solutions to problems already addressed.

The exploration of deep learning techniques for breast cancer diagnosis through histopathological image analysis has catalyzed remarkable progress in developing advanced computer-aided diagnosis (CAD) systems. However, challenges related to

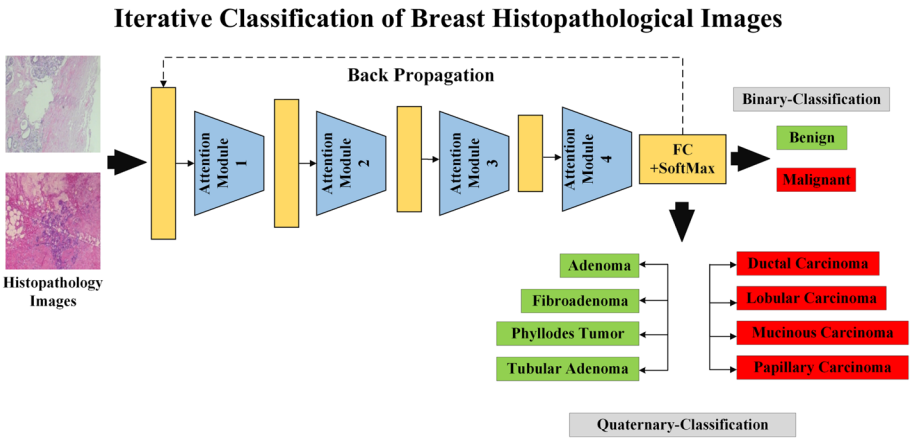

Fig. 1 Iterative Classification of Breast Histopathological Images

Table 1 A summary of Literature studies on the recent work on histopathology, deep learning and potential other methods that applicable to classification of cancer from breast Histopathology images

Research Question answered	AI Model Used	Dataset(s) Used for research	Accuracy of the best performing model	Shortcomings of the work	Authors Comments
Classify breast cancer histopathological images into 8 tumor types and also binary classify as carcinoma vs. non-carcinoma using multi-scale image features [16].	Self-attention random forest (SARF) with pyramid gray level co-occurrence matrix (PGLCM) feature extraction	BreaKHis (7909 histopathology images) and MIAS (322 mammography images) datasets	94.77% accuracy for 8-class classification and 97.98% accuracy for binary classification on BreaKHis dataset. 98.79% accuracy for 3-class classification on MIAS dataset	Did not explore other ensemble classifier architectures besides random forest. Not used pruning.	Use of self-attention mechanism and multi-scale feature extraction for breast cancer image classification.
Classify breast cancer histopathology images using modified pretrained CNN models augmented with convolutional block attention modules. Compare models on accuracy, precision, recall, F1 score [20].	Xception, VGG16, ResNet50, MobileNet, DenseNet121 pretrained models + convolutional and attention layers	BreakHis dataset (7909 images across different magnifications)	99.5% accuracy with Xception + attention model	No action was taken to reduce overfit problem.	Paper uses transfer learning combined with attention for improved breast cancer image classification.
Classify breast cancer images as benign or malignant using transfer learning models ResNet, DenseNet, and CNN [25].	Transfer learning with ResNet, DenseNet, and CNN	Breast Cancer Wisconsin (Diagnostic) dataset and Breast Cancer Histopathological Database (BreakHis)	98.3% accuracy with TLBCM model	Did not explore other neural network architectures. Experiments was not robust and lots of overfit.	Transfer learning used across multiple datasets
Classify breast cancer histology images as benign or malignant using lightweight deep transfer learning models [22].	MobileNet, DenseNet121, InceptionV3 transfer learning models + Support Vector Machine (SVM)	BreakHis v1 400x dataset (1819 images)	91.3% accuracy with MobileNet + SVM model	Lots of false positive problems.	lightweight deep transfer learning models for breast cancer classification in IoMT systems.

Table 1 (continued)

Research Question answered	AI Model Used	Dataset(s) Used for research	Accuracy of the best performing model	Shortcomings of the work	Authors Comments
Classify breast cancer histopathology images using reduced deep convolutional activation features (R-DeCAF) extracted from pretrained CNNs [19].	AlexNet, VGG-16, VGG-19 for feature extraction + SVM classifier	BreakHis dataset, ICIAR 2018 dataset	91.13% on 400X magnification BreakHis images using AlexNet + PCA	Very low Sensitivity and MCC. Lot of malignant detection miss.	Not a standard approach to get efficient and clinical results.
Classify breast cancer histopathology images into malignancy grades 1, 2 or 3 using multi-stage transfer learning and CNNs. Evaluate magnification-dependent and magnification-independent performance [18].	InceptionV3, Xception CNNs with multistage transfer learning	BreaKHis dataset, Databiox breast cancer grading dataset	97.67% ± 1.09% on Databiox dataset (magnification-dependent)	No action was taken to reduce overfit problem.	multistage transfer learning and CNNs for breast cancer grading on multiple datasets.
Classify breast histopathology images using CNN model with multi-dimensional feature fusion [23].	MDFF-Net − 1D and 2D CNNs with feature fusion network	BreakHis dataset, BACH dataset	98.86% on BreakHis dataset	Large model size and long training time compared to other CNNs. Low sensitivity and MCC.	Improved accuracy over other CNN models by fusing 1D and 2D features.
Classify breast cancer from histopathology images using convolutional neural networks with novel angular margin loss function [24].	BreastNet CNN architecture + Boosted Additive Angular Margin (BAM) loss	BreaKHis dataset	99.96% on BreaKHis 200X magnification images	Overfitting problem. Low sensitivity, specificity and MCC.	Accuracy over softmax and other angular margin losses.

Table 1 (continued)

Research Question answered	AI Model Used	Dataset(s) Used for research	Accuracy of the best performing model	Shortcomings of the work	Authors Comments
Classify breast cancer histopathology images into benign vs. malignant and into 4 subtypes using deep learning. Evaluate image-level and patient-level performance [13].	Inception Recurrent Residual Convolutional Neural Network (IRRCNN)	BreakHis dataset and 2015 Breast Cancer Classification Challenge dataset	97.65% accuracy on BreakHis dataset, 99.05% on 2015 Challenge dataset	Data growth study was not performed, coss validations were not proper and due that overfit a lot. Very low Sensitivity and MCC.	Demonstrates IRRCNN model for breast cancer classification across multiple datasets and metrics. .
Classify breast cancer histology images as benign or malignant using lightweight deep transfer learning models [22].	MobileNet, DenseNet121, InceptionV3 transfer learning models + Support Vector Machine (SVM)	BreakHis v1 400x dataset (1819 images)	91.3% accuracy with MobileNet + SVM model	Low specificity.	Not a standard approach to get efficient and clinical results.
Classify breast cancer images as benign or malignant using transfer learning models ResNet, DenseNet, and CNN [25].	Transfer learning with ResNet, DenseNet, and CNN	Breast Cancer Wisconsin (Diagnostic) dataset and Breast Cancer Histopathological Database (BreakHis)	98.3% accuracy with TLBCM model	Did not explore other neural network architectures.	Not a standard approach to get efficient and clinical results.
Classify breast cancer histopathology images using modified pretrained CNN models augmented with convolutional block attention modules. Compare models on accuracy, precision, recall, F1 score [20].	Xception, VGG16, ResNet50, MobileNet, DenseNet121 pretrained models + convolutional and attention layers	BreakHis dataset (7909 images across different magnifications)	99.5% accuracy with Xception + attention model	Did not explore other attention mechanisms besides CBAM.	transfer learning combined with attention for improved breast cancer image classification.

Table 1 (continued)

Research Question answered	AI Model Used	Dataset(s) Used for research	Accuracy of the best performing model	Shortcomings of the work	Authors Comments
Classify breast cancer histopathological images into 8 tumor types and also binary classify as carcinoma vs. non-carcinoma using multi-scale image features [16].	Self-attention random forest (SARF) with pyramid gray level co-occurrence matrix (PGLCM) feature extraction	BreaKHis (7909 histopathology images) and MIAS (322 mammography images) datasets	94.77% accuracy for 8-class classification and 97.98% accuracy for binary classification on BreaKHis dataset. 98.79% accuracy for 3-class classification on MIAS dataset.	Did not explore other ensemble classifier architectures besides random forest.	-Method used self-attention mechanism and multi-scale feature extraction for breast cancer image classification.
Classify H&E stained breast biopsy images into 4 classes - normal tissue, benign lesion, carcinoma in situ, and invasive carcinoma. Also do binary classification of carcinoma vs. non-carcinoma [8].	Convolutional neural network (CNN) and CNN + support vector machine (SVM)	New breast cancer image dataset from Bioimaging 2015 challenge, 249 images for training and 36 images for testing.	83.3% accuracy for CNN + SVM with majority voting for binary carcinoma vs. non-carcinoma classification. 77.8% accuracy for 4-class classification.	Small dataset size compared to other CNN image classification tasks. No localization of abnormal regions.	Not a standard approach to get efficient and clinical results.
Develop hybrid techniques combining deep learning and handcrafted features to accurately classify histopathological images of multi-class breast cancer [17].	Artificial neural network (ANN) with features from VGG-19, ResNet-18, and handcrafted (FCH, LBP, DWT, GLCM)	BreakHis histopathology image dataset	99.7% accuracy with VGG-19 + handcrafted features on 400x magnification images for binary classification; 97.3% accuracy with same model for multi-class classification	Did not explore other neural network architectures besides ANN.	Utilizes fusing of deep learning and handcrafted features for improved breast cancer subtype classification from histopathology images.

Table 1 (continued)

Research Question answered	AI Model Used	Dataset(s) Used for research	Accuracy of the best performing model	Shortcomings of the work	Authors Comments
Develop an automated breast cancer classification method using histopathology images that can classify into binary (benign vs. malignant) and multi-class subtypes [14].	Convolutional neural network (CNN) with fused mobile inverted bottleneck convolutions (FMB-Conv), mobile inverted bottleneck convolutions (MBConv), and dual squeeze & excitation (DSE) blocks	BreakHis histopathology image dataset	Up to 99.37% F1 score for binary classification and 91.9% F1 score for multi-class classification on 400x magnification images	Did not explore other attention mechanisms besides DSE.	Effective use of DSE blocks to improve feature learning in CNNs for histopathology image classification.
Classify histopathology images as cancerous or non-cancerous using feature engineering and harmonization techniques [21].	Feature extraction + Multi-Layer Perceptron classifier	IDC and BreakHis datasets - histopathology images of breast tumors	Up to 95% testing accuracy on BreakHis, 94.7% on IDC	Did not compare to deep learning models. Results analysis was not promising and robust.	Harmonization techniques are used to handle variability in histopathology data.
Classify histopathology images into benign and malignant breast cancer subtypes using a hybrid CNN-LSTM model [15].	Hybrid CNN-LSTM model using transfer learning with ImageNet weights	BreakHis dataset – 2480 benign and 5429 malignant breast cancer histopathology images	99% accuracy for binary classification, 92.5% accuracy for multi-class classification	Did not compare to other hybrid deep learning models.	The work shows reasonable results for breast cancer subtype classification using CNN-LSTM model. The model shows potential for automated diagnosis of cancer subtypes.
Classify breast cancer histology images into different types using deep learning approaches [7].	Pre-trained CNNs (ResNet50, InceptionV3, VGG16) for feature extraction + Gradient Boosted Trees classifier	400 H&E stained breast histology images, with 4 balanced classes - normal, benign, in situ carcinoma, invasive carcinoma	93.8% for 2-class classification (non-carcinoma vs. carcinoma), 87.2% for 4-class classification	Did not explore fine-tuning CNNs or training CNNs end-to-end.	Promising results for breast cancer classification on a small dataset by using transfer learning.The work demonstrates the potential of deep learning for histology image analysis.

Table 1 (continued)

Research Question answered	AI Model Used	Dataset(s) Used for research	Accuracy of the best performing model	Shortcomings of the work	Authors Comments
Classify breast cancer images into benign and malignant tumors using deep learning and transfer learning approaches [4].	Pre-trained deep neural networks - ResNet18, InceptionV3Net, ShuffleNet	BreakHis dataset – 7909 microscopic images of breast tumor tissue	ResNet18–99.7% binary classification, 97.81% multi-class classification	Did not compare performance to other traditional machine learning models.	The work demonstrates the potential of using deep transfer learning for accurate classification of breast cancer images.

model optimization, interpretability, and clinical integration persist. Recent research reflects a multifaceted effort to enhance the accuracy and efficiency of AI-assisted diagnostics by addressing the various facets of the diagnostic workflow.

A key aspect is the creation of a tailored convolutional neural network (CNN) architectures for enhanced feature learning and discriminative capabilities. Approaches range from recurrent CNNs [13], dual-attention CNNs [14], to hybrid CNN-LSTM networks [15], which demonstrate reasonable accuracy for malignancy grading and subtype classification on the tested datasets. However, issues like opacity and sensitivity to hyperparameters motivate hybrid techniques fusing deep learning with complementary models. Rakhlin et al. [7] combined CNN feature extraction with gradient boosted trees, achieving multi-class accuracy above 87% and improved interpretability. Similarly, Li et al.'s [16] Self-Attention Random Forest model leverages multi-scale fusion features for enhanced representation. Al-Jabbar et al. [17] fused deep learning with handcrafted features, balancing efficiency and benchmark performance. One area for active future research could be hybrid models.

Strategic transfer learning has also been utilized to work around data limitations. Fine-tuning models pretrained on natural images provides performance boosts over training from scratch in some cases, as evidenced by Mudeng et al. [18] Morovati et al. [19] and Aljuaid et al. [4]. Augmentation (Abdulla et al.) [9] and attention mechanisms (Ashurov et al.) [20] assist in maximizing learning from scarce datasets. An additional hurdle that researchers in this have to deal with is, the paucity of well annotated datasets. In this context, novel semi-supervised and self-supervised techniques that utilize unlabeled data during training (Nassib Abdallah et al.) [21] represent areas for further investigation.

Beyond accuracy, optimizing efficiency is vital for real-world deployment. MobileNet-SVM (Ogundokun et al.) [22] enables analysis under resource constraints while retaining accuracy. Enhancing model interpretability is also critical for clinical integration. In this regard, Xu et al.'s [23] MDFF-Net addresses the opacity limitations of CNNs through multi-dimensional feature fusion. Novel losses like BAM loss (Alirezazadeh & Dornaika) [24] optimize angular space to achieve remarkable accuracy despite data limitations. These are some novel approaches to address persistent problems in utilizing Deep Learning for breast histopathology classification, by need continued research and validation to translate gains in accuracy and efficiency into enhanced reliability, accessibility, and clinical utility.

As evidenced by the quited works, the presenct landscape of breast cancer diagnosis through histopathological image analysis is one of appreciable advances from its early days. Tailored CNN architectures, hybrid systems, transfer learning techniques, attention mechanisms, optimized loss functions, and lightweight models each address different facets of the complex diagnostic workflow. While substantial interdisciplinary progress has been achieved, realizing the full potential of deep learning in this domain remains an ongoing pursuit. Persisting challenges related to model optimization, feature representation, interpretability, and clinical integration continue to shape promising research directions.

Building upon these advances, in the current work we propose an attention-based approach tailored for the domain of breast histopathology image classification. A summary of recent studies on histopathology, deep learning and potential other methods that applicable to classification of breast cancer images is shown in Table 1.

Specifically, in Section 3 we describe our methodology for evaluating attention models on a publicly available breast cancer histopathology dataset. As the first step towards model development and evaluation, we require a standardized database. Therefore, in Section 3.1 we discuss our acquisition and preparation of the publicly available BreakHis

dataset [12]. This workflow lays the foundation for systematically assessing attention models, as described next in Section 3.2.

3 Methodology and experiments

3.1 Database acquisition and preparation

We have used a publicly available BreakHis database [12] in our study. This database consists of a total of 7909 images of 40X, 100X, 200X and 400X magnifications. The duration of the database collection was approximately one year from January to December 2014. The database was collected by surgical open biopsy (SOB) method at P&D laboratory, Brazil. A standard paraffin procedure was adapted for slide preparation stained with hematoxylin and eosin. All images were labelled by the pathologists. The image database was collected from 4 types of benign lesions (A, FA, TA, PT) and 4 types of malignant lesions (DC, LC, MC, PC). We considered only 40X magnification images of benign (625) and malignant (1370) classes. Thus, for each class in benign we have A = 114, FA = 253, TA = 109, PT = 149 and in malignant DC = 864, LC = 156, MC = 205, PC = 145 images. For training purposes, we resized all images into 224×224 sizes. Further, we applied random image augmentation to all images to be used in a few of our experiments.

The Breast Cancer Histopathological Image Classification (BreakHis) dataset was selected for this research due to its comprehensive and diverse representation of breast tumor tissue images, making it highly suitable for training and evaluating our proposed deep learning model. This dataset was collected from 82 patients, with 2,480 benign samples and 5,429 malignant samples across various magnification factors.

The key advantages of the BreakHis dataset that reinforce its suitability for our work are three-fold:

1. Comprehensive benign and malignant categories - The images are meticulously categorized as benign or malignant based on expert pathologist annotations. This provides a robust binary ground truth for evaluating model performance on benign vs. malignant classification.
2. Diverse histological subtypes - The dataset encompasses various histological subtypes including adenosis, fibrosis, tubular adenoma, phyllodes tumor, ductal carcinoma, lobular carcinoma, mucinous carcinoma, and papillary carcinoma. This heterogeneity in morphological subtypes within each binary class enables more rigorous evaluation of model generalization.
3. Multi-scale magnification - With images scanned at 40X, 100X, 200X and 400X magnifications, the dataset allows an assessment of model robustness across various scales and tissue resolutions. Models can be tested for consistency in predictions independent of magnification factors.

The scale, diversity, and expert annotations of the BreakHis dataset provide an ideal benchmark for developing and evaluating breast cancer histopathology classification systems. The variety of magnification, morphology, and breast tissue types represented in this dataset lend statistical power and clinical relevance to the experiments conducted in our research. Our model's strong performance on this rigorously curated

Table 2 Distribution of benign and malignant patients' images at 40X resolution

Benign			Malignant		
Tumour	#Images	#Patients	Tumour	#Images	#Patients
Adenosis	114	4	Ductal Carcinoma	864	38
Fibroadenoma	253	10	Lobular Carcinoma	156	5
Phyllodes Tumour	109	3	Mucinous Carcinoma	205	9
Tubular Adenoma	149	7	Papillary Carcinoma	145	6
Total	625	24		1370	58

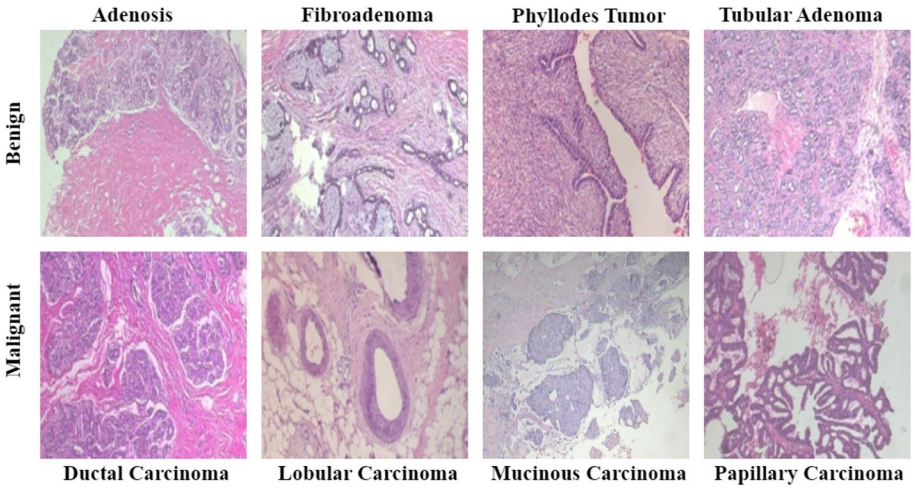

Fig. 2 Sample images from each class at 40X magnification

dataset provides evidence of its viability as a generalizable tool for distinguishing benign and malignant breast tumors in real-world applications.

3.2 Class imbalance problem

Looking at Table 2 we can simply observe a class imbalance problem between benign:625, and malignant:1370 images (for binary classification) and a similar problem between their subclasses (for quaternary classification).

To overcome the class imbalance problem, we used a random replication algorithm [26]. Further, to avoid the similarity and intra class imbalance in the database we applied data augmentation. The augmentation includes image rotation [-15^0 to $+15^0$], reflection, translation in the x-axis [30 pixels], translation in the y-axis [30 pixels], and scaling [0.85 1.25]. All augmentations are random and applied to random images in the pool. Figure 2 shows the sample images from each class at 40X magnification.

3.3 Proposed Attention-VGG16 model

We added four attention modules (shown by gray blocks) in VGG16 network to enhance the feature learning ability of the VGG16 network (see Fig. 3). The backpropagation algorithm runs in parallel to the VGG16 network and trains it through a chain rule. The weights of the VGG16 network are updated through this algorithm, and the model loss is reduced during backpropagation. Those are the gradients at each node which are responsible for change in weights. These gradients are calculated using the chain rule at each node. As we keep moving towards the initial layers of the network the number of local gradient increases. As the number of gradients increases, it results in a very small gradient value (vanishing gradient) which finally makes a very small change in weights. The above additive attention module uses only one hidden layer feed forward network to calculate the channel weights. This hidden layer network reduces the local gradient thereby avoiding the vanishing gradient problem. The current VGG16 module is modified by deploying four additive attention modules each after 4th, 7th, 10th, and 13th convolutional layers. A detailed description of the attention module is given in the next subsection. Both fully connected (FC) layer size is reduced from 4096 to 1024. The class size is chosen depending on binary or quaternary classification. The model is equipped with ADAM optimizer to minimize the focal loss function.

Feature extraction: The convolutional layer in a VGG16 network extracts hierarchical features from an input image. These convolutional features have sizes of 224, 112, 56, 28, and 14 and associated features depths (based on the number of filters in convolutional layers). By adding the attention module between the two layers (224 and 112, or 112 and 56 likewise) we can leverage its capabilities to emphasize relevant ROI from image during feature extraction. Deep neural networks can be computationally expensive and/or may over fit the data for some specific task such as histopathology image classification. To maintain the trade-off between the computational complexity and the overfitting, some novelty in the model is required. Therefore, addition of attention module in between the convolutional layers maintains the balance between the overfitting and the complexity of the model.

3.4 The attention module

Figure 4 represents a channel attention module (CAM) (expanded gray block from Fig. 3) which is designed in such a way as to utilize the channel features efficiently. Shallow

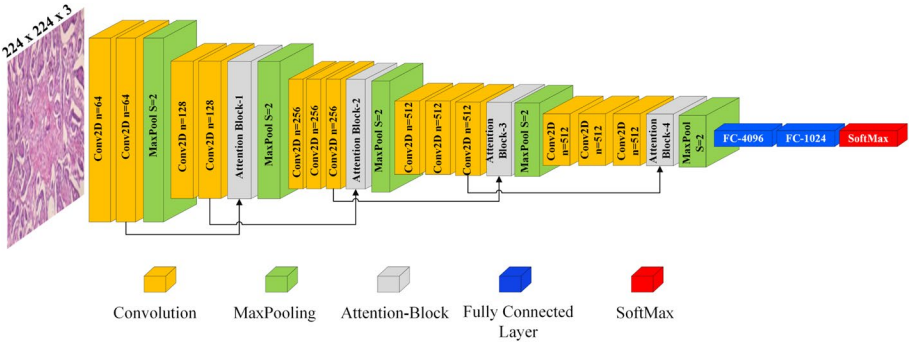

Fig. 3 FA-VGG16 architecture with four Attention modules

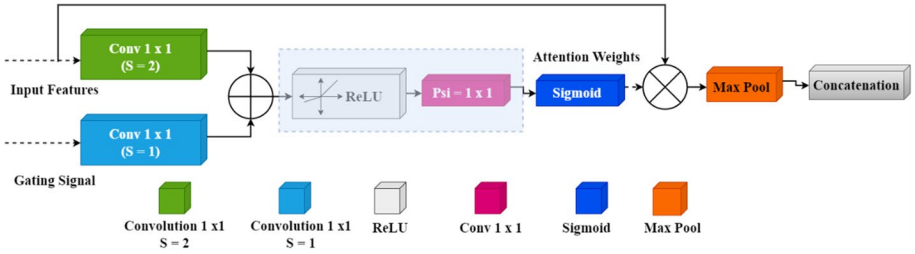

Fig. 4 Expanded Channel Attention module (CAM)

features (h_1 x w_1 x c_1) from previous VGG16 model layer (via arrow) are fed as input to the conv 1 × 1 (S = 2) layer, and deep features (h_2 x w_2 x c_2) enter from the adjacent layer to the conv 1 × 1 block. Global 1 × 1 convolution (S = 2) on shallow features scale downs the dimensions and the resultant features are added elementwise to the 1 × 1 conv on deep features. The combined signal passes through ReLU and a 1 × 1 convolution (PSI) to acquire the channel attention weights of the features map. These channel weights are finally multiplied elementwise with shallow features. Thus, shallow features are enhanced by the attention channel weights which are further concatenated to the features after MaxPooling. Convolution of shallow features in above picture are represented as (1).

$$C_{shallow}(X,Y,k) = \sum_{i=0}^{m-1}\sum_{j=0}^{m-1}\sum_{c=0}^{K-1}\left[f(i,j,c) * I(Sx+i-P, Sy+j-P, c)\right] \tag{1}$$

Similarly, convolution of deep features is represented as (2):

$$C_{Deep}(X,Y,k) = \sum_{i=0}^{m-1}\sum_{j=0}^{m-1}\sum_{c=0}^{K-1}\left[f(i,j,k) * I(x+i-P, y+j-P, c)\right] \tag{2}$$

In the above equations m×m, is the dimension of filter 'f', x, y and k are spatial and channel dimension of the image 'I'. P and S are padding, and strides used in the convolution. X, Y, and k are output spatial and channel dimensions. In the same picture ReLU is represented mathematically as (3):

$$\text{ReLU}\,[f(x)] = \text{Max}\,(0, x) \tag{3}$$

After passing through ReLU, a 1 × 1 filter is used to produce a linear projection of a feature map. The 1 × 1 convolution is also called as channel wise pooling and is represented by (4):

$$\Psi(X,Y,k) = \sum_{c=1}^{k}\left[f(1,1,c) * I(x,y,c)\right] \tag{4}$$

After aggregating (1)–(4) we acquire the $C_{Attention}$ feature map, which is modification of weights of the shallow features by attention effect.

$$C_{Attention}(X,Y,k) = \Psi\left[ReLU\left\{C_{shallow}(X,Y,k) \oplus C_{Deep}(X,Y,k)\right\}\right] \odot C_{shallow}(X,Y,k) \tag{5}$$

Since we have already considered stride in (2) the dimensions of $C_{shallow}(X,Y,k)$ and $C_{Deep}(X,Y,k)$ becomes identical for element wise addition. Finally, Attention feature map

in (1) is maxpooled and concatenated with deep feature represented by (3). In (5) '⊕' and ⊙ represents element wise addition and multiplication. This complete process is attention module shown in Fig. 4 and represented as (6):

$$C_{Module} = Concatenation\left[Maxpool\{C_{Attention}(X,Y,k)\}, C_{Deep}(X,Y,k)\right] \tag{6}$$

3.5 Loss function

Previous studies used cross-entropy loss function which is very hard to optimize below certain level for this database. Therefore, we used the focal loss function (FL) for all our experiments which penalizes the minority class loss. This loss function is an extension of the cross-entropy loss function and represented mathematically as (7). The purpose of the function is to down-weight the larger class data and focus on smaller class data. Therefore, to achieve this original cross-entropy loss function is modified with a tunable focusing parameter (Gamma) $\gamma \geq 0$ and a balancing or weighting parameter (Alpha) $\alpha = 0.25$.

$$FL = -\alpha_i(1-p_i)^{\gamma} \log \log p_t \tag{7}$$

We used the default values of $\gamma = 0.25$, and $\alpha = 2$ for all our experiments. For $\gamma = 0$ focal loss is equivalent to the cross-entropy loss. Therefore, the focal loss function is designed to penalize predictions that are false and confident by down weighting these examples. This way the model learns from the difficult examples and overall model performance improves.

3.6 Hyperparameters

Hyperparameters setting is a vital part of DL system design. In our experiments, we compared our experiments on the basis of fixed values of epochs, learning rate, batch size, optimizer, and loss function. Table 3 below shows a list of experiments along with the hyperparameters and their values.

3.7 K-fold partition

In the current experiments, we used K = 5-fold partition methods to partition the database into training and testing pools. None of the images from the training parts participated in the testing.

3.8 Experiments

Table 3 column 1 shows the exhaustive list of experiments performed in this study. The rest of the columns show hyperparameters used in the respective experiment. We focused our experiments on our proposed model Attention-VGG16 while comparing it with other models and change in the hyperparameters of the same model. In this experiment, we selected focal loss and ADAM optimizer (keeping the rest of the hyperparameters same) to train our model on 1596 images and test on 399 images. In experiment 4, we changed our loss function and optimizer with cross-entropy loss function and SGDM optimizer. These experiments suffer from a class imbalance problem as the malignant class has 1370 images and

Table 3 List of experiments and hyperparameters

Sr#	Experiment	Loss Function	#TR (%)	#TE (%)	Partition method
1	VGG16	FL	1596	399	K5
2	VGG16 (Class Balance)	FL	2192	548	K5
3	FA-VGG16 (LF1 + O1)	FL	1596	399	K5
4	FA-VGG16 (LF2 + O2)	CE	1596	399	K5
5	FA-VGG16 (Class Balance)	FL	2192	548	K5
6	FA-VGG16 (Patient-Wise)	FL	1596	399	K5
7	FA-VGG16 (Ablation-1)	FL	1596	399	K5
8	FA-VGG16 (Ablation-2)	FL	1596	399	K5
9	FA-VGG16 (Ablation-3)	FL	1596	399	K5
10	FA-VGG16 Changing Data	FL	1596	399	80:20
11	FA-VGG16 (Ben-4-Class)	FL	1596	399	K5
12	FA-VGG16 (Mal-4-Class)	FL	1596	399	K5
13	GoogleNet	FL	1596	399	K5
14	Efficient NetB0	FL	1596	399	K5
15	XceptionNet	FL	1596	399	K5

Epochs = 100; Learning Rate = 10^{-4}; Optimizer : ADAM; and BS: 8 were used

FL Fractal Loss, *TR* training Images, *TE* Test Images, *LF1* Loss function 1 (Cross Entropy), *LF2* Loss Function 2 (Dice Similarity Coeffcient)

the benign class has 625 images. Therefore, we adapted an oversampling method by randomly replicating the inferior class images [26].

The BreakHis database contains images that were captured from the same slides, leading to a high data-similarity bias. This bias is reflected in the model's overlapping or generalization bias when using images from the same patient for both training and testing. In order to address this issue, we conducted experiment 6, where the model was trained using patient-wise images. The study followed a patient-wise division of images, with 65 patients' images allocated for training and 19 for testing. The patient selection process was random. In a subsequent experiment, the database size was varied by starting with a random selection of 10% of images from both classes for training (80%) and testing (20%), then gradually increasing the number of images by 10% until 90% of the data pool was used. In all experiments involving changing data, 5-fold cross-validation was not employed. All previous experiments belong to the binary classification. We have intraclass images in benign (A, FA, TA, PT) and malignant (DC, TC, MC, PC) sections. Thus, we used these intraclass images for quaternary (4-class) classification. Experiments 12 and 13 belong to quaternary classification for benign and malignant intraclass images. Since a class imbalance problem exists in these two experiments, we used random oversampling of minority class images and applied random augmentation as explained earlier. Here we have outlined our comprehensive experimentation approach, which involved addressing biases through patient-wise division, varying database size, and exploring intra-class classification. This systematic methodology allows for a robust evaluation of our proposed FA-VGG16 model under different conditions.

We are now poised to analyze the results of our experimental investigations in Section 4. As summarized in Table 4, a variety of performance metrics are used to assess and compare the different experiments enumerated in Table 2. At the broadest level, we find

Table 4 Classification parameters for the list of experiments

Experiment	ACC	Error	Sens	Spec	Prec	FPR	F1	MCC	Kappa
VGG16	89.3	10.7	82.2	92.5	83.9	7.5	82.5	75.3	74.8
VGG16 (CI)	91.6	8.0	88.9	94.3	93.4	6.0	90.3	84.1	83.2
FA-VGG16 (LF1 + O1)	90.4	9.6	88.3	91.4	83.0	8.6	85.4	78.6	78.3
FA-VGG16 (CI)	**97.7**	**2.3**	**98.4**	**96.9**	**97.0**	**3.1**	**97.7**	**95.4**	**95.3**
FA-VGG16 (PW)	79.4	20.6	56.4	90.6	73.3	9.4	62.0	50.7	48.9
FA-VGG16 (90%data)	92.8	7.2	75.7	100	100	0.0	86.2	82.8	81.4
FA-VGG16 (4-CO) Benign	77.8	22.2	77.8	92.6	78.6	7.4	77.2	70.7	40.7
FA-VGG16 (4-CO) Malignant	92.4	7.6	92.4	97.5	92.7	0.0	92.2	90.0	79.7

CI Class Imbalance, *PW* Patient-Wise, *CO* Class Overall, *DB* Database, *ACC* Accuracy, *E* Error, *Sens* Sensitivity, *Spec* Specificity, *Prec* Precision, *FPR* False Positive Rate, *MCC* Mathew's Correlation Coefficient

that our FA-VGG16 model achieves an accuracy of 90.43% for binary classification, outperforming the basic VGG16 model. Further gains are observed when tackling class imbalance, with our approach attaining 97.66% accuracy in this setting.

This initial overview establishes that our methodology, as described in Section 3, has thus far shown promise. In the forthcoming analysis, we will drill down to compare specific experiments and parameters in finer detail. This will provide deeper insights into the effectiveness and robustness of our proposed FA-VGG16 approach under diverse data conditions.

4 Experimental results

Table 4 shows the results of all experiments mentioned in Table 3. We evaluated our experiments based on multiple classification parameters such as, accuracy, sensitivity, specificity, precision, false positive rate (FPR), F1 score, Mathew's Correlation Coefficient (MCC), and Kappa. Experiment wise values of these parameters are compared in this table. With our proposed model FA-VGG16 model we achieved 90.43% accuracy compared to 89.27% for basic VGG16 model for binary classification. Similarly, when we solved class imbalance problem in the database we acquired 97.66% accuracy for Attention-VGG16 model compared to VGG16 model.

4.1 Changing database experiment results

Table 5 shows the results of changing database experiments. Each column represents the percentage value of classification parameter and row represents percentage of changing data. As can be seen from the table all parameters are increasing while we add more databases to the training and testing pool. Also, the increment in the classification parameters with increasing database is also visible in the graph shown in Fig. 5. It is clear from the table and the graph that the database has a direct impact on the classification performance of the proposed model. Also, we achieved the highest accuracy of 92.76% with 90% training and testing database.

Table 5 Classification parameters for Changing data size experiments

% DB	ACC	Error	Sens	Spec	Prec	FPR	F1	MCC	Kappa
10%	62.5	37.5	0.00	92.6	0.0	7.4	0.0	15.9	8.7
20%	81.3	18.8	44.4	100.0	100.0	0.0	61.5	58.9	51.5
30%	81.7	18.3	55.6	92.9	76.9	7.1	64.5	53.9	52.6
40%	83.8	16.3	80.0	85.2	67.9	14.8	73.5	62.3	61.9
50%	88.5	11.5	70.0	96.4	89.4	3.6	78.5	71.8	70.8
60%	90.0	10.0	85.9	91.7	81.3	8.3	83.6	76.4	76.3
70%	92.5	7.5	87.5	94.5	86.4	5.5	87.0	81.7	81.7
80%	92.2	7.8	85.4	95.4	89.8	4.6	87.6	81.9	81.9
90%	92.8	7.2	75.7	100.0	100.0	0.0	86.2	82.8	81.4

Table 6 Individual class output parameters for quaternary-class (Benign) experiment

Class	ACC	Error	Sens	Spec	Prec	F1	MCC	Kappa
A	77.1	22.9	77.1	95.9	86.7	80.3	54.9	75.9
FA	62.0	37.9	62.0	92.2	72.9	66.5	57.6	57.4
PT	83.4	16.6	83.4	90.8	75.9	79.2	50.9	72.1
TA	88.5	11.5	88.5	91.4	78.8	82.8	49.5	77.3

Table 7 Individual class output parameters quaternary-class (Malignant) experiment

Class	ACC	Error	Sens	Spec	Prec	F1	MCC	Kappa
DC	77.2	22.8	77.2	98.5	94.6	84.9	56.4	81.4
LC	98.3	1.7	98.3	95.1	86.9	92.3	48.1	89.8
MC	96.8	3.2	96.8	98.1	94.4	95.5	50.1	94.1
PC	97.3	2.7	97.3	98.2	94.9	96.1	49.9	94.8

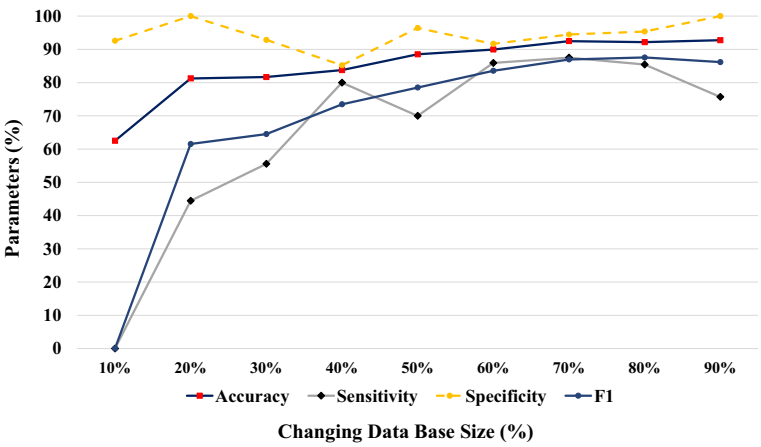

Fig. 5 Classification parameters variation with changing database

4.2 Quaternary class results

Tables 6 and 7 represent the classification performance of the intraclass distribution of the benign and malignant images. As discussed earlier benign and malignant sections have 4 intraclass images. Also, these subsections have class imbalance problems. After resolving the class imbalance problem, we achieved overall accuracy value of 77.8% for benign and 92.4% for malignant sections using our proposed model. For individual classes in benign section, we achieved accuracy of 77.1%, 62.0%, 83.4%, and 88.5% for Adenosis, Fibroadenoma, Phyllodes Tumour, and Tubular Adenoma respectively. Similarly, for individual classes in malignant section we achieved 77.2%, 98.3%, 96.8% and 97.3% accuracies for ductal carcinoma, lobular carcinoma, mucinous carcinoma, and papillary carcinoma respectively. Thus, we can conclude that our model is able to generalize the malignant class images easily.

Our study validates the proposed model and hyperparameters, but we performed additional performance evaluation tests to confirm the results. Specifically, we used the receiver operating characteristics (ROC) curve and calculated the area under the curve (AUC).

4.3 Comparison with state-of-the art models

We compared our proposed model resultswith state-of-the art models GoogleNet, XceptionNet and EfficientNetB0 for binary classifcation. Our model outperforms these SOTA models in terms of accuacy, specificity, precision, sensitivity (recall) and F1-score as can be seen from Table 8. Our proposed model with class imbalance solution shows improvement in accuracy by 3.8, 25.5, and 15.7 (%) against GoogleNet [33, 34], Efficeint NetB0 [35, 36], XceptionNet [37], Inception Recurrent Residual Convolutional Neural Network (IRRCNN) [13], BAM [24], MDFF-Net [23], and R-DeCAF [19]. Also, we can see an improvement in 10.2, 57.4, and 66.2 (%) improvement in Sensitvity, 0.8, 11.3, and 4.6 (%) improvement in Specificity against first three models.

4.4 Statistical analysis

4.4.1 ROC curve and area under the ROC curve

We used ground truth labels and the predicted scores of the benign and malignant classes and their subclasses to draw ROC curves. Using our proposed model we found classification

Table 8 Comparison of proposed model against state-of-the art models

	ACC	Sensitivity	Specificity	Precision	F1	MCC	Kappa
GoogleNet	94.0	89.3	96.2	92.2	90.2	86.4	86.0
EfficientNet B0	77.8	57.4	87.1	68.7	61.7	47.3	46.4
XceptionNet	84.4	66.2	92.7	83.5	72.0	63.7	61.6
IRRCNN	77.4	74.4	84.0	91.0	81.9	54.64	66.2
BAM	81.2	80.2	83.2	91.3	85.4	60.2	70.5
MDFF-Net	81.8	78.1	89.8	94.3	85.4	63.7	74.6
R-DeCAF	68.9	58.4	92.0	94.1	72.1	47.3	63.7
FA-VGG16 CI)	**97.7**	**98.4**	**96.9**	**97.0**	**97.7**	**97.7**	**95.4**

scores for binary classes and draw ROC curves. Figure 6a shows the ROC curves and AUC values for different experiments (binary classification) mentioned in Table 3. The AUC value for the proposed model with ADAM optimizer is 95.6%, w/ SGDM optimizer 89.6%, w/ class imbalance solution 99.2%, w/ patient wise classification 77.12%. We also draw ROC curves for quaternary classes. Figure 6b and c shows ROC curves for intraclass classification experiments for benign and malignant classes respectively. We acquire the AUC values of the proposed model for adenosis: 95.7%; fibroadenoma: 91.2%; phyllodes tumour: 94.5%; and tubular adenoma: 97.4% (all benign subclasses). Similarly, AUC values for ductal carcinoma: 97.8%; lobular carcinoma: 99.3%; mucinous carcinoma: 99.7% and papillary carcinoma: 99.7% (all malignant subclasses).

4.4.2 Paired-t-Test

Table 9 presents the statistical parameters of a paired t-test performed on different variations of the VGG16 model, including one with an attention module. The mean difference, standard deviation of differences, standard error of mean difference, 95% confidence interval, test statistic t, and degrees of freedom are provided for each variation. The two-tailed probability is also presented for each test, indicating the significance of the result.

Based on the results, the FA-VGG16 outperformed the other variations in terms of accuracy and other performance parameters. The test statistic t for the FA-VGG16 was highest among all the variations, with a value of 7.142, and a corresponding two-tailed probability of less than 0.0001, indicating that the results were statistically significant.

The results reflect the model's performance across a spectrum of experiments designed to test its robustness, accuracy, and generalizability.

The FA-VGG16 model's performance was assessed using a comprehensive set of metrics, including accuracy, sensitivity, specificity, precision, false positive rate (FPR), F1 score, Mathew's Correlation Coefficient (MCC), and Kappa index. These metrics offer a holistic view of the model's classification efficacy.

To summarize,:

1. **Accuracy**: Our model achieved a high accuracy rate of 97.7% post addressing the class imbalance problem. This signifies that the model correctly identified the vast majority of the breast histopathology images as benign or malignant. The improvement in accuracy from 89.3% (basic VGG16) to 97.7% (FA-VGG16 with class balancing) underscores the efficacy of our proposed model and methodologies in dealing with class imbalances.
2. **Sensitivity and Specificity**: These metrics evaluate the model's ability to correctly identify positive (malignant) and negative (benign) cases, respectively. Our proposed model exhibited high sensitivity (98.4%) and specificity (96.9%), indicating a strong capability to identify true malignant cases and correctly dismiss non-malignant ones.
3. **Precision and FPR**: Precision measures the proportion of true positive predictions in all positive predictions made by the model, while FPR assesses the likelihood of false alarms. Our FA-VGG16 model demonstrated high precision (97.0%) with a low FPR (3.1%), reflecting its precision in classification and a low rate of misclassification.
4. **F1 Score and MCC**: The F1 score is the harmonic mean of precision and sensitivity, providing a balance between the two in cases of uneven class distributions. The MCC is a correlation coefficient that gives a high-quality measure of the model's performance. Our model's F1 score (97.7%) and MCC (95.4%) are indicative of its exceptional balance and correlation between all aspects of binary classification.

Table 9 Paired t-Test statistical parameters for FA-VGG16

	VGG	VGG (Class balance)	ADAM	SGDM	Class Balance	Patient wise
Mean difference	0.004	0.027	-0.023	0.048	-0.007299	0.072
SD of differences	0.328	0.289	0.309	0.381	0.153	0.448
SE of mean difference	0.007	0.006	0.007	0.009	0.003	0.010
95% CI	-0.010 to 0.018	0.016 to 0.038	-0.036 to -0.009	0.031 to 0.064	-0.013 to -0.002	0.052 to 0.091
Test statistic t	0.547	4.9	-3.264	5.573	-2.502	7.142
DF	1994	2739	1994	1994	2739	1994
Two-tailed probability	$P=0.5846$	$P<0.0001$	$P=0.0011$	$P<0.0001$	$P=0.0124$	$P<0.0001$

SD Standard deviation, *DF* Degrees of Freedom, *SE* Standard Error

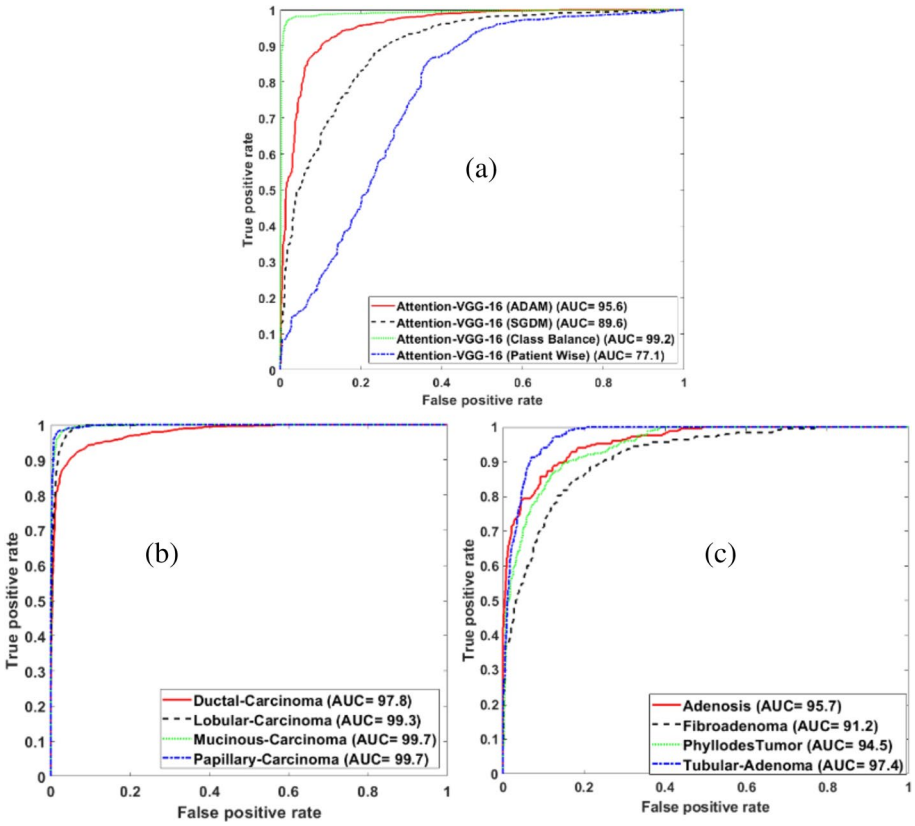

Fig. 6 Receiver operating Characteristics for the proposed model **a** with different hyperparameters for binary classification. **b** Quaternary classification benign **c** Quaternary classification malignant

5. **Kappa**: This statistic measures the agreement between predicted and actual classifications, correcting for chance agreement. A high Kappa value (95.3%) for our model suggests a strong agreement beyond chance.

Additionally, the paired t-test results with significant p-values ($p < 0.05$) for our model across various experimental setups provide statistical evidence supporting the superiority of our model over the basic VGG16 and other variations tested.

The results of our research are indicative of the FA-VGG16 model's superior performance in classifying breast histopathology images. The integrative analysis, including the ablation study and the robust statistical tests, provide compelling evidence of the model's potential for clinical application in aiding pathologists with diagnostic decision-making processes.

The results provide validation of the FA-VGG16 methodology across diverse experimental protocols. Comparisons against established models and comprehensive hyperparameter assessments substantiate the benefits of incorporating attention mechanisms within the classification architecture. While the findings demonstrate promise for facilitating automated breast cancer diagnosis, a detailed analysis of specific technical components comprising the research design warrants consideration. In the subsequent section, we offer a

rigorous exploration of salient elements incorporated in the study. This encompasses addressing class imbalance via resampling techniques, conducting ablation experiments to establish robustness, and evaluating generalizability on varied datasets. Further progress in this domain of significant clinical importance may be achieved by comprehensively investigating how prior gaps are addressed.

5 Discussion

So far we present a novel method of classification of breast cancer histopathological images known as FA-VGG16. In this method, we used 4 channel attention modules in a basic VGG-16 network. These attention modules provide a remedy to the vanishing gradient problem as well as improve the feature extraction capability of the model. In this section, we'll discuss some important aspects of our study, which fill the gap from previous studies.

5.1 Class imbalance solution

Previous studies suffered from a class imbalance problem. The database has a high class-imbalance (1:2) between benign and malignant classes. The effect of class imbalance can be seen in the performances of the models as well. The accuracy of the VGG model improves from 89.27 to 91.61% with class balance. With our proposed model (FA-VGG16) this improvement is much higher i.e. from 90.43 to 97.66%. Various class imbalance solutions have been used by researchers in previous studies such as SMOTE (Synthetic Minority Oversampling Technique) used by [27]. for multiclass coronary artery disease prediction and by [28] for Melanoma detection. A variation of this technique SMOTE-NC is used by the [29] for Covid-19 severity prediction using blood sample data. Authors in [9] used the SMOTE technique for breast cancer mammography image classification. Similarly, [30] used SMOTE for multiple data classifications including Wisconsin Breast cancer data classification. However, their approach is not empirical. Thus we identify the problem of class imbalance that exists in breast cancer classification and use the oversampling technique in the minority class.

Focal loss is a type of loss function that is commonly used in deep learning models to address the problem of class imbalance by down-weighting the contribution of easy examples (i.e., those where the model is confident in its prediction and the prediction is correct) and up-weighting the contribution of hard examples (i.e., those where the model is uncertain in its prediction or the prediction is incorrect).

This weighting, (also refered to as focal weight), modulates the standard cross-entropy loss based on the model's predicted probability for the true class. The focal weight is higher for examples that the model is less confident in (i.e., where the predicted probability is low), and lower for examples that the model is more confident in (i.e., where the predicted probability is high). By using the focal loss function, the model is encouraged to focus more on the examples that are most difficult to classify, which can lead to improved performance on imbalanced datasets. Additionally, the focal loss function can help the model to achieve a better balance between precision and recall, especially for the minority class, leading to improved overall performance.

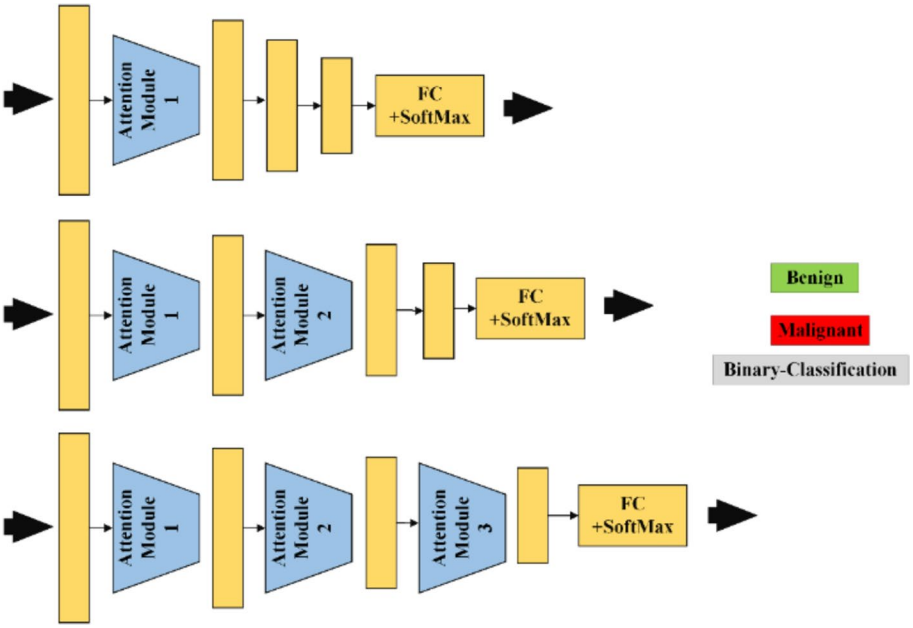

Fig. 7 Ablation study block diagram. Successive ablation of attention modules in different experiments (from top to down)

Table 10 Ablation study results

Ablation study	ACC	Error	Sens	Spec	Prec	FPR	F1	MCC	Kappa
Attention-VGG16 (Ablation-1)	88.7	11.3	78.6	93.4	86.0	6.6	81.2	74.2	73.2
Attention-VGG16 (Ablation-2)	88.8	11.2	74.6	95.3	89.0	4.7	80.2	73.8	72.6
Attention-VGG16 (Ablation-3)	90.8	9.2	78.7	96.4	90.9	3.7	84.3	78.3	77.9
Attention-VGG16 (LF2 + O2)	92.6	7.4	89.9	93.8	87.0	6.2	88.4	83.0	82.9
Attention VGG (FL + SGDM)	85.2	14.8	68.8	92.7	81.5	7.3	74.4	64.7	64.1
Attention-VGG16(LF2 + O2 + FC1024)	91.5	8.5	75.2	98.9	96.9	1.1	84.7	80.1	78.9

O2 = SGDM, FL2 = Cross Entropy Loss

5.2 Ablation study

In the ablation study, we sequentially removed the attention modules from our proposed model to evaluate their impact. The schematic ablation study is shown in Fig. 7. The results of all ablation studies are presented in Table 10. In the ablation-1 study, we removed 3 attention blocks (#2, #3, #4), in the ablation-2 study we removed 2 attention blocks (#3, #4) and in the ablation-3 study, we removed only 1 attention block (#4). The results of these ablated attention modules are shown in Table 8. Table 8 displays the progressive improvement in accuracy and other performance parameters with the addition of an ablation module to the basic VGG16. Similarly, the bar chart in Fig. 8 illustrates the performance

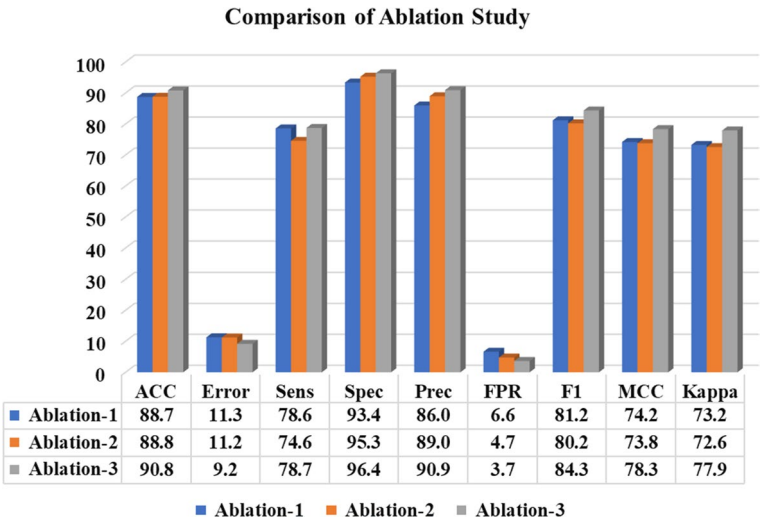

Fig. 8 Ablation study analysis using a bar chart

parameter increments. We conducted additional ablation experiments by changing the loss function and optimizer, yielding an accuracy of 92.58%. The accuracy dropped to 91.48% by decreasing the weight of the last fully connected layer to 1024, while the loss significantly decreased.

The present research presents a novel application of the attention mechanism in the VGG16 classification network. The strength of our study is that we have covered various aspects of our deep learning model such as ablation study, changing database experiment, and generalizability. We have also taken care of the class imbalance problem by oversampling the minority class data. To provide the variation in the database we augmented it using the standard techniques of translation, reflection, rotation, and scaling. Thus, we have overcome the bias in our study by varying the degree of the database. We have also tried to overcome the bias related to the similarity in the database by using all images of the same patients for training only i.e., patient-wise training and testing of the model. Moreover, we have tested our model not only for binary classification i.e., benign, malignant images, but we have also deployed the same model for quaternary classification of intraclass benign and malignant images. The results of intraclass quaternary classification are also encouraging.

Our experiments were validated through performance evaluation and benchmarking with previous studies. While our study presents a novel application of attention in the VGG16 classification network, a limitation is that we used only one classification network and applied an attention module to it. However, the potential of other DL models with parallel classification modules remains unexplored. Future research will examine the impact of attention on these models. Additionally, reducing bias can be achieved by including diverse databases from different geographic areas and ethnicities. We would like to increase the explainablity of FA-VGG [31] and explore more on self-supervise localization [32] of anomaly of pathology images.

We have also compared the running time of the model with other SOTA models and recent methods (See Fig. 9). The model does not outperform other existing models when

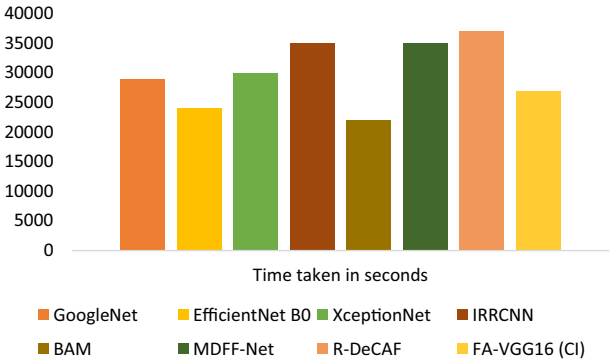

Fig. 9 Time taken in seconds for GoogleNet, EfficientNet B0, XceptionNet, IRRCNN, BAM, MDFF-Net, R-DeCAF, and the proposed FA-VGG16 (CI) on BreakHis dataset

trained on the BreakHis dataset under the same conditions in terms of FLOPS and hence the total time taken in seconds. Our proposed model is not the best in terms of FLOPS performing but better than the other five comparable methods shown in Fig. 9. We have not worked much to reduce FLOPS as our goal in this paper was to focus more on optimizing accuracy rather than lightweight model. The Intel(R) Core (TM) i9-10900 K CPU @ 3.70 GHz 3.70 GHz with NVIDIA RTX A4000 80 GB high-performance computers, and matlab2023a and python were used for this study. The time fluctuates is majorly due to the reduction of node by fast forwarding the input for specific layers which reduces the time taken and makes the model little faster.

6 Conclusion

This study presents a clinical decision support system for breast cancer classification. An iterative classification into benign and malignant subcategories is proposed, enhancing pathologists' decision-making. Performance evaluation and statistical tests confirm the model's effectiveness, indicating a significant clinical impact. The attention module integrated with VGG16 model has better feature extraction and generalization capability than other models. The experimental results are free from data selection, data exclusion, and data sampling bias. Results and analysis suggest broad clinical significance for our proposed model. Moving forward, several prospective extensions building directly upon this research warrant consideration. Validation on multi-institutional datasets could assess generalizability across patient cohorts and protocols. Exploring joint image-clinical models may reveal predictive synergies. Leveraging attention for classification-localization may provide interpretable localization assistance. Adapting the approach for survival analytics and incremental learning schemes could evaluate ability to predict prognosis and facilitate continuous performance updates utilizing emerging cases. Undertaking such pragmatic future work serves to demonstrate translational readiness, addressing real-world demands through rigorous testing of computational pathology advances.

Funding This research work was supported by the RFIER-Jio Institute "CVMI-Computer Vision in Medical Imaging" research project (RFIER-Jio Institute, Grant # 2022/33185004) fund under the "AI for ALL" research center. The funding is used to design the study and collection, analysis, and interpretation of data and in writing the manuscript as well.

Data availability Database used in this work is a public dataset and can be accessed via the link mentioned in [8]. For ease of readers, link is given : http://web.inf.ufpr.br/vri/breast-cancer-database.

GitHub Source code link:

Source code for this work is available publicly and can be accessed via the link below.

https://github.com/labsroy007/Attention-VGG16.

Declarations

Competing interests Authors have no competing interest with any organization or person.

References

1. Giaquinto AN et al (2022) Breast Cancer Statistics, 2022. CA Cancer J Clin 72(6):524–541. https://doi.org/10.3322/caac.21754
2. Beňačka R, Szabóová D, Guľašová Z, Hertelyová Z, Radoňák J (2022) Classic and New Markers in Diagnostics and classification of breast Cancer. Cancers 14(21):5444. https://doi.org/10.3390/cancers14215444
3. Smolarz B, Nowak AZ, Romanowicz H (2022) Breast Cancer—epidemiology, classification, Pathogenesis and treatment (review of literature). Cancers 14(10):2569. https://doi.org/10.3390/cancers14102569
4. Aljuaid H, Alturki N, Alsubaie N, Cavallaro L, Liotta A (2022) Computer-aided diagnosis for breast cancer classification using deep neural networks and transfer learning. Comput Methods Programs Biomed 223:106951. https://doi.org/10.1016/j.cmpb.2022.106951
5. Liu P et al (2022) Artificial intelligence to detect the femoral intertrochanteric fracture: the arrival of the intelligent-medicine era. Front Bioeng Biotechnol 10:927926. https://doi.org/10.3389/fbioe.2022.927926
6. Roy S, Whitehead TD, Li S, Ademuyiwa FO, Wahl RL, Dehdashti F, Shoghi KI (2022) Co-clinical FDG-PET radiomic signature in predicting response to neoadjuvant chemotherapy in triple-negative breast cancer. Eur J Nucl Med Mol Imaging 1–13. https://doi.org/10.1007/s00259-021-05489-8
7. Roy S, Whitehead TD, Quirk JD, Salter A, Ademuyiwa FO, Li S, An H, Shoghi KI (2020) Optimal co-clinical radiomics: Sensitivity of radiomic features to tumour volume, image noise and resolution in co-clinical T1-weighted and T2-weighted magnetic resonance imaging. EBioMedicine 59. https://doi.org/10.1016/j.ebiom.2020.102963
8. Sudipta R, Shoghi KI (2019) Computer-aided tumor segmentation from T2-weighted MR images of patient-derived tumor xenografts. In: Image Analysis and Recognition: 16th International Conference, ICIAR 2019, Waterloo, ON, Canada, August 27–29, 2019, Proceedings, Part II 16. Springer International Publishing, pp 159–171
9. Roy S, Bhattacharyya D, Bandyopadhyay SK, Tai-Hoon K (2017) An improved brain MR image binarization method as a preprocessing for abnormality detection and features extraction. Front Comput Sci 11:717–727
10. Roy S, Bhattacharyya D, Bandyopadhyay SK, Tai-Hoon K (2017) An iterative implementation of level set for precise segmentation of brain tissues and abnormality detection from MR images. IETE J Res 63(6):769–783
11. Roy S, Bandyopadhyay SK (2016) A new method of brain tissues segmentation from MRI with accuracy estimation. Procedia Comput Sci 85:362–369
12. Spanhol FA, Oliveira LS, Petitjean C, Heutte L (2016) A dataset for breast Cancer histopathological image classification. IEEE Trans Biomed Eng 63(7):1455–1462. https://doi.org/10.1109/TBME.2015.2496264
13. Alom MZ, Yakopcic C, Nasrin MS, Taha TM, Asari VK (2019) Breast cancer classification from histopathological images with inception recurrent residual convolutional neural network. J Digit Imaging 32(4):605–617. https://doi.org/10.1007/s10278-019-00182-7
14. Sarker MMK, Akram F, Alsharid M, Singh VK, Yasrab R, Elyan E (2022) Efficient breast Cancer classification network with dual squeeze and excitation in histopathological images. Diagnostics 13(1):103. https://doi.org/10.3390/diagnostics13010103
15. Srikantamurthy MM, Rallabandi VPS, Dudekula DB, Natarajan S, Park J (2023) Classification of benign and malignant subtypes of breast cancer histopathology imaging using hybrid CNN-LSTM based transfer learning. BMC Med Imaging 23(1):19. https://doi.org/10.1186/s12880-023-00964-0

16. Li J, Shi J, Chen J, Du Z, Huang L (2023) Self-attention random forest for breast cancer image classification. Front Oncol 13:1043463. https://doi.org/10.3389/fonc.2023.1043463
17. Al-Jabbar M, Alshahrani M, Senan EM, Ahmed IA (2023) Analyzing histological images using hybrid techniques for early detection of multi-class breast Cancer based on Fusion features of CNN and Handcrafted. Diagnostics 13(10):1753. https://doi.org/10.3390/diagnostics13101753
18. Mudeng V, Farid MN, Ayana G, Choe S (2023) Domain and histopathology adaptations–based classification for Malignancy Grading System. Am J Pathol 193(12):2080–2098. https://doi.org/10.1016/j.ajpath.2023.07.007
19. Morovati B, Lashgari R, Hajihasani M, Shabani H (2023) Reduced deep convolutional activation features (R-DeCAF) in histopathology images to improve the classification performance for breast cancer diagnosis. J Digit Imaging 36(6):2602–2612. https://doi.org/10.1007/s10278-023-00887-w
20. Ashurov A, Chelloug SA, Tselykh A, Muthanna MSA, Muthanna A, Al-Gaashani MSAM (2023) Improved Breast Cancer Classification through Combining Transfer Learning and Attention Mechanism. Life 13(9):1945. https://doi.org/10.3390/life13091945
21. Abdallah N, Marion J-M, Tauber C, Carlier T, Hatt M, Chauvet P (2023) Enhancing histopathological image classification of invasive ductal carcinoma using hybrid harmonization techniques. Sci Rep 13(1):20014. https://doi.org/10.1038/s41598-023-46239-0
22. Ogundokun RO, Misra S, Akinrotimi AO, Ogul H (2023) MobileNet-SVM: a lightweight deep transfer learning model to diagnose BCH scans for IoMT-Based imaging sensors. Sensors 23(2):656. https://doi.org/10.3390/s23020656
23. Xu C, Yi K, Jiang N, Li X, Zhong M, Zhang Y (2023) MDFF-Net: a multi-dimensional feature fusion network for breast histopathology image classification. Comput Biol Med 165:107385. https://doi.org/10.1016/j.compbiomed.2023.107385
24. Alirezazadeh P, Dornaika F (2023) Boosted additive angular margin loss for breast cancer diagnosis from histopathological images. Comput Biol Med 166:107528. https://doi.org/10.1016/j.compbiomed.2023.107528
25. Jakkaladiki SP, Maly F (2023) An efficient transfer learning based cross model classification (TLBCM) technique for the prediction of breast cancer. PeerJ Comput Sci 9:e1281. https://doi.org/10.7717/peerj-cs.1281
26. Kabiraj A, Meena T, Reddy PB et al (2024) Multiple thoracic diseases detection from X-rays using CX-Ultranet. Health Technol 14:291–303. https://doi.org/10.1007/s12553-024-00820-3
27. Raghav S, Suri A, Kumar D, Aakansha A, Rathore M, Roy S (2023) A hierarchical clustering approach for identification of colorectal cancer molecular subtypes from gene expression data. Intell Med. https://doi.org/10.1016/j.imed.2023.04.002
28. Chang C-C, Li Y-Z, Wu H-C, Tseng M-H (2022) Melanoma Detection using XGB Classifier combined with feature extraction and K-Means SMOTE techniques. Diagnostics 12(7):1747. https://doi.org/10.3390/diagnostics12071747
29. Chakraborty S, Kumar K, Tadepalli K et al (2023) Unleashing the power of explainable AI: sepsis sentinel's clinical assistant for early sepsis identification. Multimed Tools Appl. https://doi.org/10.1007/s11042-023-17828-y
30. Wang S, Dai Y, Shen J, Xuan J (2021) Research on expansion and classification of imbalanced data based on SMOTE algorithm. Sci Rep 11(1):24039. https://doi.org/10.1038/s41598-021-03430-5
31. Roy S, Pal D, Meena T (2024) Explainable artificial intelligence to increase transparency for revolutionizing healthcare ecosystem and the road ahead. Netw Model Anal Health Inf Bioinforma 13:4. https://doi.org/10.1007/s13721-023-00437-y
32. Kumar K, Chakraborty S, Roy S (2023) Self-supervised Diffusion Model for Anomaly Segmentation in Medical Imaging. In: Maji P, Huang T, Pal NR, Chaudhury S, De RK (eds) Pattern Recognition and Machine Intelligence. PReMI 2023, vol 14301. Springer, Cham. https://doi.org/10.1007/978-3-031-45170-6_37
33. Lai ZF, Zhang G, Zhang XB, Liu HT (2022) High-resolution histopathological image classification model based on fused heterogeneous networks with self-supervised feature representation. Biomed Res Int 2022:8007713. https://doi.org/10.1155/2022/8007713
34. Szegedy C, Ioffe S, Vanhoucke V, Alemi AA (2017) Inception-v4, inception-resnet and the impact of residual connections on learning. Proc AAAI Conf Artif Intell 31(1). https://doi.org/10.1609/aaai.v31i1.11231
35. Rahaman MM, Millar EKA, Meijering E (2023) Breast cancer histopathology image-based gene expression prediction using spatial transcriptomics data and deep learning. Sci Rep 13:13604. https://doi.org/10.1038/s41598-023-40219-0

36. Venkatesh RK, Sheela Y, Nagaraju, Sahu DA (2022) Histopathological image classification of breast cancer using EfficientNet. 2022 3rd International Conference for Emerging Technology (INCET), Belgaum, pp 1–8. https://doi.org/10.1109/INCET54531.2022.9824351
37. Lu X, Firoozeh Abolhasani Zadeh YA (2022) Deep learning-based classification for Melanoma Detection using XceptionNet. J Healthc Eng 2022:2196096. https://doi.org/10.1155/2022/2196096