scientific reports

Check for updates

OPEN

An explainable AI-driven deep neural network for accurate breast cancer detection from histopathological and ultrasound images

Md. Romzan Alom[1], Fahmid Al Farid[2], Muhammad Aminur Rahaman[1]✉, Anichur Rahman[3,4]✉, Tanoy Debnath[5], Abu Saleh Musa Miah[6] & Sarina Mansor[2]✉

Breast cancer represents a significant global health challenge, which makes it essential to detect breast cancer early and accurately to improve patient prognosis and reduce mortality rates. However, traditional diagnostic processes relying on manual analysis of medical images are inherently complex and subject to variability between observers, highlighting the urgent need for robust automated breast cancer detection systems. While deep learning has demonstrated potential, many current models struggle with limited accuracy and lack of interpretability. This research introduces the Deep Neural Breast Cancer Detection (DNBCD) model, an explainable AI-based framework that utilizes deep learning methods for classifying breast cancer using histopathological and ultrasound images. The proposed model employs Densenet121 as a foundation, integrating customized Convolutional Neural Network (CNN) layers including GlobalAveragePooling2D, Dense, and Dropout layers along with transfer learning to achieve both high accuracy and interpretability for breast cancer diagnosis. The proposed DNBCD model integrates several preprocessing techniques, including image normalization and resizing, and augmentation techniques to enhance the model's robustness and address class imbalances using class weight. It employs Grad-CAM (Gradient-weighted Class Activation Mapping) to offer visual justifications for its predictions, increasing trust and transparency among healthcare providers. The model was assessed using two benchmark datasets: Breakhis-400x (B-400x) and Breast Ultrasound Images Dataset (BUSI) containing 1820 and 1578 images, respectively. We systematically divided the datasets into training (70%), testing (20%,) and validation (10%) sets, ensuring efficient model training and evaluation obtaining accuracies of 93.97% for B-400x dataset having benign and malignant classes and 89.87% for BUSI dataset having benign, malignant, and normal classes for breast cancer detection. Experimental results demonstrate that the proposed DNBCD model significantly outperforms existing state-of-the-art approaches with potential uses in clinical environments. We also made all the materials publicly accessible for the research community at: https://github.com/romzanalom/XAI-Based-Deep-Neural-Breast-Cancer-Detection.

Keywords Breast cancer, XAI, DNBCD, CNN, Transfer learning, Breakhis-400x, BUSI, Grad-CAM

Cancer remains a critical public health challenge worldwide, and Breast Cancer (BC) stands as the most common malignancy globally, representing approximately 12.4% of all new cancer cases each year[1,2]. In the United States alone, 2024 saw an estimated 313,520 new cases of invasive BC and 310,720 new cases of non-

[1]Department of Computer Science and Engineering, Green University of Bangladesh (GUB), Purbachal American City, Kanchon, Dhaka 1460, Bangladesh. [2]Faculty of Artificial Intelligence and Engineering, Multimedia University, 63100 Cyberjaya, Malaysia. [3]Department of Computer Science and Engineering, National Institute of Textile Engineering and Research (NITER), Constituent Institute of the University of Dhaka, Savar, Dhaka 1350, Bangladesh. [4]Department of Computer Science and Engineering, Mawlana Bhashani Science and Technology University, Tangail, Bangladesh. [5]Department of Computer Science, Stony Brook University, Stony Brook, NY, USA. [6]Department of Computer Science and Engineering, Bangladesh Army University of Science and Technology (BAUST), Nilphamari, Bangladesh. ✉email: aminur@cse.green.edu.bd; anis_cse@niter.edu.bd; sarina.mansor@mmu.edu.my

invasive BC among women, with an additional 2,800 new cases of invasive BC diagnosed in men[3]. Breast cancer is one of the most commonly diagnosed cancers among American women, accounting for about 30% of all new cancer diagnoses in this population[4]. Early detection is essential for improving survival rates and reducing the morbidity associated with BC, given its diverse nature comprising various entities with distinct biochemical, histological, and clinical characteristics[5,6].

There are two major categories of high-risk breast cancer: benign and malignant. Benign breast cancer refers to non-cancerous growths in breast tissue[7]. These tumors do not invade surrounding tissues or spread to other parts of the body, and while they may cause discomfort or concern, they are typically not life-threatening. Common benign breast conditions include fibroadenomas and cysts[8]. On the other hand, malignant breast cancer involves cancerous growths that can invade surrounding breast tissue and metastasize to other parts of the body. Malignant tumors are aggressive and require prompt treatment to prevent spread and reduce the risk of mortality. Types of malignant breast cancer include ductal carcinoma in situ (DCIS), invasive ductal carcinoma (IDC), and invasive lobular carcinoma (ILC)[9,10].

Traditional diagnostic approaches, including mammography, ultrasound imaging, and magnetic resonance imaging (MRI)[11], form the first line of clinical screening[12,13]. However, these non-invasive techniques sometimes fail to detect malignant lesions accurately, necessitating a more definitive diagnosis through biopsy. The biopsy process involves collecting tissue samples, preparing them on glass slides, staining, and examining them under a microscope by pathologists to identify cancerous cells[14]. While this method is accurate, it is time-consuming and heavily reliant on the expertise of pathologists, highlighting a significant gap in the current diagnostic landscape. This gap underscores the urgent need for automated, accurate, and efficient diagnostic systems that can enhance user experience and provide clearer insights into the diagnostic process.

Nowadays, machine learning has advanced healthcare by improving disease diagnosis and patient monitoring[15,16]. Traditional methods rely on manual feature extraction, which is time-consuming and requires domain expertise, while deep learning automates this process, enabling direct learning from raw data[17,18]. Convolutional Neural Networks (CNNs), in particular, are popular deep learning approaches for image-based cancer detection due to their ability to learn intricate patterns within complex data[19]. However, deep learning techniques face challenges like the need for big labeled datasets, high computational demands, and limited accessibility in resource-constrained settings[20]. To address these issues, transfer learning[21,22] leverages pre-trained models like Densenet121, Mobilenet, Resnet50 and VGG19, reducing data requirements and improving performance[23–25]. However, despite their predictive accuracy[20], traditional deep learning models often operate as "black boxes," providing limited insight into the decision-making process and failing to show which areas of the image are most influential in reaching a diagnosis[26]. This lack of interpretability poses a challenge in clinical settings, as clinicians need transparency and clarity in diagnostic tools to trust and effectively integrate them into practice[27]. Furthermore, many models focus solely on feature extraction without adding interpretative layers, which restricts their utility in real-world applications where the reasoning behind predictions is critical.

Motivated by the need for early and precise BC detection, our work introduces the Deep Neural Breast Cancer Detection (DNBCD) system. This system leverages advances in deep learning, specifically Densenet121 and Convolutional Neural Networks (CNNs) with Transfer Learning, to automate the diagnostic process. We developed an advanced DNBCD to accurately classify benign, malignant, and normal breast tissue cases using the Breakhis-400x dataset (B-400x)[28], which provides 400×-resolution[29] histopathological images, and the Breast Ultrasound Images Dataset (BUSI)[30], which provides breast ultrasound images[31]. Our system aims not only to improve the accuracy and efficiency of BC detection but also to enhance the interpretability of the model's outputs. To this end, we integrated Gradient-weighted Class Activation Mapping (Grad-CAM)[32], which provides visual explanations by highlighting and marking regions of the input image that are most influential in the model's decision-making process, thereby offering transparency and trust in the automated diagnosis. Our contributions are multifaceted:

- In this study, we present the development of the DNBCD model, a combination of Densenet121 with additional CNN layers and transfer learning for BC detection with a focus on classifying benign, malignant, and normal cases with high accuracy.
- Secondly, we implement Grad-CAM to ensure that the model's predictions are interpretable, allowing clinicians to see which parts of the tissue images are being used to make diagnostic decisions.
- Thirdly, we address class imbalance in the datasets using class weighting, improving detection performance for both benign, malignant, and normal cases.
- Fourthly, we conduct a statistical analysis of our model and prove its significance with other state-of-the-art models.
- Finally, we conduct a thorough evaluation of our model's performance, comparing it against other existing methods and state-of-the-art techniques to demonstrate its superiority in accuracy 93.97% for B-400x[28] and 89.87% for BUSI[30] datasets.

The goals of this research align closely with several key United Nations Sustainable Development Goals (SDGs). Our work supports Goal 3: Good Health and Well-being, which advocates for reducing premature mortality from non-communicable diseases through early diagnosis and improved treatment. By advancing an automated, precise detection system, we aim to facilitate timely intervention in breast cancer cases, potentially reducing mortality rates. Moreover, the DNBCD system contributes to Goal 9: Industry, Innovation, and Infrastructure by exemplifying the integration of cutting-edge technology in healthcare to enhance accessibility, affordability, and effectiveness in cancer diagnostics. By developing a tool that can be implemented in resource-constrained environments, we further address Goal 10: Reduced Inequalities, ensuring that populations in low-resource areas gain equitable access to life-saving diagnostic technologies.

The structure of this paper is as follows: Section "Literature review" reviews the existing literature on breast cancer detection, focusing on studies using the B-400x, BUSI and others datasets. Section "Proposed methodology of DNBCD" outlines the methodology employed in developing the DNBCD system, covering the dataset characteristics, preprocessing methods, and DNBCD model architecture. Section "Result analysis" details the experimental results and evaluates the performance of the DNBCD system. Also, Section "Limitations and future works" presents the limitations and future scopes of the presented work. Finally, Section "Conclusion" provides an analysis of the findings, examining the system's potential impact on clinical decision-making and patient outcomes.

Literature review

Breakhis-400x dataset (B-400x)

Ogundokun et al.[33] proposed a hybrid CNN-ANN modelB-400x[28] dataset. Their model achieved an overall accuracy of 89.47% and an accuracy of 89.15% at the 400x magnification level. By combining CNN for feature extraction and ANN for classification, the hybrid model improved performance over standalone CNN and ANN models. However, the limitations of this approach include relatively low recall and the potential for overfitting due to the complexity of the model, which could be mitigated by optimizing hyperparameters and adjusting the architecture.

Ahmed et al.[34] introduced the Quantum-Optimized AlexNet (QOA) model to enhance breast cancer detection using the B-400x dataset[28]. This model combines AlexNet's feature extraction with a quantum layer functioning as a linear layer, achieving an accuracy of 93.67% at 400× magnification. The research highlights the potential of quantum computing to improve deep learning performance in medical imaging, particularly for breast cancer diagnosis. However, it does not fully address practical limitations, such as the scalability of quantum layers and their applicability in clinical settings, which may require further investigation.

Gupta et al.[35] proposed a hybrid deep transfer learning model that combines Xception with Support Vector Classifier (XSV) and Random Forest (XRF) to improve breast cancer tumour classification using histopathological images. The models were compared with classifiers such as Random Forest, Logistic Regression, Support Vector Classifier, K-Nearest Neighbors, and AdaBoost and evaluated on the BreakHis dataset[28] at various magnifications. At 400× magnification, the XSV model achieved an accuracy of 88.98%, while the XRF model reached 87.61%. The overall accuracy of the XSV model across all magnifications was 90.17%. Limitations include dataset imbalance, particularly fewer malignant images, which may affect performance. Future work could focus on enhancing generalization and addressing data limitations.

Ogundokun et al.[36] proposed a lightweight deep transfer learning model, Mobilenet-SVM, designed to diagnose breast cancer histology (BCH) images for Internet of Medical Things (IoMT) applications. The model combines Mobilenet and Support Vector Machine (SVM) to classify BCH images. Tested on the BreakHis dataset[28], the model achieved a test accuracy of 91.50%, along with an F1-score of 91.35%. The approach balances computational efficiency and high accuracy, making it suitable for IoMT-based imaging sensors in resource-constrained environments.

Yamlome et al.[37] proposed a CNN-based classifier to improve breast cancer histopathology image classification using transfer learning and data augmentation. The model was trained on high-resolution whole images from the BreakHis dataset[28], achieving an average magnification-level accuracy of 91.60%. Their approach outperformed prior methods by up to 6% in magnification accuracy and 2% in patient scores by utilizing whole images instead of patches, allowing the network to capture more global features. However, for images at 400× magnification, the model achieved an accuracy of 89.15%. This technique enhances the potential for more accurate breast cancer diagnosis through improved image classification.

Gupta et al.[38] proposed a partially independent framework for classifying breast cancer histopathology images using deep multi-layered features from a fine-tuned Resnet model. Their approach integrates features from multiple layers of the network, incorporating both low- and mid-level features along with high-level ones. Tested on the BreaKHis dataset[28], the method achieved an accuracy of 90.85% at 400× magnification. This framework outperformed traditional single-layer approaches and demonstrated the effectiveness of utilizing multi-layered features for improved classification performance.

Breast ultrasound images dataset (BUSI)

Munteanu et al.[39] proposed an end-to-end deep learning model for breast cancer detection using ultrasound images. Their model integrates GAN-based data augmentation, UNet for segmentation, and CNN for classification. The key contribution of their work is addressing data limitations, enhancing the training set with synthetic images, and achieving 86% accuracy. However, the limitation lies in their use of a single public dataset, which limits the model's generalizability and involved high computation cost.

Pacal et al.[40] proposed a deep learning approach for classifying breast cancer using ultrasound images. They evaluated several models, including AlexNet, VGG16, Resnet, GoogleNet, EfficientNet, and Vision Transformer. The Vision Transformer achieved the highest accuracy at 88.6%, outperforming other CNN models. However, the study noted limitations due to the small size of the BUSI dataset[30], which constrained the performance of deeper models. Future work could benefit from larger datasets and advanced data augmentation techniques.

Alotaibi et al.[41] proposed a breast cancer classification method utilizing convolutional neural networks (CNN) and image fusion techniques on ultrasound images. Their approach included a three-step preprocessing scheme: speckle noise filtering, region of interest (ROI) highlighting, and RGB fusion. When tested on the BUSI[30] and KAIMRC datasets, the VGG19 model achieved accuracies of 87.8% and 85.2%, respectively. However, the study faced limitations due to the relatively small size of the BUSI[30] dataset, which impacts the model's generalization ability, high computation cost and the imbalance among benign, malignant, and normal classes, which challenges consistent performance.

Isik et al.[42] proposed a meta-learning-based model for few-shot classification of the BUSI[30] dataset, utilizing Prototypical Networks (ProtoNet) and Model-Agnostic Meta-Learning (MAML). The highest accuracy achieved was 88.9% using ProtoNet with a Resnet50 backbone in a 10-shot setting, significantly surpassing the baseline accuracy of 83.1% with transfer learning. However, the model's limitations include dependency on dataset similarity for cross-domain training and high computational demands with deeper backbones like Resnet50.

Gheflati et al.[43] investigated the application of Vision Transformer (ViT) models for classifying breast ultrasound (US) images using the BUSI dataset[30], which comprises 780 images. The ViT B/32 model achieved the highest accuracy of 86.7% and an AUC of 0.95, surpassing traditional CNNs like Resnet50, which had an accuracy of 85.3%. However, the small size of the BUSI[30] dataset limited model generalizability, and standard data augmentation techniques, such as cropping and rotation, had minimal effect on improving accuracy due to the specific characteristics of ultrasound images.

Tagnamas et al.[44] proposed SCA-InceptionUNeXt, a lightweight U-shaped network for medical image segmentation. It integrated a modified InceptionNeXt block for efficient feature extraction and a Spatial-aware Channel Attention (SCA) module for enhanced feature fusion. The model outperformed SOTA methods on four datasets, achieving 81.66% Dice on BUSI while using 26.11M fewer parameters than U-Net. However, it struggled with low performance in certain imaging modalities and lacked interpretability in model decisions.

Sirjani et al.[45] developed a deep learning model based on an enhanced InceptionV3 architecture for classifying breast lesions in ultrasound images. Key improvements included converting InceptionV3 modules to residual inception modules and optimizing hyperparameters. Trained on five datasets, including three public and two from imaging centers, the model was compared with 24 CNN architectures, achieving an accuracy of 0.81, precision of 0.83, recall of 0.77, F1 score of 0.80, AUC of 0.81, and RMSE of 0.18. While effective for classification, the study notes limitations in model generalization and the need for further clinical validation.

Zhang et al.[46] introduced the Hierarchical Attention-guided U-Net (HAU-Net), a hybrid CNN-transformer framework for breast lesion segmentation in ultrasound images. The model combines CNNs for local detail and transformers for long-range dependencies, integrating an L-G transformer block into U-Net skip connections and a cross-attention block (CAB) in the decoder. HAU-Net achieved Dice coefficients of 83.11% on the BUSI dataset[30], 88.73% on UDIAT, and 89.48% on BLUI, surpassing state-of-the-art segmentation methods.

Others dataset

Bilal et al.[47] introduced the Improved Quantum-Inspired Binary Grey Wolf Optimizer (IQI-BGWO) to optimize the Support Vector Machine (SVM) for breast cancer diagnosis. The study is limited to the Mammographic Image Analysis Society (MIAS) dataset, restricting generalizability. Despite this, IQI-BGWO-SVM improves classification accuracy and feature selection. This work advances medical imaging with quantum-inspired optimization.

Zeng et al.[48] proposed a deep learning-based Raman spectroscopy model for diagnosing HER2-positive and triple-negative breast cancer. The study achieved high accuracy (CNN: 91.11%); however, limitations include a small sample size (75 samples), lack of external validation, no direct comparison with clinical methods, and potential spectral variability due to external factors. Additionally, the approach may require further optimization for real-world clinical integration and robustness across diverse patient populations.

Ma et al.[49] developed a surface-enhanced Raman spectroscopy (SERS)-based approach using a composite Ag NPs PSi Bragg reflector SERS substrate for early breast cancer detection. The study reported high diagnostic accuracy (95%), specificity (96.7%), and sensitivity (93.3%) and a low-cost. However, it was constrained by a small dataset (60 serum samples), absence of external validation, and potential spectral variability. Further research is needed to validate its clinical applicability across larger and more diverse populations.

Vulli et al.[50] proposed a fine-tuned Densenet-169 model utilizing FastAI and the 1-Cycle policy for automated breast cancer metastasis detection. Trained on a refined PatchCamelyon (PCam) dataset from Camelyon16, the model achieved 97.4% accuracy, surpassing existing methods. A mobile application was introduced to support early diagnosis. While the model enhances sensitivity and specificity, challenges such as high computation cost, overfitting risks, extensive data augmentation, and manual hyperparameter tuning remain, potentially limiting usability for non-experts in clinical settings.

Srinivasu et al.[51] developed a CatBoost+MLP model for breast cancer diagnosis, integrating explainable AI (SHAP) and ANOVA-based feature selection. They used the Breast Cancer Wisconsin dataset (569 records) and achieved an accuracy of 99.3%. The model effectively handled categorical data and reduced overfitting. Performance was superior to conventional techniques. However, the study was limited by dataset size, which affected generalizability. The authors suggested future improvements, including stacking and voting techniques, to enhance robustness and interpretability. Their work contributed to advancing AI-driven diagnostic models for more accurate and transparent breast cancer prediction.

In this study we explored various deep learning and machine learning approaches for breast cancer diagnosis across histopathological, ultrasound, and others datasets. Table 1 provides a comprehensive overview of our explored work in this domain. Hybrid models combining feature extraction with classifiers have shown improvements, such as Ogundokun et al.[33] CNN-ANN, Ahmed et al.[34] Quantum-Optimized AlexNet (QOA), and Gupta et al.[35] Xception-SVM/Random Forest models. Similarly, Ogundokun et al.[36] introduced Mobilenet-SVM optimized for IoMT applications, while Yamlome et al.[37] applied transfer learning and data augmentation to enhance classification. Deep transfer learning approaches have been explored to improve classification, with Gupta et al.[38] utilizing multi-layered Resnet features, while Sirjani et al.[45] and Tagnamas et al.[44] enhanced InceptionV3 for ultrasound image classification. Pacal et al.[40] and Gheflati et al.[43] investigated Vision Transformer-based models, and Isik et al.[42] introduced meta-learning (ProtoNet-Resnet50) for few-shot classification. Segmentation-focused models have been studied, including Zhang et al.[46] HAU-Net, which integrates CNNs and transformers for breast lesion segmentation, and Munteanu et al.[39], who combined GAN-

Author	Model	Dataset	Accuracy (%)	Limitations
Ogundokun et al.[33]	Hybrid CNN-ANN	B-400x[28]	89.47	Overfitting problem
Ahmed et al.[34]	QOA	B-400x[28]	93.67	Scalability of quantum layers
Gupta et al.[35]	XSV and XRF	B-400x[28]	90.17	Dataset imbalance
Ogundokun et al.[36]	Mobilenet-SVM	BreakHis	91.50 (400x)	Complexity of model
Yamlome et al.[37]	CNN-based Classifier	BreakHis	91.60 (400x)	Accuracy drops at higher magnifications
Gupta et al.[38]	Fine-tuned Resnet	B-400x[28]	90.85	Complexity in multi-layer feature integration
Munteanu et al.[39]	UNet+CNN	BUSI[30]	86	Small dataset size and high computation cost
Pacal et al.[40]	Vision Transformer	BUSI[30]	88.6	Small dataset size
Alotaibi et al.[41]	VGG19	BUSI[30], KAIMRC	87.8 (BUSI[30])	Class imbalance and high computation cost
Isik et al.[42]	ProtoNet+Resnet50	BUSI[30]	88.9	Dependency on dataset similarity
Gheflati et al.[43]	ViT B/32	BUSI[30]	86.7	Small dataset affects generalizability
Tagnamas et al.[44]	SCA-InceptionUNeXt	Multiple datasets	81.66 (BUSI[30])	Low performance and lacked interpretability
Sirjani et al.[45]	Enhanced InceptionV3	Multiple datasets	81 (BUSI[30])	Model generalization issues
Zhang et al.[46]	HAU-Net	Multiple datasets	83.11 (BUSI[30])	Model generalization issues
Bilal et al.[47]	IQI-BGWO+SVM	MIAS	99.25	Model complexity constraints
Zeng et al.[48]	CNN	75 serum samples	91.11	Small dataset affects generalizability
Ma et al.[49]	SERS	60 serum samples	95	Small dataset affects generalizability
Vulli et al.[50]	Fine-tuned Densenet	PatchCamelyon	97.40	Model complexity constraints
Srinivasu et al.[51]	CatBoost+MLP	Wisconsin	99.3	Small dataset affects generalizability

Table 1. Summary of related research on breast cancer detection and classification.

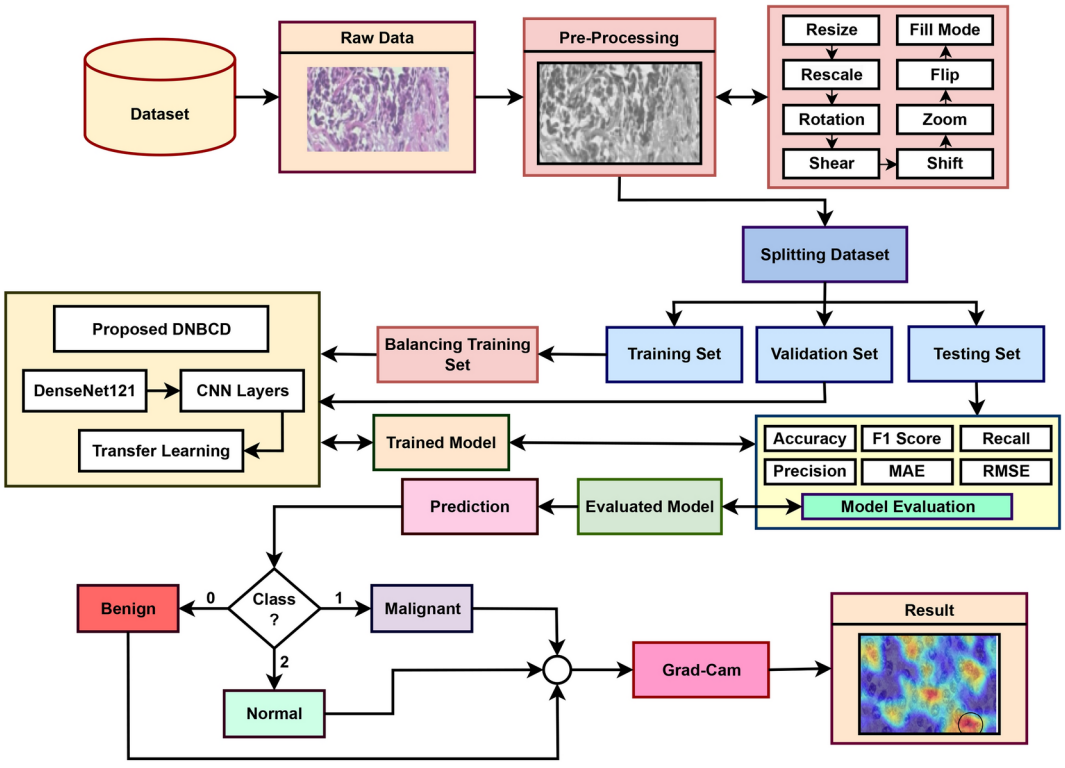


Fig. 3. Proposed methodology of DNBCD system.

based augmentation, UNet segmentation, and CNN classification. Alotaibi et al.[41] used image fusion and CNNs for ultrasound-based diagnosis. Raman spectroscopy and alternative approaches have also been investigated. Zeng et al.[48] used CNNs for HER2-positive and triple-negative breast cancer detection, while Ma et al.[49] employed SERS-based Ag NPs for early-stage diagnosis. Bilal et al.[47] applied IQI-BGWO for SVM optimization, improving feature selection and classification. Explainability and model interpretability have gained attention, with Srinivasu et al.[51] integrating CatBoost+MLP and SHAP-based explainability for breast cancer diagnosis. Vulli et al.[50] fine-tuned Densenet-169 using FastAI and the 1-Cycle policy, also developing a mobile application

for early detection. Despite these advancements, challenges still exist, like limitations in datasets such as class imbalance and small sample sizes, which affect how well the model can generalize. Our proposed model aims to tackle these issues by using strategies to improve performance and address data imbalance, ensuring that all classes are represented fairly. We also focus on making our model interpretable and transparent by using Grad-CAM techniques. These techniques help us understand how the model makes its decisions, making it easier for clinical adoption. In the future, research should keep looking into hyperparameter tuning, feature selection, scalable architectures, and explainable AI to further enhance transparency and reliability in breast cancer diagnostics.

Proposed methodology of DNBCD

As illustrated in Fig. 3, the proposed system's methodology is structured to effectively analyze and classify breast cancer images. Initially, the system takes input images from the dataset, which serves as the primary data source for the analysis. These images undergo preprocessing to enhance their quality and ensure suitability for subsequent stages. The dataset is then partitioned into training, testing, and validation sets using either a random or stratified splitting strategy. Subsequently, various models, including Densenet121[52], Mobilenet[53], Resnet50[54], VGG19[55], Alexnet[56], EfficientNetB0[57] and others are trained. Among those model we select Densenet121, Mobilenet, Resnet50, and VGG19 models for building our proposed model. Next, we modify these models by incorporating CNN layers and employing transfer learning techniques. The best-performing configuration combines Densenet121 with additional CNN layers and transfer learning, resulting in our proposed DNBCD model. After that we use this model for breast cancer classification. Finally, the system provides interpretable explanations of the classification results. The key steps of the proposed methodology are outlined as follows:

Dataset

The datasets utilized in this study are sourced from publicly available repositories and hospital records, containing labeled data for both benign and malignant cases. The quality and diversity of the images are critical in ensuring robust model performance. For both datasets, we partition the data into three subsets: 70% for training, 10% for validation and 20% for testing. Because in BUSI[30] dataset already splitting into 70-20-10 and most of the papers[39–43,45,46] are used this splitting on Breakhis-400x B-400x[28] and BUSI[30] dataset. So we used this splitting. Also we used effective score and variance for selecting this splitting.

- **Variance:** Variance measures how much data points differ from the mean[58]. It is crucial for dataset splitting because maintaining similar variance in both training and testing sets ensures accurate model evaluation and prevents overfitting. It defined as shown in Eq. (1).

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2 \tag{1}$$

Where σ^2 represents the variance, N is the number of data points, x_i denotes each data point, and μ is the mean of the data points.

- **Effective score:** The effective score accounts for the number of independent observations in correlated datasets[58]. It is essential for dataset splitting to accurately reflect the dataset's information content, especially in cases where data points are not independent. It defined as shown in Eq. (2).

$$T_{\text{eff}} = \frac{1}{\sigma_{\text{train}} + \sigma_{\text{test}}} \tag{2}$$

Where T_{eff} represents the effective score, σ_{train} is the variance of the training set, and σ_{test} is the variance of the test set.

Both the variance and effective score are very important for creating balanced datasets that help in assessing model performance reliably. In this study, Table 2 shows a comparison of effective scores and variance across different dataset splits. Here, σ_{train}, σ_{test}, and σ_{val} represent the variances of the training, testing, and validation sets, respectively. Although the best variance for the validation set is 0.0524 for the B-400x dataset and 0.0563 for the BUSI dataset when using the 60-20-20 ratio, the effective scores, which we refer to as $\mathrm{T_{eff}}$, are not the highest for this split. On the other hand, the 70-20-10 ratio gives us much higher effective scores, reaching 12.43 for the B-400x dataset and 11.56 for the BUSI dataset. These numbers indicate a better balance between how well the model performs and its ability to generalize. So, we chose the 70-20-10 ratio for our analysis because it offers a good trade-off between training efficiency and validation reliability, making the model's performance assessment stronger.

This split ensures effective model training, unbiased performance evaluation, and fine-tuning of hyperparameters, contributing to the overall reliability and accuracy of the system. And the splitting can be mathematically represented by Eqs. (3), (4), and (5).

For Training Set (S_{train}):

$$S_{\text{train}} = S_{\text{train, benign}} \cup S_{\text{train, malignant}} \cup S_{\text{train, normal}} \tag{3}$$

For Testing Set (S_{test}):

Dataset	Split ratio	Train set	Test set	Validation set	σ_{train}	σ_{test}	σ_{val}	T_{eff}
B-400x[28]	90-05-05	1638	91	91	0.0247	0.1048	0.1048	7.73
	80-10-10	1456	182	182	0.0262	0.0741	0.0741	9.95
	70-20-10	1274	364	182	0.0280	0.0524	0.0741	**12.43**
	70-15-15	1274	273	273	0.0280	0.0605	0.0605	11.30
	60-20-20	1092	364	364	0.0302	0.0524	0.0524	12.11
BUSI[30]	90-05-05	1420	78	78	0.0265	0.1132	0.1132	7.15
	80-10-10	1262	157	157	0.0281	0.0798	0.0798	9.25
	70-20-10	1104	315	157	0.0301	0.0563	0.0798	**11.56**
	70-15-15	1104	236	236	0.0301	0.0651	0.0651	10.51
	60-20-20	946	315	315	0.0325	0.0563	0.0563	11.29

Table 2. A detailed analysis of dataset splitting for B-400x and BUSI datasets was performed to compare the effective score and variance across different sets. Significant values are in bold.

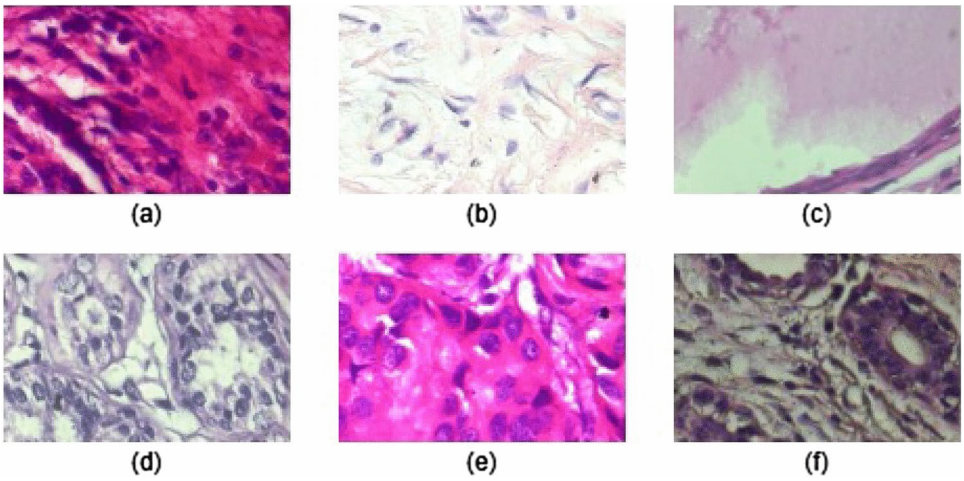


Fig. 1. Sample of data in BreakHis-400x dataset where (**a**), (**b**) and (**c**) represent benign breast cancer, and (**d**), (**e**) and (**f**) represent malignant breast cancer.

$$S_{\text{test}} = S_{\text{test, benign}} \cup S_{\text{test, malignant}} \cup S_{\text{test, normal}} \tag{4}$$

For Validation Set (S_{val}):

$$S_{\text{val}} = S_{\text{val, benign}} \cup S_{\text{val, malignant}} \cup S_{\text{val, normal}} \tag{5}$$

- **BreakHis-400x (B-400x) and BreakHis-100x (B-100x)[28] Datasets:** The B-400x dataset is a subset of the BreakHis dataset and it comprises 1,820 microscopic images of breast tumor tissue collected from 82 patients. The dataset is divided into two classes: benign, which contains 588 images, and malignant, with 1,232 images. Each image is captured at 400× magnification, making this dataset a valuable resource for microscopic-level breast cancer analysis. This dataset is publicly available and can be accessed via B-400x. Figure 1 shows the sample of dataset. Where (a and b) represent affected Benign, (c and d) represent affected malignant. In this study, Fig. 4 represents the class distribution of different category where (a) represents dataset class distribution of different category, (b) represents class distribution of train set, (c) represents class distribution of test set, (d) represents class distribution of validation set. We used $|S_{\text{train, benign, malignant}}| = 1274$ images for train set using Eq. (3), $|S_{\text{test, benign, malignant}}| = 364$ images for test set using Eq. (4), and $|S_{\text{val, benign, malignant}}| = 182$ images for validation set using Eq. (5). Additionally, the B-100x dataset is indeed a subset of the BreakHis dataset, comprising a total of 3,051 images of breast tumor tissue collected from 82 patients. This dataset is also publicly available and can be accessed through the similar link.
- **BUSI[30] Dataset:** This dataset includes breast ultrasound images from 600 female patients, aged between 25 and 75 years. It consists of 1578 ultrasound images, each with an average resolution of 500x500 pixels in PNG format. The dataset also includes corresponding ground truth images, which provide additional annotations for accurate analysis. The images are categorized into three classes: normal (266 images), benign (891 images), and malignant (421 images), offering a comprehensive overview of breast tissue conditions for

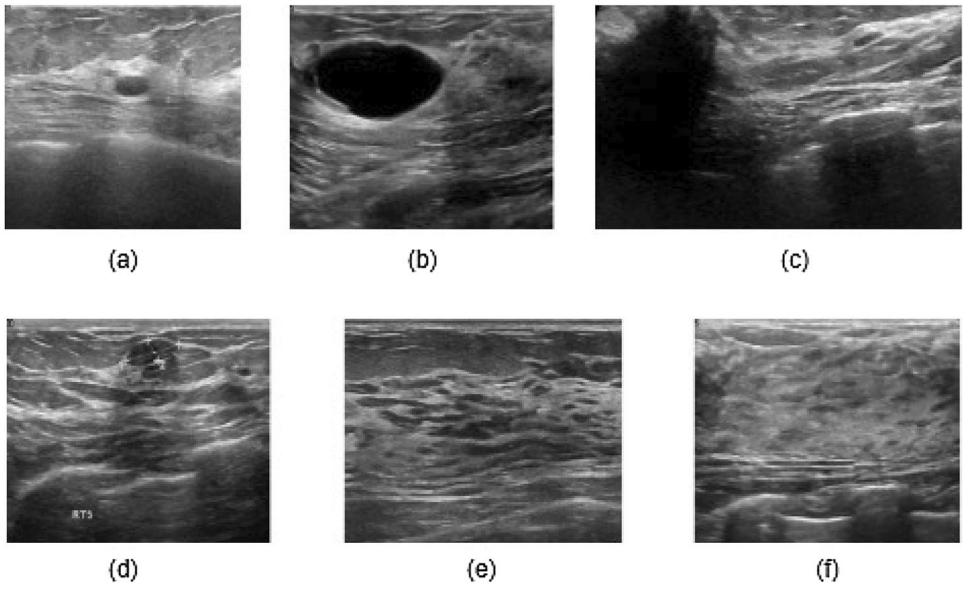


Fig. 2. Sample of data in BUSI dataset where (**a**) and (**b**) representing benign breast cancer, (**c**) and (**d**) representing malignant breast cancer, and (**e**) and (**f**) representing normal breast images.

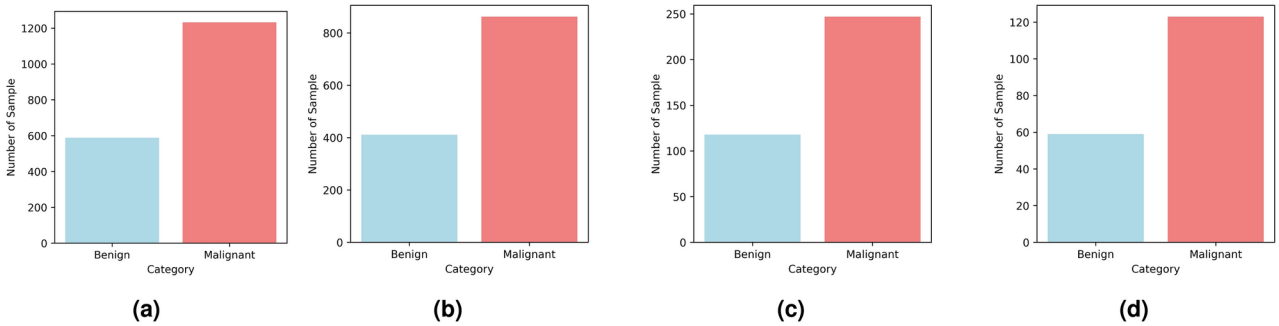


Fig. 4. Class distribution of different category from the B-400x dataset where (**a**) represents dataset class distribution of different category, (**b**) represents class distribution of train set, (**c**) represents class distribution of test set, (**d**) represents class distribution of validation set.

cancer detection research. This dataset is publicly available and can be accessed via BUSI. Figure 2 shows the sample of dataset. Where (a and b) represent affected Benign, (c and d) represent affected malignant and (e and f) represent affected normal. In this study, Fig. 5 represents the class distribution of different category where (a) represents dataset class distribution of different category, (b) represents class distribution of train set, (c) represents class distribution of test set, (d) represents class distribution of validation set. We used $|S_{\text{train, benign, malignant, normal}}|$= 1104 images for train set using Eq. (3), $|S_{\text{test, benign, malignant, normal}}|$= 316 images for test set using Eq. (4), and $|S_{\text{val, benign, malignant, normal}}| = 158$ images for validation set using Eq. (5).

Preprocessing techniques

In this study, we utilized the B-400x[28] and BUSI[30] datasets, which present several challenges, including variations in image dimensions (heights and widths), pixel quality discrepancies, limited diverse imaging conditions, and class imbalance. Figure 1 highlights the problems prevalent in the B-400x[28] dataset and Fig. 2 details the challenges associated with the BUSI[30] dataset. Additionally, Fig. 4a depicts the class imbalance in the B-400x[28] dataset, while Fig. 5a illustrates the imbalanced nature of the BUSI[30] dataset. To ensure the consistency and quality of the data, we employed several preprocessing techniques, which are as follows:

- **Resize:** The input image I is preprocessed through resizing to align with the input requirements of the DN-BCD model. Initially, the image is resized to a standardized dimension of $I_{\text{resize}} = 128 \times 128$ pixels with three color channels (RGB), a conventional input size that enhances consistency across samples and optimizes the computational efficiency of deep learning models. It is essential because it standardizes all images to the

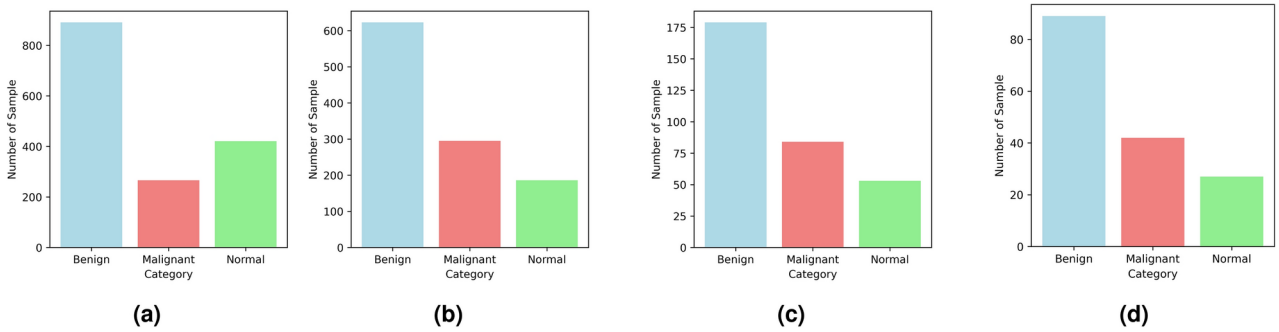


Fig. 5. Class distribution of different category and sets from the BUSI where (**a**) represents class distribution of different category, (**b**) represents class distribution of train set, (**c**) represents class distribution of test set, and (**d**) represents class distribution of validation set.

model's required dimensions, which significantly enhances the training process. This resizing operation is mathematically represented in Eq. (6).

$$I_{\text{resized}} = \text{resize}(I, (128, 128)) \tag{6}$$

- **Normalization:** The rescaling factor of $I_{\text{scaled}} = 1.0/255.0$ normalizes the pixel values of the images to a range of [0, 1]. This ensures consistency in input data, which enhances the stability and performance of neural networks. By reducing the scale of input values, normalization helps mitigate the risk of large gradients that can destabilize training, resulting in a smoother and more efficient learning experience. This rescaling operation is mathematically represented in Eq. (7)

$$I_{\text{scaled}} = \frac{I_{\text{array}}}{255.0} \tag{7}$$

- **Data augmentation:** Data augmentation techniques are used to diversify the training data by creating new image variations, simulating real-world scenarios. It is essential for enhancing the model's robustness and generalization capabilities. These include: 1. Rotation: Images are randomly rotated up to 20°, making the model resilient to slight rotations. 2. Width and Height Shift: Images are translated horizontally and vertically by up to 20%, helping the model recognize objects regardless of position changes. 3. Shear: A shear transformation of 15° distorts the image, enabling the model to handle skewed perspectives. 4. Zoom: Random zooming by up to 15% helps the model detect features at various scales. 5. Horizontal Flip: Random flipping improves robustness against mirrored images. 6. Fill Mode: Nearest pixel values fill gaps caused by transformations, preserving image continuity[59].
- **Dataset balancing:** To address the issue of data imbalance, we apply class weights to the dataset. Data imbalance occurs when the number of instances across different classes is significantly unequal, which can lead to biased models and inaccurate predictions[60,61]. This issue is especially critical in breast cancer classification, where underrepresented classes may be more challenging for the model to identify accurately. To address this, we calculate class weights to emphasize minority classes by assigning them higher weights that ensure the model gives appropriate attention to these classes, making errors on these classes more impactful during training. The class weight for each class i is computed as shown in Eq. (8).

$$W_i = \frac{S_{\text{total}}}{N \times S_i} \tag{8}$$

Where W_i represents the weight for class i, S_{total} is the total number of samples, N is the number of classes, and S_i is the number of samples in class i.

This formula ensures that the model places greater emphasis on the minority classes, thereby reducing the risk of biased predictions and enhancing classification accuracy across all classes. For example, in the B-400x[28] dataset, the computed weights are $W_{\text{benign}} = 1.5476$ and $W_{\text{malignant}} = 0.7386$. Similarly, for the BUSI[30] dataset, the weights are $W_{\text{benign}} = 0.5903$, $W_{\text{malignant}} = 1.2494$, and $W_{\text{normal}} = 1.9774$. Figure 6a represents the class weight using bar chart for B-400x[28] dataset and Figure 6b represents the class weight using bar chart for BUSI[30] dataset.

In this study, we applied various preprocessing techniques to enhance image quality and augment the dataset for improved model performance. Figure 7 and Table 3 represent the effect of each preprocessing. Where Fig. 7 illustrates the effect of these techniques on a sample image from the B-400x[28] Dataset. The original image is shown in (a), while (b) presents the resized version standardized to 128×128 pixels for uniformity. (c) depicts

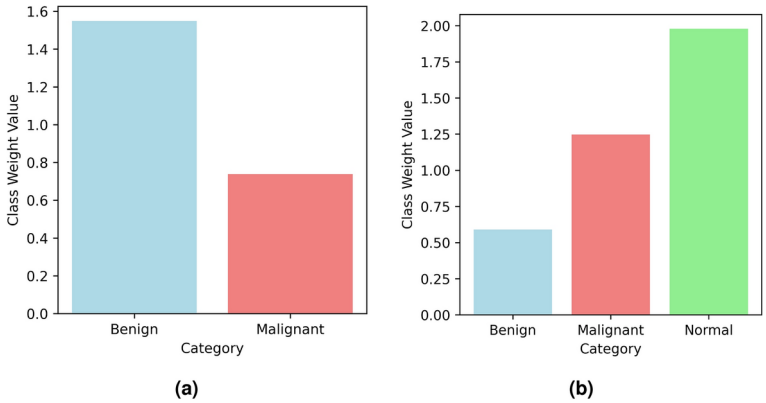


Fig. 6. Class weight of different category for handling class imbalance from the B-400x and BUSI dataset where (**a**) represents the class weight of different category for B-400x, and (**b**) represents the class weight of different category for BUSI.

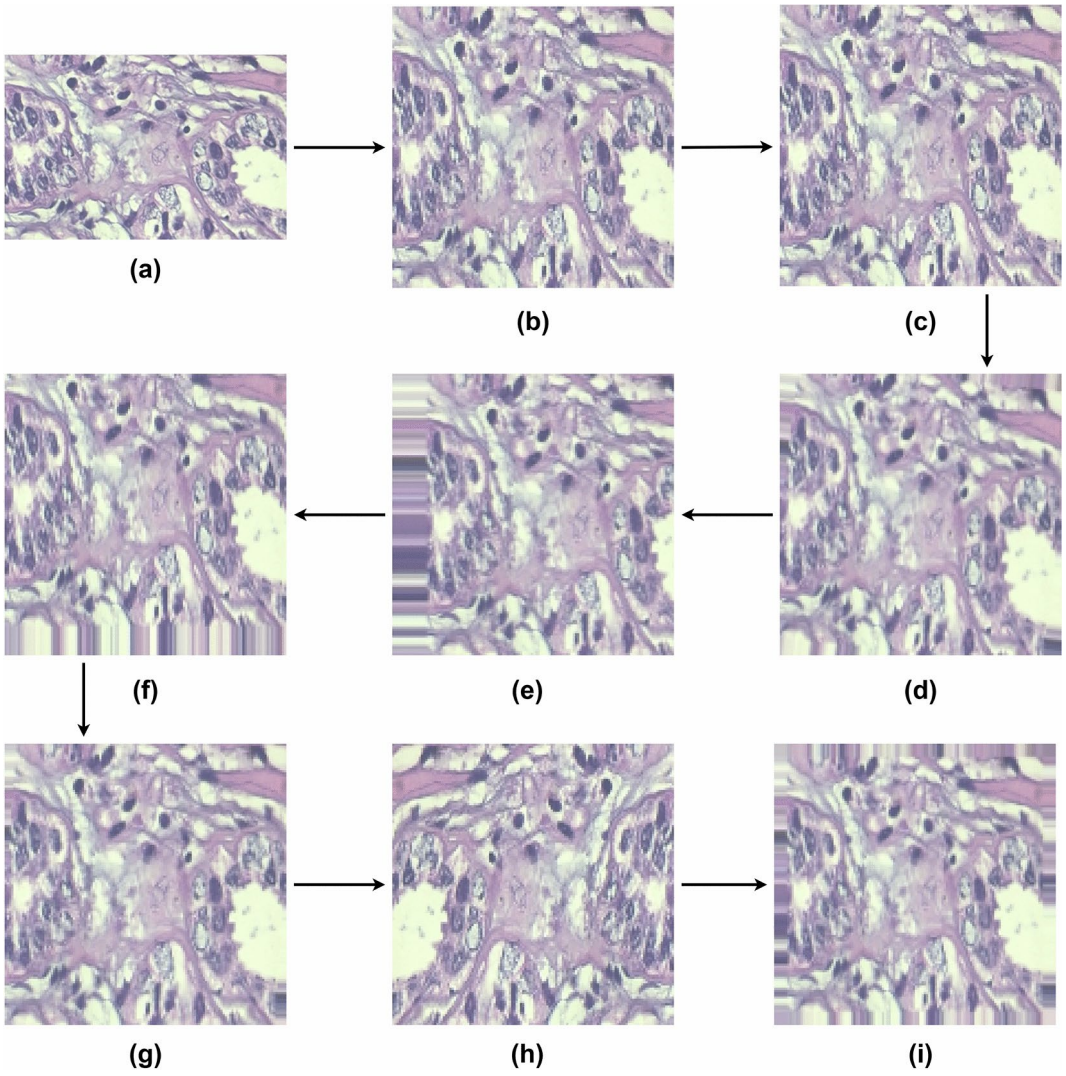


Fig. 7. Effect of the preprocessing processes applied to the B-400x dataset where (**a**) represents the original image, (**b**) is the resized version, (**c**) shows the normalized image, (**d**) illustrates the image after applying a 20° rotation, (**e**) depicts a height shift of 20%, (**f**) corresponds to a width shift of 20%, (**g**) represents a 15° shear transformation, (**h**) displays the horizontally flipped version and (**i**) shows a 15% zoomed image.

Dataset	Image status	Original	Resized	Normalized	Augmented
B-400x[28]	Minimum value	80	87	0.34	91
	Maximum value	255	255	1.0	255
	Mean value	187.66	187.55	0.74	188.36
BUSI[30]	Minimum value	0	13	0.05	19
	Maximum value	255	255	1.0	255
	Mean value	127.88	127.73	0.50	105.80

Table 3. Statistical overview of image characteristics before and after each preprocessing step for B-400x and BUSI datasets.

the normalized image with pixel values scaled to the range [0,1] for consistency. To introduce variations and improve model robustness, (d) applies a 20° rotation, (e) shifts the height by 20%, and (f) shifts the width by 20%. Additionally, (g) illustrates a 15° shear transformation, (h) presents a horizontally flipped version to enhance variability, and (i) shows a 15% zoomed-in image to help the model learn scale-invariant features. Similarly, we also applied this technique on the BUSI[30] dataset. During the training, a total of nine types of variations are applied to each image using different preprocessing techniques and those steps collectively enhance dataset diversity, mitigate overfitting, and contribute to training a more generalized and resilient deep learning model. Furthermore, Table 3 presents the statistical differences in key parameters, including minimum, maximum, and mean pixel values before and after each preprocessing step. These changes highlight the impact of resizing, normalization, and augmentation on image characteristics, ensuring better standardization and improved feature extraction.

In this study, we improve the consistency and quality of the data by applying these pre-processing approaches, which opens the door to more accurate and dependable testing and modeling.

Proposed model creation and training

In this training phase, the system utilizes separate training and validation sets to train and validate the our proposed model. In this study we created a deep neural model specifically designed for breast cancer detection. During the model creation step we examine different architecture like Densenet, Mobilenet, Resnet50 and VGG19. After that we applied transfer learning and cnn layers including Conv2D, Maxpooling, Batch Normalization, Zero Padding, Global Average Pooling, Dense, Dropout layers and activation functions such as ReLU and sigmoid or softmax[62].

- The Conv2D operation serves as a fundamental building block in convolutional neural networks (CNNs) for processing image data. It performs a convolution operation between a set of learnable filters, or kernels, and the input image or feature map[63]. This process involves sliding the filters over the input and computing element-wise multiplications followed by summation, as expressed mathematically in Eq. (9).

$$C[i,j] = \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} (I[i+m, j+n] \cdot K[m,n]) + b \tag{9}$$

Where $C[i,j]$ is the output value at position (i,j) in the output feature map, I is the input image or feature map, K is the convolutional filter (kernel) being applied, M and N are the dimensions of the filter, and b is the bias term associated with the filter.

- Max pooling is a downsampling operation commonly employed in CNNs to reduce the spatial dimensions of feature maps while retaining critical features[64]. Given an input feature map F with dimensions $H_{\text{in}} \times W_{\text{in}}$(height H_{in} and width W_{in}), and a pooling operation with a pool size of $P \times P$, the output feature map G will have dimensions $H_{\text{out}} \times W_{\text{out}}$, defined as shown in Eq. (10).

$$H_{\text{out}} = \frac{H_{\text{in}}}{P}, \quad W_{\text{out}} = \frac{W_{\text{in}}}{P} \tag{10}$$

The max pooling operation for a specific output position (i,j) is defined as shown in Eq. (11).

$$G[i,j] = \max_{m=0}^{P-1} \max_{n=0}^{P-1} F[i \cdot P + m, j \cdot P + n] \tag{11}$$

Where $G[i,j]$ is the value at position (i,j) in the output feature map, F is the input feature map, and P is the pool size (typically 2×2).

- The ReLU (Rectified Linear Unit) activation function is extensively utilized in deep learning models due to its ability to introduce non-linearity[65]. It is defined as shown in Eq. (12).

$$\mathrm{ReLU}(x) = \max(0, x) \tag{12}$$

Where $\mathrm{ReLU}(x)$ outputs the input directly if it is positive; otherwise, it outputs zero. This function is computationally efficient and helps mitigate the vanishing gradient problem, thereby accelerating the training process. However, it may lead to the "dying ReLU" problem, where a significant portion of neurons output zero, potentially slowing learning in those regions.

- Zero Padding is a technique used to add extra pixels around the border of an image, which helps preserve spatial dimensions during convolution operations[66]. This is particularly useful in deep learning models to prevent the reduction of feature map size after successive convolutions. The output of a Zero Padding layer can be represented by Eq. (13).

$$Z = \mathrm{pad}(X, p) \tag{13}$$

Where Z is the padded output, X is the original input, and p is the number of padding pixels added to each side of the input.

- Batch Normalization is a technique that normalizes the inputs of each layer to improve training speed and stability[67]. It mitigates issues related to internal covariate shift by maintaining the mean and variance of layer inputs. The output Y after Batch Normalization can be expressed in Euation.14.

$$Y = \gamma \frac{X - \mu}{\sqrt{\sigma^2 + \epsilon}} + \beta \tag{14}$$

Where Y is the normalized output, X is the input vector, μ is the mean of the input, σ^2 is the variance of the input, γ and β are learnable parameters for scaling and shifting the normalized output, and ϵ is a small constant to prevent division by zero.

- The Dense layer is critical for learning complex relationships among input features, as each neuron in this layer is connected to every neuron in the preceding layer[68]. The output vector Y for a Dense layer with N neurons, given an input vector X of length M, is computed as shown in Eq. (15).

$$Y = \sigma(W \cdot X + B) \tag{15}$$

Where Y is the output vector of the Dense layer, X is the input vector, W is the weight matrix of shape $N \times M$, B is the bias vector of length N, and σ is the activation function applied element-wise to the output.

- Dropout is a regularization technique used to prevent overfitting by randomly setting a fraction of the input units to zero during training[69]. This encourages the model to learn robust features. The output Y of a Dropout layer can be mathematically described in Eq. (16).

$$Y = X \cdot D \tag{16}$$

Where Y is the output after applying Dropout, X is the input vector, and D is a binary mask that randomly sets elements to zero with a probability p.

- The Sigmoid activation function is commonly used in neural networks to introduce non-linearity into the model. It maps any input value to a range between 0 and 1, making it particularly useful for binary classification tasks[70]. The output Y of the Sigmoid function given an input X can be expressed in Eq. (17).

$$Y = \sigma(X) = \frac{1}{1 + e^{-X}} \tag{17}$$

Where Y is the output of the Sigmoid function, X is the input to the function, and e is the base of the natural logarithm.

- The softmax activation function is employed in multi-class classification tasks, transforming a vector of real-valued inputs into a vector of probabilities, where each element corresponds to the likelihood of a particular class[71]. The softmax function is defined in Eq. (18).

$$\mathrm{Softmax}(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \tag{18}$$

Where z_i is the input to the softmax function for the i-th class, K is the total number of classes, and e is Euler's number. This function ensures that the output probabilities sum to 1, enabling effective classification.

In this study, Fig. 8 represents the model architecture of the proposed DNBCD model. The DNBCD model leverages a deep convolutional neural network with transfer learning to improve breast cancer classification. The core architecture consists of the Densenet121 backbone, additional CNN layers, and a fully connected network optimized for histopathological and ultrasound image analysis. The model design ensures efficient feature extraction, propagation, and classification, while reducing computational overhead and enhancing robustness against variations in imaging conditions.

Densenet121 as the backbone model

Densenet121 is a deep convolutional neural network that utilizes Conv2D, max pooling, zero padding, batch normalization, dense, and dropout layers to achieve efficient feature extraction and reuse. It begins with an initial Conv2D layer with a 7×7 kernel, followed by zero padding to maintain spatial dimensions, batch normalization for stable training, ReLU activation to introduce non-linearity, and max pooling to reduce dimensionality. The network consists of four dense blocks, each containing multiple Conv2D layers with a 3×3 kernel, interleaved with batch normalization and ReLU activation to ensure smooth gradient flow. Between dense blocks, transition layers apply 1×1 Conv2D for feature compression and 2×2 average pooling for downsampling. The final classification layer employs global average pooling, a dense layer, and a dropout layer to prevent overfitting before applying the softmax activation for output classification. Densenet121 employs a densely connected structure, where each layer receives feature maps from all preceding layers, leading to improved gradient propagation and feature reuse. In the proposed DNBCD model, Densenet121 is utilized as a feature extractor to enhance feature reuse and mitigate the vanishing gradient problem. By establishing $L(L+1)/2$ direct connections across L layers, it ensures efficient gradient flow, reducing the risk of degradation in deep networks. Each layer receives feature maps from all preceding layers, facilitating improved gradient propagation and feature reuse. The transformation at layer is mathematically represented by Eq. (19).

$$x_l = H_l([x_0, x_1, ..., x_{l-1}]) \tag{19}$$

Where x_l denotes the feature map at layer , H_l represents the transformation function consisting of Batch Normalization, ReLU Activation, and Convolution, and $[x_0, x_1, ..., x_{l-1}]$ indicates the concatenated feature maps from all previous layers. This connectivity enables hierarchical feature learning, allowing the model to extract low-level details such as edges and textures, as well as high-level structural attributes crucial for identifying complex patterns, such as those associated with breast cancer.

Transfer learning and task-specific adaptation

To enhance generalization and reduce training complexity, the DNBCD model employs transfer learning by initializing Densenet121 with pre-trained ImageNet weights and fine-tuning it for breast cancer classification[72]. The training process consists of two phases. In the feature extraction phase, the first k layers of Densenet121 are frozen, retaining low-level visual feature representations to preserve essential image characteristics such as texture, shape, and intensity variations. The extracted feature map at layer is given by Eq. (20).

$$F_l = f_{\theta_l}(I) \tag{20}$$

Where f_{θ_l} represents the transformation at layer with frozen parameters θ_l.

In the fine-tuning phase, the final layers of the model are unfrozen and retrained using breast cancer images from the B-400x[28] and BUSI[30] datasets. The model is optimized using categorical cross-entropy loss, as defined by Eq. (30).

Custom CNN layers for enhanced feature representation

Following feature extraction from Densenet121, additional task-specific layers are incorporated to refine classification accuracy. The Global Average Pooling (GAP) layer converts high-dimensional feature maps into compact representations[73,74], mathematically expressed by Eq. (21).

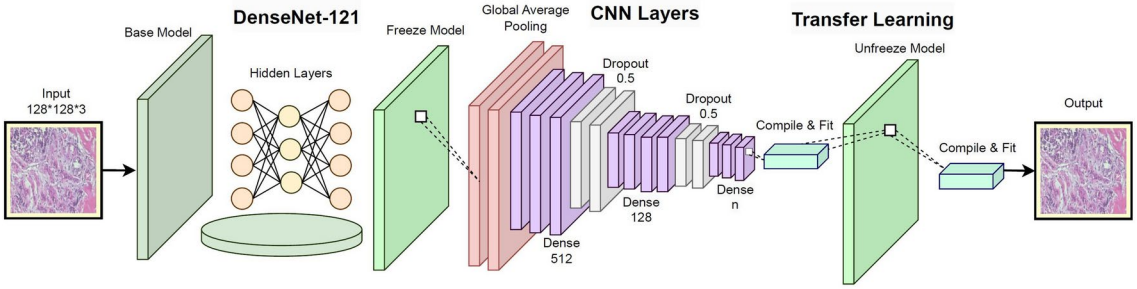


Fig. 8. Proposed architecture of DNBCD model.

$$GAP(x) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} x_{i,j} \tag{21}$$

Where H and W denote the spatial dimensions of the feature map.

Impact on feature representation and classification
The DNBCD model enhances classification reliability and feature discrimination through multiple mechanisms. Optimized gradient flow is achieved through the dense connectivity of Densenet121, mathematically formulated in Eq. (22).

$$\nabla W_l = \nabla W_{l+1} + \sum_{i=1}^{l-1} \nabla W_i \tag{22}$$

Where ∇W_l is the gradient at layer , ∇W_{l+1} is the gradient at the next layer, and ∇W_i represents the gradients of the preceding layers.

The network is structured to progressively extract features from the input images. Next we developed a modified Densenet121 architecture for breast cancer detection, utilizing cnn layers and transfer learning to boost model efficacy while reducing the dependence on large training datasets. The backbone of our model is the Densenet121 where has zero padding, conv2D, max pooling and batch normalization layers, which has been pre-trained on the ImageNet dataset. We excluded the top classification layers to tailor the architecture with the input shape set to $128 \times 128 \times 3$ to accommodate color images of this dimension. To preserve the pre-trained features, we initially froze the layers of the base model during training, allowing us to concentrate on the newly added layers. A Global Average Pooling layer follows the base model, condensing the spatial dimensions of the feature maps into a single vector for each feature map, thereby summarizing the information for the subsequent dense layers. The output from this pooling layer is processed through two fully connected (Dense) layers, where the first layer consists of 512 neurons with a ReLU activation function, introducing non-linearity and enabling the model to capture complex representations. This is succeeded by a Dropout layer with a rate of 0.5 to help prevent overfitting. The second Dense layer, containing 128 neurons and also employing ReLU activation, further refines the features before another Dropout layer with a rate of 0.5 is applied. The final output is produced by a Dense layer with N neuron which outputs a probability score reflecting the likelihood of breast cancer presence. In the final output the N is a single neuron with a sigmoid activation function for B-400x[28] dataset, specifically designed for binary classification and the N is a three neurons with a softmax activation function for BUSI[30] dataset, specifically designed for multi-class classification. In this study, we also explored three additional model combinations similar to our proposed architecture, utilizing Mobilenet, Resnet50, and VGG19. Our proposed model integrates Densenet121, CNN layers, and transfer learning, which we have described in detail above. Specifically, we implemented alternative configurations: the first model, named T_Mobilenet, replaces Densenet121 with Mobilenet; the second model, T_Resnet50, substitutes Densenet121 with Resnet50; and the third model, T_VGG19, employs VGG19 in place of Densenet121. Table 4 represents the overall architectural structures of trained models. Densenet121 is characterized by its unique architecture that employs dense connections, allowing each layer to receive input from all preceding layers. This facilitates easier gradient flow during training and helps the model leverage previously learned features, promoting efficient feature reuse and reducing the risk of overfitting. Therefore, we selected Densenet121 as our benchmark model for its effective balance of complexity, performance, and superior feature reuse in image classification tasks.

Model	Details
Proposed DNBCD	The DNBCD model is built upon Densenet121, employing transfer learning to leverage pre-trained weights. It incorporates custom CNN layers, a Global Average Pooling layer, and Dense layers for final classification. To mitigate overfitting, dropout layers are integrated, and Grad-CAM is utilized to enhance interpretability of the model's predictions. The output shape is (None, 3), representing three classes of breast tissue.
T_Mobilenet	T_Mobilenet is developed similarly to DNBCD, utilizing Mobilenet as the backbone. It employs depthwise separable convolutions to maintain efficiency while minimizing parameters. The architecture includes custom CNN layers and a Global Average Pooling layer, resulting in a classification output of shape (None, 3).
T_Resnet50	The T_Resnet50 model is developed similarly to DNBCD, utilizing transfer learning with Resnet50 as the backbone. It incorporates residual connections that facilitate the training of very deep networks by improving the flow of gradients. The model outputs a classification shape of (None, 3).
T_VGG19	The T_VGG19 model leverages transfer learning with VGG19 as the backbone, following a structure akin to DNBCD. It consists of multiple convolutional layers and max pooling layers designed to capture intricate patterns in image data. The output shape is (None, 3) for classification.
Densenet[52]	The Densenet121 model comprises an input layer for image data, followed by a series of layers including ZeroPadding2D, Conv2D, BatchNormalization, and ReLU activation. It features dense blocks and transition layers that enhance feature reuse and facilitate gradient flow, culminating in a classification output with shape (None, 3).
Mobilenet[53]	The Mobilenet model features an input layer for image data, followed by Conv2D and BatchNormalization layers. It utilizes depthwise separable convolutions to optimize performance while maintaining accuracy, culminating in a classification output with shape (None, 3).
Resnet50[54]	The Resnet50 model includes an input layer for images, followed by a series of convolutional layers and residual blocks. These residual connections enable the training of deep networks by ensuring effective gradient flow, leading to an output shape of (None, 3).
VGG19[55]	The VGG19 model is structured with several convolutional layers followed by max pooling layers, aimed at extracting deep hierarchical features from images. It concludes with fully connected layers that output three classes, with an output shape of (None, 3).

Table 4. Overview of trained models, detailing their architectural structures and key features.

After training the model, its performance is evaluated using the test set, which provides an independent assessment of the model's ability to predict breast cancer on unseen data. Key evaluation metrics, including accuracy, F1 score, recall, precision, MAE, RMSE and auc score, are calculated to quantify the model's performance. These metrics guide further improvements and refinements, ensuring greater accuracy and reliability in detecting and interpreting breast cancer.

Model selection for detecting breast cancer

Following the evaluation phase using various performance metrics like accuracy, precision, recall, f1 score, MAE, RMSE and auc score, the system employs a model selection process to identify the most accurate and reliable model for breast cancer detection. The model demonstrating superior performance is chosen for deployment in subsequent detection stages, ensuring optimal accuracy and robustness in the system's output[75].

Post-classification output explanation

To enhance the interpretability and transparency of the system, an explainable AI (XAI) approach is applied to the classification results. Techniques such as Grad-CAM are employed to highlight the regions in the ultrasound images that influenced the model's decision. This step helps in providing a visual explanation to radiologists or healthcare professionals, offering insights into how the model arrived at its classification. The integration of Grad-CAM enhances the interpretability of the DNBCD model by visually highlighting regions in ultrasound images that contribute to classification decisions. This method enables clinicians to validate whether the model focuses on diagnostically relevant features rather than irrelevant artifacts, ensuring transparency in medical decision-making. By generating heatmaps that overlay key areas of interest, Grad-CAM provides a visual confirmation of the model's decision-making process, making it more suitable for clinical use. This visualization is crucial for increasing trust in deep learning models, as it allows radiologists and healthcare professionals to interpret the model's predictions in a meaningful way. The following section details the Grad-CAM methodology, including preprocessing, gradient calculation, and heatmap generation.

- **Grad-CAM (Gradient-weighted Class Activation Mapping)**[32]: In this research, we employ Gradient-weighted Class Activation Mapping (Grad-CAM) to visualize the regions of an image that the DNBCD model prioritizes when making predictions. Grad-CAM is particularly effective for identifying the spatial locations within an input that contribute most significantly to the model's classification decision[76]. This section details the Grad-CAM process, which incorporates preprocessing, gradient calculation, and heatmap generation.
- **Image preprocessing:** The input image I is first resized to a standard dimension of 128×128 pixels to match the DNBCD model's input requirements using Eq. (6). The resized image is converted to a numerical array and normalized by scaling pixel values between 0 and 1 through division by 255. This normalization standardizes input intensities, which improves model performance and stabilizes gradient-based computations using Eq. (7).
- **Model prediction and class identification:** Once preprocessed, the image I_{scaled} is input into the DNBCD model to obtain a prediction vector p, where each element corresponds to the model's confidence in a given class. The predicted class c is then identified as the index with the highest probability using Eq. (23).

$$c = \arg\max(\mathrm{p}) \tag{23}$$

Where p represents the model's probability outputs for all potential classes.

- **Gradient computation:** To generate the Grad-CAM heatmap, we construct an auxiliary model that outputs both the final convolutional feature maps A_k and the prediction scores p. Let A_k denote the k-th feature map in the last convolutional layer.
  Using automatic differentiation via TensorFlow's `GradientTape`, the gradient of the target class score y^c with respect to each feature map A_k is calculated using Eq. (24).

$$\frac{\partial y^c}{\partial A_k} \tag{24}$$

Where y^c denotes the score of the predicted class c.

- **Calculation of feature map weights:** To emphasize the importance of each feature map, the gradients are spatially averaged across the feature map dimensions to produce weights α_k for each A_k. This weighting effectively reflects the feature map's influence on the class score using Eq. (25).

$$\alpha_k = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{k,i,j}} \tag{25}$$

Where $Z = H \times W$ is the total number of pixels in the feature map (height H and width W).

- **Generation of the Grad-CAM heatmap:** The Grad-CAM heatmap $L^c_{\text{Grad-CAM}}$ is produced by computing a weighted combination of the feature maps A_k using the previously calculated weights α_k, as shown in Eq. (26).

$$L^{c}_{\text{Grad-CAM}} = \text{ReLU}\left(\sum_{k} \alpha_k A_k\right) \tag{26}$$

The ReLU (Rectified Linear Unit) function ensures only positive activations are retained, focusing the heatmap on the most relevant spatial regions.

- **Normalization and resizing:** The resulting heatmap $L^{c}_{\text{Grad-CAM}}$ is normalized by dividing by its maximum value to ensure that values are scaled between 0 and 1, as shown in Eq. (27).

$$L_{\text{Grad-CAM}^c,\text{normalized}} = \frac{L_{\text{Grad-CAM}^c}}{\max(L_{\text{Grad-CAM}^c})} \tag{27}$$

To align with the dimensions of the original image, the heatmap is resized as shown in Eq. (28).

$$L_{\text{Grad-CAM}^c,\text{resized}} = \text{resize}\left(L_{\text{Grad-CAM}^c,\text{normalized}}, (H_{\text{image}}, W_{\text{image}})\right) \tag{28}$$

Where $L_{\text{Grad-CAM}^c,\text{resized}}$ is the resized heatmap, $L_{\text{Grad-CAM}^c,\text{normalized}}$ is the normalized Grad-CAM heatmap, and $(H_{\text{image}}, W_{\text{image}})$ are the height and width of the original image, respectively.

- **Gaussian smoothing and maximum activation highlighting:** To enhance interpretability, Gaussian smoothing is applied to the resized heatmap. Additionally, the location of maximum activation in the heatmap, which corresponds to the region of greatest relevance, is identified by finding the highest intensity point, as shown in Eq. (29).

$$(x_{\max}, y_{\max}) = \arg\max\left(L_{\text{Grad-CAM}^c_{\text{resized}}}\right) \tag{29}$$

Where $(x_{\max}, y_{\max})$ are the coordinates of the point of maximum intensity in the resized Grad-CAM heatmap $L_{\text{Grad-CAM}^c_{\text{resized}}}$.

- **Heatmap overlay on original image:** Finally, the smoothed and resized Grad-CAM heatmap is overlaid on the original image, with an adjustable transparency setting to visually illustrate the model's focus. The area of maximum activation is marked with a circular indicator, highlighting the specific region that contributed most significantly to the model's classification.

By following this methodology, the Deep Neural Breast Cancer Detection System is designed to deliver an accurate and interpretable solution for detecting breast cancer, leveraging advanced machine learning techniques and explainable AI for enhanced trust in medical diagnosis.

Result analysis

Experimental setup

The experiment was conducted on Kaggle[77] using a system equipped with P100 GPUs, each with 8 GB of RAM, and an Intel Xeon Platinum 8259CL processor. This setup provided robust computational power for efficient deep learning model training. The system was built using Python 3.7, with TensorFlow and Keras as the primary deep learning frameworks. Table 5 presents the hyperparameters used in our experiments, providing insight into the configuration settings that guided the training process.

Performance measurement parameters

Consider the DNBCD system that predicts the probabilities for a set of classes, denoted by $P(y|x)$, given an input image x. The true label for this image is denoted as y. The cross-entropy loss measures the dissimilarity between the predicted probabilities and the true labels. Mathematically, the cross-entropy loss equation is given by Eq. (30).

$$\text{Loss} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c}\log(P(y_{i,c}|x_i)) \tag{30}$$

Where N is the number of samples in the dataset or batch, C is the number of classes and $y_{i,c}$ is the true label for the i-th sample and c-th class. $P(y_{i,c}|x_i)$ is the predicted probability of the i-th sample belonging to the c -th class.

The loss equation sums over all the samples and classes, comparing the true labels with the predicted probabilities. The term $-\log(P(y_{i,c}|x_i))$ penalizes the model more when the predicted probability of the true class is low, and less when the predicted probability is high[78].

The accuracy is calculated as the ratio of correct predictions to the total number of samples as shown in Eq. (31).

Serial No	Parameter/Technique	B-400x[28]	BUSI[30]
1	Optimizer	Adam	Adam
2	Input size	$128 \times 128 \times 3$	$128 \times 128 \times 3$
3	Stride	1	1
4	Kernel size	3×3	3×3
5	Tensor shape	(None, 128, 128, 3)	(None, 128, 128, 3)
6	No. of epochs	20	20
7	Activation	Relu	Relu
8	Output activation	Sigmoid	Softmax
9	Initial learning rate	1.0×10^{-3}	1.0×10^{-3}
10	Final learning rate	1.0×10^{-4}	1.0×10^{-4}
11	Batch size	32	32
12	Class mode	Binary	Categorical
13	Metrics	Accuracy	Accuracy
14	Loss	binary_crossentropy	categorical_crossentropy

Table 5. Hyperparameter values for our proposed DNBCD model.

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{N} \tag{31}$$

And the Number of correct predictions (NCP) can be expressed as shown in Eq. (32).

$$\text{NCP} = \sum_{i=1}^{N} \delta(y_i, \text{round}(\hat{y}_i)) \tag{32}$$

Where $\delta(y_i, \text{round}(\hat{y}_i))$ is the Kronecker delta function, which returns 1 if y_i is equal to the rounded predicted value of $\hat{y}_i$, and 0 otherwise.

In this study, Fig. 9 presents the performance of various trained state-of-the-art models in terms of loss and accuracy curves. Panel (a) illustrates the train accuracy curves and (c) represents the validation accuracy curves, where the x-axis represents the number of epochs and the y-axis represents the model losses, while panel (b) shows the trained loss curves and (d) represents validation loss curves, with the x-axis denoting the number of epochs and the y-axis indicating the accuracy values. The performance metrics are evaluated using the B-400x[28] dataset. The curve uses distinct colors to represent the loss curves of the models for clear differentiation. DNBCD is represented in blue, T_Mobilenet in orange, T_Resnet50 in green, and T_VGG19 in red. The standard Densenet121 is shown in purple, Mobilenet in brown, Resnet50 in pink, and VGG19 in gray. The losses are computed using Eq. (30). It is observed that the model losses decrease with increasing epochs, with the DNBCB model consistently outperforming the others. Notably, the Resnet50 model exhibits the highest losses, likely due to its residual features. Nevertheless, all models show satisfactory performance after using our proposed system. We can see all modified models are worked well than the base models. Similarly, Fig. 10 displays the performance curves for loss and accuracy, where Panel (a) illustrates the train accuracy curves and (c) represents the validation accuracy curves, where the x-axis represents the number of epochs and the y-axis represents the model losses, while panel (b) shows the trained loss curves and (d) represents validation loss curves, with the x-axis denoting the number of epochs and the y-axis indicating the accuracy values. The performance metrics are evaluated using the BUSI[30] dataset. As in Fig. 9, the color scheme follows the same pattern: blue for DNBCB, orange for VGG16, green for Resnet50, red for Mobilenet, and purple for Densenet The losses are determined using Eq. (30), while the accuracy values are derived from Eq. (31), with Eq. (32) used to calculate the number of correct predictions. It is evident that the accuracy increases with the number of epochs, with the DNBCB model again demonstrating superior performance compared to the other models. The Resnet50 model shows the lowest accuracy, likely due to the impact of its residual features. Despite this, all models perform well, exhibiting reliable results across the evaluated datasets.

Confusion matrix for our system: The confusion matrix is a matrix that summarizes the predicted and actual class labels. It consists of four values: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN)[79]. The primary role of the confusion matrix is to provide a detailed breakdown of the model's performance by illustrating how well the classifier distinguishes between different classes. It helps in assessing classification errors, identifying patterns of misclassification, and understanding model biases. Figures 12 and 11 show the confusion matrix of different trained state-of-art models from B-400x[28] and BUSI[30] where the x-axis contains the predicted labels and the y-axis contains the true label. To evaluate the performance of our proposed DNBCD model, we utilize several metrics including recall, precision, f1-score, mean absolute error (MAE), root mean squared error (RMSE), and area under the curve score (AUC) derived from the confusion matrix. The following formulas, as presented in Eqs. (34), (33), (35), (36), (37) and (38) are employed to compute these metrics.

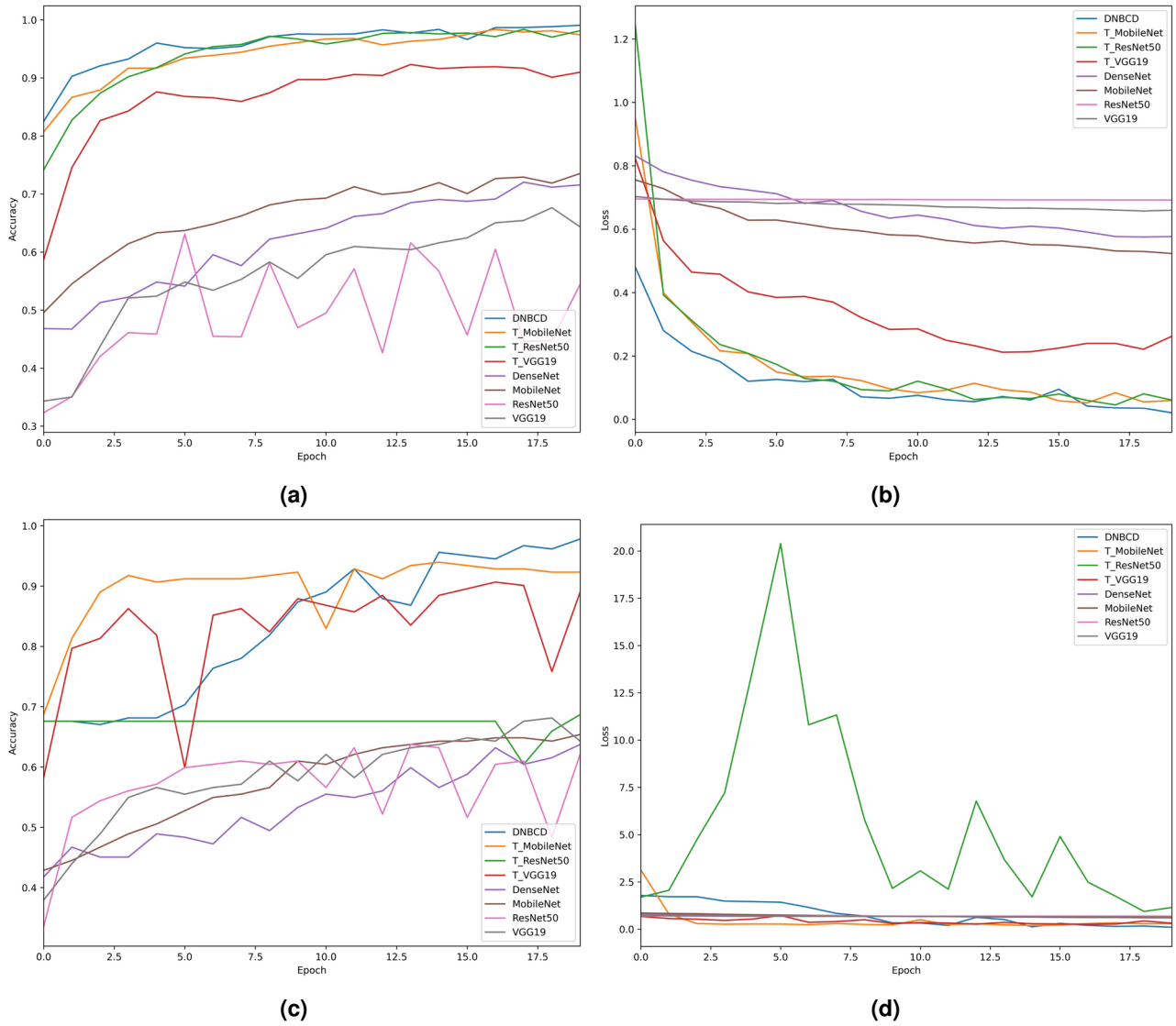


Fig. 9. Comparative performance of loss curve and accuracy curve for different systems using B-400x dataset where (**a**) represents training accuracy curve, (**b**) represents training loss curve, (**c**) represents validation accuracy curve and (**d**) represents validation loss curve.

Recall: Recall quantifies the model's ability to identify all relevant instances, calculated as the ratio of true positives to all actual positives. High recall indicates that the model effectively identifies most actual positive cases, which is particularly important in situations where missing a positive instance could have serious implications. It is given by Eq. (33).

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{33}$$

Precision: Precision measures the accuracy of positive predictions, calculated as the ratio of true positives to all predicted positives. A high precision means that when the model predicts a positive case, it's usually correct, which helps reduce the chances of false alarms. It is calculated using Eq. (34).

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{34}$$

F1 score: The F1 Score is the harmonic mean of precision and recall, providing a balance between the two metrics. This balance is important for ensuring the model accurately identifies cases without generating too many false alarms. It is calculated using Eq. (35).

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{35}$$

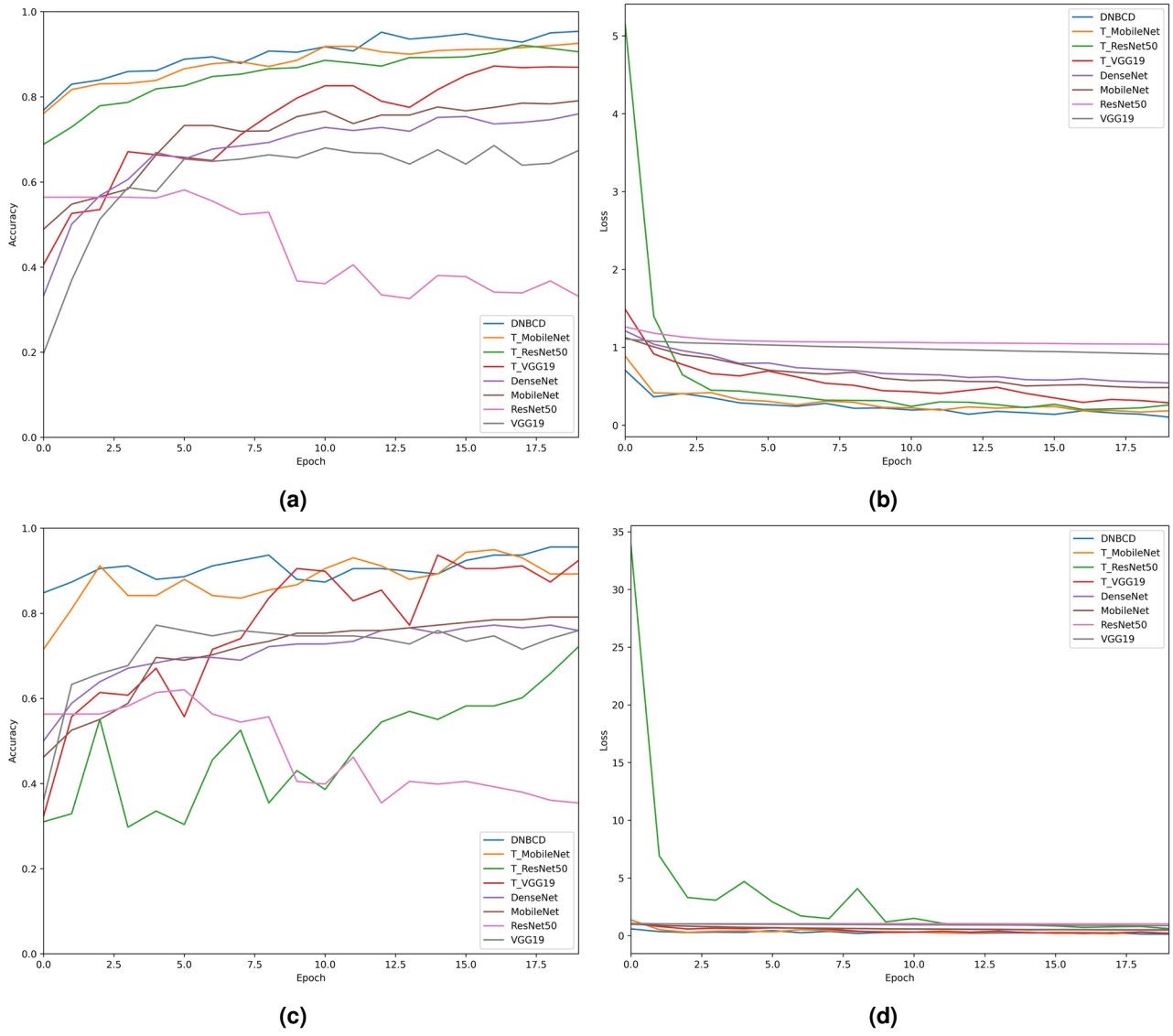


Fig. 10. Comparative performance of loss curve and accuracy curve for different systems using BUSI dataset where (**a**) represents training accuracy curve, (**b**) represents training loss curve, (**c**) represents validation accuracy curve and (**d**) represents validation loss curve.

Mean absolute error (MAE): MAE measures the average magnitude of errors in predictions, calculated as the ratio of the sum of false positives and false negatives to the total number of instances. This metric plays a crucial role in assessing the accuracy of a model's predictions, helping to understand how close the predicted values are to the actual outcomes. It is defined by Eq. (36).

$$\mathrm{MAE} = \frac{\mathrm{FP} + \mathrm{FN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{36}$$

Root mean squared error (RMSE): RMSE measures the square root of the average squared differences between predicted and actual values, highlighting larger errors more than smaller ones. This metric is vital for assessing model performance, as it highlights significant discrepancies in predictions. It is expressed in Eq. (37).

$$\mathrm{RMSE} = \sqrt{\frac{\mathrm{FP} + \mathrm{FN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}} \tag{37}$$

Area under the curve score (AUC): The AUC represents the area under the Receiver Operating Characteristic (ROC) curve, which plots the True Positive Rate (TPR) against the False Positive Rate (FPR) at various threshold settings. It is crucial for assessing a model's ability to differentiate between classes across various thresholds. It is calculated in Eq. (38).

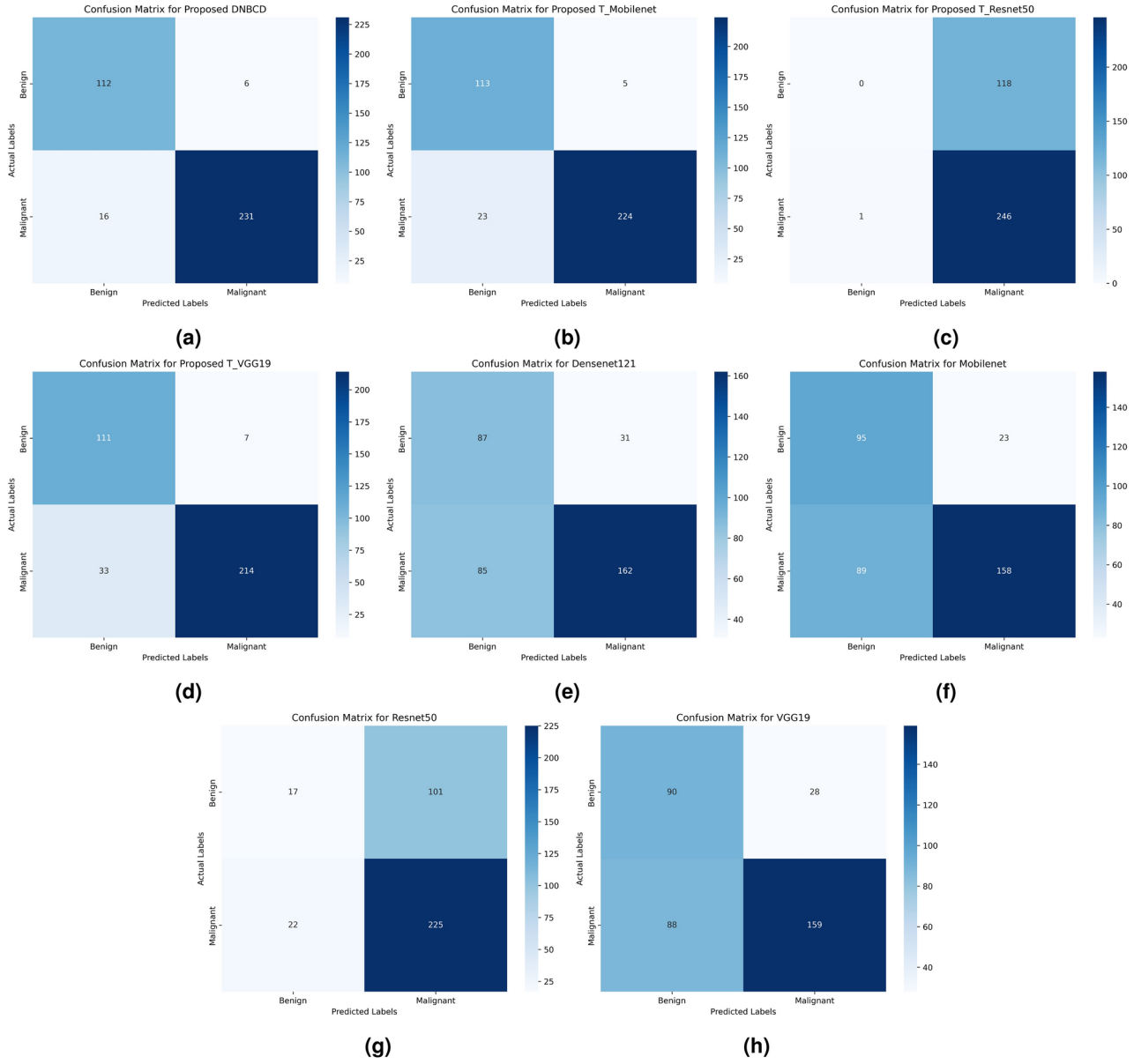


Fig. 11. Confusion matrix for different trained models using Breakhis-400x dataset where (**a**) represents confusion matrix of DNBCD, (**b**) represents confusion matrix of T_Mobilenet, (**c**) represents confusion matrix of T_ResnetNet50, (**d**) represents confusion matrix of T_VGG19, (**e**) represents confusion matrix of Densnet121, (**f**) represents confusion matrix of Mobilenet, (**g**) represents confusion matrix of Resnet50, and (**h**) represents confusion matrix of VGG19.

$$\mathrm{AUC} = \int_{-\infty}^{\infty} \mathrm{TPR}(x)\, d\mathrm{FPR}(x) \tag{38}$$

Where TPR (True positive rate) is also known as recall, and FPR (False Positive Rate) is the proportion of false positives out of all actual negatives. FPR can be expressed as shown in Eq. (39).

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}} \tag{39}$$

In this study, we employed statistical significance tests to compare the performance of DNBCD with various state-of-the-art models. Two key parameters, the T-statistic and the P-value, play crucial roles in determining the significance of these comparisons.

T-statistic: The T-statistic measures the size of the difference between two sample means relative to the variation in the sample data. A higher absolute value of the T-statistic indicates a greater difference between the groups being compared[80]. The T-statistic is calculated using Eq. (40).

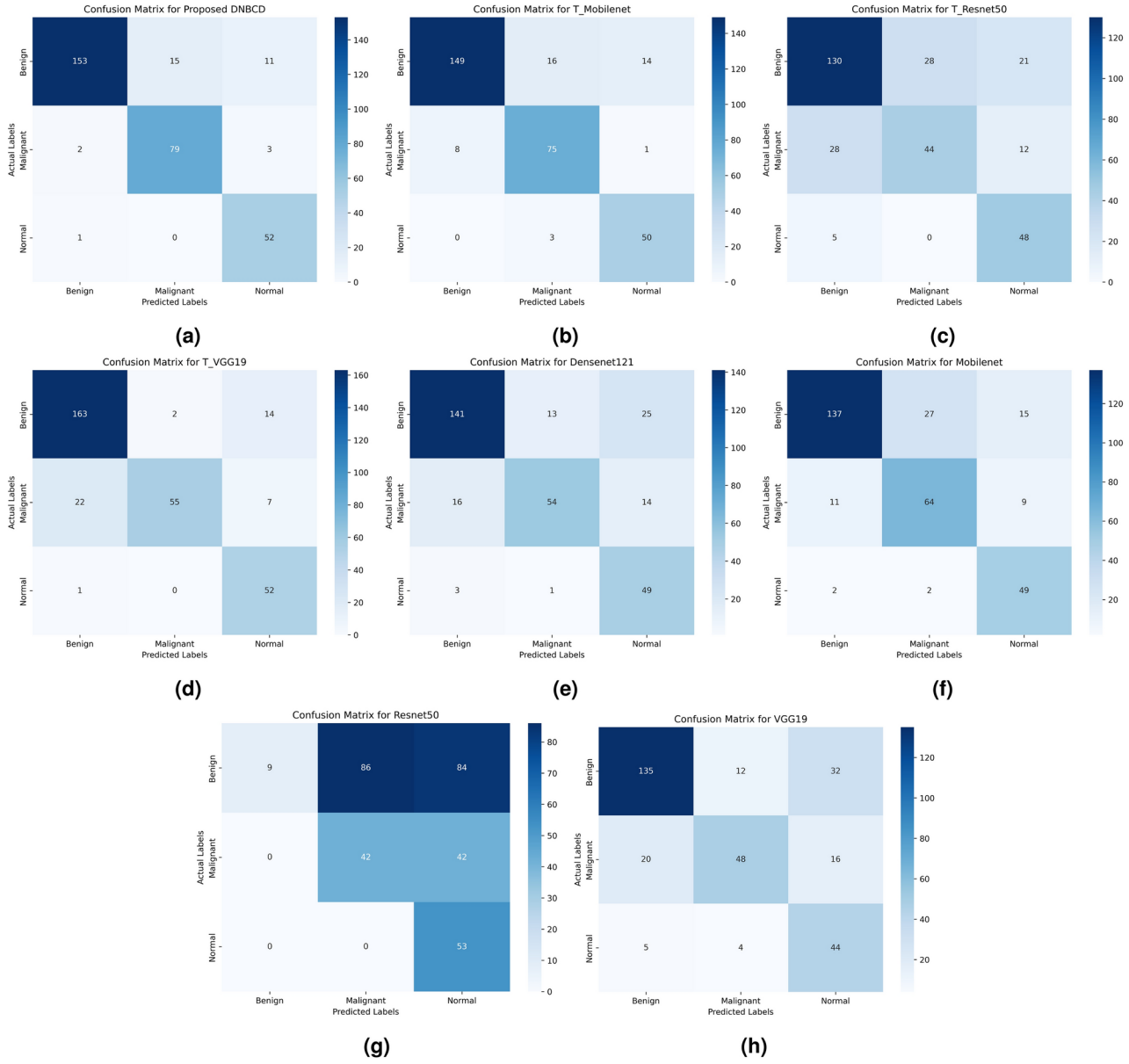


Fig. 12. Confusion matrix for different trained models using Breakhis-400x dataset where (**a**) represents confusion matrix of DNBCD, (**b**) represents confusion matrix of T_Mobilenet, (**c**) represents confusion matrix of T_ResnetNet50, (**d**) represents confusion matrix of T_VGG19, (**e**) represents confusion matrix of Densnet121, (**f**) represents confusion matrix of Mobilenet, (**g**) represents confusion matrix of Resnet50, and (**h**) represents confusion matrix of VGG19.

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\frac{s}{\sqrt{n}}} \tag{40}$$

Where $\bar{X}_1$ and $\bar{X}_2$ are the sample means, s is the pooled standard deviation, and n is the sample size.

***P*-value**: The *P*-value indicates the probability of observing the data, or something more extreme, given that the null hypothesis is true. A low *P*-value ($typically \leq 0.05$) suggests that the observed difference is statistically significant, leading to a rejection of the null hypothesis[80]. The *P*-value can be expressed in relation to the T-statistic in Eq. (41).

$$P\text{-value} = P(T \geq t \mid H_0) \tag{41}$$

Where H_0 is the null hypothesis (e.g., no difference between models) and t is the computed T-statistic.

These metrics help determine whether the differences in performance between DNBCD and the other models are statistically significant.

Comparative analysis of the performances of the proposed system with state of the arts models

The current investigation, we trained several state-of-the-art deep learning models using two datasets. Overall, the models achieved strong results, with our proposed DNBCD model demonstrating the highest performance. Table 6 presents the performance metrics for all trained models, including accuracy, precision, recall, F1 score, mean absolute error (MAE), and root mean square error (RMSE). These metrics were calculated using the following equations: Eq. (31) for accuracy, Eq. (34) for precision, Eq. (33) for recall, Eq. (35) for F1 score, Eq. (36) for mean absolute error, Eq. (37) for root mean square error and Eq. (38) for auc score. For the B-400x[28] dataset, our proposed DNBCD model achieved an impressive accuracy of 93.97%, alongside a remarkable F1 score of 95.45% and an auc score of 99.24%. Additionally, the model recorded a low loss of 0.1607, a mae of 0.0603, and a rmse of 0.2455. While DNBCD secured the second-highest precision at 97.47% and recall at 93.52%, these metrics still reflect its exceptional overall performance in cancer detection. Conversely, in the BUSI[30] dataset, the DNBCD model excelled by achieving the highest performance across all metrics. This includes an accuracy of 89.87%, a remarkable F1 score of 90.00%, an auc score of 97.75%, a precision of 91.11%, and a recall of 89.87%. The model also recorded the lowest mae of 0.1392 and a rmse of 0.4639. Notably, the loss of our proposed DNBCD model was the third lowest at 0.4095. These results underscore the robustness and effectiveness of the DNBCD model, highlighting its potential for accurate cancer detection across diverse datasets. Together, these metrics provide a comprehensive evaluation of the models' predictive capabilities, highlighting the superior performance of our proposed DNBCD model.

In this research, Figs. 13, barspschartspsloss, barspschartspsf1spsscore, barspschartspsrecall, barspschartspsprecision, barspschartspsmae, barspschartspsrmse and barspschartspsauc display the performance metrics for the trained state-of-the-art models on both datasets. The x-axis represents the different trained models, while the y-axis indicates accuracy, loss, F1 score, recall, precision, MAE, and RMSE for each model where use distinct colors to represent each model: DNBCD is represented by dark red, T_Mobilenet is represented by pinkish-red (Figs. 14, 15, 16, 17, 18, 19, 20), T_Resnet is represented by orange-yellow, T_VGG19 is represented by bright blue, Densenet is represented by green, Mobilenet is represented by teal-blue, Resnet is represented by purple, and VGG19 is represented by magenta-pink. The results indicate that the DNBCD model consistently outperforms the other models across all metrics. However, it's important to highlight that on the B-400x[28] dataset, the T_Mobilenet model achieved the highest F1 score, while the T_Resnet model demonstrated the best precision. In contrast, the T_VGG19 model had the lowest loss on the BUSI[30] dataset. Nonetheless, our proposed DNBCD model demonstrates overall superior performance on both datasets. This comparison further illustrates that the DNBCD model outperforms other models in terms of all evaluated metrics.

In this study, Table7 presents the performance metrics of the proposed DNBCD model across multiple runs for the B-400x[28] and BUSI[30] datasets, with results shown as averages along with their standard deviations (indicated by ±). This representation allows us to assess the model's consistency and reliability. For the B-400x[28] dataset, the average accuracy was 93.90 with a standard deviation of 0.4061, indicating that the model performed consistently well across different runs, showing minimal fluctuation in performance. Similarly, for the BUSI[30] dataset, the average accuracy is 89.87 with a standard deviation of 1.0329. While this value is slightly lower than that of the B-400x[28] dataset, the relatively low standard deviation suggests that the model maintains stable performance on this dataset. Overall, the DNBCD model demonstrates higher performance metrics compared to existing works shown in Table 10, showcasing its effectiveness and reliability in classification tasks across both

Dataset	Model	Accuracy	Loss	F1 Score	Precision	Recall	MAE	RMSE	AUC
B-400x[28]	**Proposed DNBCD**	**93.97**	**0.1607**	**95.45**	97.47	93.52	**0.0603**	**0.2455**	**99.24**
	T_Mobilenet	92.33	0.2453	94.12	**97.82**	90.69	0.0767	0.2770	98.23
	T_Resnet	67.40	1.2267	80.52	67.58	**99.60**	0.3260	0.5710	62.80
	T_VGG19	89.04	0.2875	91.45	96.83	86.64	0.1096	0.3310	96.76
	Densenet[52]	68.22	0.5796	73.64	83.94	65.59	0.3178	0.5637	77.99
	Mobilenet[53]	69.32	0.6062	73.83	87.29	63.97	0.3068	0.5539	77.21
	Resnet50[54]	66.30	0.6881	78.53	69.02	91.09	0.3370	0.5805	52.62
	VGG19[55]	68.22	0.6553	73.27	85.03	64.37	0.3178	0.5637	76.68
BUSI[30]	**Proposed DNBCD**	**89.87**	0.4095	**90.00**	**91.11**	**89.87**	**0.1392**	**0.4639**	**97.75**
	T_Mobilenet	86.71	0.3654	86.85	87.87	86.71	0.1772	0.5156	97.13
	T_Resnet	70.25	0.6740	70.07	71.36	70.25	0.3797	0.7378	88.64
	T_VGG19	85.44	**0.3306**	85.17	87.24	85.44	0.1930	0.5366	96.64
	Densenet[52]	77.22	0.5110	77.67	80.37	77.22	0.3165	0.7026	92.89
	Mobilenet[53]	79.11	0.4804	79.44	81.29	79.11	0.2627	0.6085	93.40
	Resnet50[54]	32.91	1.0398	23.62	70.33	32.91	0.9367	1.2118	73.12
	VGG19[55]	71.84	0.8764	72.54	75.75	71.84	0.3987	0.7956	89.17

Table 6. Performance comparison of different trained state-of-the-art models with varying parameters for B-400x and BUSI datasets. Significant values are in bold.

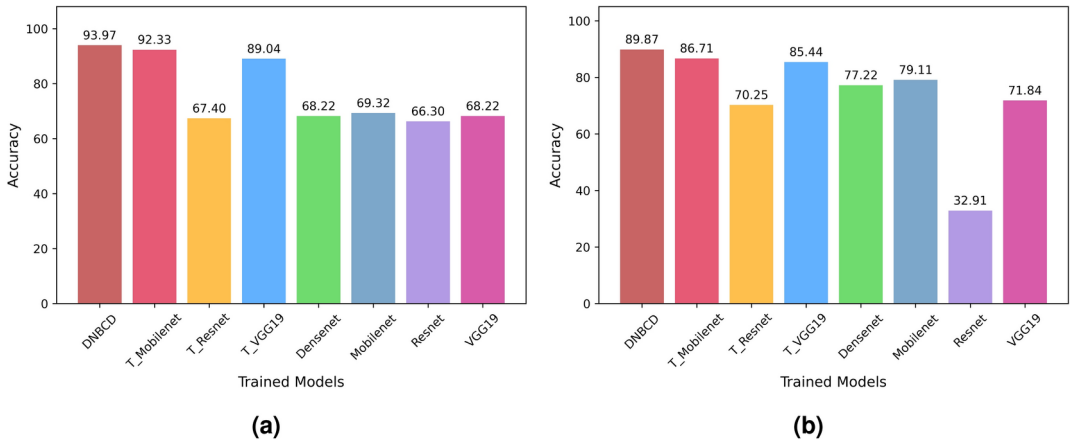


Fig. 13. Accuracy comparison of different trained state-of-the-art models for B-400x and BUSI datasets using bar charts, where (**a**) represents accuracy comparison for B-400x dataset, and (**b**) represents accuracy comparison for BUSI dataset.

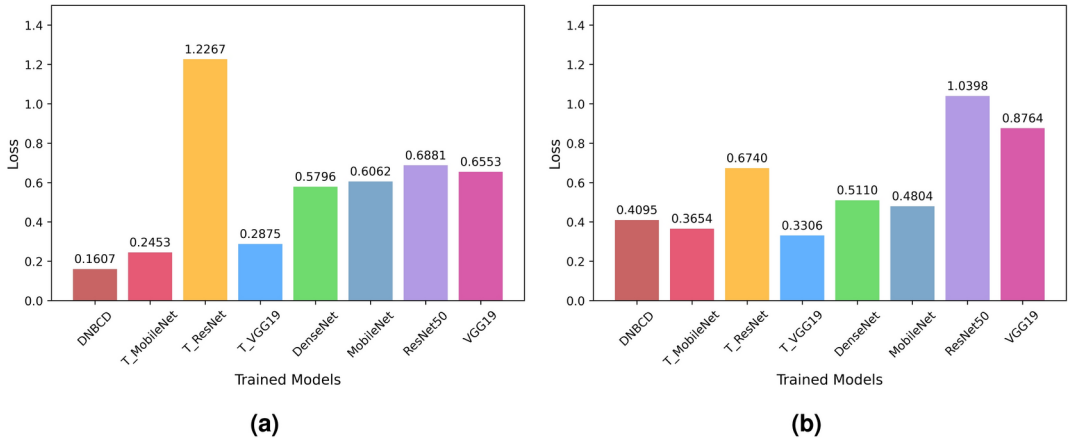


Fig. 14. Loss comparison of different trained state-of-the-art models for B-400x and BUSI datasets using bar charts, where (**a**) represents loss comparison for B-400x dataset, and (**b**) represents loss comparison for BUSI dataset.

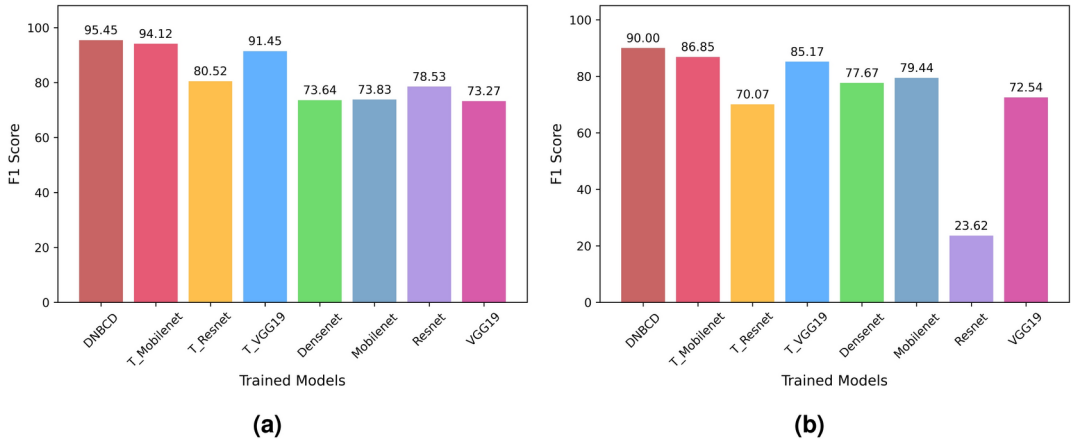


Fig. 15. F1-score comparison of different trained state-of-the-art models for B-400x and BUSI datasets using bar charts, where (**a**) represents F1-score comparison for B-400x dataset, and (**b**) represents F1-score comparison for BUSI dataset.

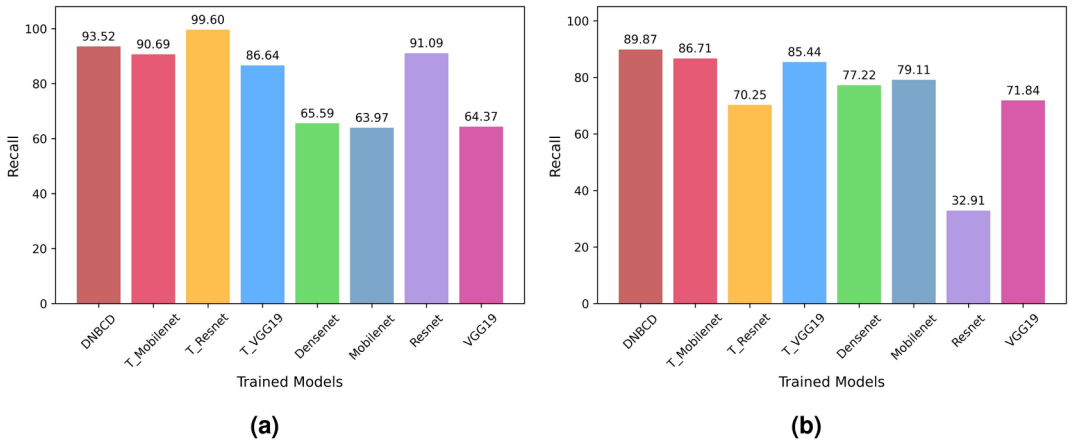


Fig. 16. Recall comparison of different trained state-of-the-art models for B-400x and BUSI datasets using bar charts, where (**a**) represents recall comparison for B-400x dataset, and (**b**) represents recall comparison for BUSI dataset.

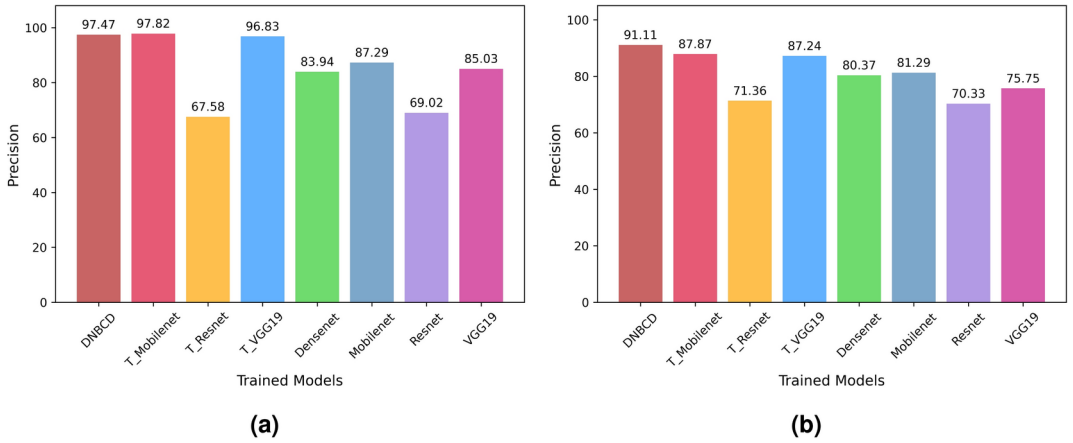


Fig. 17. Precision comparison of different trained state-of-the-art models for B-400x and BUSI datasets using bar charts, where (**a**) represents precision comparison for B-400x dataset, and (**b**) represents precision comparison for BUSI dataset.

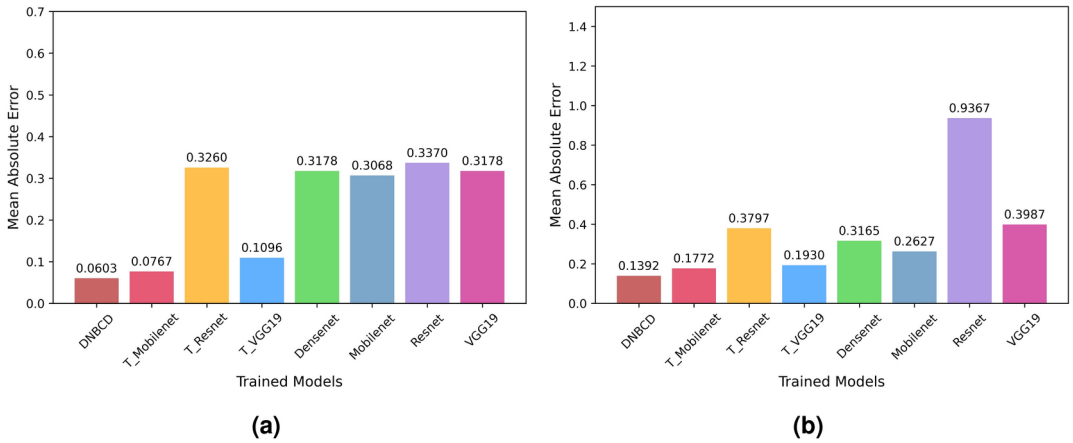


Fig. 18. Mean Absolute Error (MAE) comparison of different trained state-of-the-art models for B-400x and BUSI datasets using bar charts, where (**a**) represents MAE comparison for B-400x dataset, and (**b**) represents MAE comparison for BUSI dataset.

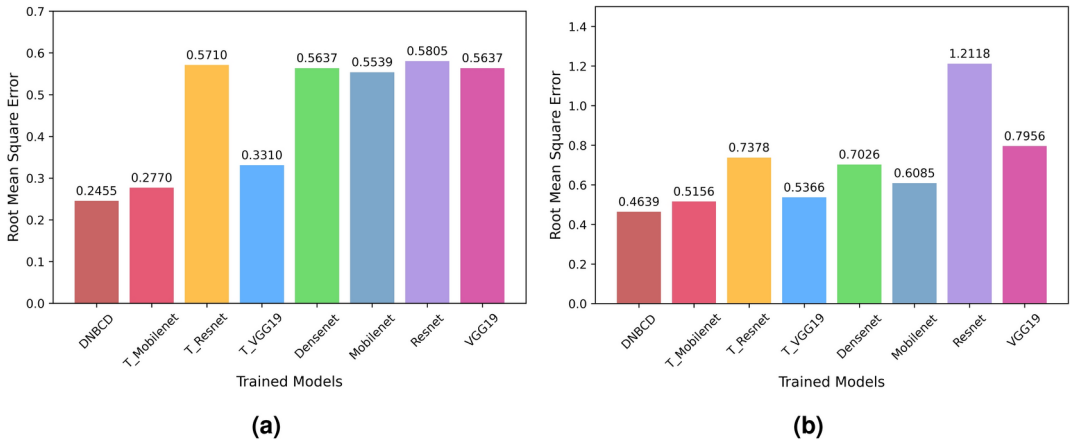


Fig. 19. Root Mean Square Error (RMSE) comparison of different trained state-of-the-art models for B-400x and BUSI datasets using bar charts, where (**a**) represents RMSE comparison for B-400x dataset, and (**b**) represents RMSE comparison for BUSI dataset.

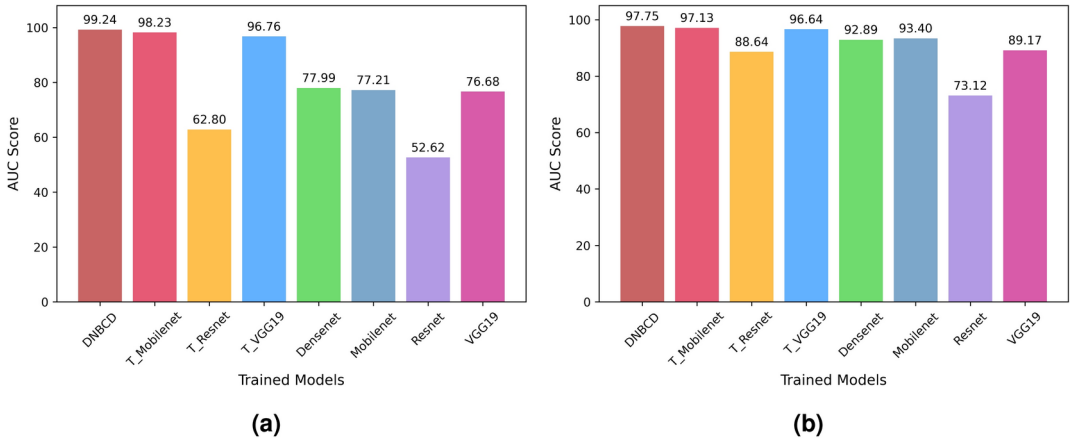


Fig. 20. AUC Score comparison of different trained state-of-the-art models for B-400x and BUSI datasets using bar charts, where (**a**) represents AUC score comparison for B-400x dataset, and (**b**) represents AUC score comparison for BUSI dataset.

Dataset	Model	Accuracy	Loss	F1 Score	Precision	Recall	MAE	RMSE	AUC
B-400x[28]	Proposed DNBCD	93.90 ± 0.4061	0.2479 ± 0.1013	95.52 ± 0.2770	95.04 ± 1.5461	96.05 ± 1.4717	0.0610 ± 0.0041	0.2468 ± 0.0083	98.58 ± 0.6701
BUSI[30]	Proposed DNBCD	89.87 ± 1.0329	0.4194 ± 0.0782	89.96 ± 0.9557	90.71 ± 0.6133	89.87 ± 1.0329	0.1361 ± 0.0221	0.4506 ± 0.0518	97.37 ± 0.4819

Table 7. Performance metrics of the proposed DNBCD model evaluated across multiple runs for the B-400x and BUSI datasets, presenting the average performance with corresponding standard deviations.

datasets. The consistency in the performance metrics indicates that the model is robust and can be expected to yield reliable results in practical applications. Figure 21 illustrates the performance metrics with error bars for our proposed DNBCD model, providing a comprehensive overview of its predictive capabilities. The dark red bar represents accuracy, highlighting the model's effectiveness in correctly classifying instances, while the pink bar denotes precision, which measures the proportion of true positive predictions relative to all positive predictions. The orange bar indicates recall, reflecting the model's ability to accurately identify all relevant instances. The light blue bar represents the F1 score, which balances precision and recall to provide a singular measure of performance. Additionally, the green bar illustrates the MAE, indicating the average magnitude of errors in predictions, while the light sky blue bar denotes the RMSE, emphasizing larger errors more significantly. The light pink bar signifies the AUC score, which evaluates the model's ability to differentiate between classes across various thresholds. Finally, the lavender bar represents model loss, indicating the average error incurred during predictions. Each bar is accompanied by black error bars that represent the standard deviation for each metric,

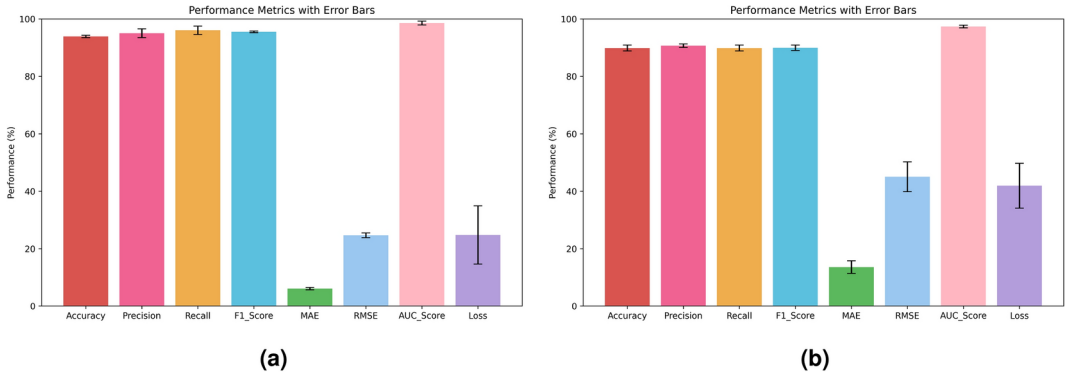


Fig. 21. Performance metrics with error bars for the B-400x and BUSI datasets where (**a**) represents performance with error bars using the B-400x dataset and (**b**) represents performance with error bars using the BUSI dataset, and Each colored bar denotes a different metric, and black error bars indicate the standard deviation across multiple runs.

Configuration	Train dataset	Test dataset	Performance (%)	Category
Configuration A	B-400x[28]	B-400x[28]	93.97	Intra dataset
Configuration B	B-400x[28]	B-100x[28]	80.32	Intra dataset
Configuration C	B-100x[28]	B-100x[28]	97.95	External dataset
Configuration D	B-100x[28]	B-400x[28]	76.44	External dataset
Configuration E	B-100x[28] + B-400x[28]	B-400x[28]	94.52	Merge dataset
Configuration F	B-100x[28] + B-400x[28]	B-100x[28]	98.47	Merge dataset
Configuration G	B-100x[28] + B-400x[28]	B-100x[28] + B-400x[28]	96.90	Merge dataset

Table 8. Performance accuracy with different dataset configurations.

indicating the variability and uncertainty of the performance across multiple runs. This detailed visualization enables a clear comparison of the DNBCD model's performance, revealing both its strengths and potential areas for enhancement.

In this research, we also evaluated our proposed model using an external set named Breakhis-100x (B-100x)[28], which is part of the Breakhis dataset[28], to assess its generalization properties. Table 8 displays the performance accuracy across different configurations. Configurations A and B utilized the internal dataset, training the model with the B-400x[28] dataset. In contrast, Configurations C and D involved the external dataset, with training conducted on the B-100x[28] dataset. Finally, Configurations E, F, and G represented merge datasets, training the model using both B-100x[28] and B-400x[28] datasets. Notably, Configuration B, which tested on the B-100x dataset, achieved an accuracy of 80.32%, demonstrating solid performance. The merge dataset configurations showed even better results: Configuration E reached an accuracy of 94.52%, Configuration F achieved an impressive 98.47%, and Configuration G recorded a performance of 96.90%. These results indicated that merging the datasets significantly enhanced model performance, suggesting that our proposed model was well-constructed and effectively performed on similar types of images from external datasets.

In this study, Table 9 represents the results of statistical significance tests comparing DNBCD with state-of-the-art trained models. Each comparison demonstrates that our proposed model, DNBCD, achieved significant T-statistic values and low *P*-values against all other models, indicating its effectiveness and superiority in performance.

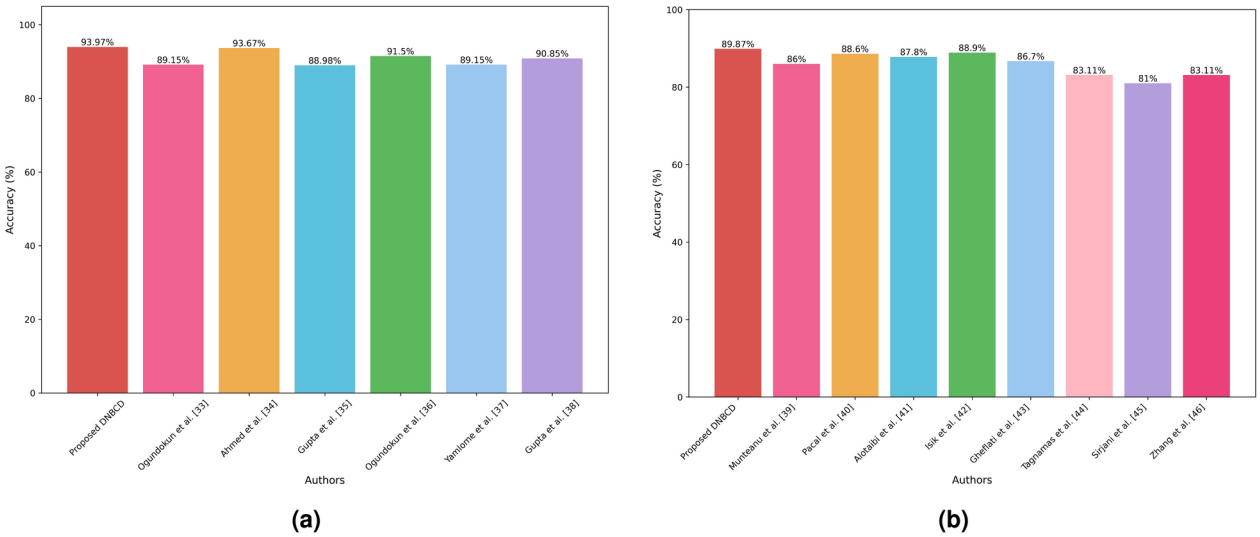
Additionally, we compared our results with existing work, as shown in Table 10, which also utilized these two datasets. Notably, our proposed DNBCD model surpassed the performance of other methods, demonstrating its superiority. Figure 22 provides a comparative analysis based on the BUSI[30] dataset and the B-400x[28] dataset[28], respectively. Where (a) represents accuracy comparison for the B-400x[28] dataset where the x-axis displays the authors, while the y-axis represents the performance metrics. The dark red bar corresponds to our proposed model, the bright pink bar represents the Ogundokun ei al.[33], the orange bar corresponds to the Ahmed ei al.[34], the blue bar represents the Gupta ei al.[35], the green bar corresponds to another Ogundokun ei al.[36], the sky blue bar represents the Yamlome ei al.[37], and the purple bar corresponds to the Gupta ei al.[38]. The results show that our proposed model performs significantly better than the others. On the other hand (b) represents accuracy comparison for the BUSI[30] dataset where the x-axis lists the authors and models used, while the y-axis represents the performance metrics. The dark red bar corresponds to our proposed model, the bright pink bar represents the Munteanu ei al.[39], the orange bar represents the Pacal ei al.[40], the blue bar corresponds to the Alotaibi ei al.[41], the green bar represents the Isik ei al.[42], the sky blue bar represents the Gheflati ei al.[43], the light pink bar represents Tagnamas ei al.[44], the purple bar corresponds to the Sirjani ei al.[45], and the dark pink bar represents

Model vs. Model	T-statistic	*P*-value	Statistical status
DNBCD vs. T_Mobilenet	4.4762	0.0065	Significant
DNBCD vs. T_Resnet	3.3723	0.0198	Significant
DNBCD vs. T_VGG19	5.8716	0.0020	Significant
DNBCD vs.Densenet[52]	5.0595	0.0039	Significant
DNBCD vs. Mobilenet[53]	6.6646	0.0011	Significant
DNBCD vs. Resnet50[54]	8.3633	0.0004	Significant
DNBCD vs. VGG19[55]	10.7553	0.0001	Significant

Table 9. Statistical significance tests of DNBCD with state-of-the-art trained models.

Dataset	Author	Model	Accuracy (%)
B-400x[28]	**Proposed**	**DNBCD**	**93.97**
	Ogundokun et al.[33]	Hybrid CNN-ANN	89.15
	Ahmed et al.[34]	Quantum-Optimized AlexNet	93.67
	Gupta et al.[35]	XSV+SRF	88.98
	Ogundokun et al.[36]	Mobilenet+SVM	91.50
	Yamlome et al.[37]	Transfer+CNN	89.15
	Gupta et al.[38]	Fine-tuned Resnet	90.85
BUSI[30]	**Proposed**	**DNBCD**	**89.87**
	Munteanu et al.[39]	UNet+CNN	86
	Pacal et al.[40]	Vision Transformer	88.6
	Alotaibi et al.[41]	VGG19	87.8
	Isik et al.[42]	ProtoNet+Resnet50	88.9
	Gheflati et al.[43]	ViTs	86.7
	Tagnamas et al.[44]	SCA-InceptionUNeXt	81.66
	Sirjani et al.[45]	InceptionV3	81
	Zhang et al.[46]	HAU-Net	83.11

Table 10. Performance comparison of different existing works with varying parameters for B-400x and BUSI dataset. Significant values are in bold.



(a) **(b)**

Fig. 22. Performance comparison of existing research with accuracy in graphical form, where (**a**) represents accuracy comparison using B-400x dataset and (**b**) represents accuracy comparison using BUSI dataset.

the Zhang ei al.[46]. The results indicate that our proposed model significantly outperformed the other models on the BUSI[30] dataset.

In this study, Table 11 presents a technical comparison of various trained state-of-the-art models evaluated on the B-400x[28] and BUSI[30] datasets, highlighting critical parameters such as trainable and non-trainable parameters, optimizers, total parameters, and training and testing times. Our proposed DNBCD model, with a total of 22,716,997 parameters, demonstrates a robust architecture designed for capturing complex features in breast cancer imaging. Despite its relatively high training time (TT: Total Training Time of 184.84 seconds for B-400x[28] and 236.45 seconds for BUSI[30]), this is justified by its ability to deliver more accurate diagnostic results. In comparison, while models like T_Resnet exhibit even greater complexity with 74002315 parameters, the DNBCD model effectively balances complexity and efficiency, ensuring that the time invested in training (FTTT: Final Total Training Time) and testing (TsT: Total Testing Time) translates into enhanced diagnostic capabilities. This analysis underscores the importance of developing models that not only excel in accuracy but also remain practical for clinical deployment, thereby contributing significantly to advancements in breast cancer diagnosis.

The current investigation, we explored the image enhancement technique known as Contrast Limited Adaptive Histogram Equalization (CLAHE). Despite its intention to improve image contrast, our results showed no performance increase for the B-400x dataset[28], where the accuracy of our proposed model was 86.85%. Similarly, for the BUSI dataset[30], the accuracy was only 73.10%. These findings suggest that CLAHE did not enhance model performance as expected. They highlight that techniques such as resizing, normalization, and data augmentation have proven to be more effective than image enhancement in our experiments. As a result, we conclude that image enhancement methods like CLAHE may not be suitable for building our proposed model.

This analysis demonstrates that the proposed model consistently outperforms all other models on both datasets, suggesting its superior capability in addressing the task at hand. According to these observations, our proposed DNBCD system exhibits highly accurate detection of breast cancer. Furthermore, we utilize the Grad-CAM technique to provide insightful visual explanations of the model's outputs. This method highlights the regions of the input images that significantly contribute to the model's decision-making process, enhancing the interpretability of the results.

Result discussion

In this study, the DNBCD model's performance analysis, several key insights emerged. The model achieved high accuracy and consistently outperformed baseline architectures, demonstrating its effectiveness in classifying breast cancer images. The Grad-CAM visualizations provided a transparent view into the model's decision-making by highlighting regions in the images that significantly influenced classifications. This interpretability is especially valuable in medical applications, as it allows healthcare professionals to validate and trust the model's predictions. Additionally, performance metrics such as accuracy, F1 score, and recall suggest the model's robustness across different datasets. Observed patterns, such as increased precision for benign cases and higher recall in malignant classifications, indicate that the model is adept at distinguishing subtle differences between categories. Overall, these results underscore the model's potential for practical application in clinical settings, offering an efficient, reliable, and interpretable tool for breast cancer diagnosis.

In this research, Figs. 23 and 24 illustrate the DNBCD model's prediction output for both datasets. Figure 23 represents the output sample for a histopathology image sample from the B-400x dataset, while Fig. 24 shows the output sample for a ultrasound image sample from the BUSI dataset. The first panel (left) presents the original tissue sample, while the second panel (center) displays the model's classification, labeled "Benign," indicating no malignant features. The third panel (right) displays the Grad-CAM heatmap, highlighting regions

Dataset	Model	Trainable	Non-trainable	Optimizer	Total params	TT	FTTT	TsT
B-400x[28]	Proposed DNBCD	7544449	83648	15088900	22716997	184.84	335.47	4.48
	T_Mobilenet	3797569	21888	7595140	11414597	155.14	177.32	1.53
	T_Resnet	24649473	53120	49298948	74001541	163.66	224.24	2.15
	T_VGG19	20352833	0	40705668	61058501	160.45	174.02	2.72
	Densenet[52]	1025	7037504	2052	7040581	176.02	–	5.02
	Mobilenet[53]	1025	3228864	2052	3231941	143.95	–	1.56
	Resnet50[54]	2049	23587712	4100	23593861	156.85	–	1.96
	VGG19[55]	513	20024384	1028	20025925	147.37	–	1.23
BUSI[30]	Proposed DNBCD	7544707	83648	15089416	22717771	236.45	397.70	8.44
	T_Mobilenet	11415371	3797827	21888	7595656	191.06	217.88	2.90
	T_Resnet	74002315	24649731	53120	49299464	209.21	263.13	3.66
	T_VGG19	61059275	20353091	0	40706184	197.11	211.81	4.74
	Densenet[52]	7046731	3075	7037504	6152	217.84	-	5.58
	Mobilenet[53]	3238091	3075	3228864	6152	186.21	-	2.36
	Resnet50[54]	23606155	6147	23587712	12296	195.19	-	2.85
	VGG19[55]	20029003	1539	20024384	3080	191.46	-	2.09

Table 11. Technical comparison of different trained state-of-the-art models with varying parameters for B-400x and BUSI datasets.

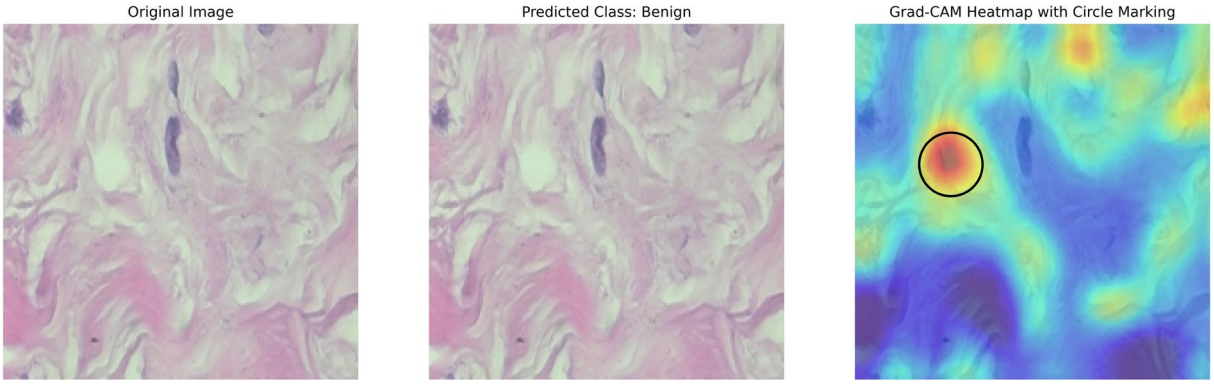


Fig. 23. Example output of DNBCD system for detecting breast cancer from the Break-400x dataset the first panel (left) representing original image having breast cancer, second panel (center) predicted class of benign from the original image, and third panel (right) GradCam heatmap image for explaining detected breast cancer region indicated with red and yellow color and marked by black circle for most affected area.

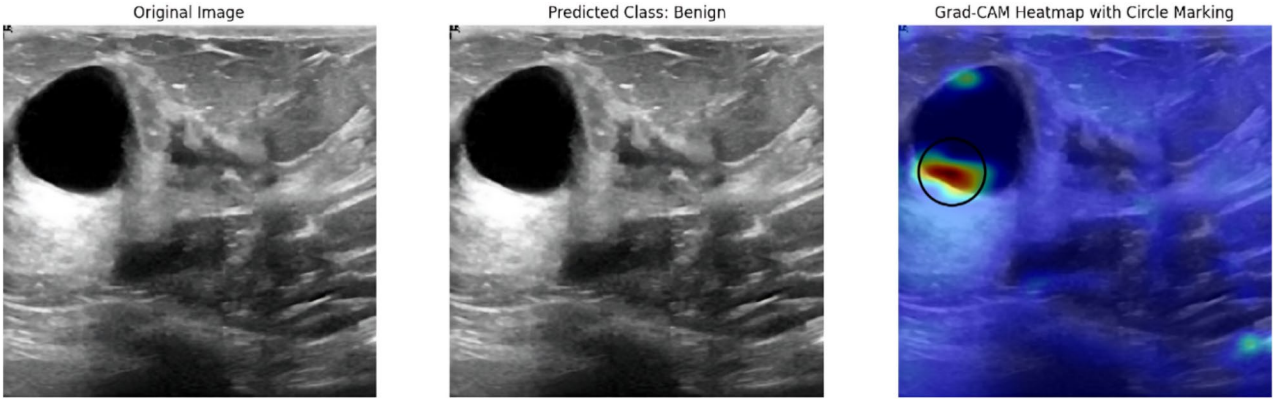


Fig. 24. Example output of DNBCD system for detecting breast cancer from the BUSI Dataset the first panel (left) representing original image having breast cancer, second panel (center) predicted class of benign from the original image, and third panel (right) GradCam heatmap image for explaining detected breast cancer region indicated with red and yellow Color and marked by black circle for most affected area.

that are significant to the model's decision. Warmer colors (red and yellow) indicate areas of interest, with red representing the highest affected regions and yellow indicating lower impact. The black circle highlights the most critical area identified by the model. Blue color areas are non-influential in this classification. On the other hand, Figs. 25 and 26 illustrate the DNBCD model's prediction output for both datasets. Figure 25 represents the output sample for a histopathology image sample from the B-400x dataset, while Fig. 26 shows the output sample for a ultrasound image sample from the BUSI dataset. The first panel (left) presents the original tissue sample, while the second panel (center) displays the model's classification, labeled "Malignant," indicating no benign features. The third panel (right) displays the Grad-CAM heatmap, highlighting regions that are significant to the model's decision. Warmer colors (red and yellow) indicate areas of interest, with red representing the highest affected regions and yellow indicating lower impact. The black circle highlights the most critical area identified by the model. Blue color areas are non-influential in this classification.

In this study, Fig. 27 illustrates the DNBCD model's prediction output for a histopathology image sample. The first panel (left) displays the original tissue sample, while the second panel (center) shows the model's classification as "Normal," indicating the absence of benign or malignant features. In contrast, Fig. 28 shows another sample where the model's prediction is "Benign." The first panel (left) presents the original tissue sample, and the second panel (center) displays the model's "Benign" classification, indicating no malignant features. The third panel (right) includes a Grad-CAM heatmap, highlighting regions the model deemed significant for its decision. Warmer colors (red and yellow) denote areas of interest, with red representing the most affected regions and yellow indicating lesser impact. The black circle marks the area of highest model focus, while blue areas are considered non-influential. However, this prediction is incorrect; the correct label is "Normal." The primary reason for this misclassification is a prominent white line in the original image, which does not indicate cancer but resembles benign features. This artifact likely led the DNBCD model to incorrectly classify the sample as "Benign."

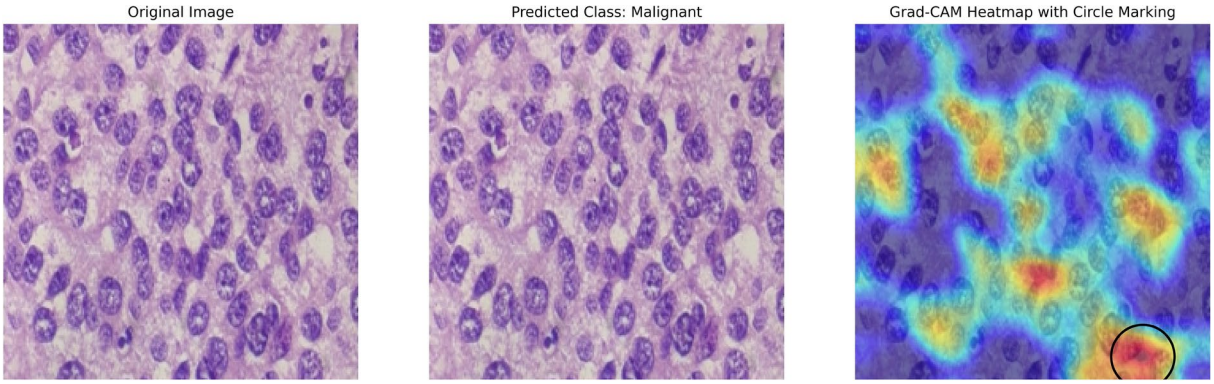


Fig. 25. Example output of DNBCD system for detecting breast cancer from the Break-400x dataset the first panel (left) representing original image having breast cancer, second panel (center) predicted class of malignant from the original image, and third panel (right) GradCam heatmap image for explaining detected breast cancer region indicated with red and yellow color and marked by black circle for most affected area.

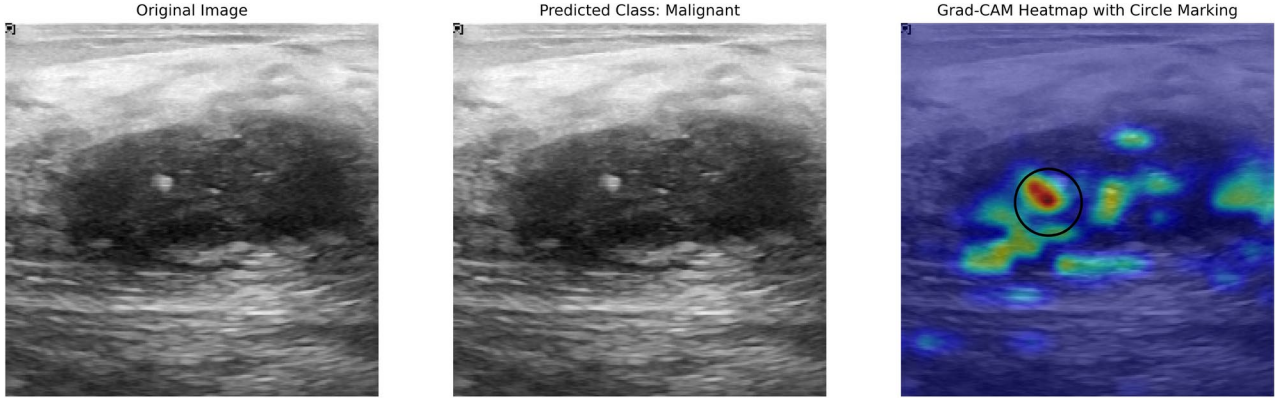


Fig. 26. Example output of DNBCD system for detecting breast cancer from the BUSI dataset the first panel (left) representing original image having breast cancer, second panel (center) predicted class of malignant from the original image, and third panel (right) GradCam heatmap image for explaining detected breast cancer region indicated with red and yellow color and marked by black circle for most affected area.

Additionally, we integrated Grad-CAM to enhance the transparency and interpretability of our DNBCD model, significantly improving the user experience. A major challenge with deep learning models is their "black-box" nature, making it difficult to understand how decisions are made. Grad-CAM addressed this by generating a heatmap that highlights the most important regions in an image, showing which areas influenced the model's classification. This was done by computing gradients of the predicted class score relative to the final convolutional layer. For better clarity, we overlaid the heatmap onto the original image, where high-intensity areas in red indicated the most relevant regions. The circle was drawn at the most activated region in the heatmap, found using `np.argmax()`, with its center being the highest intensity pixel. After determining the center, a 20-pixel radius circle was placed around this area, visually marking the model's key decision zone. This transparency allowed radiologists to cross-check AI predictions with medical expertise. By improving visual explanations, Grad-CAM increased trust in deep learning models, making AI-based breast cancer detection more explainable, reliable, and suitable for real-world clinical applications.

Limitations and future works

Limitation

While the DNBCD model shows promise, several limitations must be acknowledged regarding its deployment. A key concern is variability in image quality across datasets, influenced by factors such as lighting and resolution, which can hinder performance and generalizability. To address this, future work should incorporate a broader range of imaging conditions during training. The model's training primarily on the Breakhis-400x and BUSI datasets may limit its generalizability to diverse breast cancer imaging data encountered in clinical environments. Validating the model on a wider array of datasets that reflect various demographics and imaging modalities is essential. Despite attempts to mitigate class imbalance through weighting, this issue remains, particularly for underrepresented classes. This imbalance can significantly affect performance metrics, such as recall and precision. For instance, in Table 6, the T_Resnet model demonstrates the highest recall on the Breakhis-400x

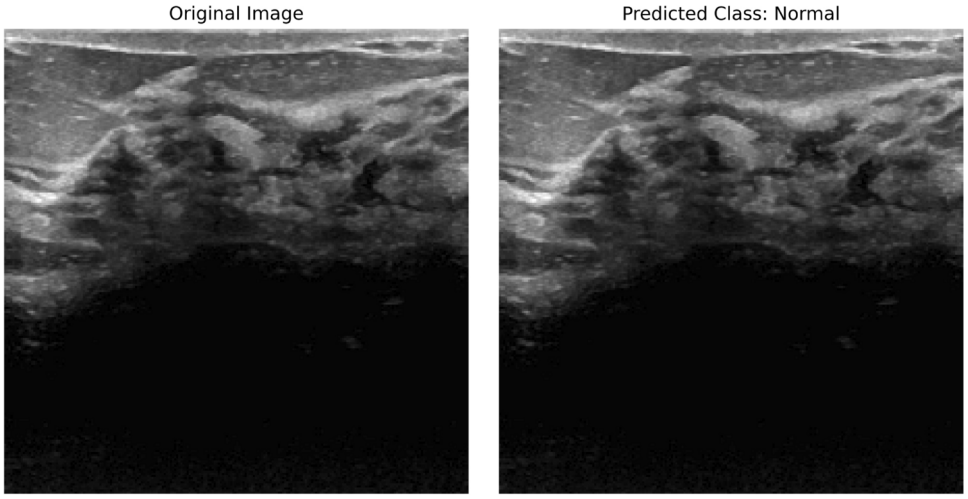


Fig. 27. Example output of DNBCD system for detecting non-breast cancer from the BUSI dataset the first panel (left) representing original image having non-breast cancer, second panel (right) predicted class of normal.

Fig. 28. Example output of DNBCD system for incorrect detecting breast cancer from the BUSI dataset the first panel (left) representing original image having non-breast cancer, second panel (center) predicted class of benign from the original image, and third panel (right) Grad Cam heatmap image for explaining detected breast cancer region indicated with red and yellow color and marked by black circle for most affected area.

dataset, but its precision is notably lower. Similarly, the BUSI dataset shows a disparity between recall and precision metrics, indicating that the model may excel in identifying certain classes while underperforming in others. Future research should explore strategies such as oversampling, synthetic data generation, or ensemble methods to better address this imbalance and improve the model's overall performance across all classes. Lastly, the model's complexity results in higher computational time compared to simpler models, posing challenges for real-time clinical application. Balancing accuracy with computational efficiency will be crucial in future iterations of the DNBCD system. By addressing these limitations, we can improve the robustness and applicability of the DNBCD model in real-world clinical settings.

Future work

This research demonstrates the potential of the DNBCD system to significantly enhance breast cancer diagnosis accuracy, thereby supporting healthcare providers in making early, informed decisions. Future work addresses identified limitations by extending the system's capability to handle more diverse and balanced datasets and progressing toward a clinical-grade diagnostic tool. Planned enhancements include fine-tuning the model for real-world deployment in clinical settings and incorporating additional multimodal imaging data, such as MRI, alongside existing ultrasound and histopathological images, to increase robustness and generalizability. Exploring advanced attention mechanisms may further improve the model's ability to focus on relevant features, leading to richer diagnostic insights. By overcoming these challenges and following a clear development timeline, we aim to create a practical, highly accurate diagnostic tool for widespread deployment, ultimately contributing to improved patient outcomes through the early and precise diagnosis of breast cancer.

Conclusion

This study presents the Deep Neural Breast Cancer Detection (DNBCD) system, a sophisticated deep learning-based framework tailored for accurate and automated breast cancer detection. The system leverages advanced preprocessing techniques and a deep convolutional neural network (CNN) architecture, complemented by interpretability , Gradient-weighted Class Activation Mapping (Grad-CAM) technique, which provides visual explanations of the model's decisions. The experimental results of the proposed model with the B-400x and BUSI datasets, demonstrate a robust capability to classify breast cancer cases with impressive accuracy higher than the existing approach for both datasets. The high performances indicate that DNBCD is highly accurate and adaptable across different imaging modalities, making it a versatile tool in breast cancer diagnostics. Importantly, DNBCD goes beyond classification by highlighting cancerous regions within images, allowing clinicians to visually interpret the areas that most influence the model's predictions. The integration of interpretability through Grad-CAM is a critical asset, as it addresses the limitations of deep learning models in medical applications, which is the lack of transparency. By visually marking affected regions, DNBCD enables healthcare professionals to understand the model's reasoning, which is essential for clinical adoption. This capability not only aids in confirming the model's findings but also serves as an educational tool for understanding subtle patterns in both cancerous and non-cancerous tissues. Overall, these results underscore DNBCD's potential as a reliable diagnostic tool, providing healthcare professionals with a resource that could significantly improve early detection and diagnosis of breast cancer. The system's high accuracy, combined with its interpretability, positions it as a promising advancement in digital health technology. In the future, our plan is to create a practical, highly accurate tool for widespread use, eventually enhancing patient outcomes by promoting early and precise breast cancer diagnosis.

Data availability

In the study, we used Breakhis-400x (B-400x), Breakhis-100x (B-100x) and Breast Ultrasound Images (BUSI) publicly available datasets, which can be downloaded from the below link: B-400x and B-100x Datasets (Publicly Available Dataset): https://web.inf.ufpr.br/vri/databases/breast-cancer-histopathological-database-breakhis/. BUSI (Publicly Available Dataset): https://www.kaggle.com/datasets/aryashah2k/breast-ultrasound-images-dataset.

Received: 8 December 2024; Accepted: 7 April 2025
Published online: 20 May 2025

References

1. Bray, F. et al. Global cancer statistics 2022: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* **74**, 229–263 (2024).
2. Organization, W. H. World health organization (who) — https://www.who.int/ (2024). [Accessed 20-08-2024].
3. Foundation, N. B. C. Breast cancer facts & stats 2024 - incidence, age, survival, & more. https://www.nationalbreastcancer.org/breast-cancer-facts/#~: (2024). [Accessed 08-08-2024].
4. Breastcancer.org. Breast cancer facts and statistics 2024 — breastcancer.org. https://www.breastcancer.org/facts-statistics (2024). [Accessed 08-08-2024].
5. for Biotechnology Information, N. C. Breast cancer early detection: a phased approach to implementation. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7237065/ (2024). [Accessed 08-08-2024].
6. Swaminathan, H., Saravanamurali, K. & Yadav, S. A. Extensive review on breast cancer its etiology, progression, prognostic markers, and treatment. *Med. Oncol.* **40**, 238 (2023).
7. Khan, S. I., Shahrior, A., Karim, R., Hasan, M. & Rahman, A. Multinet: A deep neural network approach for detecting breast cancer through multi-scale feature fusion. *J. King Saud Univer.-Comput. Inform. Sci.* **34**, 6217–6228 (2022).
8. Stachs, A., Stubert, J., Reimer, T. & Hartmann, S. Benign breast disease in women. *Dtsch. Arztebl. Int.* **116**, 565 (2019).
9. Zhao, H. The prognosis of invasive ductal carcinoma, lobular carcinoma and mixed ductal and lobular carcinoma according to molecular subtypes of the breast. *Breast Cancer* **28**, 187–195 (2021).
10. Khan, M. S. I. et al. Accurate brain tumor detection using deep convolutional neural network. *Comput. Struct. Biotechnol. J.* **20**, 4733–4745 (2022).
11. Bayram, B., Kunduracioglu, I., Ince, S. & Pacal, I. A systematic review of deep learning in MRI-based cerebral vascular occlusion-based brain diseases. *Neuroscience* https://doi.org/10.1016/j.neuroscience.2025.01.020 (2025).
12. Modiri, A., Goudreau, S., Rahimi, A. & Kiasaleh, K. Review of breast screening: Toward clinical realization of microwave imaging. *Med. Phys.* **44**, e446–e458 (2017).
13. Abhisheka, B., Biswas, S. K., Purkayastha, B., Das, D. & Escargueil, A. Recent trend in medical imaging modalities and their applications in disease diagnosis: A review. *Multim. Tools Appl.* **83**, 43035–43070 (2024).
14. Iacob, R. et al. Evaluating the role of breast ultrasound in early detection of breast cancer in low-and middle-income countries: A comprehensive narrative review. *Bioengineering* **11**, 262 (2024).
15. Zou, Y. *et al.* Precision matters: The value of pet/CT and pet/MRI in the clinical management of cervical cancer. *Strahlentherapie und Onkologie* 1–12 (2024).
16. Pacal, I. Maxcervixt: A novel lightweight vision transformer-based approach for precise cervical cancer detection. *Knowl.-Based Syst.* **289**, 111482 (2024).
17. Gardezi, S. J. S., Elazab, A., Lei, B. & Wang, T. Breast cancer detection and diagnosis using mammographic data: Systematic review. *J. Med. Internet Res.* **21**, e14464 (2019).
18. Rahman, A. et al. Machine learning and deep learning-based approach in smart healthcare: Recent advances, applications, challenges and opportunities. *AIMS Public Health* **11**, 58 (2024).
19. Dar, R. A. et al. Breast cancer detection using deep learning: Datasets, methods, and challenges ahead. *Comput. Biol. Med.* **149**, 106073 (2022).
20. Sharafaddini, A. M., Esfahani, K. K. & Mansouri, N. Deep learning approaches to detect breast cancer: A comprehensive review. *Multim. Tools Appl.* https://doi.org/10.1007/s11042-024-20011-6 (2024).
21. Bilal, A. et al. Bc-qnet: A quantum-infused elm model for breast cancer diagnosis. *Comput. Biol. Med.* **175**, 108483 (2024).
22. COŞKUN, D. *et al.* A comparative study of yolo models and a transformer-based yolov5 model for mass detection in mammograms. *Turkish J. Electr. Eng. Comput. Sci.* **31**, 1294–1313 (2023).
23. Zhuang, F. et al. A comprehensive survey on transfer learning. *Proc. IEEE* **109**, 43–76 (2020).

24. Yao, X. et al. Fusion of shallow and deep features from 18f-FDG PET/CT for predicting EGFR-sensitizing mutations in non-small cell lung cancer. *Quant. Imaging Med. Surg.* **14**, 5460 (2024).
25. Pacal, I. A novel swin transformer approach utilizing residual multi-layer perceptron for diagnosing brain tumors in MRI images. *Int. J. Mach. Learn Cybernet.* **15**(9), 3579–3597 (2024).
26. Hassija, V. et al. Interpreting black-box models: a review on explainable artificial intelligence. *Cogn. Comput.* **16**, 45–74 (2024).
27. Cheng, Z. et al. Application of serum sers technology based on thermally annealed silver nanoparticle composite substrate in breast cancer. *Photodiagn. Photodyn. Ther.* **41**, 103284 (2023).
28. Spanhol, F. A., Oliveira, L. S., Petitjean, C. & Heutte, L. A dataset for breast cancer histopathological image classification. *IEEE Trans. Biomed. Eng.* **63**, 1455–1462. https://doi.org/10.1109/TBME.2015.2496264 (2015).
29. Xu, X. et al. Large-field objective lens for multi-wavelength microscopy at mesoscale and submicron resolution. *Opto-Electr. Adv.* **7**(6), 230212–1 (2024).
30. Al-Dhabyani, W., Gomaa, M., Khaled, H. & Fahmy, A. Dataset of breast ultrasound images. *Data Brief* **28**, 104863. https://doi.org/10.1016/j.dib.2019.104863 (2020).
31. Chen, L. et al. Hpda/zn as a creb inhibitor for ultrasound imaging and stabilization of atherosclerosis plaque. *Chin. J. Chem.* **41**, 199–206 (2023).
32. Selvaraju, R. R. et al. Grad-cam: Visual explanations from deep networks via gradient-based localization. *Int. J. Comput. Vision* **128**, 336–359 (2020).
33. Ogundokun, R. O. *et al.* Hybrid deep learning for breast cancer diagnosis: Evaluating CNN and ann on breakhis_v1_400x. in *2024 International Conference on Science, Engineering and kausiness for Driving Sustainable Development Goals (SEB4SDG)*, 1–6 (IEEE, 2024).
34. Ahmed, H. K., Tantawi, B., Magdy, M. & Sayed, G. I. Quantum optimized alexnet for histopathology breast image diagnosis. in *International Conference on Advanced Intelligent Systems and Informatics*, 348–357 (Springer, 2023).
35. Gupta, M. et al. Deep transfer learning hybrid techniques for precision in breast cancer tumor histopathology classification. *Health Inform. Sci. Syst.* **13**, 20 (2025).
36. Ogundokun, R. O., Misra, S., Akinrotimi, A. O. & Ogul, H. Mobilenet-svm: A lightweight deep transfer learning model to diagnose bch scans for iomt-based imaging sensors. *Sensors* **23**, 656 (2023).
37. Yamlome, P., Akwaboah, A. D., Marz, A. & Deo, M. Convolutional neural network based breast cancer histopathology image classification. in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 1144–1147 (IEEE, 2020).
38. Gupta, V. & Bhavsar, A. Partially-independent framework for breast cancer histopathological image classification. in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, (2019).
39. Munteanu, B. -Ş, Murariu, A., Nichitean, M., Pitac, L.-G. & Dioşan, L. Value of original and generated ultrasound data towards training robust classifiers for breast cancer identification. *Inform. Syst. Front.* **27**(1), 75–96 (2024).
40. Pacal, İ. Deep learning approaches for classification of breast cancer in ultrasound (us) images. *J. Instit. Sci. Technol.* **12**, 1917–1927 (2022).
41. Alotaibi, M. et al. Breast cancer classification based on convolutional neural network and image fusion approaches using ultrasound images. *Heliyon* **9**(11), 22406 (2023).
42. Işık, G. & Paçal, İ. Few-shot classification of ultrasound breast cancer images using meta-learning algorithms. *Neural Comput. Appl.* **36**(20), 12047–12059 (2024).
43. Gheflati, B. & Rivaz, H. Vision transformers for classification of breast ultrasound images. in *2022 44th annual international conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 480–483 (IEEE, 2022).
44. Tagnamas, J., Ramadan, H., Yahyaouy, A. & Tairi, H. Sca-inceptionunext: A lightweight spatial-channel-attention-based network for efficient medical image segmentation. *Knowl.-Based Syst.* https://doi.org/10.1016/j.knosys.2025.11316 (2025).
45. Sirjani, N. et al. A novel deep learning model for breast lesion classification using ultrasound images: A multicenter data evaluation. *Physica Med.* **107**, 102560 (2023).
46. Zhang, H. et al. Hau-net: Hybrid cnn-transformer for breast ultrasound image segmentation. *Biomed. Signal Process. Control* **87**, 105427 (2024).
47. Bilal, A. et al. Breast cancer diagnosis using support vector machine optimized by improved quantum inspired grey wolf optimization. *Sci. Rep.* **14**, 10714 (2024).
48. Zeng, Q. et al. Serum raman spectroscopy combined with convolutional neural network for rapid diagnosis of her2-positive and triple-negative breast cancer. *Spectrochim. Acta Part A Mol. Biomol. Spectrosc.* **286**, 122000 (2023).
49. Ma, X. et al. Detection of breast cancer based on novel porous silicon Bragg reflector surface-enhanced Raman spectroscopy-active structure. *Chin. Opt. Lett.* **18**, 051701 (2020).
50. Vulli, A. et al. Fine-tuned densenet-169 for breast cancer metastasis prediction using fastai and 1-cycle policy. *Sensors* **22**, 2988 (2022).
51. Srinivasu, P. N. et al. Xai-driven catboost multi-layer perceptron neural network for analyzing breast cancer. *Sci. Rep.* **14**, 28674 (2024).
52. Samudrala, S. & Mohan, C. K. Semantic segmentation of breast cancer images using densenet with proposed pspnet. *Multim. Tools Appl.* **83**, 46037–46063 (2024).
53. Dafni Rose, J., VijayaKumar, K., Singh, L. & Sharma, S. K. Computer-aided diagnosis for breast cancer detection and classification using optimal region growing segmentation with mobilenet model. *Concurr. Eng.* **30**, 181–189 (2022).
54. Behar, N. & Shrivastava, M. Resnet50-based effective model for breast cancer classification using histopathology images. *CMES-Computer Modeling in Engineering & Sciences* **130** (2022).
55. Albashish, D., Al-Sayyed, R., Abdullah, A., Ryalat, M. H. & Almansour, N. A. Deep CNN model based on vgg16 for breast cancer classification. in *2021 International conference on information technology (ICIT)*, 805–810 (IEEE, 2021).
56. Ahmad, J., Akram, S., Jaffar, A., Rashid, M. & Bhatti, S. M. Breast cancer detection using deep learning: An investigation using the ddsm dataset and a customized alexnet and support vector machine. *IEEE Access* (2023).
57. Folorunso, S. O., Awotunde, J. B., Rangaiah, Y. P. & Ogundokun, R. O. Efficientnets transfer learning strategies for histopathological breast cancer image analysis. *Int. J. Model., Simul., Scient. Comput.* **15**, 2441009 (2024).
58. Joseph, V. R. & Vakayil, A. Split: An optimal method for data splitting. *Technometrics* **64**, 166–176 (2022).
59. Debnath, T. et al. Four-layer convnet to facial emotion recognition with minimal epochs and the significance of data diversity. *Sci. Rep.* **12**, 6991 (2022).
60. Spelmen, V. S. & Porkodi, R. A review on handling imbalanced data. in *2018 international conference on current trends towards converging technologies (ICCTCT)*, 1–11 (IEEE, 2018).
61. Islam, S. et al. Sgbba: An efficient method for prediction system in machine learning using imbalance dataset. *Int. J. Adv. Comput. Sci. Appl.* **12**(3), 1–12 (2021).
62. Kundu, D. *et al.* Federated deep learning for monkeypox disease detection on gan-augmented dataset. *IEEE Access* (2024).
63. Agarwal, M., Gupta, S. & Biswas, K. K. A new conv2d model with modified relu activation function for identification of disease type and severity in cucumber plant. *Sustain. Comput.: Inform. Syst.* **30**, 100473 (2021).
64. Brutzkus, A. & Globerson, A. An optimization and generalization analysis for max-pooling networks. in *Uncertainty in Artificial Intelligence*, 1650–1660 (PMLR, 2021).

65. Mesran, M. et al. Investigating the impact of relu and sigmoid activation functions on animal classification using CNN models. *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)* **8**, 111–118 (2024).
66. Islam, M. A., Kowal, M., Jia, S., Derpanis, K. G. & Bruce, N. D. Position, padding and predictions: A deeper look at position information in CNNS. *Int. J. Comput. Vision* **132**(9), 3889–3910 (2024).
67. Segu, M., Tonioni, A. & Tombari, F. Batch normalization embeddings for deep domain generalization. *Patt. Recogn.* **135**, 109115 (2023).
68. Kayadibi, İ & Güraksın, G. E. An explainable fully dense fusion neural network with deep support vector machine for retinal disease determination. *Int. J. Comput. Intell. Syst.* **16**, 28 (2023).
69. Meyerowitz-Katz, G. et al. Rates of attrition and dropout in app-based interventions for chronic disease: Systematic review and meta-analysis. *J. Med. Internet Res.* **22**, e20283 (2020).
70. Szandała, T. Review and comparison of commonly used activation functions for deep neural networks. *Bio-inspired neurocomputing* 203–224 (2021).
71. Iwana, B. K., Kuroki, R. & Uchida, S. Explaining convolutional neural networks using softmax gradient layer-wise relevance propagation. in *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, 4176–4185 (IEEE, 2019).
72. Hossain, M. M. *et al.* Crime text classification and drug modeling from bengali news articles: A transformer network-based deep learning approach. in *2023 26th International Conference on Computer and Information Technology (ICCIT)*, 1–6 (IEEE, 2023).
73. Muntaqim, M. Z. et al. Eye disease detection enhancement using a multi-stage deep learning approach. *IEEE Access* https://doi.org/10.1109/ACCESS.2024.3476412 (2024).
74. Rahim, M. A. et al. An enhanced hybrid model based on CNN and BILSTM for identifying individuals via handwriting analysis. *CMES-Comput. Model. Eng. Sci.* **140**(2), 1689–1710 (2024).
75. Rahman, A. et al. Federated learning-based AI approaches in smart healthcare: Concepts, taxonomies, challenges and open issues. *Clust. Comput.* **26**, 2271–2311 (2023).
76. Raghavan, K. Attention guided grad-cam: An improved explainable artificial intelligence model for infrared breast cancer detection. *Multim. Tools Appl.* **83**, 57551–57578 (2024).
77. Kaggle: Your Machine Learning and Data Science Community — kaggle.com. https://www.kaggle.com/. [Accessed 25-09-2024].
78. Alom, M. R. *et al.* Enhanced road lane marking detection system: A cnn-based approach for safe driving. in *2023 5th International Conference on Sustainable Technologies for Industry 5.0 (STI)*, 1–6, https://doi.org/10.1109/STI59863.2023.10464405 (IEEE, 2023).
79. Gutta, S. Machine Learning Metrics in simple terms — medium.com. https://medium.com/analytics-vidhya/machine-learning-metrics-in-simple-terms-d58a9c85f9f6. [Accessed 15-03-2025].
80. Kim, T. K. T test as a parametric statistic. *Korean J. Anesthesiol.* **68**, 540–546 (2015).

Author contributions

Conceptualization: MRA, MAR, and AR; Methodology: MRA, MAR, AR, TD, and ASMM; Implementation: MRA and MAR; Writing-Original draft preparation: MRA, MAR, AR, and TD; Writing-review and Editing: MRA, MAR, AR, TD, FAF, ASMM, and SM; Supervision: MAR and AR; Funding Acquisition: FAF, ASMM, and SM. All authors have read and agreed to the published version of the manuscript.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to M.A.R., A.R. or S.M.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Terms and Conditions

Springer Nature journal content, brought to you courtesy of Springer Nature Customer Service Center GmbH (“Springer Nature”).
Springer Nature supports a reasonable amount of sharing of research papers by authors, subscribers and authorised users (“Users”), for small-scale personal, non-commercial use provided that all copyright, trade and service marks and other proprietary notices are maintained. By accessing, sharing, receiving or otherwise using the Springer Nature journal content you agree to these terms of use (“Terms”). For these purposes, Springer Nature considers academic use (by researchers and students) to be non-commercial.
These Terms are supplementary and will apply in addition to any applicable website terms and conditions, a relevant site licence or a personal subscription. These Terms will prevail over any conflict or ambiguity with regards to the relevant terms, a site licence or a personal subscription (to the extent of the conflict or ambiguity only). For Creative Commons-licensed articles, the terms of the Creative Commons license used will apply.
We collect and use personal data to provide access to the Springer Nature journal content. We may also use these personal data internally within ResearchGate and Springer Nature and as agreed share it, in an anonymised way, for purposes of tracking, analysis and reporting. We will not otherwise disclose your personal data outside the ResearchGate or the Springer Nature group of companies unless we have your permission as detailed in the Privacy Policy.
While Users may use the Springer Nature journal content for small scale, personal non-commercial use, it is important to note that Users may not:

1. use such content for the purpose of providing other users with access on a regular or large scale basis or as a means to circumvent access control;
2. use such content where to do so would be considered a criminal or statutory offence in any jurisdiction, or gives rise to civil liability, or is otherwise unlawful;
3. falsely or misleadingly imply or suggest endorsement, approval , sponsorship, or association unless explicitly agreed to by Springer Nature in writing;
4. use bots or other automated methods to access the content or redirect messages
5. override any security feature or exclusionary protocol; or
6. share the content in order to create substitute for Springer Nature products or services or a systematic database of Springer Nature journal content.

In line with the restriction against commercial use, Springer Nature does not permit the creation of a product or service that creates revenue, royalties, rent or income from our content or its inclusion as part of a paid for service or for other commercial gain. Springer Nature journal content cannot be used for inter-library loans and librarians may not upload Springer Nature journal content on a large scale into their, or any other, institutional repository.
These terms of use are reviewed regularly and may be amended at any time. Springer Nature is not obligated to publish any information or content on this website and may remove it or features or functionality at our sole discretion, at any time with or without notice. Springer Nature may revoke this licence to you at any time and remove access to any copies of the Springer Nature journal content which have been saved.
To the fullest extent permitted by law, Springer Nature makes no warranties, representations or guarantees to Users, either express or implied with respect to the Springer nature journal content and all parties disclaim and waive any implied warranties or warranties imposed by law, including merchantability or fitness for any particular purpose.
Please note that these rights do not automatically extend to content, data or other material published by Springer Nature that may be licensed from third parties.
If you would like to use or distribute our Springer Nature journal content to a wider audience or on a regular basis or in any other manner not expressly permitted by these Terms, please contact Springer Nature at

onlineservice@springernature.com