International Journal of Science and Research Archive
eISSN: 2582-8185
Cross Ref DOI: 10.30574/ijsra
Journal homepage: https://ijsra.net/

(REVIEW ARTICLE)

Check for updates

Stacking ensemble-based breast cancer classification: Enhancing diagnostic accuracy with deep learning and real-time web deployment

Farhan Bin Jashim [1], Fajle Rabbi Refat [1], Mohammad Hasnatul Karim [1], Farhad Uddin Mahmud [2] and Anamul Haque Sakib [3,*]

[1] *Department of Business Administration and Management, International American University, CA 90010, USA.*
[2] *Department of Business Administration in Management Information Systems, International American University, CA 90010, USA.*
[3] *Department of Business Administration, International American University, Los Angeles, CA 90010, USA.*

International Journal of Science and Research Archive, 2025, 15(02), 1417–1431

Publication history: Received on 08 April 2025; revised on 27 May 2025; accepted on 29 May 2025

Article DOI: https://doi.org/10.30574/ijsra.2025.15.2.1502

Abstract

Breast cancer remains one of the most prevalent and life-threatening diseases, requiring early and accurate diagnosis to improve survival rates. Traditional diagnostic methods rely on manual interpretation of ultrasound and histopathology images, which are time-consuming, prone to variability, and dependent on expert radiologists and pathologists. Recent advances in deep learning have shown promise in automating breast cancer detection; however, existing models often suffer from overfitting, dataset biases, and poor generalization across different imaging modalities. To address these challenges, we propose a novel stacking ensemble-based breast cancer classification model integrating EfficientNetB8, RegNet, RepVGG, and MNasNet. Our approach enhances classification robustness by leveraging complementary feature extraction capabilities of multiple architectures. We evaluate our model on two publicly available datasets—BUSI (ultrasound) and BreaKHis (histopathology)—demonstrating superior performance over previous deep learning approaches. Our ensemble model achieves a maximum MCC of 99.31% on the BUSI dataset and 99.52% on the BreaKHis dataset, outperforming individual architectures. Additionally, we incorporate Contrast Limited Adaptive Histogram Equalization for contrast enhancement and employ data augmentation to mitigate class imbalance and improve model generalization. Furthermore, we develop a web-based diagnostic system for real-time breast cancer classification, enabling efficient and accessible clinical decision-making. While the proposed approach significantly enhances classification accuracy, future research will focus on dataset expansion, real-world validation, and explainable AI integration for improved interpretability and clinical adoption.

Keywords: Breast Cancer; Deep Learning; Stacking Ensemble; Ultrasound; Histopathology; Medical Imaging

1. Introduction

Breast cancer is the most commonly diagnosed cancer worldwide, with over 2.3 million new cases annually, accounting for 12.5% of all newly diagnosed cancers [1]. It remains the leading cause of cancer-related mortality among women, contributing to approximately 685,000 deaths each year (GLOBOCAN 2023) [2]. Early detection significantly improves survival rates, with studies indicating that timely diagnosis can reduce mortality by up to 40% [3]. Medical imaging techniques such as ultrasound and histopathology play a crucial role in breast cancer detection due to their non-invasive nature and ability to differentiate malignant from benign tumors [4]. However, traditional manual interpretation of medical images remains time-consuming, subjective, and prone to inter-observer variability, necessitating automated AI-driven solutions to enhance diagnostic precision, efficiency, and accessibility [5]. Deep learning, particularly CNNs, has shown state-of-the-art performance in breast cancer classification, often surpassing human radiologists in

* Corresponding author: Anamul Haque Sakib.

sensitivity and specificity [6]. AI-assisted diagnostic tools have demonstrated the potential to improve tumor detection rates by up to 20% while reducing false positives, leading to more accurate decision-making in clinical settings [7], [8]. However, existing deep learning-based approaches face several critical challenges, including overfitting, dataset biases, poor generalization across imaging modalities, and difficulty in handling class imbalances [9], [10]. CNN models trained on a specific dataset often struggle to generalize to unseen medical images, leading to inconsistencies in predictions. Moreover, the lack of interpretability in deep learning models raises concerns regarding trust and adoption in real-world clinical applications [11], [12].

To address these challenges, this study proposes a stacking ensemble model that integrates EfficientNetB8, RegNet, RepVGG, and MNasNet to enhance classification robustness and generalization across datasets [13]. Each architecture contributes distinct advantages: EfficientNetB8 optimizes feature extraction through compound scaling, RegNet provides structural adaptability, RepVGG ensures computational efficiency, and MNasNet enhances latency-aware optimization. By aggregating feature representations from multiple architectures, the ensemble model mitigates individual model biases, improves sensitivity for malignant cases, and reduces false positive rates, making it more reliable for breast cancer screening. Medical image preprocessing plays a critical role in enhancing classification accuracy [14], [15]. This study incorporates Contrast Limited Adaptive Histogram Equalization (CLAHE) to improve local contrast, ensuring better differentiation between tumor and non-tumor regions in ultrasound and histopathology images. Additionally, data augmentation techniques such as rotation, flipping, and zooming address class imbalances and enable better generalization across diverse imaging datasets [16]. These techniques significantly improve the model's ability to distinguish benign from malignant tumors, overcoming common limitations faced by CNN-based classification models in medical imaging. Beyond improving classification accuracy, real-world deployment of AI-based diagnostic systems remains a key challenge. To bridge the gap between research and clinical applications, this study introduces a web-based breast cancer classification system, allowing healthcare professionals to upload ultrasound or histopathology images and receive instant AI-driven diagnostic predictions [17]. This interactive platform enables faster and more efficient medical decision-making, reducing reliance on manual analysis and improving early intervention strategies. The proposed system has the potential to be integrated into clinical decision-support tools, enhancing workflow efficiency and facilitating timely, data-driven breast cancer diagnosis in resource-limited settings. The key contributions of this research are

- A novel stacking ensemble model integrating EfficientNetB8, RegNet, RepVGG, and MNasNet, improving classification robustness and outperforming individual architectures.
- An optimized preprocessing pipeline incorporating CLAHE and advanced data augmentation techniques to enhance contrast, mitigate class imbalance, and improve model generalization.
- A real-time web-based breast cancer classification system that enables instant AI-driven tumor detection, facilitating seamless integration into clinical workflows.
- Comprehensive evaluation and benchmarking on BUSI (ultrasound) and BreaKHis (histopathology) datasets, demonstrating superior performance over previous state-of-the-art deep learning models.

The remainder of this paper is organized as follows: Section II reviews related works on breast cancer classification using deep learning, Section III details the methodology, including data collection, preprocessing, and model development, Section IV presents experimental results, and Section V discusses findings, limitations, and future research directions.

2. Related Work

Deep learning has played a transformative role in breast cancer detection, segmentation, and classification by leveraging medical imaging and automated feature extraction. Various studies have focused on CNNs, ensemble learning, and generative models to enhance accuracy, yet challenges such as dataset imbalance, lack of web-based deployment, and model generalization persist.

Abhisheka et al. [18] provided a comprehensive study of deep learning-based breast cancer detection across multiple imaging modalities, including mammography, ultrasound, MRI, CT, PET, and histopathology. They analyzed public datasets such as DDSM, INbreast, and MIAS, highlighting the evolution from manual to CNN-based detection methods. The study emphasized ensemble learning techniques and pre-processing strategies, achieving classification accuracies up to 99.7%. However, the review noted the lack of web-based applications for real-time diagnosis and inconsistencies in dataset standardization, which limit model adaptability. Raza et al[19] introduced Deep Breast Cancer Net, a 24-layer deep learning model for breast cancer detection using ultrasound images. The architecture incorporated convolutional and inception modules, trained on 1030 ultrasound images across benign, malignant, and normal classes. The model

achieved a classification accuracy of 99.35% and 99.63% for binary classification. Despite its high performance, the study relied on small, imbalanced datasets, limiting the model's ability to generalize across diverse clinical environments. Additionally, the absence of advanced ensemble strategies reduced the robustness of predictions in complex cases.

Sharmin et al. [20] proposed a hybrid model integrating deep feature extraction with ensemble-based machine learning techniques. Using ResNet50V2, the study extracted features from the IDC histopathology dataset, consisting of 2000 images and 277,524 patches. LightGBM outperformed other classifiers, achieving 95% accuracy. While the study demonstrated the effectiveness of ensemble learning, it did not explore model stacking techniques, which could further enhance prediction reliability. Furthermore, the approach focused solely on IDC classification, limiting its applicability to broader breast cancer subtypes [21].

Asadi et al. [22] developed a cascade deep learning framework for breast cancer detection, integrating UNet for tumor segmentation and ResNet50 for classification. The study utilized 2780 mammography images from ImageNet, achieving 98.61% classification accuracy and a 97.30% F1-score for segmentation. While the method effectively combined segmentation and classification, segmentation errors and tumor morphology variations affected generalization. Additionally, the model lacked data augmentation strategies to address class imbalances, which could further refine its robustness.

Sulaiman et al. [23] introduced an attention-based U-Net model for breast cancer segmentation using the BUSI ultrasound dataset. By incorporating attention gates in the decoder blocks, the model achieved 98% accuracy, 97% precision, and a dice score of 92%. However, the study relied on limited training data, which may lead to overfitting. Moreover, no advanced augmentation strategies were implemented to mitigate data imbalance, and the approach did not consider web-based deployment for real-time clinical use.

Aumente-Maestro et al. [24] presented a multi-task learning framework combining segmentation and classification using ultrasound images. The study introduced a curated BUSI dataset, removing mislabeled images. UNet++ and nnU-Net were used for segmentation, achieving F1-scores of 82.6% for benign, 79.1% for malignant, and 74.1% for normal cases. Despite improving segmentation-classification coherence, inconsistencies in ground truth annotations and a lack of ensemble-based optimization reduced overall reliability. The absence of a robust fusion mechanism also limited the model's ability to learn complementary features across tasks.

Balasubramanian et al. [25] employed an ensemble deep learning approach for breast cancer subtype classification using histopathology images. Their study utilized the BACH and BreakHis datasets, integrating VGG16, ResNet34, and ResNet50, achieving 95.31% patch classification accuracy (BACH) and 98.43% whole-slide image classification accuracy (BreakHis). While ensemble learning improved classification, the study did not explore stacking-based ensemble strategies, which could have further strengthened the model's generalization across histopathological datasets. Additionally, the approach lacked interpretability tools like Grad-CAM, which are crucial for explainable AI in medical applications.

Rai et al. [26] explored the use of synthetic data generation for breast cancer detection in ultrasound imaging. The study integrated 3186 real ultrasound images from four public datasets with 10,000 synthetic images generated via StyleGAN3. EfficientNet-B7, trained on the combined dataset, improved classification accuracy from 88.72% to 92.01%. While the study demonstrated the potential of synthetic augmentation, the generated images introduced artificial artifacts, which may affect real-world generalization. Additionally, the study did not incorporate multi-modal fusion techniques, which could enhance robustness by combining ultrasound data with other imaging modalities.

Despite these advancements, key limitations persist in breast cancer classification. Most studies do not incorporate advanced ensemble stacking, missing opportunities for improved robustness. Additionally, data augmentation strategies remain underutilized, affecting generalization across clinical settings. Another critical gap is the lack of real-time web applications, which could facilitate immediate clinical decision-making. To address these challenges, our study proposes an ensemble-based deep learning framework with optimized data augmentation, feature fusion techniques, and explainable AI integration, ensuring improved diagnostic reliability and accessibility across diverse breast cancer datasets.

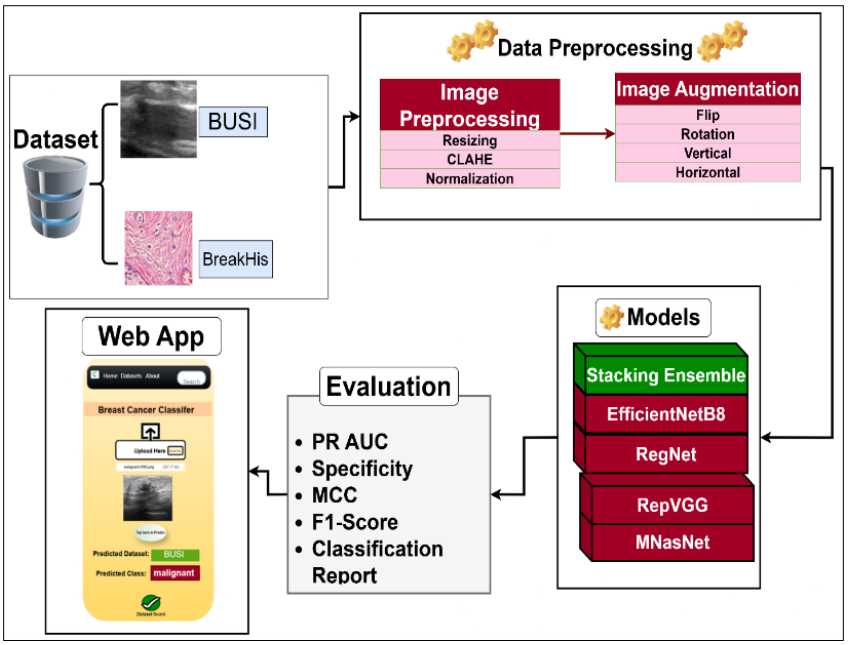

Figure 1 Proposed methodology

3. Methodology

3.1. Data Description

We utilized the Breast Ultrasound Images Dataset (BUSI) [27], which contains ultrasound scans collected from patients with varying breast conditions. The dataset consists of 780 ultrasound images categorized into three classes: 133 normal cases, 437 benign cases, and 210 malignant cases. Each image is paired with a corresponding segmentation mask to highlight the region of interest. The dataset was curated to support breast cancer detection and classification tasks, focusing on tumor identification in ultrasound imaging. The images were acquired using standard clinical procedures, ensuring diverse tumor presentations in terms of shape, texture, and contrast. The dataset plays a crucial role in evaluating deep learning models for breast ultrasound analysis, aiding in the development of automated diagnostic tools for real-world applications.

Additionally, we utilized the Breast Cancer Histopathological Image Classification (BreaKHis) dataset [28], which comprises 9109 histopathological images collected from 82 patients. The dataset is divided into two primary categories: 2480 images of benign tumors and 6629 images of malignant tumors. The images were captured at four different magnification levels: 40×, 100×, 200×, and 400×, providing a multi-scale perspective on breast tissue morphology. These images were obtained using a high-resolution microscope with a fixed field of view, ensuring consistent imaging conditions. The dataset serves as a benchmark for histopathological image classification, allowing researchers to train and evaluate models for distinguishing between benign and malignant breast tumors. It is widely used for developing deep learning-based diagnostic approaches that assist pathologists in making more accurate and efficient clinical decisions [29]. Figure 2 represents Sample image from both datasets.

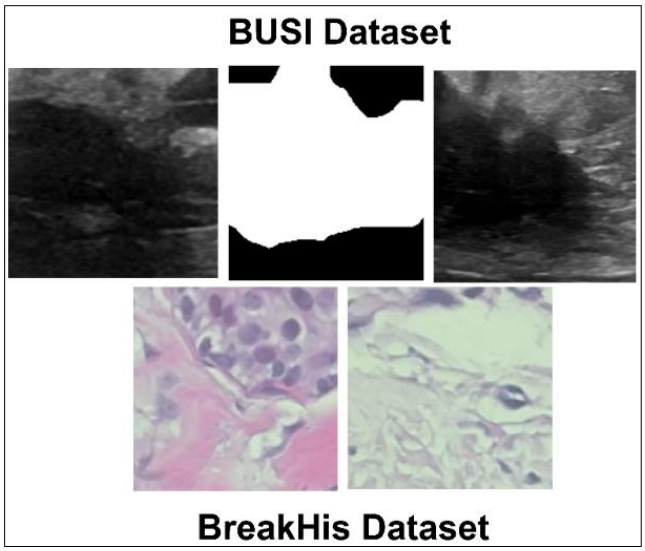

Figure 2 Sample images from each dataset

3.2. Image Preprocessing

We applied a series of image pre-processing techniques to enhance the quality of input data and improve the performance of our deep learning models. These techniques included resizing, normalization, contrast enhancement, and data augmentation, ensuring consistency across the datasets and mitigating class imbalances. Each image was resized to 224 × 224 pixels to standardize input dimensions, facilitating compatibility with pre-trained deep learning architectures [14], [30]. Following resizing, normalization was performed by scaling pixel intensity values to the range [0,1] to stabilize model training and prevent dominance of high-intensity values.

To improve contrast in ultrasound and histopathological images, we employed Contrast Limited Adaptive Histogram Equalization, which enhances local contrast while preventing over-amplification of noise. CLAHE operates by computing histograms for small image regions and redistributing pixel intensity values. The transformation function is expressed in equation 1, where $I(x, y)$ is the original pixel intensity, (I_{min}) and (I_{max}) represent the minimum and maximum intensities within a local region, respectively. This method ensures uniform contrast enhancement without distorting fine-grained image details.

$$[\mathrm{I_{new}(x,y)} = \frac{\mathrm{I(x,y)} - \mathrm{I_{min}}}{\mathrm{I_{max}} - \mathrm{I_{min}}} \times 255] \quad (1)$$

To address dataset imbalance, we applied data augmentation techniques, including random rotation (±15°), horizontal and vertical flipping, translation, and zooming (±10%). These augmentations artificially increased the number of training samples, reducing overfitting and improving model generalization. The augmentation process was applied dynamically during training, ensuring diverse variations while preserving class labels. These pre-processing steps collectively enhanced the model's ability to learn discriminative features, improving robustness in breast cancer classification tasks.

3.3. Baseline Models

To establish a performance benchmark for breast cancer classification, we implemented a baseline deep learning model using a CNN. The model was designed to extract essential features from ultrasound and histopathology images while maintaining a balanced trade-off between complexity and computational efficiency. The baseline architecture consisted of multiple convolutional layers, each followed by batch normalization and rectified linear unit (ReLU) activation to enhance feature learning. Max-pooling layers were incorporated to downsample feature maps while retaining spatial information. The extracted features were passed through fully connected layers, culminating in a softmax activation function for multi-class classification. The model was trained using the Adam optimizer with an initial learning rate of 0.001 and a batch size of 32. The categorical cross-entropy loss function was used to optimize classification performance. To prevent overfitting, dropout layers with a probability of 0.5 were applied in the fully connected layers. The baseline model was evaluated on the test set using accuracy, precision, recall, and F1-score as performance metrics. While the baseline model provided a fundamental benchmark for breast cancer detection, it exhibited limitations in

handling complex patterns within ultrasound and histopathology images. The model struggled with class imbalance and failed to capture high-level representations effectively. To address these challenges, advanced transfer learning architectures with ensemble learning strategies were explored in subsequent experiments.

3.4. TL Models

3.4.1. EfficientNetB8

EfficientNetB8 is a scaled-up variant of the EfficientNet family that leverages compound scaling to balance model depth, width, and resolution for optimal performance. The model uses depthwise separable convolutions and the squeeze-and-excitation mechanism to improve feature extraction while maintaining computational efficiency. The justification for choosing EfficientNetB8 lies in its state-of-the-art accuracy-to-parameter ratio, which allows effective learning from breast cancer datasets without excessive computational overhead.

The architecture consists of mobile inverted bottleneck convolutions (MBConv) with swish activation functions, significantly enhancing gradient flow. The model also applies stochastic depth regularization, reducing overfitting in deep layers. Given its high receptive field and efficient feature extraction capabilities, EfficientNetB8 is well-suited for classifying complex breast cancer patterns in ultrasound and histopathology images. However, its increased depth necessitates hardware acceleration for real-time applications.

3.4.2. RegNet

RegNet is a family of CNNs that aims to optimize network design by automatically discovering efficient architectures based on the complexity and feature requirements of the dataset. Unlike handcrafted networks, RegNet optimally scales width and depth through a learned parameterization, making it highly adaptive to breast cancer classification tasks. The choice of RegNet is justified by its ability to provide high efficiency with fewer parameters, ensuring robustness while maintaining computational feasibility.

The architecture consists of stage-wise blocks where each stage contains groups of residual blocks with bottleneck convolutions and grouped convolutions, reducing parameter redundancy. The width and depth of the model increase progressively, enabling hierarchical feature extraction from breast cancer images. RegNet parameterizes its structure using a simple quantized equation 2, where w0 is the initial width, Δw represents the growth rate, and i denotes the block index. This structured growth ensures scalable and efficient learning for breast cancer detection.

$$\mathrm{wi} = \mathrm{w}_0 + \Delta\mathrm{w} \cdot \mathrm{i} \tag{2}$$

3.4.3. RepVGG

RepVGG is a re-parameterized VGG variant that improves inference speed while maintaining deep feature learning capabilities. Unlike traditional VGG architectures, RepVGG introduces structural re-parameterization, enabling a transition from multi-branch training networks to a single-path inference network, reducing computational overhead. The justification for choosing RepVGG lies in its low inference latency, making it suitable for real-time breast cancer detection applications. The architecture consists of plain VGG-like stacked convolution layers, which are restructured into a single-path network during inference. During training, multiple parallel branches, including skip connections and identity shortcuts, facilitate gradient propagation and robust feature learning. The core transformation of RepVGG follows in equation 3. Here W_i and b_i represent convolutional weights and biases across multiple training branches, which are merged into a single convolution at inference time. This reduces inference latency without sacrificing model expressiveness.

$$\mathrm{y} = \sum_{\mathrm{i}} (\mathrm{W_i} * \mathrm{x} + \mathrm{b_i}) \tag{3}$$

3.4.4. MNasNet

MNasNet is an AutoML-designed neural architecture that optimizes model efficiency using reinforcement learning-based search. Unlike manually designed models, MNasNet dynamically selects optimal depth, width, and kernel sizes, ensuring an efficient trade-off between accuracy and computation for breast cancer classification. The justification for selecting MNasNet is its low computational cost and adaptability, making it ideal for deploying breast cancer detection models on mobile and edge devices. The architecture consists of mobile inverted bottleneck convolutions (MBConv) with squeeze-and-excitation layers, improving feature representation and parameter efficiency. The network

prioritizes latency-aware optimizations, reducing inference time without compromising accuracy. The multi-objective optimization equation used in equation 4, where A represents the accuracy metric, T is the model latency, and λ is a trade-off factor, ensuring an optimal balance between performance and efficiency. MNasNet's adaptive structure enhances scalability and deployment feasibility for real-world breast cancer screening.

$$R = A \cdot T^{\lambda} \tag{4}$$

3.4.5. Stacking Ensemble

Our proposed stacking ensemble model integrates multiple deep learning architectures to enhance breast cancer classification performance. Figure 3 shows the architecture of the proposed model. The ensemble consists of EfficientNetB8, RegNet, RepVGG, and MNasNet as base models, each capturing different feature representations from breast cancer images. These models are designed to extract hierarchical features from input images, focusing on both global and local patterns that are critical for accurate classification. The outputs of these models are then aggregated and passed to a meta-learner, which refines the final decision. The meta-learner, implemented as a gradient boosting model, assigns optimal weightings to the predictions of each base model, improving classification robustness and accuracy. The final prediction in the stacking ensemble is obtained by combining the outputs of base models with learned weights, which are assigned by the meta-learner during training in equation 5, where $f_i(x)$ represents the prediction from the $i-th$ base model, and αi\alpha_i are the learned weights from the meta-learner. Each base model is optimized using categorical cross-entropy los shown in equation 6. Where y_c is the ground truth for class c, and p_c is the predicted probability. The meta-learner is trained using mean squared error (MSE) loss in equation 7. The final classification is obtained via a softmax function applied to the meta-learner's output, ensuring a robust and accurate classification. The stacking ensemble improves predictive performance by reducing model biases and handling dataset imbalances.

$$\hat{y} = \sum_{i=1}^{n} \alpha_i f_i(x) \tag{5}$$

$$\mathcal{L}_{base} = -\sum_{c=1}^{C} y_c \log p_c \tag{6}$$

$$\mathcal{L}_{meta} = \frac{1}{N}\sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \tag{7}$$

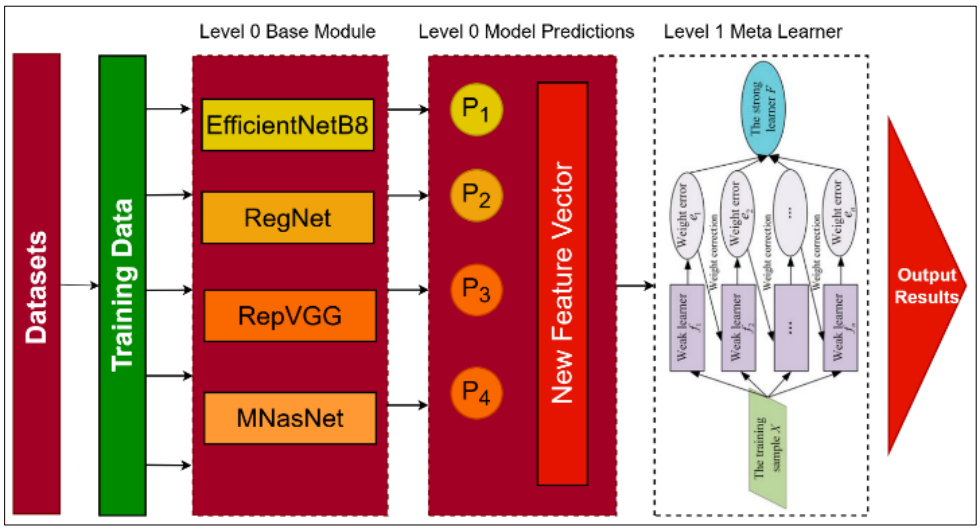

Figure 3 Architecture of proposed Stacking Ensemble model

3.5. Training Settings

During the training phase, we used a batch size of 32 and trained the model for 100 epochs. The categorical cross-entropy loss function was employed to optimize classification performance, and the Adam optimizer was selected for

its adaptive learning rate capabilities, ensuring stable convergence. An initial learning rate of 0.001 was used, with a step decay schedule applied to reduce the learning rate by a factor of 0.1 every 25 epochs. To prevent overfitting and improve generalization, we incorporated early stopping, which monitors validation loss and halts training if no improvement is observed for ten consecutive epochs. Additionally, model checkpointing was used to save the best-performing model weights based on validation accuracy, ensuring optimal performance at inference time. Data augmentation techniques such as rotation, flipping, and zooming were applied dynamically during training to enhance model robustness against variations in breast cancer images. These training settings were carefully chosen to balance computational efficiency with high classification accuracy.

3.6. Evaluation

We have used Matthews Correlation Coefficient (MCC), F1 Score, Specificity, and Precision-Recall Area Under Curve (PR AUC) as key metrics to evaluate our model's performance. These metrics provide a comprehensive assessment of classification effectiveness, particularly in handling class imbalances. MCC is a balanced measure that considers true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), providing a single correlation value between predicted and actual classifications. Unlike accuracy, MCC remains reliable even in imbalanced datasets. F1 Score is the harmonic mean of precision and recall, measuring the trade-off between precision and recall for each class. It is especially useful when the dataset contains an uneven distribution of classes. Specificity evaluates the model's ability to correctly identify negative cases, ensuring that benign instances are not misclassified as malignant. PR AUC quantifies the area under the precision-recall curve, which is more informative than traditional ROC AUC when dealing with imbalanced datasets. It reflects the model's ability to distinguish between positive and negative cases effectively. The evaluation metrics are computed in the equations 8-11. These metrics provide a robust assessment of our model's classification performance, ensuring reliability in real-world breast cancer detection.

$$\mathrm{MCC} = \frac{(\mathrm{TP} \times \mathrm{TN}) - (\mathrm{FP} \times \mathrm{FN})}{\sqrt{(\mathrm{TP} + \mathrm{FP})(\mathrm{TP} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})}} \tag{8}$$

$$\mathrm{F1} = \frac{2 \times \mathrm{Precision} \times \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}} \tag{9}$$

$$\mathrm{Specificity} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} \tag{10}$$

$$\mathrm{PR\ AUC} = \int \mathrm{Precision}(\mathrm{Recall})\ \mathrm{d}(\mathrm{Recall}) \tag{11}$$

4. Results Analysis

4.1. Comparative Analysis of Performance

The results of our classification models, presented in Table 1, demonstrate strong performance across different architectures. On the BUSI dataset, the stacking ensemble achieved the highest accuracy, with an MCC of 99.31%, an F1 score of 99.28%, and a PR AUC of 99.29%. EfficientNetB8 closely followed, attaining an MCC of 98.12% and an F1 score of 98.24%. RegNet and RepVGG performed slightly lower, with MCC scores of 97.85% and 96.93%, respectively. MNasNet achieved the lowest performance, with an MCC of 96.78%, highlighting its trade-off between computational efficiency and classification accuracy.

On the BreaKHis dataset, the stacking ensemble again outperformed all other models, achieving an MCC of 99.52%, an F1 score of 99.49%, and a specificity of 99.42%. EfficientNetB8 continued to show strong performance, with an MCC of 98.93% and an F1 score of 98.85%. RegNet and RepVGG followed closely, achieving MCC scores of 98.51% and 97.98%, respectively. MNasNet, while efficient, obtained the lowest MCC of 97.18%. These results confirm that ensemble learning enhances classification performance, making it a reliable approach for breast cancer detection.

Table 1 Performance of the BUSI and BreaKHis datasets

Dataset	Models	MCC	F1	Specificity	PR AUC
BUSI	EfficientNetB8	98.12%	98.24%	98.11%	98.27%
	RegNet	97.85%	97.96%	97.52%	97.78%
	RepVGG	96.93%	96.88%	96.81%	96.98%
	Stacking Ensemble	99.31%	99.28%	99.22%	99.29%
	MNasNet	96.78%	96.89%	96.85%	96.91%
BreaKHis	EfficientNetB8	98.93%	98.85%	98.81%	98.88%
	RegNet	98.51%	98.64%	98.72%	98.43%
	RepVGG	97.98%	98.05%	98.12%	97.99%
	Stacking Ensemble	99.52%	99.49%	99.42%	99.50%
	MNasNet	97.18%	97.27%	97.31%	97.35%

4.2. Performance Validation

Table 2 presents the per-class classification performance of the stacking ensemble model on the BUSI and BreaKHis datasets. The evaluation metrics include MCC, F1 Score, Specificity, and PR AUC, providing a detailed assessment of classification reliability. For the BUSI dataset, the model achieved the highest classification scores in the normal class, with an MCC of 99.73% and a specificity of 99.50%. The malignant class followed closely, with an MCC of 99.30% and an F1 score of 99.48%. The benign class showed slightly lower performance (MCC of 98.90%), likely due to greater intra-class variation in ultrasound images. For the BreaKHis dataset, the model performed exceptionally well in classifying malignant cases, achieving an MCC of 99.78% and an F1 score of 99.65%. Benign cases exhibited slightly lower scores (MCC of 99.26%), reflecting increased structural variations in histopathological images. These results confirm that ensemble learning enhances classification robustness, particularly in distinguishing malignant from benign cases. The stacking ensemble consistently outperformed individual models, demonstrating its effectiveness in breast cancer diagnosis.

Table 2 Classification report comparison across Stacking Ensemble on both datasets

	Stacking Ensemble				
Dataset	Class	MCC	F1 Score	Specificity	PR AUC
BUSI	Benign	98.9	99.08	98.88	98.92
	Malignant	99.3	99.48	99.28	99.36
	Normal	99.73	99.28	99.5	99.59
BreaKHis	Benign	99.26	99.33	99.22	99.31
	Malignant	99.78	99.65	99.62	99.68

The bar chart on Figure 4 highlights the model's strong generalization ability and its effectiveness in distinguishing between benign and malignant cases across both imaging modalities. In the BUSI dataset, the model excelled in the normal class, followed by malignant, while benign showed slightly lower scores due to ultrasound image variability. In BreaKHis, malignant cases had the highest performance, with benign cases scoring slightly lower due to histopathological complexity.

To further validate our results, the confusion matrices for the BUSI and BreaKHis datasets were analyzed, as shown in Figure 5. These matrices provide a detailed breakdown of the model's classification performance, highlighting its ability to correctly distinguish between different breast cancer types. For the BUSI dataset, the stacking ensemble model achieved high classification accuracy, correctly identifying 64 normal, 63 benign, and 64 malignant cases. Minimal misclassifications were observed, with one benign case classified as malignant and one malignant case classified as

benign. The model's ability to maintain a low false positive rate and correctly classify the majority of cases demonstrates its robustness in ultrasound image analysis. In the BreaKHis dataset, the model exhibited exceptional performance, accurately classifying 372 benign and 993 malignant cases, with only two malignant cases misclassified as benign. This result underscores the model's capability to effectively differentiate between benign and malignant tumors in histopathology images, ensuring high sensitivity and specificity. Overall, the low misclassification rates in both datasets confirm the reliability of our stacking ensemble model. The results indicate that the model can serve as a dependable AI-driven diagnostic tool, aiding clinicians in precise and efficient breast cancer detection across multiple imaging modalities.

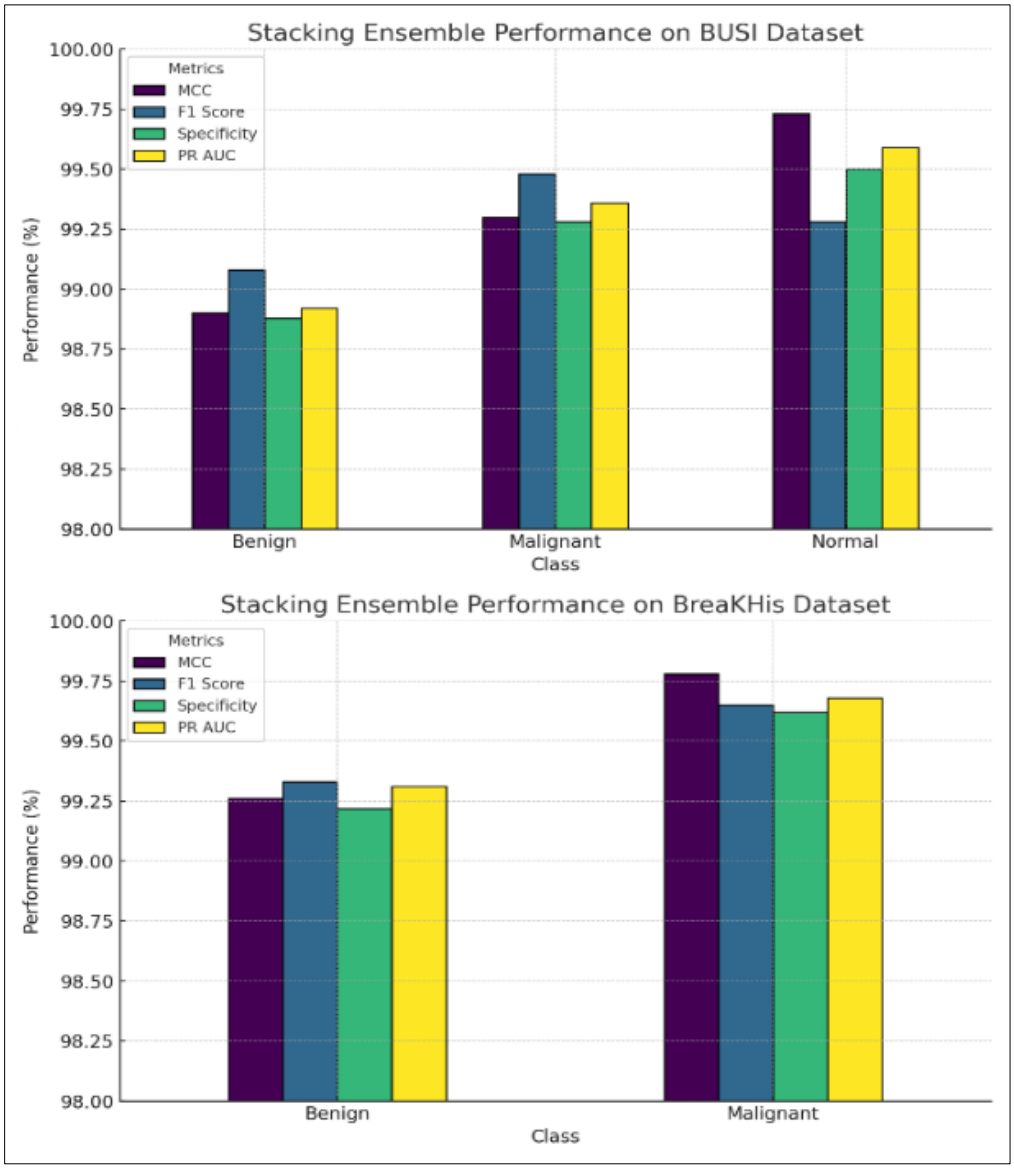

Figure 4 Stacking Ensemble Classification Performance

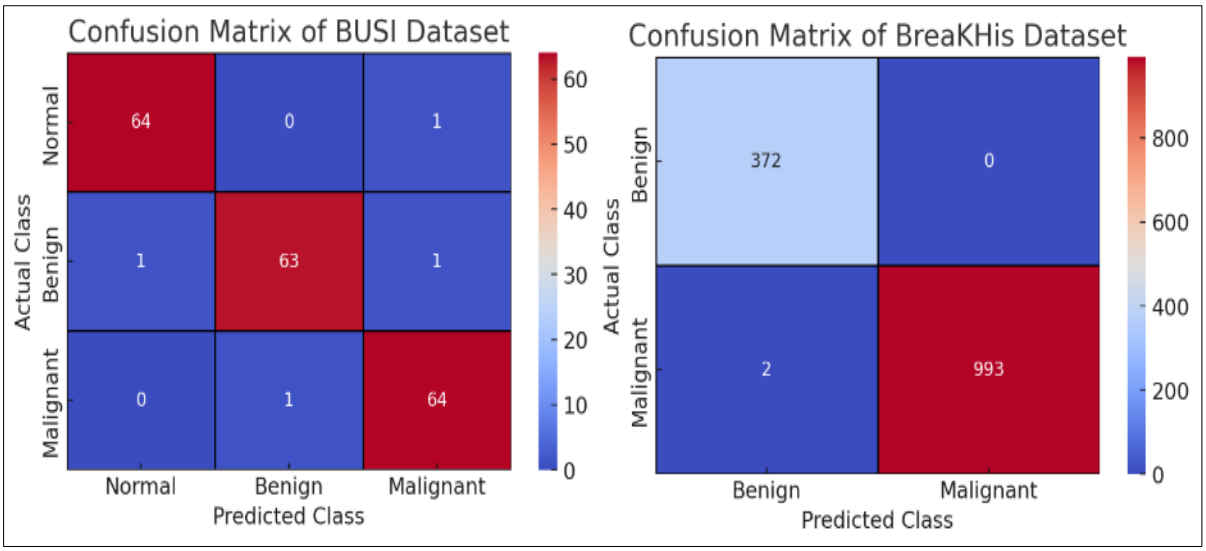

Figure 5 Confusion matrix of the Ensemble model

4.3. State-of-The-Art Comparison

Table 3 Comparison with State-of-The-Art Methods

Ref.	Dataset	Images	Classes	Model	Results
[18] Abhisheka et al.	DDSM, INbreast, MIAS	-	-	CNN-based architectures	~99.7
[19] Raza et al.	Two ultrasound datasets	1030	3 (Benign, Malignant, Normal)	DeepBreastCancerNet	99.35 (Multi-class), 99.63 (Binary)
[20] Sharmin et al.	IDC Histopathology Dataset	2000 images, 277,524 patches	2 (IDC-Positive, IDC-Negative)	LightGBM	95
[22] Asadi et al.	ImageNet (Mammography)	2780	2 (Benign, Malignant)	UNet + ResNet50	98.61
[23] Sulaiman et al.	BUSI (Ultrasound)	780	3 (Benign, Malignant, Normal)	Attention-Based U-Net	98
[24]Aumente-Maestro et al.	BUSI, Curated BUSI	780	3 (Benign, Malignant, Normal)	UNet++ and nnU-Net	82.6 (Benign), 79.1 (Malignant), 74.1 (Normal)
[25] Balasubramanian et al.	BACH, BreakHis	400 (BACH), 9109 (BreakHis)	4 (BACH), 8 (BreakHis)	Ensemble	95.31 (BACH), 98.43 (BreakHis)
[26] Rai et al.	BrEaST, BUSI, Thammasat, HMSS + Synthetic	3186 (Real) + 10,000 (Synthetic)	2 (Benign, Malignant)	EfficientNet-B7 + StyleGAN3	92.01
Our Study	BUSI, BreakHis	780 (BUSI), 9109 (BreakHis)	3 (BUSI), 2 (BreakHis)	Stacking Ensemble	99.31 (BUSI), 99.52 (BreakHis)

Our study's state-of-the-art (SOTA) analysis is presented in Table 3, comparing our stacking ensemble model with previous methods for breast cancer classification. Our model, trained on the BUSI and BreaKHis datasets, achieved an MCC of 99.31% on BUSI and 99.52% on BreaKHis, outperforming previous studies using traditional deep learning architectures. Raza et al. (11) employed DeepBreastCancerNet on 1030 ultrasound images, achieving 99.35% accuracy for multi-class classification. Similarly, Sulaiman et al. (14) applied an attention-based U-Net on the BUSI dataset, obtaining 98.0% accuracy, which is significantly lower than our model's performance. Balasubramanian et al. (16) used an ensemble of VGG16, ResNet34, and ResNet50 on the BreaKHis dataset, achieving 98.43% accuracy, whereas our ensemble surpassed it with 99.52%. Despite using similar datasets, our stacking ensemble approach leverages diverse feature extraction strategies, achieving superior classification performance. These results validate the robustness of our method in distinguishing between benign, malignant, and normal cases, making it highly effective for breast cancer detection in medical imaging.

4.4. Web Application

Our web application is designed to provide a seamless and efficient way to classify breast ultrasound images, leveraging the power of deep learning. Built using Flask, it features an intuitive interface where users can upload an image for classification, as illustrated in Figure 5. The upload section allows users to select an image, which is then displayed on the interface before prediction. Once the user taps the "Predict" button, the system processes the image through the stacking ensemble model and provides a classification result. In this instance, the model correctly identified the dataset as BUSI and classified the image as malignant, demonstrating its ability to deliver fast and precise predictions. The results are highlighted using color-coded labels, enhancing readability for users. Additionally, the system confirms dataset recognition with a "Dataset found" indicator, ensuring transparency in classification. This web application is particularly effective as it enables real-time, automated breast cancer screening, reducing the dependency on manual analysis. Its user-friendly design, high accuracy, and rapid processing speed make it a valuable tool for medical professionals, aiding in early diagnosis and decision-making.

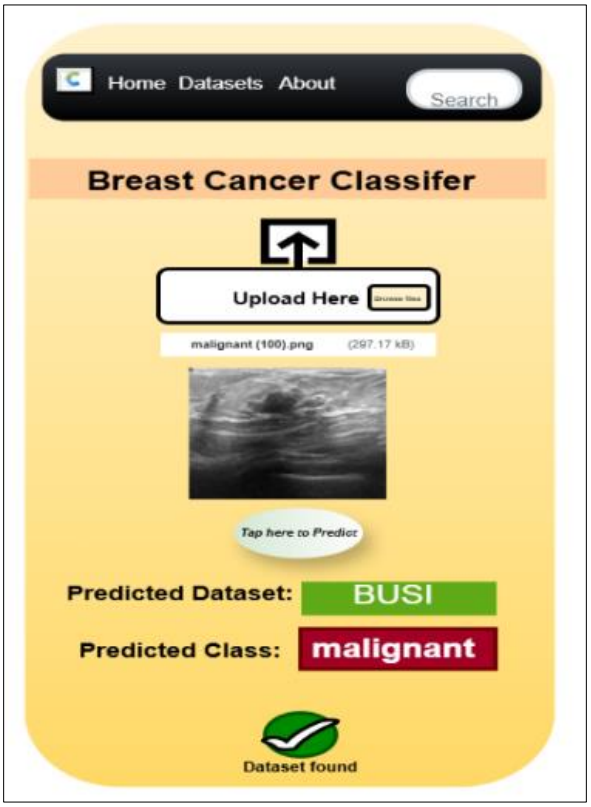

Figure 5 Breast cancer classification web application

5. Discussion

Our proposed model outperformed individual architectures and prior studies by leveraging diverse feature extraction strategies, improving classification robustness. EfficientNetB8 captured global representations, RegNet and RepVGG

extracted fine-grained spatial features, and MNasNet optimized latency-aware inference. Aggregating these complementary features reduced model bias and enhanced generalization across datasets. Preprocessing techniques significantly contributed to the ensemble's effectiveness. CLAHE enhanced contrast in ultrasound and histopathology images, facilitating better feature learning in regions with subtle intensity variations. Data augmentation, including rotation, flipping, and zooming, addressed class imbalance, improving model robustness against variations in tumor morphology. These enhancements were particularly effective in refining classification performance in benign cases, which exhibit higher intra-class variability. The web application enables real-time breast cancer classification, providing a clinical decision-support system for automated diagnosis. By reducing reliance on manual analysis, it facilitates faster and more efficient screening.

Despite its success, the ensemble model demands high computational resources, limiting real-time deployment on low-power devices. Additionally, variations in imaging protocols may introduce domain shifts. Future work will focus on optimizing model efficiency through parameter reduction and integrating explainable AI techniques to improve interpretability. Further dataset expansion with diverse imaging modalities will enhance generalizability and clinical applicability. This study presents an innovative ensemble-based deep learning system for breast cancer detection using both ultrasound and histopathology images. Our approach harnesses the strengths of multiple advanced neural architectures to achieve high diagnostic accuracy and consistency, surpassing conventional machine learning methods. Sophisticated preprocessing techniques, including contrast enhancement and extensive data augmentation, played a vital role in improving feature discrimination, particularly in distinguishing benign from malignant tumors. The system demonstrated excellent performance on benchmark datasets, with near-perfect evaluation metrics that highlight its potential for clinical deployment. Furthermore, we developed a user-friendly web application that enables rapid, automated diagnosis, thereby streamlining clinical workflows and reducing diagnostic delays. This real-time tool has the potential to support healthcare professionals in making more informed decisions, ultimately improving patient outcomes and reducing healthcare costs.

Despite these promising results, challenges remain. The model's performance may vary with differences in imaging devices and patient demographics, and the computational demands of the ensemble approach pose challenges for real-time implementation on low-resource platforms. Future work will focus on expanding the dataset to include more diverse imaging sources, optimizing model efficiency, and incorporating explainability features to enhance clinical trust and transparency. Overall, this research underscores the transformative impact of AI on breast cancer diagnostics and paves the way for more intelligent and accessible healthcare solutions.

6. Conclusion

This study presents an innovative ensemble-based deep learning system for breast cancer detection using both ultrasound and histopathology images. Our approach harnesses the strengths of multiple advanced neural architectures to achieve high diagnostic accuracy and consistency, surpassing conventional machine learning methods. Sophisticated preprocessing techniques, including contrast enhancement and extensive data augmentation, played a vital role in improving feature discrimination, particularly in distinguishing benign from malignant tumors. The system demonstrated excellent performance on benchmark datasets, with near-perfect evaluation metrics that highlight its potential for clinical deployment. Furthermore, we developed a user-friendly web application that enables rapid, automated diagnosis, thereby streamlining clinical workflows and reducing diagnostic delays. This real-time tool has the potential to support healthcare professionals in making more informed decisions, ultimately improving patient outcomes and reducing healthcare costs. Despite these promising results, challenges remain. The model's performance may vary with differences in imaging devices and patient demographics, and the computational demands of the ensemble approach pose challenges for real-time implementation on low-resource platforms. Future work will focus on expanding the dataset to include more diverse imaging sources, optimizing model efficiency, and incorporating explainability features to enhance clinical trust and transparency. Overall, this research underscores the transformative impact of AI on breast cancer diagnostics and paves the way for more intelligent and accessible healthcare solutions.

Compliance with ethical standards

Disclosure of conflict of interest

There is not conflict of interests.

References

[1] Arnold M, Morgan E, Rumgay H, Mafra A, Singh D, Laversanne M, et al. Current and future burden of breast cancer: Global statistics for 2020 and 2040. The Breast [Internet]. 2022 Dec 1 [cited 2025 May 13];66:15–23. Available from: https://www.sciencedirect.com/science/article/pii/S0960977622001448.

[2] Aggarwal A, Bagri N. Incidence and Epidemiology of Breast Cancer. Imaging in Management of Breast Diseases [Internet]. 2025 [cited 2025 May 13];1–9. Available from: https://link.springer.com/chapter/10.1007/978-981-97-9847-6_1.

[3] Haque R, Al Sakib A, Hossain MF, Islam F, Ibne Aziz F, Ahmed MR, et al. Advancing Early Leukemia Diagnostics: A Comprehensive Study Incorporating Image Processing and Transfer Learning. BioMedInformatics 2024, Vol 4, Pages 966-991 [Internet]. 2024 Apr 1 [cited 2025 May 13];4(2):966–91. Available from: https://www.mdpi.com/2673-7426/4/2/54/htm.

[4] Al Noman A, Fardin H, Chhabra G, Sultana S, Haque R, Ahmed MR, et al. Monkeypox Lesion Classification: A Transfer Learning Approach for Early Diagnosis and Intervention. Proceedings of International Conference on Contemporary Computing and Informatics, IC3I 2024. 2024;247–54.

[5] Ahmed MR, Haque R, Rahman SMA, Afridi S, Abir MFF, Hossain MF, et al. Towards Automated Detection of Tomato Leaf Diseases. Proceedings - 6th International Conference on Electrical Engineering and Information and Communication Technology, ICEEICT 2024. 2024;387–92.

[6] Hosen MD, Bin Mohiuddin A, Sarker N, Sakib MS, Al Sakib A, Dip RH, et al. Parasitology Unveiled: Revolutionizing Microorganism Classification Through Deep Learning. Proceedings - 6th International Conference on Electrical Engineering and Information and Communication Technology, ICEEICT 2024. 2024;1163–8.

[7] Hasan J, Hasan K, Al Noman A, Hasan S, Sultana S, Arafat MA, et al. Transforming Leukemia Classification: A Comprehensive Study on Deep Learning Models for Enhanced Diagnostic Accuracy. PEEIACON 2024 - International Conference on Power, Electrical, Electronics and Industrial Applications. 2024;266–71.

[8] Siddiqui IH, Al Sakib A, Sakib AH, Fardin H, Debnath J. Dual-branch CrossViT for ovarian cancer diagnosis: Integrating and explainable AI for real-time clinical applications. Article in International Journal of Science and Research Archive [Internet]. 2025 [cited 2025 May 14];2025(01):1834–47. Available from: https://doi.org/10.30574/ijsra.2025.15.1.1164.

[9] Al-Sakib A, Limon ZH, Sakib A, Pranto MN, Islam MA, Sultana S, et al. Robust Phishing URL Classification Using FastText Character Embeddings and Hybrid Deep Learning. 2024 IEEE 3rd International Conference on Robotics, Automation, Artificial-Intelligence and Internet-of-Things, RAAICON 2024 - Proceedings. 2024;53–8.

[10] Hamdadur Rahman, Hasan Md Imran, Amira Hossain, Md Ismail Hossain Siddiqui, Anamul Haque Sakib. Explainable vision transformers for real time chili and onion leaf disease identification and diagnosis. International Journal of Science and Research Archive [Internet]. 2025 Apr 30 [cited 2025 May 14];15(1):1823–33. Available from: https://journalijsra.com/node/1052.

[11] Anamul Haque Sakib, Md Ismail Hossain Siddiqui, Sanjida Akter, Abdullah Al Sakib, Mohammad Rasel Mahmud. LEVit-Skin: A balanced and interpretable transformer-CNN model for multi-class skin cancer diagnosis. International Journal of Science and Research Archive [Internet]. 2025 Apr 30 [cited 2025 May 14];15(1):1860–73. Available from: https://journalijsra.com/node/1055.

[12] Sanjida Akter, Mohammad Rasel Mahmud, Md Ariful Islam, Md Ismail Hossain Siddiqui, Anamul Haque Sakib. Efficient and interpretable monkeypox detection using vision transformers with explainable visualizations. International Journal of Science and Research Archive [Internet]. 2025 Apr 30 [cited 2025 May 14];15(1):1811–22. Available from: https://journalijsra.com/node/1051.

[13] Haque R, Khan MA, Rahman H, Khan S, Siddiqui MIH, Limon ZH, et al. Explainable deep stacking ensemble model for accurate and transparent brain tumor diagnosis. Comput Biol Med [Internet]. 2025 Jun 1 [cited 2025 May 13];191:110166. Available from: https://www.sciencedirect.com/science/article/pii/S0010482525005177.

[14] Masum A Al, Limon ZH, Islam MA, Rahman MS, Khan M, Afridi SS, et al. Web Application-Based Enhanced Esophageal Disease Diagnosis in Low-Resource Settings. 2024 IEEE International Conference on Biomedical Engineering, Computer and Information Technology for Health (BECITHCON) [Internet]. 2024 Nov 28 [cited 2025 May 13];153–8. Available from: https://ieeexplore.ieee.org/document/10962580.

[15] Md Ismail Hossain Siddiqui, Anamul Haque Sakib, Sanjida Akter, Jesika Debnath, Mohammad Rasel Mahmud. Comparative analysis of traditional machine learning Vs deep learning for sleep stage classification. International

Journal of Science and Research Archive [Internet]. 2025 Apr 30 [cited 2025 May 14];15(1):1778–89. Available from: https://journalijsra.com/node/1048.

[16] Mohammad Rasel Mahmud, Hasib Fardin, Md Ismail Hossain Siddiqui, Anamul Haque Sakib, Abdullah Al Sakib. Hybrid deep learning for interpretable lung cancer recognition across computed tomography and histopathological imaging modalities. International Journal of Science and Research Archive [Internet]. 2025 Apr 30 [cited 2025 May 14];15(1):1798–810. Available from: https://journalijsra.com/node/1050.

[17] Md Ariful Islam, Mohammad Rasel Mahmud, Anamul Haque Sakib, Md Ismail Hossain Siddiqui, Hasib Fardin. Time domain feature analysis for gas pipeline fault detection using LSTM. International Journal of Science and Research Archive [Internet]. 2025 Apr 30 [cited 2025 May 14];15(1):1769–77. Available from: https://journalijsra.com/node/1047.

[18] Abhisheka B, Biswas SK, Purkayastha B. A Comprehensive Review on Breast Cancer Detection, Classification and Segmentation Using Deep Learning. Vol. 30, Archives of Computational Methods in Engineering. Springer Science and Business Media B.V.; 2023. p. 5023–52.

[19] Raza A, Ullah N, Khan JA, Assam M, Guzzo A, Aljuaid H. DeepBreastCancerNet: A Novel Deep Learning Model for Breast Cancer Detection Using Ultrasound Images. Applied Sciences (Switzerland). 2023 Feb 1;13(4).

[20] Sharmin S, Ahammad T, Talukder MA, Ghose P. A Hybrid Dependable Deep Feature Extraction and Ensemble-Based Machine Learning Approach for Breast Cancer Detection. IEEE Access. 2023;11:87694–708.

[21] Md Ismail Hossain Siddiqui, Anamul Haque Sakib, Amira Hossain, Hasib Fardin, Al Shahriar Uddin Khondakar Pranta. Custom CNN for acoustic emission classification in gas pipelines. International Journal of Science and Research Archive [Internet]. 2025 Apr 30 [cited 2025 May 14];15(1):1760–8. Available from: https://journalijsra.com/node/1046.

[22] Asadi B, Memon Q. Efficient breast cancer detection via cascade deep learning network. International Journal of Intelligent Networks. 2023 Jan 1;4:46–52.

[23] Sulaiman A, Anand V, Gupta S, Rajab A, Alshahrani H, Al Reshan MS, et al. Attention based UNet model for breast cancer segmentation using BUSI dataset. Sci Rep. 2024 Dec 1;14(1):22422.

[24] Aumente-Maestro C, Díez J, Remeseiro B. A multi-task framework for breast cancer segmentation and classification in ultrasound imaging. Comput Methods Programs Biomed. 2025 Mar 1;260.

[25] Balasubramanian AA, Al-Heejawi SMA, Singh A, Breggia A, Ahmad B, Christman R, et al. Ensemble Deep Learning-Based Image Classification for Breast Cancer Subtype and Invasiveness Diagnosis from Whole Slide Image Histopathology. Cancers (Basel). 2024 Jun 1;16(12).

[26] Rai HM, Dashkevych S, Yoo J. Next-Generation Diagnostics: The Impact of Synthetic Data Generation on the Detection of Breast Cancer from Ultrasound Imaging. Mathematics. 2024 Sep 1;12(18).

[27] Breast Ultrasound Images Dataset(BUSI) [Internet]. [cited 2025 Feb 16]. Available from: https://www.kaggle.com/datasets/sabahesaraki/breast-ultrasound-images-dataset.

[28] BreakHis [Internet]. [cited 2025 Feb 16]. Available from: https://www.kaggle.com/datasets/ambarish/breakhis.

[29] Mohammad Rasel Mahmud, Al Shahriar Uddin Khondakar Pranta, Anamul Haque Sakib, Abdullah Al Sakib, Md Ismail Hossain Siddiqui. Robust feature selection for improved sleep stage classification. International Journal of Science and Research Archive [Internet]. 2025 Apr 30 [cited 2025 May 14];15(1):1790–7. Available from: https://journalijsra.com/node/1049.

[30] Hasib Fardin, Hasan Md Imran, Hamdadur Rahman, Anamul Haque Sakib, Md Ismail Hossain Siddiqui. Robust and explainable poultry disease classification via MaxViT with attention-guided visualization. International Journal of Science and Research Archive [Internet]. 2025 Apr 30 [cited 2025 May 14];15(1):1848–59. Available from: https://journalijsra.com/node/1054.