Ensemble Learning-Based Risk Assessment for Early Detection of Breast Cancer from Medical Images

Khushi Agarwal, Praveena K R, Akshara Gattupalli, Yanaparati Venkata Thanmayi

School of Computer Science and Engineering, Vellore Institute of Technology, Chennai 600127, India

ABSTRACT

Breast cancer remains a leading cause of mortality among women worldwide, emphasizing the urgent need for accurate, early, and automated diagnostic systems. Traditional diagnostic methods often face challenges of inconsistency and high computational complexity, while individual deep learning models—despite strong feature extraction capabilities—frequently lack robustness and generalization across diverse histopathological datasets. To address these limitations, this study proposes a high-performance ensemble deep learning framework for reliable breast cancer classification using histopathological images. The proposed system employs a Soft Voting Ensemble that integrates three state-of-the-art Convolutional Neural Network (CNN) architectures: VGG16, DenseNet121, and Xception. Each contributes distinct advantages—VGG16 offers fine-grained texture extraction, DenseNet121 enables efficient feature reuse and mitigates vanishing gradients, and Xception enhances discriminative learning through depth wise separable convolutions. Through transfer learning and optimized feature fusion, the ensemble achieves superior accuracy, generalization, and robustness compared to standalone models. Experimental validation using the BreaKHis dataset demonstrates the model's effectiveness, achieving an overall accuracy of 95.33%, with balanced precision, recall, and F1-scores of 0.95 across benign and malignant classes. Furthermore, ROC and Precision-Recall analysis (AUC = 0.989, AP = 0.990) confirm the model's strong discriminative capability and reliability under class imbalance. By leveraging complementary strengths of multiple CNNs, the ensemble effectively mitigates bias and variance issues inherent in individual models, resulting in stable and interpretable predictions. Designed for computational efficiency and clinical scalability, this framework provides a reliable and adaptable solution for automated breast cancer detection, holding strong potential to enhance diagnostic precision, support clinical decision-making, and contribute to improved patient outcomes.

Keywords: *Ensemble Learning, Breast Cancer Detection, Risk Assessment, Medical Imaging, Machine Learning, Soft Voting Classifier.*

I. INTRODUCTION

A. Context and Background on Breast Cancer Detection

Breast cancer continues to be one of the most significant global health concerns, representing the most frequently diagnosed malignancy and the leading cause of cancer-related mortality among women worldwide. According to recent epidemiological studies, approximately one in eight women are likely to develop breast cancer during their lifetime. The prognosis and survival rate of patients depend heavily on the stage at which the disease is detected—emphasizing the critical importance of early and accurate diagnosis.

Traditional imaging-based diagnostic modalities such as mammography, ultrasound, and Magnetic Resonance Imaging (MRI) play a pivotal role in the screening and identification of breast abnormalities. However, each of these modalities has inherent limitations. Mammography, while being the clinical gold standard, exhibits reduced sensitivity in women with dense breast tissue, leading to missed detections or false negatives. Ultrasound, on the other hand, is heavily operator-dependent and prone to subjective interpretation errors, while MRI, though highly sensitive, remains cost-prohibitive and less accessible in resource-constrained clinical settings. These challenges collectively highlight the pressing need for automated, objective, and computationally efficient diagnostic systems that can complement radiologists by improving diagnostic accuracy and consistency.

With the rapid advancement of artificial intelligence (AI) and deep learning (DL), Computer-Aided Diagnosis (CAD) systems have emerged as a promising solution for medical image interpretation. Deep learning models, especially Convolutional Neural Networks (CNNs), have demonstrated remarkable capabilities in feature extraction and classification from histopathological and radiographic images. Despite these successes, the majority of existing systems still face issues related to generalization, interpretability, and robustness when applied to diverse patient datasets.

B. Problem Statement and Motivation

Existing CAD systems often rely on single deep learning architectures such as VGG, ResNet, or MobileNet for classification tasks. Although these architectures achieve high accuracy on individual datasets, their performance tends to degrade when tested on unseen data due to overfitting, limited adaptability, and domain-specific bias. Moreover, medical diagnostics demand an exceptional balance between sensitivity (true positive rate) and specificity (true negative rate). High sensitivity is critical to avoid false negatives, which could delay life-saving treatment, while high specificity reduces unnecessary biopsies and patient anxiety. Achieving this balance remains a persistent challenge in breast cancer detection.

Furthermore, no single deep learning architecture can universally capture all aspects of medical imaging data. For

instance, some CNNs are adept at extracting fine-grained spatial features, while others excel in capturing hierarchical texture or contextual information. This limitation motivates the development of an ensemble learning-based strategy capable of integrating multiple feature extraction paradigms to deliver more reliable, generalized, and clinically interpretable outcomes.

C. Proposed Solution and Main Contributions

To overcome the limitations of existing CAD frameworks, this paper proposes an Optimized Ensemble Learning Model for automated breast cancer detection using medical imaging data. The proposed approach integrates three state-of-the-art deep learning architectures like VGG19, DenseNet, and Xception within a Soft Voting Classifier framework. Each model contributes unique feature representations derived from its architectural characteristics. VGG19 efficiently captures fine-grained local features through deep hierarchical convolutional layers. DenseNet promotes feature reuse and gradient propagation, enabling a rich and compact feature space. Xception enhances depthwise spatial feature extraction, providing improved discriminative capability.

The Soft Voting Classifier aggregates the probabilistic outputs of these models, yielding a unified prediction that leverages the complementary strengths of all three architectures. This fusion approach ensures improved stability, enhanced feature diversity, and reduced overfitting, leading to a more robust diagnostic framework suitable for real-world medical environments.

The main contributions of this work are summarized as follows:

1. Novel Ensemble Architecture: Development of a hybrid deep ensemble combining VGG19, DenseNet, and Xception using a Soft Voting strategy to achieve superior diagnostic precision and reliability.
2. Feature-Level Complementarity: Integration of multi-perspective feature extraction mechanisms to capture both local and global spatial information from medical images.
3. Improved Robustness and Generalization: Design of an ensemble pipeline capable of maintaining consistent performance across varied imaging datasets and clinical conditions.
4. Clinical Applicability: Creation of a computationally efficient and scalable framework that can serve as a reliable decision-support tool in real-world healthcare environments.

D. Paper Organization

The remainder of this paper is structured as follows. Section II reviews related literature, focusing on recent advances in deep learning-based breast cancer detection and ensemble learning approaches. Section III elaborates on the proposed methodology, including dataset details, preprocessing techniques, model architectures, and ensemble integration strategy. Section IV presents experimental design, performance evaluation metrics, and comparative analysis with existing models. Section V concludes the paper and outlines potential directions for future research, emphasizing clinical translation and deployment possibilities.

II. LITERATURE SURVEY

EMT-Net is a new, efficient deep learning model for breast cancer detection that addresses the limitations of traditional, computationally expensive systems. It tackles two tasks at once-tumor classification and segmentation—using an efficient MobileNetV1 backbone. The model uses a weighted binary cross-entropy loss to better balance the trade-off between sensitivity and specificity. Evaluated on 1,511 breast ultrasound images, EMT-Net achieved a high sensitivity of 94.1% and specificity of 85.3%, with an overall accuracy of 88.6%. Critically, it processes each image in just 0.35 seconds on a simulated mobile device, making it a viable solution for real-time clinical use on resource-constrained hardware. [1]

A study aiming to improve breast cancer prediction used the Breast Cancer Surveillance Consortium dataset to analyze demographic and clinical risk factors with various machine learning models. The researchers found that the Random Forest model was the most effective, achieving the highest accuracy of 75.2%, slightly outperforming the ensemble model. The analysis also identified hormone replacement therapy and the 60–64 age group as key risk factors significantly associated with a higher occurrence of breast cancer. This highlights the value of machine learning in identifying high-risk individuals and improving the accuracy of traditional prediction methods. [2]

This study introduced a new method for early breast cancer detection using a combination of Fuzzy C-Means (FCM) and a UNET-based deep CNN to overcome the limitations of traditional methods like SVM. By first segmenting breast tissue images with FCM and then extracting features with a UNET architecture, the model was able to classify cells as either benign or malignant with high accuracy. The approach demonstrated superior performance on the BreakHis dataset, achieving 99.23% accuracy and strong scores across multiple metrics, including precision, recall, and F-score, confirming its potential for accurate and efficient cancer diagnosis. [3]

A study addressed the limitations of traditional mammography and computationally intensive deep learning models for breast cancer detection by using an EfficientNet architecture. By pre-training the model on ImageNet and fine-tuning it on mammography datasets with image preprocessing and augmentation, the researchers created a highly accurate yet efficient system. The EfficientNet model outperformed other popular models like ResNet and VGG, achieving a remarkable 97.5% accuracy, with a precision of 96.3% and a recall of 95.8%. This demonstrated its potential for practical and real-time clinical deployment. [4]

A study addressed the limitations of traditional, single-modality breast cancer diagnostics by introducing a Multi-Modal Radiomics and Deep CNN (MMRC) approach. This method integrates ultrasound, MRI, and mammogram images to improve diagnostic accuracy. By using optimized

segmentation and a deep CNN for feature extraction and classification, the MMRC approach achieved an impressive average accuracy of 96%. This confirms that combining multiple imaging modalities with deep learning is a highly effective and robust strategy for early and accurate breast cancer diagnosis. [5]

A study proposed a DenseNet-based model for automated breast cancer diagnosis from mammograms to overcome the high false positive and negative rates of traditional methods. The model leverages dense connectivity and transfer learning to improve feature extraction and handle small datasets. The DenseNet model proved superior to other architectures, including ResNet and VGG, achieving higher accuracy, precision, recall, sensitivity, and specificity. Its robustness and efficiency make it suitable for real-time clinical applications. [6]

A study addressed the limitations of breast cancer detection systems by introducing M2Net, a two-stage, multi-label detection model built on Faster R-CNN. This new framework is designed to overcome the shortcomings of conventional models that handle lesion detection and BI-RADS classification separately. M2Net uses YOLOv7 for initial region of interest extraction and a unique MammogramSlidingWindows strategy to mimic a radiologist's workflow. Its core innovation is the use of dual classifier heads that simultaneously detect lesion types and their corresponding BI-RADS categories. The model significantly outperformed traditional single-label models, achieving impressive results on the CBIS-DDSM dataset, with the ResNeXt101 backbone providing the best performance, including an 85.54% mAP for BI-RADS classification. [7]

Breast cancer is a leading cause of death among women, making early detection crucial. Deep learning models like DenseNet show strong potential in classifying histopathological images, yet prior studies often overlooked optimizer and learning rate effects, limiting reliability. Using the BreakHis dataset (40X, 2,022 images), transfer learning with DenseNet201 was applied under various optimizers (Adam, SGD, RMSProp) and learning rates (10^{-2}, 10^{-3}, 10^{-4}), with a 70-15-15 train-validation-test split. DenseNet201 achieved 100% accuracy with SGD (10^{-2}) and RMSProp (10^{-3}), and 99% with Adam (10^{-4}), showing that optimized hyperparameters can significantly enhance early detection.[8]

A study addressed the limitations of breast cancer detection from mammograms, which are prone to false positives and negatives, by using a transformer-based model instead of conventional CNNs. The model was trained on a large, diverse collection of datasets, including DDSM, MIAS, INbreast, and VinDr-Mammo, with standard image preprocessing. By leveraging patch embeddings and self-attention, the transformer architecture was able to capture complex spatial patterns more effectively than CNNs like ResNet and Xception. The model demonstrated superior performance, achieving 95.9% accuracy, 94.9% recall, and 97.1% precision, along with an AUC-ROC of 0.977, proving its potential for more reliable and automated breast cancer detection. [9]

Breast cancer is a leading cause of death among women, and reliable early detection is crucial, though conventional methods like mammography and histopathology are limited by variability and expert dependence. Many deep learning models struggle with generalization, highlighting the need for automated and robust approaches. Using the BreakHis dataset (7,909 histopathological images) with preprocessing, normalization, and augmentation, a fine-tuned ResNet50V2 model with added dense and dropout layers achieved an AUC of 0.84413, 79.6% accuracy, and 0.93 average precision, showing strong classification ability despite higher false negatives.[10]

Breast cancer is the most common cancer worldwide, causing over 670,000 deaths in 2022, and early detection via mammography is vital, though manual interpretation limits accuracy. Using the RSNA screening mammography dataset (~20,000 patients) with preprocessing (ROI cropping, windowing, rescaling, padding), ConvNeXT-S and EfficientNetV2-S were trained with four-fold cross-validation and balanced sampling. ConvNeXT-S outperformed EfficientNetV2-S, achieving an AUC of 0.9433, 93.36% accuracy, 93.21% precision, 95.24% recall, and 95.13% F1, demonstrating superior feature extraction and generalization.[11]

Breast cancer is a major global health issue, and early detection via mammography and ultrasound is vital. Conventional imaging and earlier CNNs often face false results, poor generalization, and difficulty with complex tumor patterns, necessitating robust AI models. Advanced architectures—Transformers, ResNet50V2, ConvNeXT, EfficientNet, and U-Net—were applied with preprocessing (normalization, cropping, resizing, augmentation, dual-phase encoding) across diverse datasets. Transformers achieved 95.9% accuracy (AUC 0.977), ResNet50V2 79.6% (AUC 0.844), ConvNeXT 93.36% (AUC 0.9433), and U-Net excelled in ultrasound segmentation with 98.87% accuracy and Dice score 0.9227, showing effectiveness for classification and segmentation.[12]

Breast cancer is a major global health concern, and early detection is crucial. Traditional methods and earlier CAD systems often lack accuracy, highlighting the need for lightweight, efficient deep learning models. Using the BUS-2 ultrasound dataset (647 images: 437 benign, 210 malignant) with preprocessing, augmentation, and normalization, a customized MobileNet with three dense layers and transfer learning was trained using Adam optimizer and cross-entropy loss. The model achieved 92.92% accuracy and 0.95 precision for malignant cases, outperforming VGG19 (86.97%) and SVM (83.3%), demonstrating strong generalization and clinical applicability.[13]

Breast cancer is the second leading cause of cancer deaths among women, and machine learning is vital for accurate prediction. Single classifiers often struggle with outliers, missing values, and large datasets, limiting reliability. The Expert System for Breast Cancer Detection (ESBCD) applied ensemble methods—Random Forest, AdaBoost, and XGBoost—on the Wisconsin Breast Cancer dataset (698 samples, 10 features) with preprocessing for missing values and

outliers. ESBCD with XGBoost performed best, achieving 97.01% accuracy, precision, recall, and F-measure, outperforming standalone classifiers and demonstrating robustness for clinical use.[14]

Breast cancer is one of the most common causes of mortality among women, where early diagnosis using machine learning can greatly improve treatment outcomes and survival. Traditional diagnostic techniques are often time-consuming and imprecise, creating a research gap in evaluating ensemble algorithms for enhanced accuracy. This study used the Wisconsin Breast Cancer dataset (569 samples, 30 features), applied preprocessing, and split the data into training and testing sets. LightGBM and XGBoost were implemented and compared, with XGBoost achieving 99% accuracy, surpassing LightGBM (98%). The findings show that XGBoost delivers stronger generalization and precision, making it a more reliable choice for medical prediction tasks.[15]

Breast cancer accounts for 15% of all cancers in women and is the second leading cause of cancer-related deaths, requiring accurate prediction systems. Using the Wisconsin Breast Cancer dataset (569 samples, 30 features), a stacked ensemble of SVC, KNN, Random Forest, and MLP with GaussianNB as meta-classifier was developed. After preprocessing (feature selection, outlier detection, normalization), the model achieved 99.41% accuracy, 98.75% MCC, and 99.42% F1, outperforming individual classifiers, showing superior generalization and reliability for clinical prediction. [16]

Breast cancer remains a major health challenge, with over 2 million cases reported in 2021, and early detection is critical to reduce mortality. Conventional diagnostic methods and traditional machine learning classifiers often struggle to classify tumors accurately, motivating lightweight CNN approaches. This study proposed a CNN with convolution, flatten, and dense layers using ReLU and sigmoid activations, trained on the Wisconsin Breast Cancer dataset (569 samples: 357 benign, 212 malignant) with Adam optimizer and binary cross-entropy loss. The model achieved 99% accuracy, 98.5% precision, and 100% recall, outperforming other optimizers and demonstrating an effective and efficient approach for breast cancer classification.[17]

Breast cancer affects 1 in 8 women globally, and early detection is critical, though traditional methods often lack sensitivity and specificity. To improve robustness, ensemble models with precedence-based algorithms were applied to two Kaggle datasets—breast cancer (569 samples, 33 features) and cervical cancer (858 samples, 36 features)—testing classifiers like Logistic Regression, Random Forest, AdaBoost, KNN, Decision Trees, and Naïve Bayes. Ranked and fused ensembles achieved 98.5% accuracy (Random Forest + AdaBoost) for breast cancer and 97.9% (Random Forest + KNN) for cervical cancer, showing better reliability than single models.[18]

Breast cancer is the most common cancer in women, with ultrasound imaging offering non-invasive diagnosis but limited by noise and poor contrast. CNN and transformer models struggle with small, overlapping lesions and lack real-time applicability, highlighting the need for a robust, interpretable model. CascadeNet combines YOLOv12 for real-time tumor localization with a modified DETR for classification, trained on the Kaggle Breast Ultrasound dataset with annotations, augmentation, and adaptive softmax loss. It achieved 96.41% mAP@50, 90.37% recall, 93.64% F1-score, and 93.76% accuracy, outperforming YOLOv12, DETR, YOLOv8, and Faster R-CNN, while running at 25 FPS, making it suitable for real-time clinical deployment.[19]

Histopathology imaging is the gold standard for breast cancer diagnosis, but manual interpretation is slow and inconsistent. While deep learning offers automation, single CNN models risk overfitting and limited generalization. To address this, an ensemble of VGG16, ResNet50, and InceptionV3 was trained on the BreakHis dataset (7,909 images at 40x, 100x, 200x, 400x) with preprocessing steps like augmentation and normalization. Using weighted averaging and stacking, the ensemble achieved 98.59% accuracy, 98.63% precision, and 98.52% recall, outperforming individual models (VGG16: 95.37%, ResNet50: 96.71%, InceptionV3: 97.42%) and demonstrating greater robustness and diagnostic reliability in breast cancer detection [20].

Reference	Methodology	Dataset	Strength	Limitation	Application
[3]	CNN and ResNet50V2 Transfer learning	BreakHis (Histopathology)	Custom CNN outperformed ResNet50v2 (87.83% vs 81.30% validation accuracy)	Moderate AUC (0.82); limited to histopathology images	Automated diagnosis from histopathological biopsy images, assisting pathologists.
[19]	YOLOv12 (Detection) + Modified DETR (Classification)	Kaggle Breast Ultrasound Images	High mAP@50 (96.41%), real-time capable (25 FPS)	Complex pipeline; requires annotated bounding boxes	Real-time ultrasound-based breast cancer detection in clinical settings
[7]	Two-stage Faster R-CNN with multi-label learning for lesion localization, type classification, and BI-RADS scoring. Includes a sliding-window approach for improved detection.	In-house dataset (353 images) and CBIS-DDSM (2,142 images).	Simultaneous detection of lesion type and BI-RADS level; outperforms single-label models	Requires manual annotation; limited to four BI-RADS levels.	Clinical mammography analysis, computer-aided diagnosis (CAD) systems.
[6]	DenseNet-based deep learning model for classifying breast tumors as benign or malignant	Not explicitly named (likely public mammography datasets)	High accuracy, precision, recall, and efficiency; robust with small datasets	Limited to binary classification; no lesion localization	Automated breast cancer screening and diagnosis
[5]	Multi-Modal Radiomics and Deep CNN (MMRC) approach using ultrasound, MRI, and mammograms	Real-time ultrasound, MRI, and mammogram images	Multi-modal fusion improves accuracy; comprehensive evaluation	Computationally intensive; requires multi-modal data	Multi-modal breast cancer diagnosis and treatment planning
[4]	Transfer learning with EfficientNet (b0–b4) for mammography classification	CBIS-DDSM	Lightweight; uses early stopping and cosine annealing for efficient training	Moderate accuracy (75%); limited to image classification only	Mammography-based breast cancer screening
[2]	Ensemble machine learning (RF, DT, NB, etc.) using risk factors.	Breast Cancer Surveillance Consortium (BCSC) risk factor dataset	Uses non-invasive risk factors; interpretable	Lower accuracy (75.2%); no imaging data	Risk assessment and preventive care for breast cancer
[1]	Lightweight multitask network (MobileNet-based) for classification and segmentation	Breast ultrasound images from multiple public datasets	Efficient; mobile-deployable; balanced sensitivity/specificity	Limited to ultrasound images; may not generalize to other modalities	Real-time breast cancer diagnosis on mobile devices
[18]	Precedence-based ensemble learning using RF, K-NN, AdaBoost, etc., with fusion based on TPR/TNR averages	Cervical cancer (858 samples) and breast cancer (569 samples) from Kaggle	High accuracy (98.5% for breast cancer); interpretable ensemble selection	Limited to tabular data; no imaging analysis	Risk prediction and early screening using clinical and demographic data
[8]	Lightweight CNN with ReLU and sigmoid activations, trained on tabular data from WBCD	Wisconsin Breast Cancer Dataset (WBCD – 569 samples)	99% accuracy; simple and efficient	Not tested on image data; limited to WBCD	Binary classification of breast cancer using clinical features
[16]	Stacking ensemble with SVC, K-NN, RF, MLP as base models and GaussianNB as meta-classifier	Wisconsin Breast Cancer Dataset (WBCD)	99.41% accuracy; robust to class imbalance	Computationally intensive; limited to tabular data	High-accuracy breast cancer prediction using ensemble ML
[15]	Comparative analysis of LightGBM and XGBoost ensemble algorithms	Wisconsin Breast Cancer Dataset (569 samples, 30 features)	XGBoost achieved 99% accuracy, outperforming LightGBM (98%)	Limited to tabular data; no image-based analysis	Clinical decision support for early breast cancer diagnosis
[14]	Ensemble learning (RF, AdaBoost, XGBoost) with preprocessing for missing values and outliers	Wisconsin Breast Cancer Dataset (preprocessed, 670 instances)	ESBCD with XGBoost achieved 97.01% accuracy; includes robust preprocessing	Increased computational cost due to ensemble methods	Healthcare informatics; automated diagnostic systems
[13]	Transfer learning with MobileNet architecture; data augmentation and hyperparameter tuning	BUS-2 Breast Ultrasound Images (647 images: 437	Achieved 92.92% accuracy; efficient and lightweight model suitable for mobile deployment	Limited to ultrasound images; may not generalize to other modalities	Point-of-care ultrasound diagnosis; telemedicine

		benign, 210 malignant)			
[12]	Custom U-Net architecture with dual-phase encoding/decoding, skip connections, dropout	Breast Ultrasound Images (780 images: 437 benign, 210 malignant, 133 normal)	High accuracy (98.87%) and Dice coefficient (92.27%); excellent segmentation performance	Requires high computational resources; complex architecture	Medical image segmentation; radiologist assistance tools
[11]	Comparative study of ConvNeXT-small and EfficientNetV2-S on mammograms	RSNA Screening Mammography Dataset (~20,000 patients, 4 images per patient)	ConvNeXT achieved 94.33% AUC and 93.36% accuracy; outperformed EfficientNet	Large dataset required; computationally intensive	Screening mammography; automated mass detection in radiology
[10]	Transfer learning with ResNet50V2; data augmentation (brightness, flip, rotation); binary classification	BreakHis (7,909 histopathological images, 40×–400× magnifications)	Achieved ROC-AUC of 0.84413; good feature extraction capabilities	High false negatives (140); accuracy only 79.63%	Histopathology image analysis
[9]	Vision Transformer (ViT) architecture; patch-based image processing; multi-head self-attention	DDSM, MIAS, INbreast, VinDr-Mammo (multiple mammography datasets)	High accuracy (95.9%) and AUC (0.977); outperformed CNNs like ResNet, EfficientNet	Computationally intensive; requires large annotated datasets	Mammography screening; real-time CAD systems

Table 1: Comparative Summary of Key Literature: Methodology, Characteristics, and Finding

III. PROPOSED METHODOLOGY:

The flowchart outlined in Figure 1 outlines the end-to-end process of a breast cancer image classification system. It begins with Data Collection, where publicly available datasets are used, and expert annotations provide accurate labeling. Next, in Data Preprocessing, the images undergo noise reduction, resizing, normalization, and data augmentation to enhance data quality and variability. Following this, Feature Extraction involves analyzing color, texture, shape, and size to capture critical image characteristics. The Model Selection & Training stage employs an ensemble learning approach using soft voting, combining base models such as Xception, DenseNet, and VGGNet, trained on labeled data to improve classification performance. The system's effectiveness is then measured during Model Evaluation using metrics like accuracy, precision, recall, F1 score, confusion matrix, and AUC curve. Once validated, the model moves to Deployment & Monitoring, where it is integrated into hospital web systems and IoT devices for real-time disease classification.

Finally, Validation & Expert Feedback ensures continuous improvement by incorporating real-time testing and expert insights to enhance model accuracy and reliability.

1. Data Collection

The first step involves gathering datasets that are essential for building and training the model. Publicly available datasets provide a solid foundation, ensuring access to diverse and standardized data. To enhance the quality of the dataset, expert annotation is used for labeling, which ensures that the data is accurately categorized and reliable for training purposes. Expert involvement in labeling is particularly crucial in medical or agricultural domains, where precise knowledge is required to distinguish between different classes. This step ensures that the dataset is both comprehensive and trustworthy.

The centroid (mean vector) of each class is calculated by averaging the feature vectors of all samples belonging to that class, as shown in equation (1).

$$\mu_l^{\rightarrow} = \frac{1}{|C_l|}\sum_{i \in C_l} x_i^{\rightarrow} \quad (1)$$

where μ_l is the centroid vector for class l, C_l is the set of indices of samples in class l (where $l \in Y$), $|C_l|$ is the number of samples in class l, x_i is the feature vector of the i-th sample in class l.

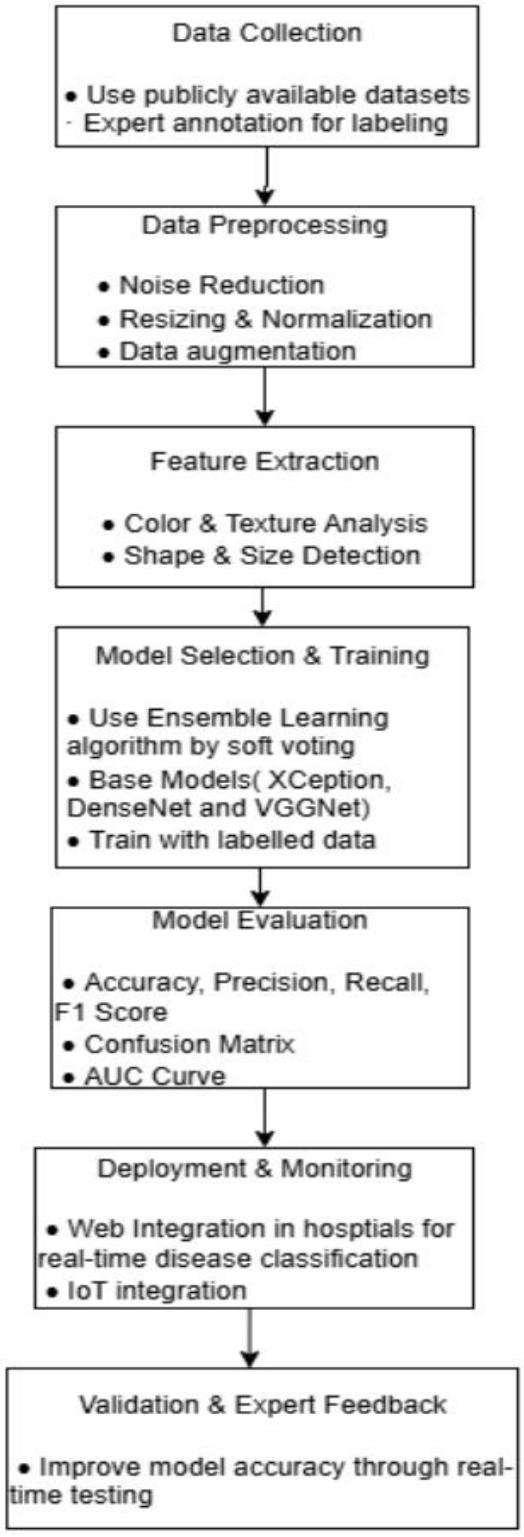

Figure 1: Workflow of the breast cancer classification system from data collection to deployment and validation.

2. Data Preprocessing

Raw data often contains inconsistencies, noise, and irrelevant information, which can reduce the efficiency of the model. Data preprocessing helps refine the dataset to make it suitable for training. This process includes noise reduction to eliminate irrelevant distortions, resizing and normalization to ensure uniformity across all images, and data augmentation to increase the diversity of the dataset by applying transformations like rotation, flipping, or scaling. These steps collectively improve the robustness and generalization capabilities of the model, ensuring better performance during real-world application.

The predicted class for an observation is assigned by finding the label whose representative vector μ_l is closest to the input vector $\hat{x}$, minimizing the distance between them, as shown in equation (2).

$$x^{\rightarrow} is\, \hat{y} = arg \min_{l \in Y} \left| |\mu_l^{\rightarrow} - x^{\rightarrow}| \right| \quad (2)$$

where $\hat{x}$ is the predicted class vector, $\hat{y}$ is the assigned class label, μ_l is the prototype (mean) vector for class l in the label set Y, arg min denotes the label l that minimizes the distance between μ_l and $\hat{x}$.

The correlation coefficient measures the strength and direction of the linear relationship between two random variables. It is calculated using covariance and standard deviations, as shown in equation (3).

$$\rho_{X_1X_2} = \frac{Cov(X_1,X_2)}{\sqrt{DX_1 * DX_2}} = \frac{EX_1X_2 - EX_1 * EX_2}{\sqrt{DX_1 * DX_2}} \quad (3)$$

where $\rho_{X_1X_2}$ is the correlation coefficient between variables X_1 and X_2, $Cov(X_1, X_2)$ is the covariance between X_1 and X_2, $E(X_1)$ and $E(X_2)$ are the expected values (means) of X_1 and X_2 respectively, $E(X_1X_2)$ is the expected value of the product of X_1 and X_2, $D(X_1)$ and $D(X_2)$ are the variances of X_1 and X_2.

3. Feature Extraction

Feature extraction focuses on identifying meaningful attributes from the preprocessed data. Important features such as color, texture, shape, and size are analyzed and extracted, as these characteristics play a vital role in distinguishing between different classes, such as diseased vs. healthy samples. For instance, texture analysis may help in identifying irregular patterns, while shape detection can highlight structural abnormalities. By extracting relevant features, the model is better equipped to learn discriminative patterns, improving classification accuracy.

The total objective function in gradient boosting combines the training loss with a regularization term to balance model accuracy and complexity. This is expressed in equation (4).

$$obj = \sum_{i=1}^{m} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \quad (4)$$

where obj is the overall objective function to be minimized, $l(y_i, \hat{y}_i)$ is the loss between the true label y_i and the predicted value $\hat{y}_i$, $\Omega(f_k)$ is the regularization function for the k-th tree, K is the total number of trees in the ensemble, m is the number of training examples.

4. Model Selection & Training

This step involves choosing suitable machine learning or deep learning models and training them with the labeled dataset. An ensemble learning approach is employed using soft voting, which combines the predictions of multiple base models to improve accuracy and reliability. Base models such as Xception, DenseNet, and VGGNet are selected due to their proven performance in image classification tasks. The models

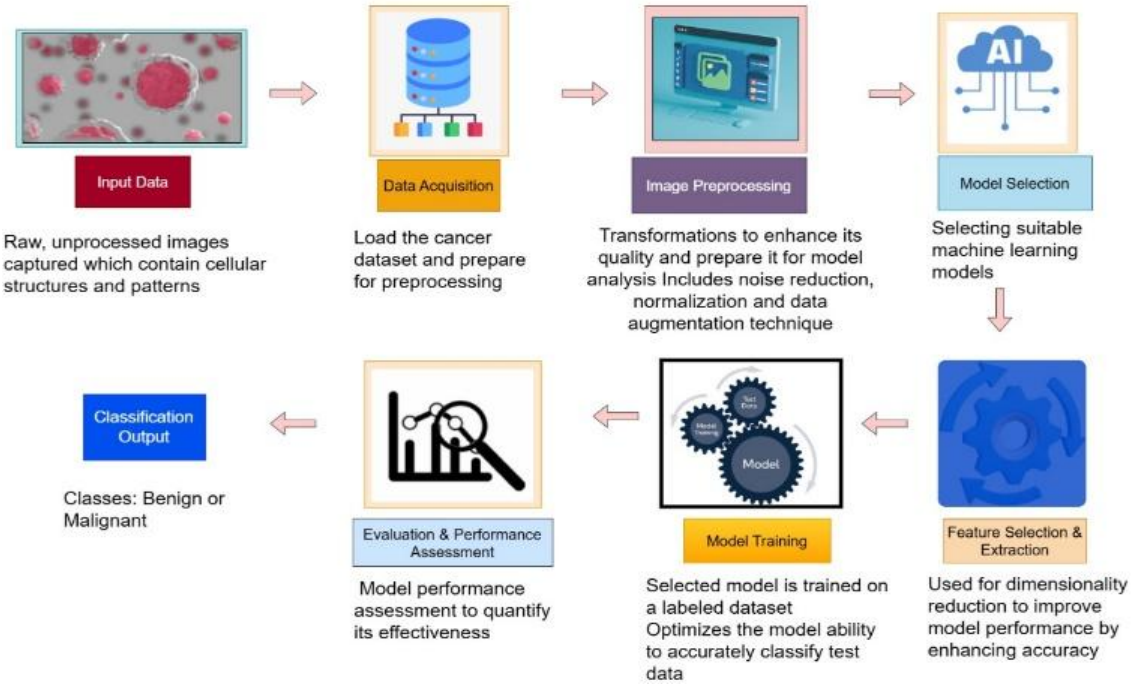

Figure 2: The System Architecture for the proposed methodology

are trained on labelled data, enabling them to learn patterns, features, and distinctions necessary for disease classification. Ensemble methods reduce bias and variance, ensuring more stable predictions.

To optimize the model in gradient boosting, a second-order Taylor expansion of the loss function is used to approximate the objective function. This allows efficient computation using both the first and second derivatives of the loss, as shown in equation (5).

$$obj = \sum_{i=1}^{m} \left[f_t(x_i) g_i + \frac{1}{2} \left(f_t(x_i) \right)^2 h_i \right] + \Omega(f_t) \quad (5)$$

where obj is the approximate objective function at the current boosting step, $f_t(x_i)$ is the prediction of the new tree being added for data point x_i, g_i is the first derivative (gradient) of the loss function with respect to the previous prediction $\hat{y}_i^{(t-1)}$, h_i is the second derivative (Hessian) of the loss function with respect to $\hat{y}_i^{(t-1)}$, $\Omega(f_t)$ is the regularization term that penalizes model complexity, m is the number of training examples.

5. Model Evaluation

After training, the model is evaluated using performance metrics to measure its effectiveness. Key metrics include accuracy, precision, recall, and F1 score, which provide insights into the balance between correctly identifying positive and negative samples. The confusion matrix is used to visualize the classification results, highlighting true positives, false positives, false negatives, and true negatives. Additionally, the AUC (Area Under the Curve) is used to evaluate the model's ability to distinguish between classes. This evaluation step ensures that the model performs reliably and is ready for real-world deployment.

The accuracy of a classifier is estimated as shown in equation (6).

$$Accuracy = \frac{tp+tn}{tp+tn+fp+fn} \quad (6)$$

where tp denotes true positives, tn denotes true negatives, fp denotes false positives, fn denotes false positives

The recall score shown in equation (7) is defined as the ratio of true positives to the total number of actual positive instances.

$$Recall = \frac{tp}{tp+fn} \quad (7)$$

where tp (True Positives) are the number of instances correctly identified as positive, fn (False Negatives) are the number of instances incorrectly identified as negative.

The precision score is defined as the ratio of true positives to the total number of positive predictions as depicted in equation (8).

$$Precision = \frac{tp}{tp+fp} \quad (8)$$

where tp (True Positives) are the number of instances correctly identified as positive, fp (False Positives) are the number of instances incorrectly identified as positive.

The F1-score is the harmonic mean of precision and recall, providing a single score that balances both metrics as shown in equation (9)

$$F = 2 * \left(\frac{precision * recall}{precision + recall} \right) \quad (9)$$

6. Deployment & Monitoring

Once validated, the trained model is deployed for real-world use. In the case of disease classification, web integration enables hospitals or institutions to use the model for real-time

diagnosis and decision-making. IoT integration allows continuous monitoring through connected devices, enhancing accessibility and efficiency. Monitoring during deployment ensures that the model performs consistently under practical conditions and provides timely feedback for further improvements. This step bridges the gap between research and practical application.

In gradient boosting, the residuals at each iteration are computed as the negative gradient of the loss function with respect to the current model's prediction. These residuals guide the next learner in minimizing the overall loss, as shown in equation (10).

$$Compute\ r_{im} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{\{F(x)=F_{(m-1)}(x)} \quad \forall i = 1,2,..,n \qquad (10)$$

where r_{im} is the pseudo-residual for the *i-th* data point at iteration *m*, $L(y_i, F(x_i))$ is the loss function, $F(x_i)$ is the model prediction for input x_i, $F_{(m-1)}(x)$ is the prediction from the previous $(m-1)^{th}$ iteration, n is the total number of data points.

After computing the residuals and fitting a weak learner (typically a decision tree), the optimal multiplier γ_m is determined by minimizing the loss function over all data points. This step fine-tunes how much influence the new learner has on the model, as shown in equation (11).

$$\gamma_m = arg\ min\ \sum_{i=1}^{n} L\big(y_i, F_{m-1}(x_i) + \gamma h_m(x_i)\big) \qquad (11)$$

where γ_m is the optimal step size (or learning rate multiplier) for the *m*-th iteration, $L(y_i)$ is the loss function, $F_{(m-1)}(x_i)$ is the model prediction from the previous iteration, $h_m(x_i)$ is the output of the newly fitted base learner on input x_i, n is the total number of training samples.

7. Validation & Expert Feedback

The final step involves validating the deployed model and gathering expert feedback to assess its real-world performance. Real-time testing with actual data helps identify gaps, errors, or biases that may not have surfaced during training. Expert feedback is crucial to refine the system and make it more accurate and reliable. Continuous validation ensures that the model evolves and improves over time, maintaining high standards of performance and relevance in dynamic environments.

The regularization function in gradient boosting helps control model complexity and prevent overfitting by penalizing the number of leaves and the magnitude of leaf weights, as shown in equation (12).

$$\Omega(f_t) = \gamma^T + \frac{1}{2}\lambda\left|\left|\omega\right|\right|^2 \qquad (12)$$

where γ is the regularization term for the number of leaves (T), λ is the L2 regularization parameter, $\|\omega\|^2$ is the squared sum of the leaf weights.

The ensembles are constructed incrementally, with each new ensemble correcting the inaccuracy of the preceding one according to the stated equation in (13).

$$f_k(x) = \sum_{m=1}^{k} y_m\, h_m(x) \qquad (13)$$

where f_k shows partial ensembles with k members for ' m 'decision trees, y_m represents the weight of individual learners h_m.

The z−score of a sample *x* is estimated as shown in equation (14).

$$z = \frac{(x-\mu)}{\sigma} \qquad (14)$$

where μ denotes the mean of the training samples, σ denotes the standard deviation of the training samples.

The Chi-squared statistic, $\chi 2$, is calculated as in equation (15). This formula is used in hypothesis testing to determine if there is a statistically significant difference between observed and expected frequencies in one or more categories of a contingency table.

$$X_c^2 = \sum((O_i - B_i)^2/B_i) \qquad (15)$$

where "c" are the degrees of freedom, "O" is your observed value, "B" is your expected value.

This study proposes an ensemble-based deep learning architecture for breast cancer histopathological image classification. Figure 2 represents the workflow of the breast cancer image classification process from data acquisition to classification output, covering preprocessing, model training, evaluation, and performance assessment. The methodology follows a series of sequential stages, from data acquisition to classification, as detailed below:

IV. RESULTS AND DISCUSSIONS

Dataset Description:

The Breast Cancer Histopathological Database (BreaKHis) was developed in collaboration with the P&D Laboratory – Pathological Anatomy and Cytopathology, Parana, Brazil, and introduced by Spanhol et al. (2016). It contains 7,909 microscopic images of breast tumor tissue collected from 82 patients via the SOB method (partial mastectomy/excisional biopsy). Images are captured at four magnifications (40X, 100X, 200X, 400X), with a resolution of 700 × 460 pixels stored in PNG format with 8-bit depth per RGB channel. Figure 3 represents a slide of breast malignant tumor (stained with HE) seen in different magnification factors.

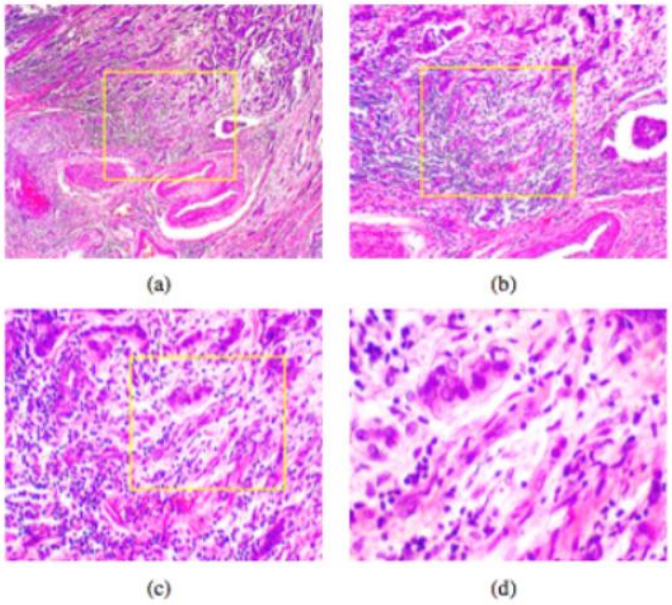

Figure 3: A slide of breast malignant tumor (stained with HE) seen in different magnification factors: (a) 40X, (b) 100X, (c) 200X, and (d) 400X.

Classes and Subtypes:

There are two classes in the dataset. *Benign*: Adenosis (A), Fibroadenoma (F), Phyllodes Tumor (PT), Tubular Adenoma (TA) and *Malignant*: Ductal Carcinoma (DC), Lobular Carcinoma (LC), Mucinous Carcinoma (MC), Papillary Carcinoma (PC)

The BreakHis 1.0 is structered as shown in table 2:

Magnification	Benign	Malignant
40X	652	1,370
100X	644	1,437
200X	623	1,390
400X	588	1,232
Total of Images	2,480	5,429

Table 2: BreakHis 1.0 Dataset Distribution by Magnification and Class

File Naming Convention: Each filename encodes biopsy method, tumor class, tumor type, patient ID, magnification, and sequence (e.g., SOB_B_TA-14-4659-40-001.png corresponds to a benign Tubular Adenoma, collected in 2014, slide ID 4659, at 40X magnification, first image in sequence).

Relevance to the Study

Breast cancer remains one of the leading causes of cancer-related deaths worldwide. Accurate diagnosis of tumor types is critical for treatment and prognosis. The BreaKHis dataset provides a large-scale, diverse, and well-annotated benchmark for developing and evaluating machine learning and deep learning models for histopathological image classification. Its inclusion of multiple magnification levels and tumor subtypes makes it particularly suitable for research in different fields like Automated cancer detection, Multi-class classification, Magnification-independent learning, and Clinical decision support systems.

1.Basic CNN

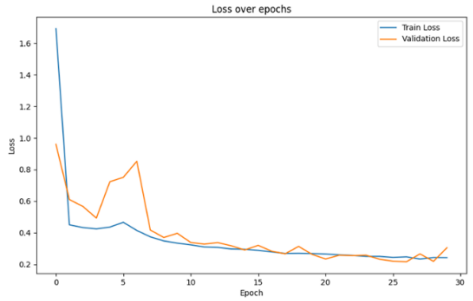

Figure 4: Graph for Loss in Basic CNN Model

Figure 4 depicts the training and validation loss over 30 epochs. Both losses start high at approximately 1.6 and 0.8 respectively, then decrease rapidly in the initial epochs before stabilizing around 0.2-0.3. The close alignment between training and validation loss throughout the training process indicates good generalization without overfitting, suggesting the model has learned meaningful patterns that transfer well to unseen data.

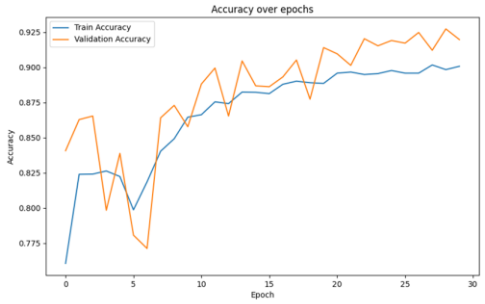

Figure 5: Graph for Accuracy in Basic CNN Model

The training and validation accuracy progression, illustrated in Figure 5, rises from around 80% to approximately 92% over 30 epochs. Both curves track closely with only minor fluctuations, reflecting a stable learning process. The parallel movement of training and validation accuracy suggests the model generalizes effectively without overfitting, which is a strong indicator of reliable real-world performance.

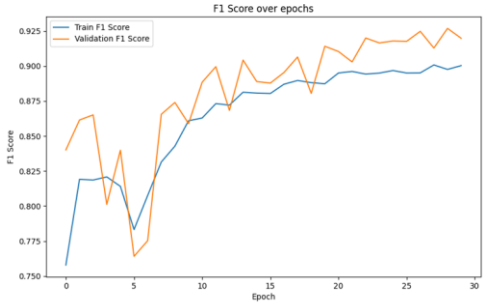

Figure 6: Graph for F1 Score in Basic CNN Model

The F1 score progression, as shown in Figure 6, improves steadily from about 80% to 92% over 30 epochs. The close alignment of training and validation F1 scores highlights balanced performance in both precision and recall, indicating

that the model effectively distinguishes between benign and malignant cases without significant bias.

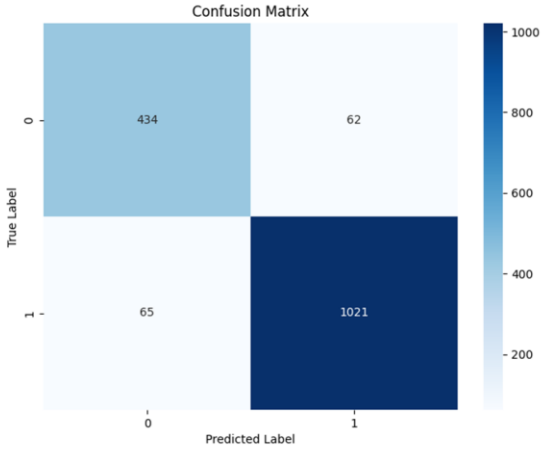

Figure 7: Confusion Matrix for Basic CNN Model

Figure 7 presents the confusion matrix revealing strong classification performance with 434 true negatives, 1021 true positives, 62 false positives, and 65 false negatives. This translates to high accuracy with relatively few misclassifications. The balanced error distribution suggests the model doesn't exhibit significant bias toward either class, making it reliable for clinical applications.

2.ResNet50

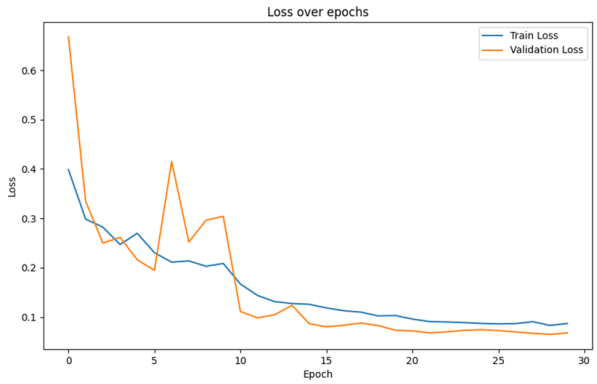

Figure 8: Graph for Loss in ResNet50 Model

As shown in Figure 8, the training and validation loss decrease steadily from about 0.7 to 0.1 across 30 epochs. The training loss exhibits a consistent downward trend, while the validation loss shows minor fluctuations between epochs 5–10 before stabilizing. Their eventual convergence at low values highlights effective model learning and strong generalization performance.

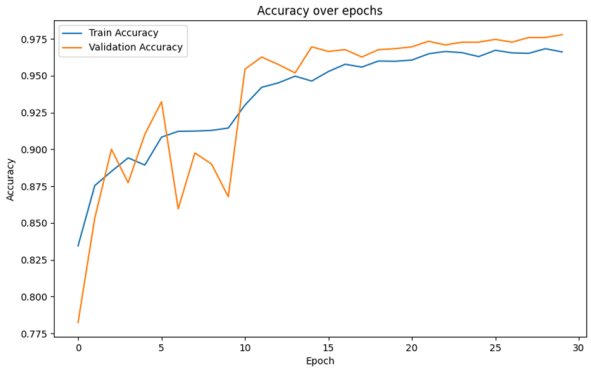

Figure 9: Graph for Accuracy in ResNet50 Model

Throughout 30 epochs, training and validation accuracy increased significantly from about 77% to 97%, as seen in Figure 9. While the validation curve displayed some volatility during the initial 10 epochs, both curves eventually converged smoothly at high accuracy levels, reflecting excellent performance and strong generalization of the model.

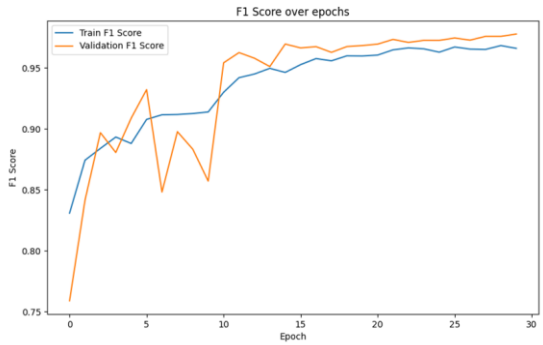

Figure 10: Graph for F1 Score in ResNet50 Model

Figure 10 illustrates the F1 score progression increasing from approximately 75% to 97% over 30 epochs. The training and validation curves show consistent improvement with closely aligned performance, demonstrating balanced precision and recall across both classes throughout the training process.

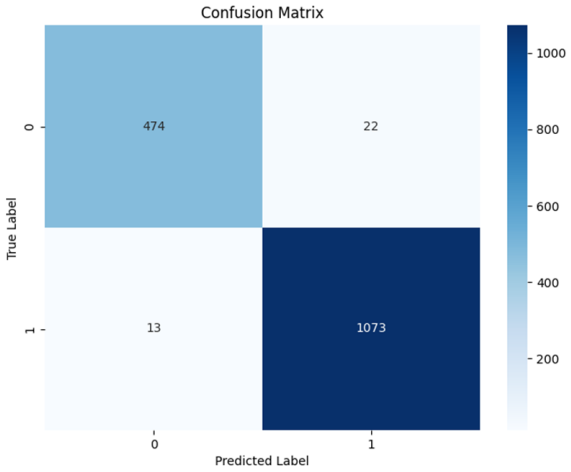

Figure 11: Confusion Matrix in ResNet50 Model

The confusion matrix in Figure 11 highlights excellent classification performance, with 474 true negatives, 1073 true

positives, 22 false positives, and 13 false negatives. The very small number of misclassifications demonstrates highly accurate model behavior with minimal error rates across both classes.

3. VGG1

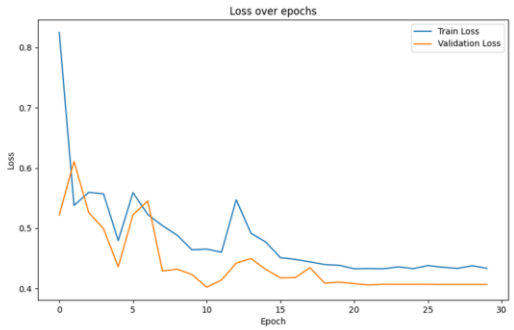

Figure 12: Graph for Loss in VGG1 Model

In Figure 12, the training and validation loss trends across 30 epochs begin near 0.8 and steadily decline to about 0.43. Noticeable fluctuations appear between epochs 5 and 15, showing periods of instability before both curves stabilize. The eventual convergence suggests that, despite the volatility, the model reached a reasonable level of performance.

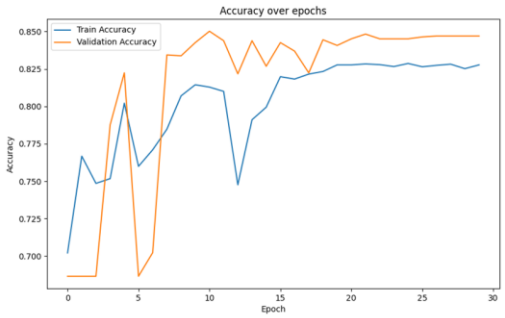

Figure 13: Graph for Accuracy in VGG1 Model

In Figure 13, the accuracy trends for both training and validation rise from nearly 70% to about 83% across 30 epochs. However, the learning process is not smooth, as sharp fluctuations appear in the initial stages. A noticeable decline occurs around epoch 12, after which the model gradually recovers. Such irregularities point to possible challenges related to tuning parameters, dataset quality, or architectural stability during training.

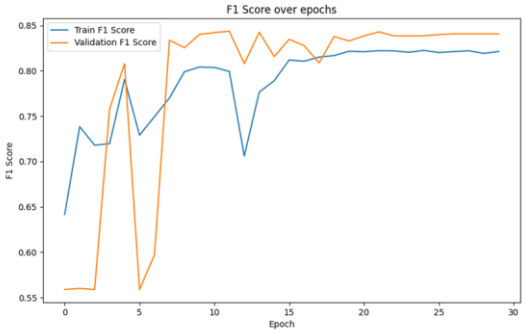

Figure 14: Graph for F1 Score in VGG1 Model

The F1 score progression, shown in Figure 14, rises from roughly 73% to 82% across 30 epochs. Much like the accuracy curves, it demonstrates instability, including a steep decline around epoch 12 before recovering steadily. This pattern reflects training challenges but also highlights that the model ultimately achieved a reasonably balanced level of performance.

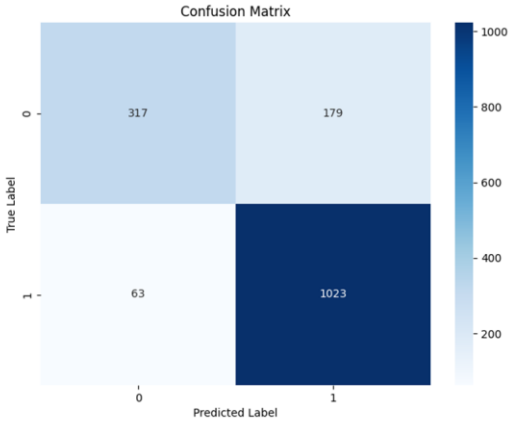

Figure 15: Confusion Matrix for VGG1 Model

The confusion matrix in Figure 15 shows 317 true negatives, 1023 true positives, 179 false positives, and 63 false negatives. While the model demonstrates decent overall accuracy, the relatively high number of false positives indicates a tendency to favor the positive class, which could be a concern in medical applications.

4. MobileNetV2

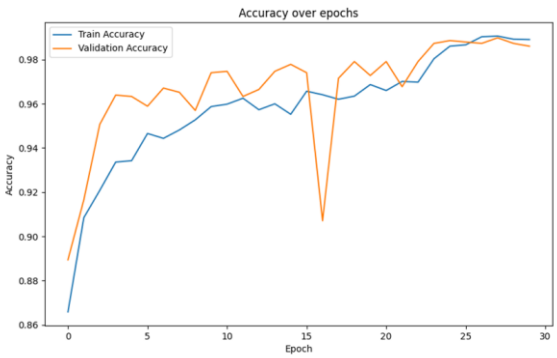

Figure 16: Graph for Accuracy in MobileNetV2 Model

Figure 16 depicts training and validation accuracy over 30 epochs, improving from around 80% to 98%. A notable performance disruption occurs around epoch 15 where both metrics drop significantly before recovering to excellent levels. This pattern suggests a training instability that was eventually

overcome.

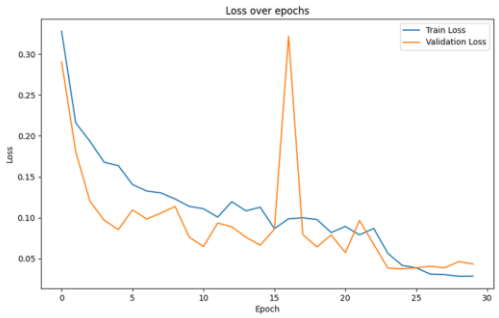

Figure 17: Graph for Loss in MobileNetV2 Model

The progression of training and validation loss, depicted in Figure 17, drops from about 0.3 to near zero over 30 epochs. Although there is a spike around epoch 15 linked to the accuracy dip, both losses ultimately stabilize at very low levels, demonstrating excellent model training.

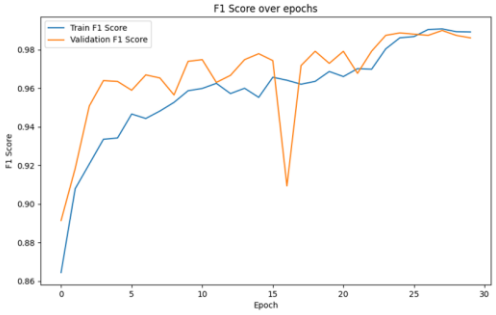

Figure 18: Graph for F1 Score in MobileNetV2 Model

F1 scores rise from around 88% to 98% over 30 epochs, with a temporary disruption around epoch 15, as illustrated in Figure 18. Despite this setback, the model ultimately achieves excellent final performance with high precision and recall.

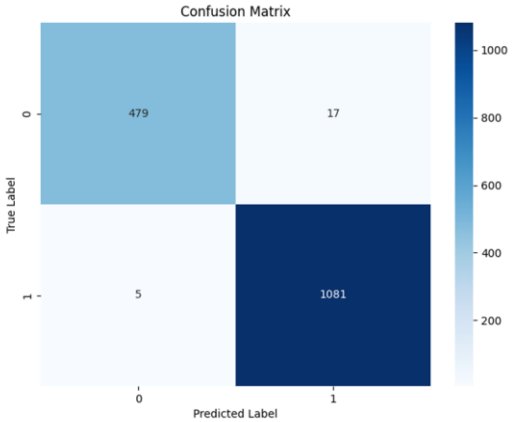

Figure 19: Confusion Matrix for MobileNetV2 Model

Figure 19 presents an exceptional confusion matrix with 479 true negatives, 1081 true positives, 17 false positives, and only 5 false negatives. This represents outstanding classification performance with minimal errors, making it highly suitable for clinical applications.

Hyperparameter Optimization:

Parameters Explored:

- Learning Rates: [0.1, 0.01, 0.001].
- Batch Sizes: [16, 32, 64].

Results:

- Best learning rate: 0.001.
- Best batch size: 16.

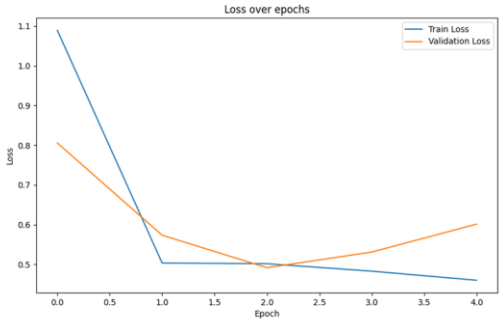

Figure 20: Graph for Loss over epochs in hyperparameter optimization

Training and validation loss over 4 epochs show different behaviours, with training loss decreasing sharply from approximately 1.1 to 0.45 and validation loss declining more gradually from 0.85 to 0.6, as depicted in Figure 20. This diverging pattern suggests potential overfitting, indicating the model may be capturing training-specific patterns rather than generalizable features.

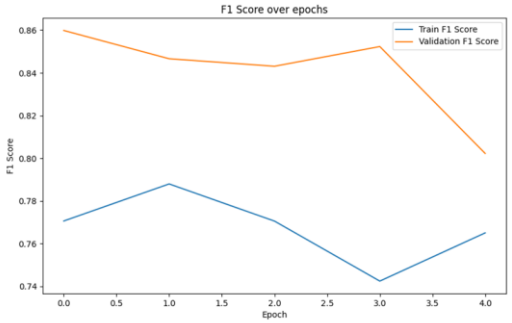

Figure 21: Graph for F1 Score over epochs in hyperparameter optimization

Over 4 epochs, both training and validation F1 scores show a declining trend, with training decreasing from about 86% to 80% and validation from 78% to 74%, as illustrated in Figure 21. This drop highlights potential learning issues, suggesting that modifications to the model architecture or training strategy may be necessary.

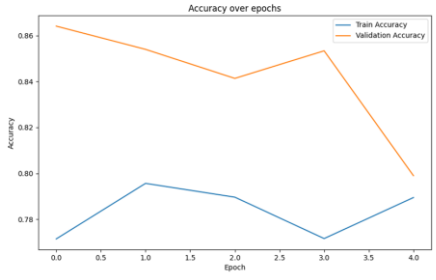

Figure 22: Model Accuracy During Hyperparameter Tuning

Figure 22 illustrates a similar declining pattern in accuracy metrics over 4 epochs. Training accuracy decreases from approximately 88% to 82%, while validation accuracy falls from 80% to 76%. This performance degradation indicates fundamental issues with the model's learning process or data quality.

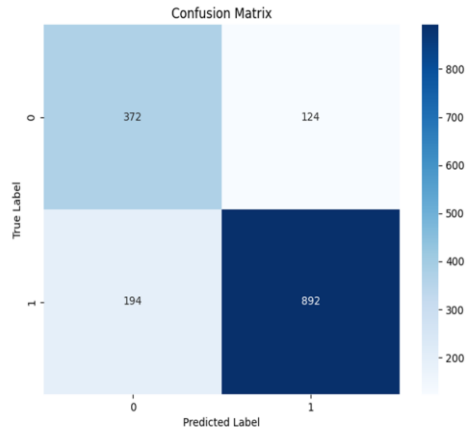

Figure 23: Confusion Matrix

A confusion matrix in Figure 23 shows 372 true negatives, 892 true positives, 124 false positives, and 194 false negatives. The high number of misclassifications, particularly the 194 false negatives, raises serious concerns for medical applications, where missing positive cases can have critical consequences.

Ensemble Deep Learning for Breast Cancer Histopathological Image Classification:

This section presents the ensemble learning framework developed to classify breast cancer histopathological images.

The approach integrates three pretrained convolutional neural networks - VGG16, DenseNet121, and Xception through a soft voting ensemble strategy to enhance predictive robustness and generalization. The ensemble achieved an overall accuracy of 95.33%, with balanced precision, recall, and F1-scores across benign and malignant classes, indicating high diagnostic reliability.

Table 3. Performance metrics of the breast cancer classification model for Benign and Malignant classes.

Metric	Benign	Malignant	Macro Average
Precision	0.95	0.96	0.95
Recall	0.96	0.95	0.95
F1-Score	0.95	0.95	0.95
Support	300	300	600

The model achieved strong and balanced performance across both classes. From table 3, it can be found that precision, recall, and F1-scores were 0.95 - 0.96 for Benign and Malignant samples, indicating reliable and consistent classification. The macro-averaged values of 0.95 confirm that the model performs uniformly well without bias toward either class.

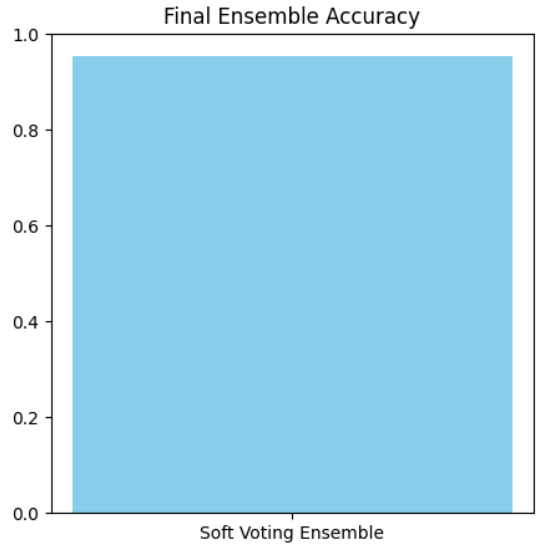

Fig 24. Accuracy of soft voting ensemble model

The graphical representation of the accuracy of the model is visualized in figure 24. The proposed Soft Voting Ensemble model achieved an overall accuracy of 95.33%, demonstrating balanced and reliable classification performance across both Benign and Malignant categories.

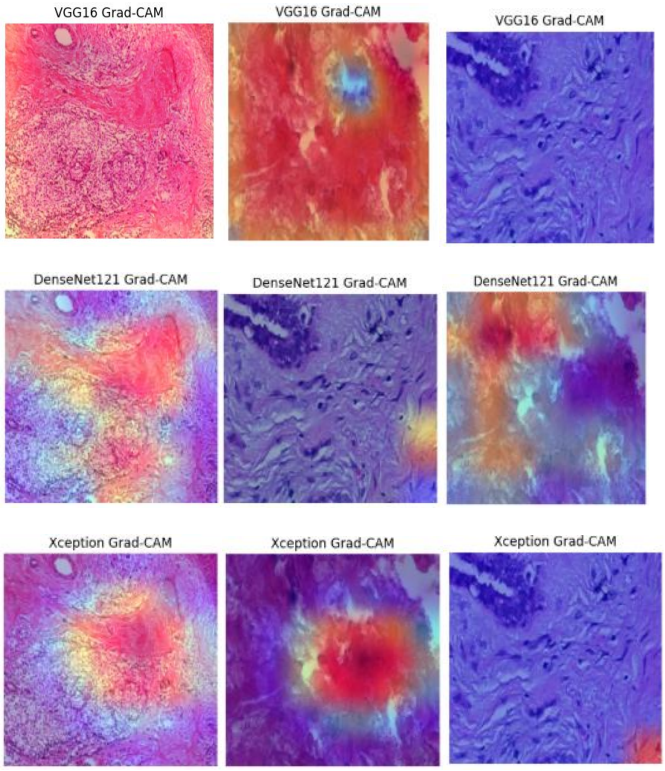

Figure 25. Grad-CAM visualizations generated from VGG16, DenseNet121, and Xception models for three representative test samples.

The heatmaps highlight discriminative regions influencing the classification of breast histopathology images. Areas in red indicate regions of higher importance contributing to the model's decision, showing that the ensemble models effectively focus on relevant tumor regions for both benign and malignant cases.

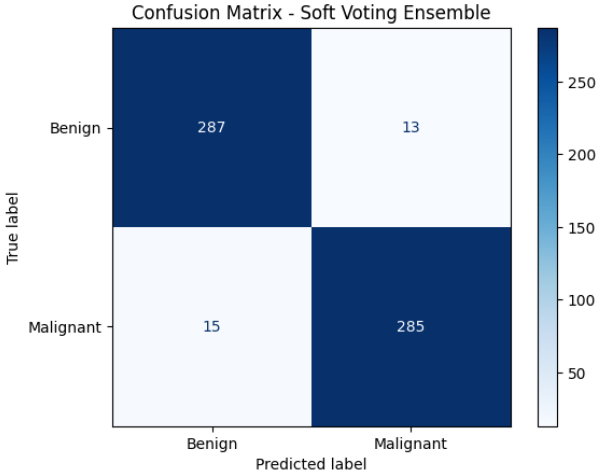

Figure 26. Confusion Matrix

The confusion matrix shown in Figure 26. illustrates the classification outcomes for the Benign and Malignant categories. Out of 300 benign samples, 287 were correctly identified, while only 13 were misclassified. Similarly, 285 malignant samples were correctly detected, with 15 false negatives. The near-diagonal dominance of the matrix indicates highly accurate and balanced predictions across both classes.

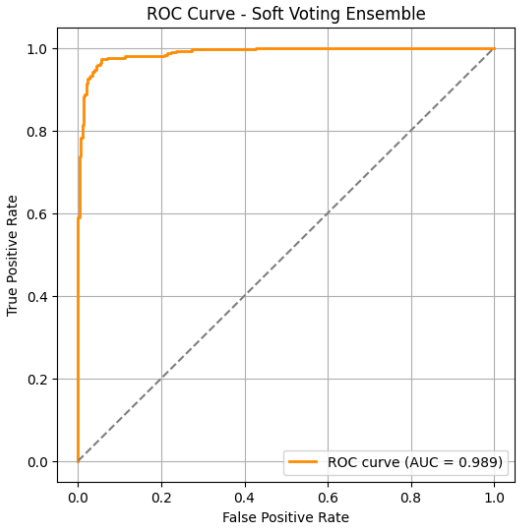

Figure 27. ROC Curve for the ensemble model

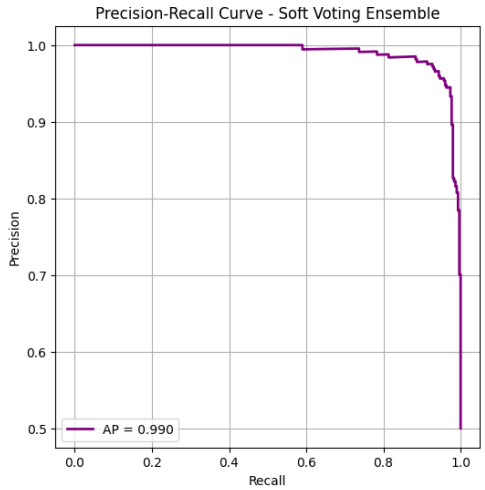

Figure 28. Precision-Recall Curve

The Receiver Operating Characteristic (ROC) curve visualized in Figure 27. evaluates the trade-off between the true positive rate and false positive rate. The Area Under the Curve (AUC) of 0.989 signifies exceptional discriminative ability, confirming that the ensemble model can effectively distinguish between benign and malignant with minimal misclassification.

The Precision-Recall (PR) curve represented in Figure 28. demonstrates the model's ability to maintain high precision even at varying recall levels. The average precision (AP) score of 0.990 indicates that the classifier sustains strong predictive reliability, especially under class imbalance conditions. The near-perfect PR curve highlights the robustness of the soft voting approach.

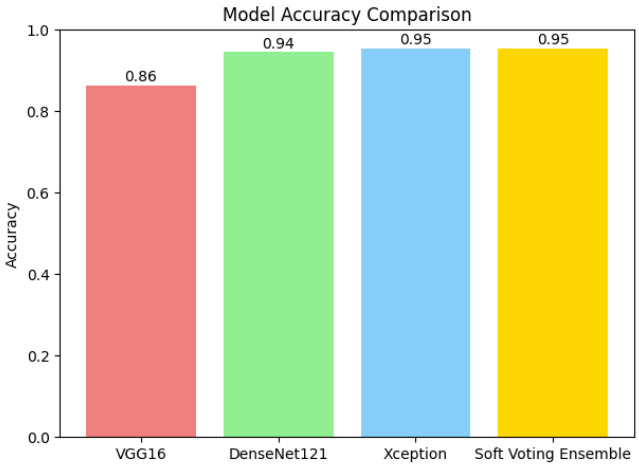

Figure 29. Model Accuracy Comparison

This bar chart depicted in Figure 28. compares the performance of individual deep learning models - VGG16, DenseNet121, and Xception - against the proposed Soft Voting Ensemble. While VGG16 achieved 86% accuracy, DenseNet121 and Xception reached 94% and 95%, respectively. The ensemble model slightly outperformed all individual architectures with an accuracy of 95.33%, confirming the advantage of model combination through soft voting.

Figure 30(a). VGG16 Training Accuracy and Loss Curve

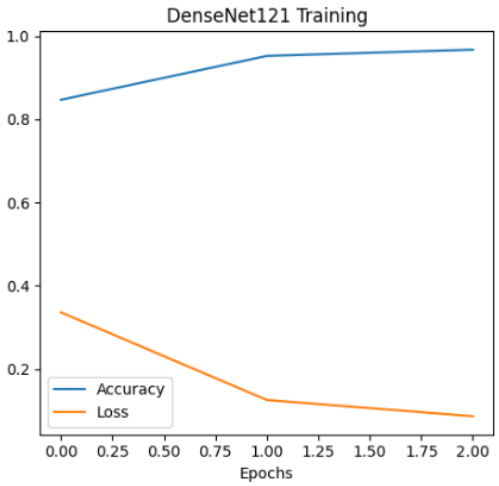

Figure 30(b). DenseNet121 Training Accuracy and Loss Curve

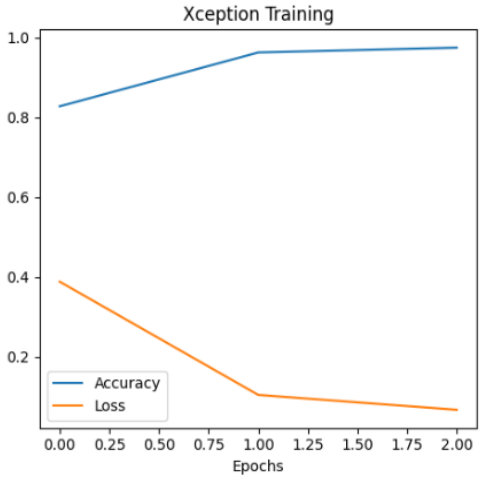

Figure 30(c). Xception Training Accuracy and Loss Curve

The training accuracy and loss plots for VGG16, DenseNet121, and Xception represented in Figure 30 (a),(b),(c) demonstrate smooth convergence over epochs. All three models show steadily increasing accuracy and decreasing loss, indicating effective learning and minimal overfitting. Among them, DenseNet121 and Xception achieved near-optimal convergence, supporting their inclusion in the ensemble for enhanced final performance.

V. Conclusion

This study proposed a robust Soft Voting Ensemble framework integrating VGG19, DenseNet121, and Xception architectures for automated breast cancer classification from histopathological images. By combining the complementary feature extraction strengths of each network - VGG19's fine-grained spatial analysis, DenseNet's feature reuse, and Xception's efficient depthwise separability, the ensemble achieved superior diagnostic reliability and generalization compared to individual models. The proposed model attained a classification accuracy of 95.33%, with balanced precision, recall, and F1-scores across benign and malignant categories, demonstrating consistent performance and reduced bias. The strong performance and explainability of the ensemble approach emphasize its potential as a decision-support tool in real-world diagnostic settings.

Future work will focus on expanding the dataset diversity, incorporating multimodal imaging inputs, and exploring lightweight deployment strategies for integration into clinical and edge-based diagnostic systems. Overall, this research establishes ensemble deep learning as a promising pathway toward accurate, interpretable, and scalable breast cancer detection solutions, supporting early diagnosis and improved patient outcomes.

REFERENCES

[1] J. Shi, A. Vakanski, M. Xian, J. Ding, and C. Ning, "EMT-NET: Efficient Multitask Network for Computer-Aided Diagnosis of Breast Cancer," *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*, pp. 1–5, Mar. 2022, doi: 10.1109/isbi52829.2022.9761438.

[2] C. Kaur and R. Madaan, "Breast Cancer Prediction from Risk Factors Using Ensemble Technique," *IEEE*, pp. 353–358, Nov. 2024, doi: 10.1109/icaiccit64383.2024.10912245.

[3] L. K. S. Potti and S. Maruthuperumal, "Breast Cancer Cell Detection using FCM and Prediction using UNET based Deep Convolutional Neural Network," *IEEE*, pp. 1–6, Oct. 2024, doi: 10.1109/gcat62922.2024.10924106.

[4] S. Gengtian, B. Bing, and Z. Guoyou, "EfficientNET-Based Deep learning approach for breast cancer detection with mammography images," *2022 7th International Conference on Computer and Communication Systems (ICCCS)*, pp. 972–977, Apr. 2023,doi: 10.1109/icccs57501.2023.10151156.

[5] Krithiga, "Improved Deep CNN Architecture based Breast Cancer Detection for Accurate Diagnosis," *IEEE*, pp. 200–205, Aug. 2023, doi: 10.1109/icaiss58487.2023.10250616.

[6] M. Renukadevi and S. Gomathi, "An Automated and Smart Breast Cancer Detection and Classification Framework using DenseNet based Deep Learning Approach," *2022 6th International Conference on Trends in Electronics and Informatics (ICOEI)*, pp. 816–820, Apr. 2025, doi: 10.1109/icoei65986.2025.11013070.

[7] "M2NET: Two-Stage Multi-Label Breast Cancer Detection Networks," *IEEE Conference Publication | IEEE Xplore*, May 27, 2024. https://ieeexplore.ieee.org/document/10635406

[8] P. V, N. Khatri, H. Sharma, and P. K. Shukla, "Empowering Early Diagnosis: Optimizers Revolutionizing Breast Cancer Detection with DenseNet," *IEEE*, pp. 582–588, Nov. 2023, doi: 10.1109/ictacs59847.2023.10389992.

[9] S. Sharma and Y. Singh, "Enhancing breast cancer detection using a Transformer-Based model," *2021 International Conference on Information Science and Communications Technologies (ICISCT)*, pp. 264–269, Nov. 2024, doi: 10.1109/icisct64202.2024.10957451.

[10] P. Kaushik and S. Choudhary, "Enhanced Breast Cancer Detection using ResNet50V2-based Convolutional Neural Networks," *IEEE*, pp. 374–379, Sep. 2024, doi: 10.1109/icscsa64454.2024.00066.

[11] M. Hasan, “Deep Learning for Breast Cancer Detection: Comparative Analysis of ConvNeXT and EfficientNet,” *IEEE*, pp. 1387–1391, Dec. 2024, doi: 10.1109/iccit64611.2024.11021905.

[12] A. C. Yadav, Z. Alam, and M. Mufeed, “U-Net-Driven Advancements in Breast Cancer Detection and Segmentation,” *IEEE*, pp. 1–6, Aug. 2024, doi: 10.1109/iceect61758.2024.10738914.

[13] N. R. Panda, D. Muduli, and S. K. Sharma, “Customized MobileNet with Transfer Learning for Enhanced Early Breast Cancer Detection: A Deep Learning Approach,” *IEEE*, pp. 1–5, Dec. 2024, doi: 10.1109/scopes64467.2024.10991324.

[14] R. Paul, S. Kr. Biswas, A. N. Boruah, A. Kr. Das, S. Reshmi, and B. Purkayastha, “Expert System for Breast Cancer Prediction using Ensemble Learning,” *2022 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COM-IT-CON)*, pp. 113–118, May 2022, doi: 10.1109/com-it-con54601.2022.9850678.

[15] R. Kumar, M. Chaudhry, H. K. Patel, N. Prakash, A. Dogra, and S. Kumar, “An analysis of ensemble Machine learning Algorithms for breast Cancer Detection: Performance and Generalization,” *2022 9th International Conference on Computing for Sustainable Global Development (INDIACom)*, pp. 366–370, Feb. 2024, doi: 10.23919/indiacom61295.2024.10498618.

[16] J. Aditya, “Optimized Ensemble Prediction Model for Breast Cancer,” *IEEE*, pp. 1–4, Nov. 2021, doi: 10.1109/itss-ioe53029.2021.9615269.

[17] M. K. Elbashir, M. Ezz, M. Mohammed, and S. S. Saloum, “Lightweight convolutional neural network for breast cancer classification using RNA-SEQ gene expression data,” *IEEE Access*, vol. 7, pp. 185338–185348, Jan. 2019, doi: 10.1109/access.2019.2960722.

[18] G. Kumari, U. V. Ramesh, P. P. Ramamohan, K. S. Pavani, and A. Lakshmanarao, “Cancer Detection with Ensemble Learning Model from Novel Precedence based Algorithms,” *IEEE*, Sep. 2023, doi: 10.1109/ic3i59117.2023.10397629.

[19] P. Kumar, S. K. Choudhary, M. Deepika, P. Vamshikrishna, and D. Karthik, “CascadeNet: A Multi-Stage Deep Learning Framework for Breast Cancer Detection in Ultrasound Imaging,” *IEEE*, pp. 177–182, Jun. 2025, doi: 10.1109/icicv64824.2025.11086032.

[20] Z. Hameed, S. Zahia, B. Garcia-Zapirain, J. J. Aguirre, and A. M. Vanegas, “Breast cancer histopathology image classification using an ensemble of deep learning models,” *Sensors*, vol. 20, no. 16, p. 4373, Aug. 2020, doi: 10.3390/s20164373