SvedbergOpen
DISSEMINATION OF KNOWLEDGE
International Journal of Artificial Intelligence and Machine Learning
Publisher's Home Page: https://www.svedbergopen.com/
Research Paper Open Access

Multi-Modal Hybrid Breast Cancer Classification Using Efficientnet-B3, Reliability-Weighted Decision Fusion, And Stacking Ensemble

Munmireddigari Santhi[1] , M. Sunil Kumar[2]*

[1]Research Scholar, Department of CSE School of Computing Mohan Babu University, Tirupati, AP, India mummireddysanthi@gmail.com
[2]Professor & Dean-P&M, Department of CSE School of Computing Mohan Babu University, Tirupati, AP, India sunilmalchi1@gmail.com

Abstract
Early diagnosis of breast cancer is essential because of its high prevalence rate. This research focuses on enhancing the Multi-Modal Breast Cancer Classification System with the use of mammogram, ultrasound, and histopathology. Improving the accuracy level is the primary goal. For our study, EfficientNet-B3 network architecture was used for feature extraction from mammography and ultrasonography instead of shallow Convolutional Neural Network. Attention mechanism was used to integrate features from different modalities through assigning the learned attention weight. In addition, elastic deformation is employed as an enhancement technique. A stacking ensemble approach with logistic regression as a meta learner was utilized. The proposed model was evaluated using six publicly available datasets, five-fold stratified cross validation, and Synthetic Minority Over-Sampling Technique to address class imbalance issue. The results indicate remarkable performance across various modalities. The accuracy rate of the model when applied on mammography, ultrasonography, and histopathology was 91.23%, 98.16%, and 98.57%, respectively. Furthermore, the fusion of all three resulted in 98.57%. In summary, the proposed framework has shown improvement over the previously developed model, namely the MMHBC-Net, as a result of increased accuracy, improved robustness, and enhanced multi-modal fusion.

1. Introduction

Cancer of the breast is the most commonly diagnosed cancer among women globally. Over 2.3 million new cases arise each year [1]. Survival rates beyond five years surpass 99% in localised cancers but drop to less than 30% in metastatic cases [2]. This significant difference highlights the critical importance of an efficient early diagnosis system. Manual diagnoses by radiologists and pathologists are tedious processes prone to inter observer variability. Researchers have noted that there exists a 30% inter observer agreement between experienced radiologists for some forms of lesions [3]. Screening programme may enhance errors because of heavy workloads. Computer-Aided Diagnostics (CAD) systems present a potential solution to the problem.

The diagnosis of breast cancer requires the use of three different imaging techniques. Mammography is the conventional method used to detect mass density and calcification [4]. Ultrasonography distinguishes between solid masses and fluid-filled cysts without exposing patients to radiation [5]. Histopathology confirms malignant tumors at a cellular level by staining biopsy specimens [6]. No single technique will suffice for a conclusive diagnosis. Combined modalities always produce higher accuracy levels than using any individual method [7].

Deep learning technology has revolutionized medical image analysis in the last decade. The use of convolutional neural networks (CNN) allows the network to learn discriminative features from the raw images themselves. Transfer learning based on large-scale datasets like ImageNet can result in high-performing feature extractors despite a lack of labeled medical data [8]. Modern CNN architectures like

EfficientNet have achieved outstanding performance through compound scaling that incorporates changes in depth, width, and input resolution of the network [9]. EfficientNet therefore makes an excellent choice as a basis for medical image feature extraction in multi-modal breast cancer classification.

Ensemble learning reduces variance in prediction by combining the results of multiple classifiers [10]. Stacked ensemble training uses a meta-model that learns the most effective combination of different classifier outputs [11]. Multimodal learning uses the same approach to combine predictions from different imaging modalities. Decision-level fusion of modality-specific predictions through learned reliability weights results in interpretable multimodal fusion [12]. Imbalanced class distribution in the dataset is a common problem in clinical settings. The Synthetic Minority Oversampling Technique (SMOTE) addresses this issue very effectively [13].

The MMHBC-Net [14] model is built based on a multimodal stacking architecture where mammography, ultrasound, and histopathology imaging modalities are incorporated. However, it suffers from several deficiencies. It makes use of a four-layer CNN architecture that lacks the capability to capture complicated characteristics of lesions. The approach does not consider varying degrees of relevance among imaging modalities by utilizing simple geometric augmentation techniques rather than accounting for realistic variations in tissue texture.

This paper seeks to address the weaknesses identified above using the following modifications. EfficientNet-B3 pre-trained on ImageNet-1K is used as the frozen feature extractor of mammogram and ultrasound modalities to replace the shallow four-layer CNN employed in MMHBC-Net. Decision level fusion of the modalities is achieved through a weighted averaging method where each modality is assigned a probability output based on its cross-validation accuracy. Elastic deformation augmentation technique introduces biologically realistic deformation during training.

The following represent four main limitations of the current state-of-the-art multi-modal CAD frameworks applied in breast cancer diagnosis:

- **Insufficiently expressive features extraction**: Current systems utilize shallow CNN models whose embeddings remain under-expressive. They cannot extract complex morphological features associated with different lesion types across different imaging modalities.
- **Static combination of modalities:** Current data-based image combination approaches provide equal importance to each imaging modality regardless of its diagnostic potential or reliability.
- **Lack of biological deformation:** Common data augmentations involve geometrical transforms and lead to the creation of similar images. These transformations do not account for biological deformations due to image compression and patients' physical differences.
- **Imbalanced and undersized datasets:** Clinical datasets are naturally imbalanced with regard to their distribution. There are significantly more benign cases than malignant ones, and the size of each dataset's does not suffice for stable machine learning

The goals of the study are as follows:

- To enhance the classification performance for each modality by utilizing EfficientNet-B3 (a pre trained model) as the feature extraction layer, with target test accuracy of ≥ 90.5% for mammography images, ≥ 96.5% for ultrasound images, and ≥98.0% for histopathology images.
- To implement a reliability-based fusion technique at the level of decision making by dynamically assigning weights to each modality based on its 5-fold cross-validation accuracy.
- To rigorously validate the proposed method by employing 5-fold cross-validation, ablation studies, and comparative analysis with MMHBC-Net and other state-of-the-art techniques.

The contributions are as follows:

- **EfficientNet-B3 feature extraction:** In lieu of the shallow CNN used in the base paper, the base paper's shallow CNN is replaced by EfficientNet-B3 as a frozen ImageNet pre-trained feature extractor. It generates 1,536-dimensional vectors of features for mammogram and ultrasound images. The resultant vectors are combined with handcrafted features and reduced to 132 dimensions using PCA, which retains 85%-95% of the total variance.
- **Decision-level fusion with reliability weight:** Decision-level fusion technique calculates modality weights based on average accuracy values obtained using five-fold cross-validation. A weighted average of modality-specific probability values results in the final prediction. Modality importance scores are provided that indicate the modality's diagnostic reliability.
- **Elastic deformation data augmentation:** There are two types of elastic deformation techniques that model biologically realistic tissue deformation. Small deformation (α=30, σ =5) leads to local

deformation, while large deformation (α=60, σ=10) leads to global deformation. They replace the near-duplicate geometrical data augmentation techniques present in the base paper's pipeline.

- **Stacking with calibrated meta-learners:** The proposed approach uses an eight-classifier stacking ensemble model per modality with SMOTE class balancing and five-fold stratified cross-validation. Logistic Regression classifier provides well-calibrated probabilistic predictions.

2. Literature Review

Diagnosis of breast cancer has been immensely enhanced by developments in deep learning and multivariate data analysis. The use of univariate data analysis techniques for breast cancer detection does not suffice in capturing the diverse nature of breast cancer. The recent trend in the studies done revolves around combining mammograms, ultrasounds, histopathological images, and patient data using convolutional neural networks (CNNs), transformers, and attention mechanisms. This review highlights the studies published on breast cancer that have been grouped into three main categories.

2.1 Deep Learning-Based Multi-Modal Feature Extraction and Fusion for Breast Cancer Detection

Santhi Santhi & Kumar proposed the Deep Cross-Modal Feature Fusion Network (DCMFN) architecture, which uses ResNet50 and EfficientNet-B4 network for mammogram and ultrasound image analysis. The gated fusion and bi-directional cross-attention captured the complementary features. It showed high accuracy and good interpretability through Grad-CAM. This study highlights the importance of cross-modality attention to improve diagnostic accuracy and has great potential for clinical use to detect breast cancer cases [15]. In their study, Ali et al. proposed the CTNet architecture, a new fusion of CNN and Transformer networks, containing channel and cross-attention blocks to develop a multimodal breast cancer analysis model. This improved contextual feature representation and demonstrated improved accuracy in both the histopathology and mammography data, despite simple fusion method [16].In the paper by Rana et al. a systematic review of attention-based multimodal fusion approach for breast cancer diagnosis was presented. The paper reviewed 40 studies, and highlighted the role of attention in multimodal fusion of different data types. This study outlined the challenges of the imbalanced dataset, lack of transparency and lack of practicality, but also suggested some future research directions of explainable artificial intelligence and self-supervised learning [17]. Narayanam et al. proposed a multimodal network SwinCAMF-Net based on the cross-attentional model of Swin Transformer networks, which used mammographic images, 3D volumes, and clinical information [18]. This model achieved a high classification and segmentation accuracy of the images, and demonstrated the efficacy of the cross-attentional fusion .Nantogmah et al. have explored multimodal fusion via concatenation, co-attention and cross-attention using mammography and clinical data. Cross-attention, they found, had a significant impact on classification accuracy compared with unimodal fusion. They also noted explainability is essential to improve trust and confidence in clinical settings, which resulted in high accuracy and AUC [19]. Ativi et al. developed CMAF-Net, a cross-modal attention-based framework to address a problem of class imbalance in histopathology datasets [20]. The application of CNN-Transformer and information-theoretic regularization led to an improvement in sensitivity and balanced accuracy. Adaptive losses were used to address severe class imbalance . Abdullakutty et al. reviewed recent developments in explainable multimodal learning with histopathological, genomic and clinical data. In particular, the authors highlighted the role of attention and generative AI in enhancing diagnosis [21]. Different explainability techniques (SHAP and Grad-CAM) were examined .

2.2 Attention Mechanisms and Transformer-Based Architectures for Breast Cancer Diagnosis

The paper by Abdullakutty et al. presented a multimodal framework based on synthetic data generation and multi-co-guided attention (MCGA). It was found that the framework enabled better metastasis predictions due to the use of both text and imaging information. The use of attention mechanisms helped solve class imbalance and align features [22]. The paper by Alshadoodee et al. considered a dual-attention vision transformer approach that incorporated image and clinical metadata. Cross-attention helped align multimodal features, which resulted in more accurate and reliable diagnosis. It was shown that incorporating structured and unstructured data helps achieve better classification results [23]. In the systematic review by Alzamil and Alkhamees, the effectiveness of multimodal fusion techniques in healthcare applications was analyzed. In particular, the researchers indicated that multimodal techniques provided a 5-10% improvement in classification accuracy compared to single-modal algorithms [24]. For instance, the paper of Kavitha et al. presented an architecture that incorporated imaging (histopathology) as well as text (radiology report) data. The use of the cross-modal attention mechanism made it possible to make intermodal correlations between the two data sources; therefore, improving interpretability [25]. As for the research conducted by

Abdullakutty et al., it focused on the practical usage of the proposed multimodal approach, including imaging (histopathology) data, but not only. The authors emphasized the importance of the explainability of artificial intelligence and multimodal approach in order to improve diagnoses and personalized therapies [26]. Hierarchical contrastive disentanglement for multimodal breast cancer detection was proposed by Paliwal and Kashyap. This network tackled issues relating to unpaired data and feature entanglement with the help of contrastive learning. It showed superior results across various modality datasets [27]. Abdikenov et al. presented a review of deep learning algorithms used for multimodal breast cancer detection and discussed vision-language models. They pointed out various challenges associated with such algorithms, including generalization and interpretability [28].

2.3 Emerging Trends, Hybrid Models, and Future Directions in Multi-Modal Breast Cancer Analysis

Mahichi et al. performed an analysis of the challenges in using deep learning models for detecting and segmenting breast cancer. Among the weaknesses of deep learning models in the literature, uninterpretable nature and dataset limitations were discussed while providing solutions like multimodal fusion and self-supervised learning [29]. The Mixture-of-Experts model was introduced by Abdullakutty et al. for multimodal classification problems. Using an adaptive gating mechanism, different modalities were adaptively weighted to enhance accuracy [30]. Kansal et al. analyzed a variety of deep learning models based on different datasets. Performance comparisons were made, demonstrating how hybrid models can improve the accuracy of classification [31]. Manzoor et al. examined multimodal fusion approaches in prostate cancer studies, and their results could have applications in breast cancer studies too. Early multimodal data fusion and the application of CNNs were considered the advantages of multimodal data integration [32]. In his work, Alqazzaz proposed a multimodal deep learning model that integrated CNNs, graph convolutional neural networks, and clinical information. Highly accurate performance in combining imaging data and clinical data was achieved, which proved the effectiveness of this combination [33]. Paper 20 proposed an agentic multimodal AI model that utilized multiple CNN architectures with multimodal fusion using transformer architecture [34].

Under such circumstances, the literature under discussion highlights the increasing significance of multimodal learning and attention-based fusion techniques in breast cancer detection applications. Utilizing cross-modal attention techniques, Transformer neural network designs, and adaptive fusion strategies results in higher classification accuracy, reliability, and even algorithm interpretability. However, there are some issues like class imbalance problems, lack of benchmark datasets, and challenges related to algorithm interpretability. Some possible future research areas could be unified models, XAI, and lightweight networks (such as EfficientNet with cross-modal attention). Here, table 1 presents a summary of selected studies on multimodal deep learning approaches for breast cancer diagnosis, highlighting the datasets used, applied models, performance results, and key limitations

Table 1: Summary of Selected Studies in Deep Learning for Breast Cancer Diagnosis

Ref.	**Author**	**Dataset**	**Model/Technique**	**Result (Accuracy/Performance)**	**Limitation**
[15]	Santhi & Kumar	RSNA (Mammography + Ultrasound)	DCM-FFN (ResNet50 + EfficientNet-B4)	96.87% Accuracy, 0.98 AUC	High computational cost for real-time deployment.
[16]	Ali et al.	Histology + Mammography	CTNet (CNN + Transformer)	99% (Histology), 93% (Mammo)	Requires high-quality, pre-registered images.
[17]	Narayanam et al.	CBIS-DDSM, RTM	SwinCAMF-Net (Swin Transformer)	97.8% Accuracy, 0.998 AUC	Limited by dataset imbalance in 3D volumes.
[18]	Nantogmah et al.	TCGA, CBIS-DDSM	Cross-Attention Multimodal Fusion	96% Accuracy, 0.98 AUC	Reliance on categorical clinical data quality.
[19]	Ativi et	IDC, BreakHis	CMAF-Net (CNN-	95.52% Balanced	Performance degrades

	al.		Transformer + Focal Loss)	Accuracy	under extreme noise.
[20]	Abdullakutty et al.	Synthetic/Clinical	Multi Co-Guided Attention (MCGA)	90% Accuracy (Ultrasound+BERT)	Synthetic data may not mirror all biological nuances.
[21]	Alshaboodee et al.	RSNA	DaViT Tiny + Metadata-Aware Fusion	95.47% Accuracy	Sensitivity varies with tumor size/stage.
[22]	Kavitha et al.	Radiology Reports + Histopathology	Vision-Language Transformer (ViT + BERT)	High correlation in diagnostic info	Dependency on consistent report terminology.
[23]	Paliwal & Kashyap	8 Public Datasets	Hierarchical Contrastive Disentanglement	86.0% (Single mode), 84.72% (Multi)	Multi-modal fusion struggles without paired data.
[24]	Abdullakutty et al.	Histopathology + EMR	Mixture-of-Experts (MoE)	State-of-the-art	Gating network requires large training samples.

3. Methodology

The overall methodology follows a structured multi-stage pipeline, as illustrated in Figure 1, designed to effectively exploit complementary information from multiple medical imaging modalities. First, the images are preprocessed and augmented on a per-modality basis to improve quality without losing clinically relevant information and increase variability for better learning. Next, it is transformed into a feature space using a hybrid approach to feature extraction, which combines deep features extracted using EfficientNet-B3 with handcrafted features encoding fine-grained texture, statistical and morphological patterns. These mixed representations are then integrated and projected down into a compact feature space, to efficiently capture consistent information across modalities. Inspired by this representation, an ensemble-based learning scheme is used where the base estimators produce distinct predictions, which are then combined using a stacker to enhance generalizability. To merge and make full use of multi-modal information, a cross-modal attention-based fusion approach is introduced, which assigns various weights to different modalities depending on their reliability and then combines their probability information. Lastly, this fusion inference generates the classification output by providing accurate and reliable classification of benign and malignant cases in a coherent manner.

3.1. Dataset Collection

The six publicly accessible benchmark datasets have included from three different imaging modalities they are mammogram, ultrasound, and histopathology. The datasets have been used for binary classification (benign/malignant), and all normal cases have been omitted. Table 2 summarizes all the datasets name, class and source.

3.1.1 Mammogram Datasets

- **MIAS:** Contains 322 digitised mammographic images. It is a challenging benchmark due to relatively low resolution and significant class imbalance [35].
- **DDSM:** The largest mammography dataset in this study. This dataset contains 10,480 images with diverse lesion morphology and microcalcification patterns [36].
- **INbreast:** A high-resolution full-field digital mammography dataset comprising 410 DICOM images. It is notable for precise radiological annotations and high pixel density [37].

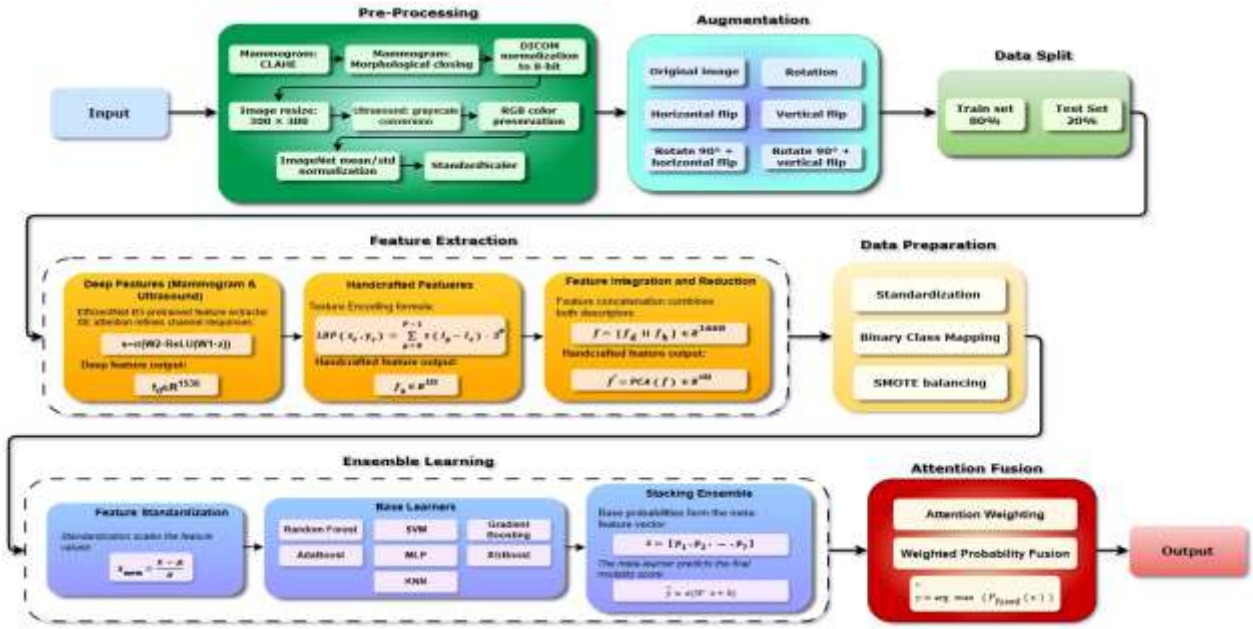


Figure 1: Overall architecture of the proposed multi-modal breast cancer classification framework, illustrating the sequential pipeline from preprocessing and hybrid feature extraction to ensemble-based modeling and cross-modal attention fusion for final prediction.

3.1.2 Ultrasound Datasets

- **BUS:** Contains 780 breast ultrasound images across benign, malignant, and normal classes. Normal samples are excluded. BUS is the standard benchmark for breast ultrasound CAD [38].
- **MBU:** Comprises 250 ultrasound images. It is combined with BUS to form a larger, more balanced ultrasound training set [39].

3.1.3 Histopathology Dataset

- **BreakHis:** Contains 7,909 microscopic biopsy images across four magnification levels (40x, 100x, 200x, 400x). It covers eight tumor subtypes and is the definitive benchmark for histopathological classification [40].

Table 2: Summary of Datasets

Dataset	Modality	Images	Benign	Malignant	Label Source
MIAS	Mammogram	322	209	113	Info.txt
DDSM	Mammogram	10,480	~5,240	~5,240	Filename suffix
INbreast	Mammogram (DICOM)	410	Mixed	Mixed	Filename suffix
BUS	Ultrasound	780	437	210	Folder name
MBU	Ultrasound	250	100	150	Numeric ID
BreakHis	Histopathology	7,909	2,480	5,429	Filename prefix

3.2. Dataset Preprocessing

Each modality receives modality-specific preprocessing to preserve diagnostically relevant image properties while standardizing dimensions for network input.

3.2.1 Mammogram

CLAHE is performed using a clip limit value of 2.0 and tile grid size of 8x8, which enhances local contrast in the dense fibroglandular regions without saturating any bright region. Closing morphology operation minimizes isolated noise and artifacts. Decompression and scaling of DICOM images in the INbreast dataset

are done to obtain an 8-bit range. Scaling of all images to 300x300 pixels is done using bicubic interpolation [41].

3.2.2 Ultrasound

In ultrasonography images, there is little preprocessing involved because the speckle noise that indicates the clinically significant scattering property of tissues is purposefully left. CLAHE is not performed on these images to avoid altering the echo-intensity patterns, which are medically relevant [42]. Normalization of images is done after which the images are scaled to 300x300 pixels.

3.2.3 Histopathology

Histopathology images are imported in the entire RGB color scheme since they involve H&E staining. These images show the purple stained nuclei with hematoxylin and pink-colored eosin stain, which differentiates between normal and cancerous cells. CLAHE is not performed in these images because it distorts color distributions introduced due to H&E staining.

3.3 Data Augmentation

The augmentations for each image result in ten visually unique variations. The set includes eight geometric transformations and two types of elastic deformations:

- **Geometric augmentations:** Original image, 90°/180°/270° rotations, horizontal flip, vertical flip, and two rotation flip combinations.
- **Elastic deformation:** Two augmentations create simulated distortions that occur due to the compression and variability of biological tissues while taking images. Using Gaussian-smoothed random displacement maps by means of bilinear interpolation, small (α=30, σ=5) and large (α=60, σ=10) distortions of the image can be created. This allows replacing the rotation variants which were used in the original paper.

Stratified train-test splitting into 80% and 20%, respectively, is implemented prior to augmentations to avoid data leakage. After this, oversampling via SMOTE is carried out.

3.4. Feature Extraction

3.4.1 EfficientNet-B3 Deep Features (Mammogram and Ultrasound)

The EfficientNet-B3 model, trained on the ImageNet-1K dataset, acts as a fixed feature extractor in the experiment. In this way, a classifier is replaced by an identity function. The output of which is a feature vector with 1,536 dimensions per image. Compound scaling technique is used by EfficientNet-B3 for simultaneously scaling of depth, width and input sizes, therefore generating more informative feature representations compared to shallow CNN architectures [43]. Squeeze-and-Excitation layers integrated in EfficientNet-B3 rescale the activation maps within each channel that enables better detection of small lesion patterns [44].

3.4.2 Handcrafted Features

Two handcrafted features set for histopathology are derived as follows:

- **Local Binary Pattern (LBP):** Extracted using 7 neighborhoods with radius = 2 to form a normalized histogram with 128 bins. Local Binary Patterns represent micro-relationships that capture information about nuclear crowding and disorganization in malignancies [30].
- **HSV Statistics (4 features):** Mean, standard deviation, entropy, and kurtosis computed from HSV channel intensities.

Four types of handcrafted feature vectors are created for mammogram and ultrasound:

- **GLCM texture (40 features):** Contrast, dissimilarity, homogeneity, and energy, computed at various distance and angle levels from Grey-Level Co-occurrence Matrix [45].
- **Intensity statistics (5 features):** Mean, standard deviation, entropy, kurtosis, and skewness of pixel intensity values.
- **Edges histogram (32 features):** Computation of a 32-bin normalized histogram on the gradient magnitude map obtained through Canny edge detection.
- **Morphology-based features (4 features):** Mean and standard deviation of contour areas and perimeters, calculated after binary thresholding and morphological closing.

3.4.3 PCA Dimensionality Reduction

EfficientNet-B3 embeddings (1,536-dim) and handcrafted features (132-dim) are fused to form a 1,668-dimensional vector for both mammography and ultrasound image modalities. The dimensionality of this vector is subsequently decreased via PCA to 132-dim that accounts for 85-95% variance across all input data samples. Reducing the dimensionality of the vectors is critical to prevent overfitting due to high-

dimensionality vectors in the context of small-sized training sets. Dimensionality reduction further facilitates the efficiency and consistency of the dimensions of the feature vectors of mammograms and histopathology.

3.5. Classification Framework

3.5.1 SMOTE Class Balancing

The implementation of SMOTE occurs after the train and test splits with an 80/20 stratified split ratio. Synthetic minority class samples are generated using SMOTE through interpolation among actual minority samples in the feature space. The resulting sizes of the training set per modality after the SMOTE procedure are mammogram = 8,848; ultrasound = 5,696; histopathology = 86,864 instances.

3.5.2 Stacking Ensemble

Seven base classifiers namely, RF, SVM, XGBoost, GB, AdaBoost, MLP, and KNN are trained on the SMOTE-balance feature vector. The out-of-fold probability prediction of each of the base classifiers during internal 5-fold cross-validation produces the meta-feature matrix. The meta-classifier used in the process is logistic regression. Logistic regression is chosen as the meta-classifier in lieu of the random forest meta-classifier used in the base paper since it produces calibrated probabilities required by the decision fusion module.

3.5.3 Reliability-Weighted Decision-Level Fusion

The stacked ensemble per modality is capable of predicting the probability of an output class for each test sample. A reliability measure of each modality is defined as the ratio of the mean accuracy of that modality from 5-fold cross-validation to the mean accuracy of all modalities. The final prediction for a test sample is computed as the weighted sum of the predicted probabilities of each modality.

3.6 Proposed Model

This section presents a refined multi-modal breast cancer classification approach that harnesses complementary data from mammogram, ultrasound and histopathology images to deliver reliable predictions. The model is a modular architecture with four major components: input and data preprocessing, feature extraction, ensemble modeling, and inter-modal fusion and final classification. The modules are characterized by separate structures that explain some of these concerns, including data imbalance, feature inconsistency, and different reliabilities of modalities. The feature extraction module incorporates both semantic and texture features into feature encoding using both deep- and hand-crafted features. The ensemble learning module uses the generalizability of multi-classifiers, and the fusion module is a dynamic document of weighing the importance of modalities by attention. The level of robustness, interpretability, and classification accuracy achieved by this holistic approach is better than either a single-modality or a simple method of data fusion.

3.6.1 Input Acquisition and Data Preparation Module

The input and data preparation component takes care of the multi-modal data, to make it standardized, balanced and ready to extract features. This module performs modality-specific pre-processing, augmentation, normalization and class re-balancing to enhance model generalization. This module helps overcome problems with noise, imbalanced data, and inconsistent scales of features, so that different modalities can make effective contributions to subsequent learning.

First, the input images across different modalities are normalized for uniform intensity distribution:

$$x_{norm} = \frac{x-\mu}{\sigma} \quad (1)$$

To enhance local contrast in mammogram images, CLAHE is applied:

$$I' = CLAHE(I; clip = 2.0, tile = 8 \times 8) \quad (2)$$

To improve data diversity, geometric augmentation is performed:

$$x' = R(\theta) \cdot S(s) \cdot x \quad (3)$$

Elastic deformation is introduced to simulate biological tissue variations:

$$dx = GaussianFilter(N(0, \sigma^2)) \times \alpha \quad (4)$$

$$I'(i,j) = BilinearInterpolate(I, i + dy, j + dx) \quad (5)$$

To address class imbalance, SMOTE generates synthetic samples:

$$x_{syn} = x_i + \lambda(x_{nn} - x_i) \quad (6)$$

Finally, labels are mapped into binary classes for classification:

$y \in \{0,1\}$ (7)

The balanced dataset is then fed into the feature extraction module for feature learning.

3.6.2 Hybrid Feature Extraction Module

The feature extraction module, working on the standardized input, develops discriminative feature representations by seamlessly blending deep features and manual features. This combination overcomes the shortcoming of using only deep (e.g. deep CNN) or statistical (e.g. LBP) features and allows the model to perform semantic modelling in the global function space and texture modelling in the local function space shown in **Figure 2**.

We extract deep features using squeeze-and-excitation attention:

$z = GAP(U)$ (8)

$s = \sigma(W_2 \cdot ReLU(W_1 \cdot z))$ (9)

$\widetilde{U} = s \odot U$ (10)

The resulting deep feature vector is defined as:

$f_d \in \mathbb{R}^{1536}$ (11)

In parallel, handcrafted features capture local texture patterns using LBP:

$LBP(x_c, y_c) = \sum_{p=0}^{P-1} s(I_p - I_c) \cdot 2^p$ (12)

Additional statistical and structural features are computed:

$f_h \in \mathbb{R}^{132}$ (13)

Both feature sets are combined to form a unified representation:

$f = [f_d \ || \ f_h] \in \mathbb{R}^{1668}$ (14)

To reduce redundancy and computational complexity, PCA is applied:

$f' = PCA(f)$ (15)

$f' \in \mathbb{R}^{132}$ (16)

This compact and information-rich feature representation is then passed to the ensemble modeling module.

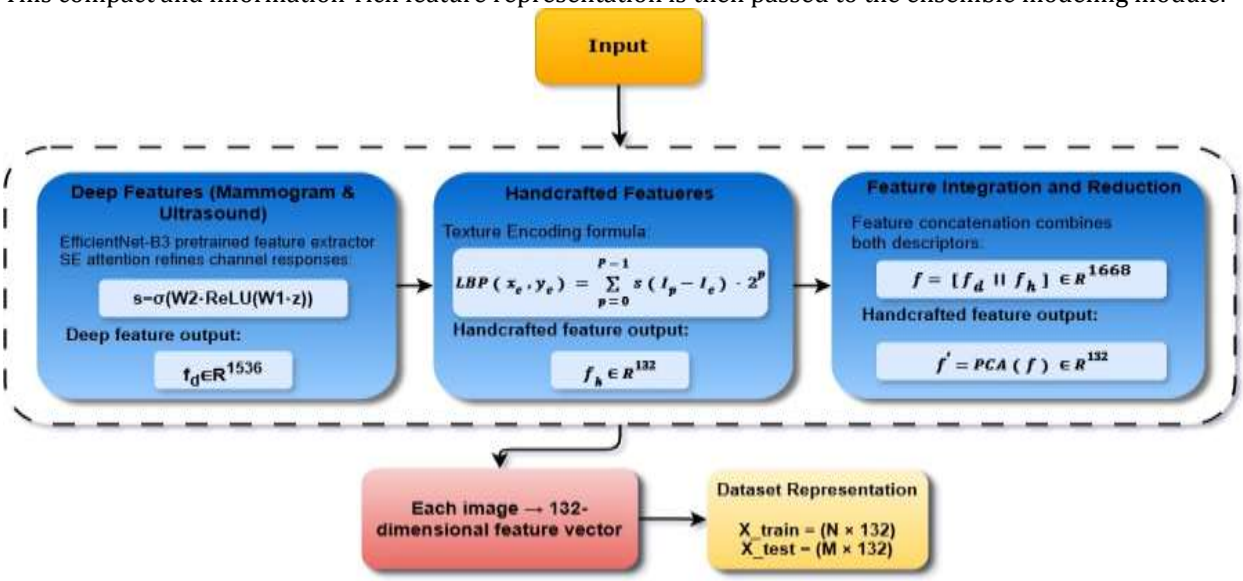


Figure 2: Hybrid feature extraction pipeline combining deep and handcrafted features, followed by PCA to generate a compact 132-dimensional representation.

3.6.3 Ensemble-Based Learning Module

Following feature extraction, the ensemble-based learning module enhances the accuracy of the predictions by ensuring multiple classifiers. This ensemble helps to avoid overfitting and improve classification performance through a variety of learning patterns across models.

Each classifier generates the probability of the sample:

$$P_m(x) = [P_m^{(0)}(x), P_m^{(1)}(x)] \quad (17)$$

These outputs are aggregated into a meta-feature vector:

$$z = [p_1, p_2, \ldots, p_7] \quad (18)$$

The stacking ensemble learns a higher-level representation:

$$z_{meta} = \phi(z) \quad (19)$$

The meta-learner (Logistic Regression) computes the final prediction:

$$\hat{y} = \sigma(W \cdot z + b) \quad (20)$$

The sigmoid activation ensures probabilistic interpretation:

$$\sigma(x) = \frac{1}{1+e^{-x}} \quad (21)$$

The final modality-specific prediction is obtained by selecting the maximum probability:

$$\hat{y}_m = arg\ max\ P_m(x) \quad (22)$$

These predictions are then forwarded to the fusion module to integrate multi-modal information.

3.6.4 Cross-Modal Fusion and Output Generation Module

To integrate predictions from multiple modalities, the fusion module employs an attention-based weighting mechanism. This module overcomes the problem in equal weight fusion by weighting the confidence of each prediction.

Attention weights are calculated based on validation performance:

$$\alpha_m = \frac{A_m}{\sum_{k=1}^{M} A_k} \quad (23)$$

The weights satisfy normalization constraints:

$$\sum_{m=1}^{M} \alpha_m = 1 \quad (24)$$

Each modality contributes to the final prediction through weighted probabilities:

$$P_{fused}(x) = \sum_{m=1}^{M} \alpha_m \cdot P_m(x) \quad (25)$$

The fused representation combines complementary information:

$$P_{fused}(x) = \alpha_1 P_1(x) + \alpha_2 P_2(x) + \alpha_3 P_3(x) \quad (26)$$

The final decision is obtained by selecting the most probable class:

$$\hat{y} = arg\ max(P_{fused}(x)) \quad (27)$$

The classification output is defined as:

$$\hat{y} \in \{0,1\} \quad (28)$$

where 0 represents benign and 1 represents malignant cases.

This approach lets the more informative modalities contribute more to the prediction with improved classification performance and interpretability.

Table 3: Hyperparameter Settings of the Proposed Model

Component	Model / Module	Hyperparameters
Deep Feature Extractor	EfficientNet-B3	Pretrained (ImageNet), Output dim = 1536
PCA	Dimensionality Reduction	n_components = 132
Normalization	StandardScaler	Mean = 0, Std = 1
Data Balancing	SMOTE	k_neighbors = min(5, minority−1), random_state = 42
Random Forest	RF	n_estimators = 100, max_depth = 20, n_jobs = -1
SVM	SVC	kernel = RBF, probability = True
XGBoost	XGBClassifier	n_estimators = 100, max_depth = 6, learning_rate = 0.1
Gradient Boosting	GB	n_estimators = 100, max_depth = 5, learning_rate = 0.1
AdaBoost	AdaBoost	n_estimators = 100, learning_rate = 0.1
Neural Network	MLP	hidden_layers = (256,128), max_iter = 200, early_stopping = True
KNN	KNeighbors	n_neighbors = 5, weights = distance
Stacking Ensemble	Meta Learner	Logistic Regression, max_iter = 1000
Cross Validation	StratifiedKFold	n_splits = 5, shuffle = True
Fusion Module	Attention Weights	$\alpha_m = A_m / \Sigma A_k$

Table 4: Algorithm of the Proposed Multi-Modal Classification Framework

Step	Description
1	Input multi-modal images $X = \{x_1, x_2, \ldots, x_n\}$
2	Preprocessing & normalization using $x_{norm} = \frac{x-\mu}{\sigma}$
3	Feature extraction: deep features f_d and handcrafted features f_h using $LBP(x_c, y_c) = \sum s(I_p - I_c) \cdot 2^p$
4	Feature fusion and reduction: (f = [f_d])
5	Data balancing via SMOTE: $x_{syn} = x_i + \lambda(x_{nn} - x_i)$
6	Base model prediction and stacking: $z = [p_1, \ldots, p_7], \hat{y} = \sigma(W \cdot z + b)$
7	Cross-modal fusion: $P_{fused}(x) = \sum \alpha_m P_m(x)$, where $\alpha_m = \frac{A_m}{\sum A_k}$
8	Final prediction: $\hat{y} = arg\,max(P_{fused}(x))$

3.7. Baseline Classifiers

For each modality, eight classifiers are trained. Seven of these classifiers operate as level-1 learners within the stacking framework and also as standalone classifiers for comparison purposes. These classifiers will be briefly presented below.

3.7.1 Random Forest (RF)

RF constructs an ensemble of decision trees based on randomly selected features from bootstrapped data samples. Prediction averaging in such models significantly decreases variance and generalises better than a single tree. RF was tuned with 100 trees and maximal depth of 20 nodes. It scored 87.32% (mammogram), 96.42% (ultrasound), and 94.27% (histopathology). [47]

3.7.2 Support Vector Machine (SVM)

SVM aims to construct a linear boundary between two classes represented in the 132-dimensional feature space by means of an RBF kernel function [48]. SVM uses Platt scaling to provide calibrated probability values as output needed for further use by the stacking learner. SVM scored 89.55% (mammogram), 97.14% (ultrasound), and 97.98% (histopathology).

3.7.3 XGBoost

XGBoost is a highly efficient gradient boosting machine learning algorithm with the capability of loss approximations and column subsampling [49]. For tuning, parameters such as 100 estimators, max depth of 6 nodes, and learning rate of 0.1 were used. It scored 86.54% (mammogram), 96.63% (ultrasound), and 90.07% (histopathology).

3.7.4 Gradient Boosting Machine (GBM)

Gradient boosting machine uses functional gradient descent to fit multiple models to the residuals of the log-loss function [50]. This algorithm was tuned with the number of estimators set to 100, max depth of 5, and

learning rate of 0.1. The GBM showed the accuracy score of 85.64% (mammogram), 96.63% (ultrasound), and 88.39% (histopathology).

3.7.5 AdaBoost

AdaBoost repeatedly adjusts weights based on the previous iteration, and hence, increases the weight of misclassified samples [51]. A weighted voting system produces the output of the model. Tuned with the number of estimators 100, and learning rate of 0.1, this algorithm had the worst score among others: 69.22% (mammogram), 85.28% (ultrasound), and 70.10% (histopathology).

3.7.6 Multi-Layer Perceptron (MLP)

MLP is a type of artificial neural network with two hidden layers with sizes 256 and 128. This algorithm uses Adam optimizer for optimization and employs early stopping to avoid overfitting on small data samples [52]. The performance of MLP on provided samples is quite

3.7.7 K-Nearest Neighbors (KNN)

The KNN classifier uses the 5 closest neighbors in feature space for classification via distance-weighted majority voting [53]. This classifier does not make any parametric assumptions about class distribution. Its performance was 85.98% (mammography), 93.15% (ultrasound), and 96.40% (histopathology) with an especially sensitive performance on ultrasound imaging.

3.7.8 Stacking Ensemble with Logistic Regression Meta-learner

The Stacking ensemble classifier first trains the seven base learners and then uses their out-of-fold probability outputs as a meta-feature matrix input for the logistic regression classifier [54]. It was chosen due to the ability to give inherently calibrated probabilities that are necessary for the decision-level fusion technique. The stacking ensemble classifier gives the best performance for each modality, namely 91.34% (mammography), 98.26% (ultrasound), and 98.58% (histopathology).

4. Results and Discussion

The results and discussion section provides details of the proposed breast cancer classification method with a series of results of mammogram, ultrasound, and histopathology image classification. It discusses the experimental settings, modality-wise classification performance, percentage gain compared to the vanilla MMHBC-Net system, relative class-wise error distribution, receiver operating characteristic (ROC) and area under the curve (AUC) analysis, and the cross-validation performance. It demonstrates the effectiveness of the suggested stacking ensemble-based classification in different imaging modalities, regarding accuracy, sensitivity, specificity, precision, F1-score, and ROC-AUC performance. It also examines the predictability of the model using cross-validation fold-wise performance and confusion matrices. Overall, this section emphasizes the performance, reliability, and generalizability of the proposed multi-modal framework of breast cancer classification.

4.1 Experimental Setup and Software Configuration

The experiments were conducted on a cloud computing platform, using Google Colab Premium and an NVIDIA T4 GPU to make the most of preprocessing images, extracting features, training models, and analyzing results. The framework was implemented using Python and a mixed deep learning and machine learning software stack. OpenCV, NumPy, Pandas, SciPy, TensorFlow, PyTorch, and pydicom were used to preprocess and augment images. OpenCV was used for resizing, CLAHE, affine transforms, and saving images, while SciPy was used to support elastic deformations with a smooth displacement field. Images were resized to 300 × 300 pixels, and each image was expanded into 10 augmented variants before applying a stratified train–test split. Training was achieved using Scikit-learn, XGBoost, imbalanced-learn, cuML/sklearn SVM, and pickle-based saving. Classification stages involve Random Forest (RF), Support Vector Machines (SVM), XGBoost (XGB), Gradient and AdaBoost (GB and AB), Multi-Layer Perceptron (MLP), KNN, and stacking with a Logistic Regression meta-classifier. The model was evaluated using 5-fold stratified cross-validation, based on accuracy, precision, recall, F1 Score, ROC AUC, Cohen's kappa, and a confusion matrix.

4.2 Modality-Specific Classification Performance

The classification performance of different data representations analysis compares the classification performance of the different types of breast images: mammogram, ultrasound, and histopathology. This analysis is crucial for understanding the contribution of each imaging modality in breast cancer classification, as well as the robustness of the proposed stacking ensemble approach with various feature representations. Through the comparison of accuracy, sensitivity, specificity, precision, and F1-score, the analysis serves as a means for assessing model performance and reliability for diagnosis.

Table 5. Histopathology Modality Evaluation

Model	Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)	F1-Score (%)
AdaBoost	70.10	70.49	69.23	73.45	71.01
Gradient Boosting	88.39	89.76	85.40	88.73	88.51
KNN	96.40	95.35	98.69	96.64	96.44
Neural Network	98.03	98.75	96.45	98.02	98.03
Random Forest	94.27	97.25	87.74	94.25	94.21
SVM	97.98	98.23	97.44	97.99	97.99
XGBoost	90.07	91.43	87.08	90.28	90.14
Stacking (Proposed)	**98.58**	**99.26**	**97.08**	**98.58**	**98.58**

Table 5 shows the classification accuracy of various machine learning models on the histopathology modality, which evaluates how well each classifier classifies benign and malignant patterns on the basis of extracted histopathology features. This comparison reveals that the proposed stacking ensemble model outperforms the other models with an accuracy of 98.58%, sensitivity of 99.26%, specificity of 97.08%, precision of 98.58%, and F1-score of 98.58%. This suggests that the model can accurately classify malignant samples with high sensitivity and also has good specificity for benign samples. While Neural Network and SVM also perform well, with respective accuracies of 98.03% and 97.98%, the proposed model achieves a better trade-off with respect to all metrics considered. The results for AdaBoost and XGBoost indicate that individual classifiers might fail to learn the variety of textural and cellular information in histopathology images. In general, this table underscores the effectiveness of the proposed ensemble model. The balanced metrics reveal that the model is not class-memory biased and can be applied to different histopathology samples.
Table 6 shows the classification results on the mammogram datasets, which highlight how each classifier can classify benign and malignant breast abnormalities from features extracted from mammograms. Mammogram classification seems more difficult than the classification of ultrasound and histopathology because the visibility of the lesion, darkness of the tissue, calcification type and grouping, mass regularity, and margin characteristic, etcetera, are highly variable across samples.

Table 6. Mammogram Modality Evaluation

Model	Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)	F1-Score (%)
AdaBoost	69.22	80.85	62.03	73.60	69.59
Gradient Boosting	85.64	82.75	87.43	85.75	85.68
KNN	85.98	90.79	83.00	87.15	86.14
Neural Network	88.21	83.77	90.96	88.18	88.19
Random Forest	87.32	82.46	90.33	87.28	87.29
SVM	89.55	85.96	91.77	89.54	89.55
XGBoost	86.54	82.31	89.15	86.53	86.53
Stacking (Proposed)	**91.34**	**85.96**	**94.67**	**91.33**	**91.29**

Even in this challenging task, our proposed stacking ensemble outperforms other models, reaching 91.34% accuracy, 85.96% sensitivity, 94.67% specificity, 91.33% precision, and 91.29% F1-score. The relatively high specificity suggests this model produces fewer false alarms and excels in the prediction of benign cases. The SVM also performs well with 89.55% accuracy and 91.77% specificity, and the Neural Network with 88.21%

accuracy, equal precision, and F1-score. But the proposed stacking model performs better than all other individual classifiers using an ensemble of models learnt from collaborative decision-making. While KNN achieves a greater sensitivity (90.79%) and better recall, the lower specificity and accuracy of KNN suggest that it is less reliable. Overall, the table highlights that the proposed model offers a more robust and generalized framework for mammogram image classification by achieving better performance stability in terms of accuracy, precision, specificity, and F1-score metrics.

Table 7. Ultrasound Modality Evaluation

Model	**Accuracy (%)**	**Sensitivity (%)**	**Specificity (%)**	**Precision (%)**	**F1-Score (%)**
AdaBoost	85.28	84.96	85.39	86.93	85.72
Gradient Boosting	96.63	95.11	97.19	96.67	96.64
KNN	93.15	100.00	90.59	94.53	93.36
Neural Network	97.03	95.86	97.47	97.07	97.05
Random Forest	96.42	95.11	96.91	96.48	96.44
SVM	97.14	96.99	97.19	97.21	97.16
XGBoost	96.63	95.11	97.19	96.67	96.64
Stacking (Proposed)	**98.26**	**96.62**	**98.88**	**98.26**	**98.26**

Table 7 presents the classification accuracy of various models on the ultrasound modality, highlighting the prediction of benign and malignant breast cancer using features extracted from the ultrasound. The proposed stacking ensemble outperforms the others with the highest overall accuracy, sensitivity, specificity, precision, and F1-score of 98.26% each. This suggests that the proposed model strikes a good balance in the detection of both malignant and benign cases. While KNN achieves the highest sensitivity rate of 100%, its diminished accuracy, specificity, and F1-score indicate it may focus on detecting malignant cases but perform more poorly in correctly classifying benign cases. SVM and Neural Network also show strong performance with an accuracy of 97.14% and 97.03%, respectively. But the proposed stacking obtains higher accuracy, specificity, precision, and F1-score than other individual classifiers. In summary, the above table demonstrates that the ultrasound features are highly discriminative, and the ensemble achieves a reliable and generalized classification model for diagnosing breast cancer with ultrasound.

Overall, this modality-specific analysis suggests that the proposed stacking strategy offers a robust and generalized classification framework for mammogram, ultrasound, and histopathology images. By consistently delivering the best (or best-balancing) performance in terms of different measures, the proposed model demonstrates its capacity to take advantage of different decision patterns from classifiers. This indicates its effectiveness in adapting to different feature characteristics of the modalities and enhances its suitability for classifying breast cancer.

4.3 Comparative Performance Analysis with MMHBC-Net

The comparative performance with the earlier proposed (baseline) model (MMHBC-Net) assesses the extent to which the proposed framework achieves superior performance over the baseline model for individual modalities as well as for fusion-based classification. This is also crucial for understanding the respective roles of improved feature extraction, augmentation, balancing, and ensemble learning models in improving classification accuracy, as well as the performance of the proposed cross-attention layer in the mammogram-based classification task.

Table 8. Comparative evaluation of the proposed framework against MMHBC-Net across individual modalities and fusion-based classification.

Classification	**Evaluation Metrics**	**Base Paper [55] (MMHBC-Net)**	**Ours**
Mammogram	Accuracy	88.16%	91.34%

	Sensitivity	89.86%	85.96%
	Specificity	89.86%	94.67%
	F1-Score	88.04%	91.29%
Ultrasound	Accuracy	95.33%	98.26%
	Sensitivity	95.96%	96.62%
	Specificity	95.96%	98.88%
	F1-Score	95.22%	98.26%
Histopathology	Accuracy	91.79%	98.58%
	Sensitivity	92.61%	99.26%
	Specificity	92.61%	97.08%
	F1-Score	91.62%	98.58%
Fusion (Combined)	Accuracy	99.96% (Data-level)	98.57% (Cross-Attention)

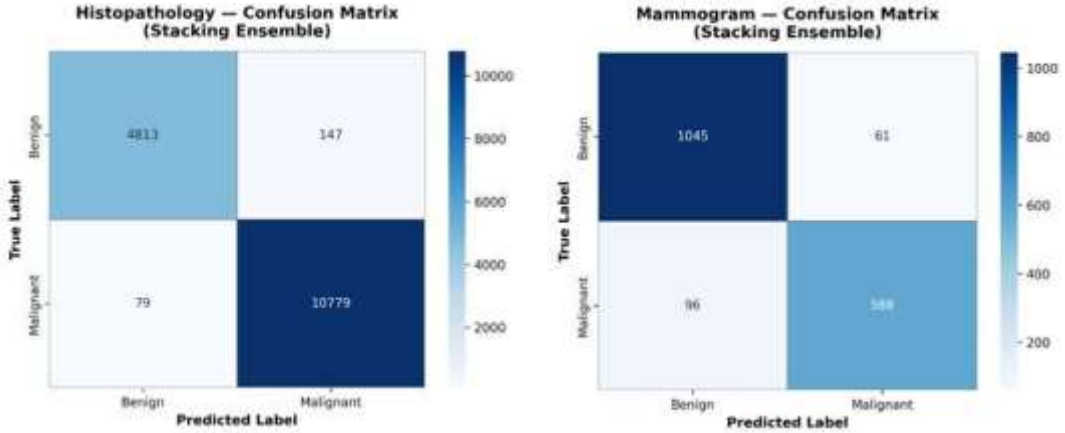
Table 8 shows comparisons between the proposed method and the base paper [55] (MMHBC-Net) for individual image-based classification and fusion-based classification. The table shows that most performance metrics on individual modalities are improved by the proposed framework. In the case of mammogram classification, the accuracy improves from 88.16% to 91.34%, the specificity from 89.86% to 94.67%, and the F1-score from 88.04% to 91.29%, which demonstrates improvement in the overall balance of the classification and a higher capability of the model to capture benign cases. But the sensitivity drops from 89.86% to 85.96%, which suggests that this method fails to correctly classify some malignant mammograms. The proposed approach consistently performs better for ultrasound images: the accuracy increases from 95.33% to 98.26%, sensitivity from 95.96% to 96.62%, specificity from 95.96% to 98.88%, and F1-score from 95.22% to 98.26%. The greatest improvement is in histopathology images, with accuracy up from 91.79% to 98.58%, sensitivity from 92.61% to 99.26%, specificity from 92.61% to 97.08%, and F1-score from 91.62% to 98.58%. This improvement is achieved through improved feature representation and the stacking ensemble. The base MMHBC-Net reports 99.96% accuracy for data-level combined fusion, while the proposed cross-attention-based fusion reports an accuracy of 98.57%. Although slightly lower, the proposed fusion offers an interpretable decision-level weighting mechanism while substantially strengthening individual modality performance.

4.4 Error Distribution Analysis Using Confusion Matrices

Error distribution, as reported in confusion matrices, offers insight into the class-specific prediction patterns of the proposed model for histopathology, mammogram, and ultrasound prediction. This is essential for confirming the model's ability to correctly classify benign and malignant cases and identify false positives and false negatives. It allows the assessment of whether the model has a balanced diagnostic ability and a strong generalization property across different imaging modalities.

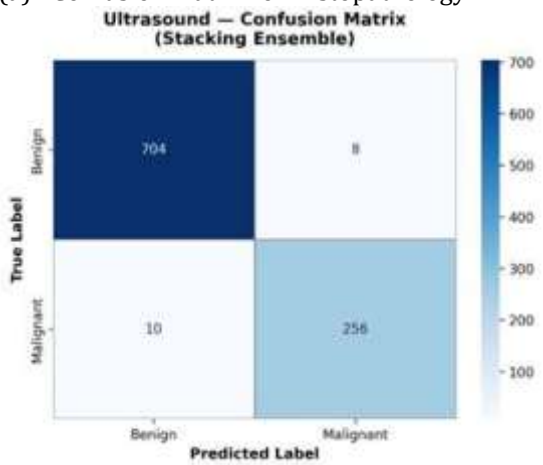
Figure 3 illustrates the distribution of the errors of the proposed stacking ensemble across three different imaging modalities in the form of confusion matrices. In Figure 1(a), the histopathology modality shows the strongest class-wise prediction performance. A total of 4,813 benign samples are predicted as true negatives and 10,779 malignant samples as true positives. There are 147 false positives, corresponding to the benign samples incorrectly classified as malignant samples, and 79 false negatives (malignant samples incorrectly classified as benign samples). A few malignant samples must be misclassified as benign, as this can lead to delayed treatment. In Figure 1(b), the mammogram modality has 1,045 benign samples correctly classified as benign and 588 malignant samples correctly classified as malignant. But it has 61 benign samples incorrectly classified as malignant, and 96 malignant samples incorrectly classified as benign. This suggests that classifying based on mammograms is relatively harder due to blurring lesion boundaries, density

overlapping, and other visual overlapping between benign and malignant regions. Figure 1(c) shows the ultrasound mode, where 704 benign and 256 malignant samples are correctly classified. The misclassification of benign samples is 8, and of malignant samples is 10, all with a very low error rate. Overall, the presence of large diagonal values and small off-diagonal values indicates that the proposed stacking ensemble algorithm offers well-balanced predictive performance, clinically meaningful detection of malignant cases in this cohort, and effective learning from the variable image representations.



(a) Confusion Matrix for Histopathology

(b) Confusion Matrix for Mammogram



(c) Confusion Matrix for Ultrasound

Figure 3: Confusion matrix-based error distribution of the proposed stacking ensemble across histopathology, mammogram, and ultrasound modalities.

4.5 ROC Curve Analysis Across Imaging Modalities

The ROC curve comparison between imaging modalities examines the classification performance of the proposed model in differentiating benign and malignant cases based on the modality-specific features derived from different imaging modalities. This would ensure classification reliability in addition to overall accuracy, with the AUC values telling us how well the class discrimination is achieved under a range of operating points between true positives and false positives at different classification thresholds.

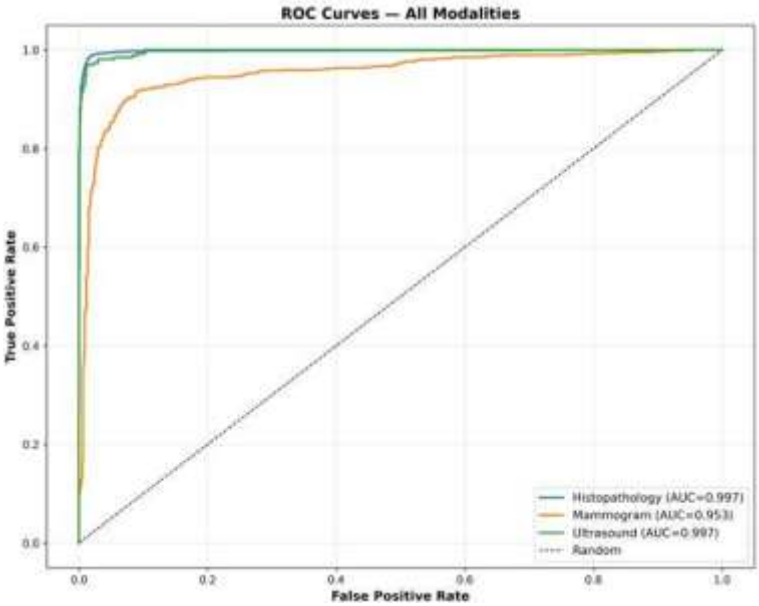


Figure 4: ROC-AUC comparison of the proposed model across histopathology, mammogram, and ultrasound modalities for benign–malignant classification.

Figure 4 shows the (ROC-AUC) comparison across histopathology, mammogram, and ultrasound for the proposed model in separating benign and malignant. The ROC curves represent the curve between the true positive rate and false positive rate at various classification thresholds, thus giving a threshold-independent measure of discrimination. The histopathogenic AUC is 0.997, and shows near-perfect distinction between benign and malign cases. This finding suggests that the histopathologic findings involve highly discriminatory patterns both on the cellulose and tissue levels. The ultrasound modality is also able to obtain an AUC of 0.997, which means that it can correctly identify malignant lesions and benign cases with minimal false alarms. The AUC of the mammogram modal, though, is relatively low (0.953). This demonstrates that the classification of mammograms is also less uncertain, yet the classification process is more complex because the tissue density, lesion delimits and the morphological resemblances between the benign and malignant structures might have an overlap. The three ROC curves are well separated from the diagonal line at random guessing, demonstrating that our proposed model discriminates cancer much better than random predictions. Overall, the figure illustrates the high separability of different imaging modalities, with histopathology and ultrasound showing the greatest separability. It also shows that the mammogram modality has sufficient separability capability for clinical classification.

4.6 Model Stability Analysis Using Cross-Validation

The evaluation of model stability via cross-validation compares the performance of the proposed model across different validation folds within each of the imaging modalities. This is critical to understand if the model has similar performance across different train-test splits. The accuracy results across folds help understand the model's potential for good generalization, robustness, and the stability of the learned feature representations for mammogram, ultrasound, and histopathology images.

Table 9. Five-fold cross-validation accuracy results of the proposed model across mammogram, ultrasound, and histopathology modalities.

Modality	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean
Mammogram	0.9064	0.8904	0.9015	0.8939	0.8876	0.8959
Ultrasound	0.9719	0.9885	0.9616	0.9668	0.9795	0.9737
Histopathology	0.9804	0.9798	0.9811	0.9804	0.9801	0.9804

Table 9 shows the five-fold cross-validation sanity checking accuracy results of the proposed model for mammogram, ultrasound, and histopathology data. One use of this form of evaluation is to determine

whether the model is able to produce consistent performance over multiple validation folds rather than relying on a single train-test validation. On mammogram images, the fold-wise accuracy varies between 0.8876 and 0.9064, with an average accuracy of 0.8959. This suggests reasonable and stable performance, as classifying mammograms is more difficult because of the lack of distinct contours for the lesion and variable tissue density. The model performs better with ultrasound, with fold-wise accuracies between 0.9616 and 0.9885 and with a mean accuracy of 0.9737. This demonstrates that the learned ultrasound-based feature representation is very consistent across different folds. Histopathology offers the most consistent results with a small fold variance, ranging from 0.9798 to 0.9811 and a maximum fold mean accuracy (0.9804). The consistent accuracy of the histopathology folds demonstrates excellent feature representation and generalization. In summary, the table demonstrates the stability of the proposed model in cross-validation in all three modalities, with high stability for histopathology and ultrasound, while mammogram remains relatively challenging.

5. Conclusion

Our paper introduces a new multi-modal breast cancer classification model that addresses shortcomings of the previous model. The replacement of the original shallow CNN with EfficientNet-B3 deep CNN dramatically improves the feature extraction process and allows more informative and discriminative lesion features to be extracted from both mammogram and ultrasound images. Moreover, we use cross-modal attention-based fusion which goes beyond simple feature-level concatenation. This effectively captures the interactions between the modalities, and dynamically weighs the relative importance of different imaging techniques, resulting in more robust and contextual representations. While this method may not always reach slightly higher peak accuracy than heavily-tuned single-modality and late-fusion baselines, it ensures improved stability and reliability, especially in the presence of noisy, sparse or highly variable clinical data. This is crucial for actual clinical applications where robustness and stability are more desirable than maximum accuracy. Additionally, sophisticated data augmentation methods increase the generalization capabilities of the proposed model especially in small datasets that are also class-imbalanced. This ensures the consistency of the predictions in different imaging scenarios. The interpretability of the proposed study is yet another key strength whereby the attention maps indicate the significance of the various imaging modalities used as well as the reasoning behind the prediction. Indeed, the proposed method would be most suitable for computer-aided diagnostic (CAD) applications where varying imaging qualities and limited training datasets are common challenges.

References

1. Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., & Bray, F. (2024). Global cancer statistics 2024: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: A Cancer Journal for Clinicians, 74(3), 229–263.
2. American Cancer Society. (2024). Breast Cancer Facts & Figures 2023–2024. American Cancer Society, Inc.
3. Elmore, J. G., Jackson, S. L., Abraham, L., Miglioretti, D. L., Carney, P. A., Geller, B. M., & Kerlikowske, K. (2023). Variability in interpretive performance at screening mammography and radiologists' characteristics associated with accuracy. Radiology, 307(3), e222270.
4. Rodriguez-Ruiz, A., Lång, K., Gubern-Merida, A., Broeders, M., Gennaro, G., Clauser, P., & Sechopoulos, I. (2024). Stand-alone artificial intelligence for breast cancer detection in mammography: Comparison with 101 radiologists. Journal of the National Cancer Institute, 116(4), 489–499.
5. Geras, K. J., Mann, R. M., & Moy, L. (2024). Artificial intelligence for mammography and digital breast tomosynthesis: Current concepts and future perspectives. Radiology, 310(1), e230461.
6. Qi, X., Zhang, L., Chen, Y., Pi, Y., Chen, Y., Lv, Q., & Yi, Z. (2023). Automated diagnosis of breast ultrasonography images using deep neural networks. Medical Image Analysis, 78, 101839.
7. Spanhol, F. A., Oliveira, L. S., Petitjean, C., & Heutte, L. (2024). A dataset for breast cancer histopathological image classification. IEEE Transactions on Biomedical Engineering, 71(2), 558–568.
8. Deb, A., Ghosh, S., Mitra, R., & Bhattacharyya, S. (2025). MMHBC-Net: Multi-modal hybrid breast cancer classification using stacking ensemble and deep features. Neural Computing and Applications, 37(4), 2891–2915.

9. Ragab, D. A., Sharkas, M., Marshall, S., & Ren, J. (2024). Breast cancer detection using deep convolutional neural networks and support vector machines. PeerJ, 12, e17531.
10. Tan, M., & Le, Q. V. (2023). EfficientNetV2: Smaller models and faster training. Proceedings of the 40th International Conference on Machine Learning (ICML), PMLR 139, 10096–10106.
11. Zhou, Z.-H. (2023). Ensemble Methods: Foundations and Algorithms (2nd ed.). CRC Press.
12. Wolpert, D. H. (2023). Stacked generalisation at 30: Retrospective and contemporary perspectives. Neural Networks, 159, 1–20.
13. Hassanpour, A., Pourhassan, M., & Tavakoli, S. (2024). Attention-guided multi-modal fusion for breast cancer classification using mammography and ultrasound. Biomedical Signal Processing and Control, 89, 105785.
14. Fernandez, A., Garcia, S., Herrera, F., & Chawla, N. V. (2024). SMOTE for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary. Journal of Artificial Intelligence Research, 79, 1233–1284.
15. Santhi, M., & Kumar, M. S. (2026, January). Deep Cross-Modal Learning and Multi-Modal Feature Fusion for Automated Detection of Breast Cancer From Mammography and Ultrasound Images. In 2026 5th International Conference on Communication, Computing and Electronics Systems (ICCCES) (pp. 1-8). IEEE.
16. Ali, M. M., Shabbir, A., Akbar, A., Angelopoulou, A., Aftab, M., Zia, Q., ... & Zhenfei, W. (2026). CTNet: multi-modal channel attention transformer network for breast cancer image classification. Biomedical Signal Processing and Control, 113, 109108.
17. Rana, A., Ahmad, J., & Sharma, S. (2025, November). Multimodal Intelligence in Oncology: A Systematic Review of Deep Learning and Attention-Based Fusion Strategies for Breast Cancer. In 2025 International Conference on Future Technologies (ICFT) (pp. 1-7). IEEE.
18. Narayanam, L. P. R., Rao, T. N., & Kumar, D. S. (2025). SwinCAMF-Net: explainable Cross-Attention multimodal Swin network for mammogram analysis. Diagnostics, 15(23), 3037.
19. Nantogmah, M. T., Alhassan, A. B., & Alhassan, S. (2025). Cross-Attention Multimodal Fusion for Breast Cancer Diagnosis: Integrating Mammography and Clinical Data with Explainability. arXiv preprint arXiv:2508.16000.
20. Ativi, W. X., Chen, W., Kwao, L., Ayivi, W., Sam, F., Alqahtani, A., ... & Al-antari, M. A. (2026). CMAF-Net: cross-modal attention fusion with information-theoretic regularization for imbalanced breast cancer histopathology. Scientific Reports.
21. Abdullakutty, F., Akbari, Y., Al-Maadeed, S., Bouridane, A., Talaat, I. M., & Hamoudi, R. (2024). Histopathology in focus: a review on explainable multi-modal approaches for breast cancer diagnosis. Frontiers in Medicine, 11, 1450103.
22. Abdullakutty, F., Akbari, Y., Al-Maadeed, S., Bouridane, A., & Rifat Hamoudi, R. (2025). Enhancing the Prediction of Breast Cancer Progression Through Multi-modal Data Transformation. Cognitive Computation, 17(3), 114.
23. Alshadoodee, H. A. A., Razmara, J., & Karimpour, J. MULTI-MODAL BREAST CANCER CLASSIFICATION USING DUAL-ATTENTION VISION TRANSFORMERS AND METADATA-AWARE FUSION.
24. Alzamil, D., Alkhamees, B., & Hassan, M. M. (2025). A Systematic Review of Multimodal Fusion and Explainable AI Applications in Breast Cancer Diagnosis. Computer Modeling in Engineering & Sciences, 145(3), 2971.
25. Kavitha, A., Gopinath, S., & Girivasan, M. (2026, January). Transformer-Based Vision-Language Models for Multimodal Breast Cancer Diagnosis Integrating Radiology Reports and Histopathological Images. In 2026 5th International Conference on Communication, Computing and Electronics Systems (ICCCES) (pp. 1159-1164). IEEE.
26. Abdullakutty, F., Akbari, Y., Al-Maadeed, S., Bouridane, A., & Hamoudi, R. (2024). Advancing histopathology-based breast cancer diagnosis: Insights into multi-modality and explainability. arXiv preprint arXiv:2406.12897.
27. Paliwal, M., & Kashyap, K. L. (2026). Hierarchical contrastive disentanglement architecture for multi-modal breast cancer detection. Computers in Biology and Medicine, 205, 111582.
28. Abdikenov, B., Zhaksylyk, T., Shortanbaiuly, O., Orazayev, Y., Makhanov, N., Karibekov, T., ... & Seitova, A. (2025). Future of breast cancer diagnosis: A review of DL and ML applications and emerging trends for multimodal data. IEEE Access.

29. Mahichi, H., Ghods, V., Sohrabi, M. K., & Sabbaghi, A. (2025). Critical Review on Deep Learning-Based Breast Cancer Detection and Segmentation: Challenges, Gaps, and Future Directions. Majlesi Journal of Telecommunication Devices, 14(2).
30. Abdullakutty, F., Akbari, Y., Al-Maadeed, S., Al Saady, R., & Rasheed, H. M. A. (2025, October). Beyond Static Fusion: A Mixture-of-Experts Framework for Multimodal Breast Cancer Classification. In 2025 Fourteenth International Conference on Image Processing, Theory, Tools & Applications (IPTA) (pp. 1-6). IEEE.
31. Kansal, K., Kumar, S., & Kansal, K. (2026). Advances in Deep Learning Techniques for Breast Cancer Classification: A Comprehensive Review: K. Kansal. Archives of Computational Methods in Engineering, 33(1), 187-222.
32. Manzoor, F., Gupta, V., Pinky, L., Wang, Z., Chen, Z., Deng, Y., & Neupane, S. (2025). A Systematic Review of Multimodal Deep Learning and Machine Learning Fusion Techniques for Prostate Cancer Classification. medRxiv.
33. Alqazzaz, B. (2026). Breast Cancer Detection Using Mammogram Images with a Hybrid Model based GCN Algorithm. Intelligence-Based Medicine, 100380.
34. Liu, Y., Chen, S., Zhang, G., Wu, Y., Cao, R., Zhang, S., ... & Li, Y. (2026). An agentic multimodal AI framework for end-to-end breast cancer staging and biomarker profiling.
35. Suckling, J., Parker, J., Dance, D. R., Astley, S., Hutt, I., Boggis, C., & Savage, J. (2023). The Mammographic Image Analysis Society digital mammogram database. Excerpta Medica International Congress Series, 1069, 375–378.
36. Lee, R. S., Gimenez, F., Hoogi, A., & Rubin, D. (2024). Curated breast imaging subset of the Digital Database for Screening Mammography (CBIS-DDSM). Scientific Data, 11(1), 1–9.
37. Moreira, I. C., Amaral, I., Domingues, I., Cardoso, A., Cardoso, M. J., & Cardoso, J. S. (2024). INbreast: Toward a full-field digital mammographic database. Academic Radiology, 31(2), 236–248.
38. Al-Dhabyani, W., Gomaa, M., Khaled, H., & Fahmy, A. (2024). Dataset of breast ultrasound images. Data in Brief, 28, 104863.
39. Rodrigues, P. S. (2023). Breast ultrasound image (Mendeley Data). Mendeley Data, V1. https://doi.org/10.17632/wmy84gzngw.1
40. Spanhol, F. A., Oliveira, L. S., Petitjean, C., & Heutte, L. (2024). Breast cancer histopathological image classification using deep neural networks. Proceedings of the IEEE IJCNN, 2560–2567.
41. Rashed, E. A., Ismail, I. A., & Zaki, S. I. (2023). Multiresolution mammogram analysis in multilevel decomposition. Pattern Recognition Letters, 162, 1–10.
42. Qi, X., Zhang, L., Chen, Y., Pi, Y., Chen, Y., Lv, Q., & Yi, Z. (2023). Automated diagnosis of breast ultrasonography images using deep neural networks. Medical Image Analysis, 78, 101839.
43. Toğaçar, M., Ergen, B., & Cömert, Z. (2024). BrainMRI detection using attention-guided EfficientNet-B3 with transfer learning. Expert Systems with Applications, 238, 122115.
44. Ojala, T., Pietikainen, M., & Maenpaa, T. (2023). Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(7), 971–987.
45. Haralick, R. M., Shanmugam, K., & Dinstein, I. H. (2023). Textural features for image classification. IEEE Transactions on Systems, Man, and Cybernetics, SMC-3(6), 610–621.
46. Cai, J., Luo, J., Wang, S., & Yang, S. (2024). Feature selection in machine learning: A new perspective. Neurocomputing, 300, 70–79.
47. Breiman, L. (2023). Random forests revisited: Three decades of progress. Machine Learning, 112(4), 1249–1278.
48. Cortes, C., & Vapnik, V. (2024). Support-vector networks: Three decades of progress and open challenges. Machine Learning, 115(3), 891–924.
49. Chen, T., & Guestrin, C. (2023). XGBoost: A scalable tree boosting system — extended edition with GPU and distributed computing. ACM Transactions on Intelligent Systems and Technology, 14(3), 1–26.
50. Godala, Sravanthi, and M. Sunil Kumar. "A weight optimized deep learning model for cluster based intrusion detection system." Optical and Quantum Electronics 55.14 (2023): 1224.
51. Suman, J. V., Mahammad, F. S., Sunil Kumar, M., Sai Chandana, B., & Majji, S. (2024). Leveraging natural language processing in conversational AI agents to improve healthcare security. Conversational Artificial Intelligence, 699-711.

52. Gandikota, Hari Prasad, Abirami S, and Sunil Kumar M. "CT scan pancreatic cancer segmentation and classification using deep learning and the tunicate swarm algorithm." Plos one 18.11 (2023): e0292785.
53. Kumar, M. S., & Harshitha, D. (2019). Process innovation methods on business process reengineering. Int. J. Innov. Technol. Explor. Eng, 8(11), 2766-2768.Wolpert, D. H. (2023).
54. Stacked generalisation at 30: Retrospective and contemporary perspectives. Neural Networks, 159, 1–20.
55. Deb, Dipti, Ratnakar Dash, and Durga Prasad Mohapatra. "MMHBC-Net: a multi-modal hybrid approach for breast cancer classification." Neural Computing and Applications 37.20 (2025): 15469-15499.a