AI-Driven Feature-Enhanced Stacking Ensemble with Global-Context Vision Transformers for Breast Cancer Classification in Ultrasound Images

Nghia Trong Vo, *Student Member, IEEE*, Hoang Phi Yen Duong, *Graduate Student Member, IEEE*, Tuan Thanh Nguyen, Nhan Duc Le, and Trung Q. Duong, *Fellow, IEEE*

***Abstract*—Breast cancer remains a leading cause of death among women worldwide. Early detection of breast cancer is a crucial step towards improving survival rates for patients affected by the disease and is typically performed with the help of ultrasound imaging. Current rapid advancements in artificial intelligence (AI) research have produced a plethora of machine learning methods that aid in building automated diagnostic assistance systems for early cancer detection, including breast cancer detection. While deep learning has shown promise in medical image analysis, most existing approaches rely on single models or simple ensemble methods that fail to fully exploit complementary feature representations across architectures. This paper introduces a novel feature-enhanced stacking ensemble framework that combines state-of-the-art global context vision transformer (GCViT) with well-established convolutional neural network (CNN) architectures (ResNet-50V2, ConvNeXt-Tiny, and EfficientNetV2-B3) for automated breast cancer classification from ultrasound images. Unlike conventional ensembles that aggregate only prediction probabilities, our approach extracts deep feature embeddings from a dedicated CNN branch and concatenates them with base model predictions as input to a meta-learner, a multi-layer perceptron (MLP), enabling the ensemble to leverage both decision-level and feature-level information. When incorporating a meta model with feature representations from a CNN-based feature extractor, we are able to produce superior performance across multiple metrics compared to prior works. We accomplish top performance of 94.23% accuracy, 95.47% AUC-ROC. To further evaluate the robustness and generalizability of our approach, we conduct additional experiments on the melanoma cancer image dataset and achieve 95.4% accuracy. We provide comprehensive explainability analysis through shapley additive explanations (SHAP) values for feature attribution, permutation importance for model contribution quantification, and saliency maps for visual interpretation from base models and the end-to-end ensemble model to explain their contributions to final predictions.**

***Index Terms*—Ensemble Learning, Vision Transformer, Breast Cancer Classification, Deep Learning.**

I. Introduction

Breast cancer is considered among the most common types of cancer diagnosed in women around the world. There is an estimation of 2.3 million new cases diagnosed around the world in 2020 [1]. Early detection for breast cancer plays a crucial step in effective breast cancer diagnosis procedure and treatment, which can significantly improve survival rates to approximately 99.8% when patients are diagnosed as early as stage I [2]. In breast cancer screening and diagnosis, there are multiple medical imaging techniques that can be applied and ultrasound imaging is typically employed due to its flexible applications in clinical practices, feasible effectiveness for early diagnosis of breast cancer under limited resource as well as offering as an alternative tool with more affordable cost than mammography [3]. Because of these benefits, there is no doubt that ultrasound can be employed as a practical imaging modality for tumor detection and classification using machine learning algorithms, especially convolutional neural networks (CNNs) and vision transformers (ViTs).

In the past few years, artificial intelligence (AI) research, especially in computer vision, has witnessed many breakthrough in deep learning architectures that significantly aid the process of medical image diagnosis for cancer patients from the traditional CNNs to the more recent ViT models. Not only these models can learn from normal images to detect and differentiate between common objects such as dogs and cars, but also understand a variety of medical imaging modalities such as X-rays, computed tomography (CT) scans, magnetic resonance imaging (MRI), ultrasound and positron emission tomography (PET) scans. An important milestone in the remarkable rise of CNN models was marked by the introduction of AlexNet, which significantly improved performance in object detection and classification [4]. The effectiveness of CNNs in medical imaging tasks has been demonstrated in healthcare applications [5], [6]. Since then, more sophisticated deep learning architectures have been introduced, including the ResNet family (e.g., ResNet-50V2) [7], EfficientNet family (e.g., EfficientNetV2-B3) [8], ConvNeXt family (e.g., ConvNeXt-Tiny) [9], and vision transformers [10], [11], [12]. Improved performance across evaluation metrics in medical image detection and classification tasks has been reported in multiple studies [13], [14], [15], [16]. Consequently, these models have great potential to build AI-assisted diagnostic tools in clinical practice. Beyond advances in CNN architectures, the integration of these models with remote networking and intelligent assistance frameworks has also been explored to support next-generation Internet of Things (IoT) systems for medical imaging and healthcare applications [17], [18], [19].

N. T. Vo, H. P. Y. Duong, and T. Q. Duong are with the Faculty of Engineering and Applied Science, Memorial University of Newfoundland, St. John's, NL A1C 5S7, Canada (e-mail: {ntvo, yhpduong, tduong}@mun.ca).

T. T. Nguyen is with the School of Computing and Mathematical Sciences, University of Greenwich, UK (e-mail: tuan.nguyen@greenwich.ac.uk).

Nhan Duc Le is with Danang Hospital, Danang 50000, Viet Nam (e-mail: drnhandanang@gmail.com).

This work was supported in part by the Canada Excellence Research Chair (CERC) Program CERC-2022-00109.

Corresponding author is Trung Q. Duong (tduong@mun.ca).

Transfer learning is a common approach to leverage the pre-trained models on large natural dataset and re-train on a more specialized medical dataset since the dataset in medical domain is typically small in size, which can prevent models from learning sufficient data and thus not being able to generalize the data distribution well as well as potentially causing overfitting. The efficiency of applying transfer learning to a brain tumor classification task has been demonstrated, where a pre-trained ResNet152V2 model was fine-tuned on a much smaller brain tumor MRI dataset compared to ImageNet [20]. Because the pre-trained model has already learned the common patterns in natural objects, it will converge faster when training further on a specialized dataset, which can come at a cost of overfitting. To address this issue, data augmentation techniques were applied to significantly increase the number of training images, leading to improved generalization and higher overall accuracy compared to training without augmented data [20]. This goes to show that by applying a combination of techniques such as transfer learning, data augmentation, early stopping during training, etc., CNN models can achieve high performance for classification tasks even with limited data and computing resources.

Following the success of CNN models, ViTs have been introduced as competitive alternatives to traditional CNN models [10]. The ViT's architecture was inspired by the transformer model, which was initially introduced for natural language processing (NLP). Unlike traditional CNN models, ViT model treats an image as a sequence of fixed-size patches. With the self-attention mechanism, it is able to effectively capture the contextual relationships between token patches and understand the spatial information of the patches by combining them with positional embeddings. In several empirical studies, ViT models have demonstrated that they can potentially outperform traditional CNN models across performance metrics for classification tasks, especially in medical imaging. A comprehensive comparison in performance between CNN and ViT models has been reported using a bladder dataset consisting of 2629 images, where 70% of the images were used to train CNN and ViT models; InceptionResNetV2 achieved the highest accuracy of 98.73% among CNN models, while ViT-B32 achieved an even higher accuracy of 99.49% [21].

In this study, we aim to take advantage of the strengths of both CNN and ViT models to leverage their complementary capabilities. While CNN models are known to be proficient in capturing different patterns and features of images, ViT models demonstrate sophisticated understanding of the context and relationships between image patches due to the power of the attention mechanism. This combination can be promising when applied with ensemble learning, which can improve the ability to distinguish between benign and malignant breast tumors. In addition, we employed some of the most recent CNN and ViT models to study whether these latest models can improve performance on classification tasks compared to older models. The main contributions of this paper are summarized as follows:

- A feature-enhanced stacking framework integrating GCViT [22] with complementary CNNs (ResNet50 [7], ConvNeXt [9], EfficientNetB3 [8]), where a meta-learner combines both class predictions and 128-dimensional deep feature embeddings to achieve 94.23% accuracy on breast ultrasound classification. To the best of our knowledge, this study presents the first integration of GCViT in a medical imaging ensemble.
- In addition to BreastMNIST, we further validate the proposed ensemble framework on the melanoma cancer image dataset from a data science platform, called Kaggle, demonstrating its robustness and generalizability across different cancer types and imaging modalities. We also conduct additional experiments to explore and analyze how different configuration for the ensemble learning's components can lead to different performance results.
- Comprehensive interpretability analysis using shapley additive explanations (SHAP) [23] values, permutation feature importance [24], and saliency maps [25] to quantify model contributions, identify critical features, and visualize decision-making processes.

The subsequent sections of this paper are arranged as follows: Section II surveys existing research relevant to this study. In Section III, we present the overview of stacked generalization for ensemble learning and our proposed architecture design. Section IV dives into the details of the employed datasets, pre-processing and data augmentation techniques and implementation details of the proposed ensemble learning with stacked generalization and explainable artificial intelligence (XAI), and evaluation metrics. Section V presents the experimental results together with comprehensive analyses, model performance comparisons and model interpretations with XAI methods. Lastly, the paper wraps up in Section VI.

II. Related Work

Numerous studies have explored the integration of deep learning models with cancer diagnosis practice to improve breast cancer detection and classification using ultrasound imaging. Recent studies have shown promising results in applying deep learning approaches to automate and enhance the clinical practice of diagnosing breast ultrasound images with the primary goal of lowering diagnostic errors by humans and improving the efficiency of cancer type assessments. [26] examined the applications of three common CNN models including VGG19, InceptionV3, and ResNet which were trained on a breast cancer histopathological dataset called BreakHis for classification tasks. the study shown good results with high prediction accuracy and it is also resource efficient due to the transfer learning approach applied here, which is suitable for environments where computing resources are constraint.

Besides applying CNN models for breast cancer classification tasks, there are some studies that have explored the hybrid approaches where they combined both CNN and ViT architectures to improve performance. For instance, study in [27] proposed a robust hybrid framework to train model on a mammogram image dataset where CNN architecture was used to extract local features of images and ViT architecture extracted global features. After combining the outputs of both models, the feature vectors were forwarded as the inputs to a multi-layer perceptron (MLP) head and then a fully-connected layer to produce output probabilities for predictions of benign and malignant classes. The evaluation performance of the hybrid approach was then compared with other CNN models. Their findings revealed that even though CNN models had higher accuracy for training and validation dataset than the hybrid CNN-ViT model, they can potentially suffer from overfitting. However, the hybrid model in the study demonstrated good generalization with the least overfitting, which suggests its ability to classify unseen data more accurately compared to other conventional CNN models.

Additionally, the feature extraction approach to train classification models from the mentioned studies can be developed further by combining with ensemble-based methods. In particular, the studies in [28] and [29] have followed a similar approach by employing pretrained CNN models to extract features from images and then feeding these feature vectors to a classifier, which could be an MLP or ensemble learning algorithms. For instance, the voting classifier in [29] demonstrates the strength of ensemble learning methods where the authors employed several different machine learning algorithms to make diverse predictions, which are then combined and forwarded to the ensemble voting model to improve generalization ability and prevent overfitting issues. Similarly, [30] employed three base CNN models to train for the classification of two benchmark datasets of cervical cytology. The authors proposed a rank-based fusion ensemble algorithm to determine the final predictions by calculating fused scores and fuzzy ranks from confidence scores produced by the base models.

Recent studies have demonstrated that integrating ensemble learning with Internet of Medical Things (IoMT) infrastructures can significantly enhance the robustness, scalability, and privacy of medical diagnosis systems. Federated and distributed ensemble frameworks have been proposed to address data isolation and privacy concerns across healthcare institutions, where dynamic model fusion strategies enable collaborative learning from heterogeneous medical imaging sources without sharing raw data, achieving improved diagnostic accuracy and communication efficiency for coronavirus disease of 2019 (COVID-19) detection [31]. Beyond imaging-centric applications, context-aware ensemble learning has also been explored in wearable IoMT environments, where selective sensor fusion across an ensemble of classifiers allows adaptive stress detection under varying noise conditions, highlighting the importance of multimodal fusion and model diversity in real-world healthcare monitoring [32]. Ensemble learning has further been combined with federated learning to support remote and underserved healthcare settings, enabling early diagnosis of pulmonary diseases through distributed IoMT devices while reducing latency, computational overhead, and communication costs [33]. Multimodal ensemble frameworks that fuse heterogeneous data sources, such as medical images and physiological or audio signals, have also been introduced to improve early disease detection, leveraging neuro-fuzzy and weighted fusion strategies to enhance diagnostic reliability and clinical interpretability in IoMT-enabled environments [34]. In parallel, cloud-based IoMT systems have adopted ensemble classifiers to deliver scalable medical diagnosis services while addressing security and privacy challenges; such approaches integrate privacy-preserving computation, access control, and lightweight ensemble inference to protect both patient data and model intellectual property [35]. Collectively, these works highlight the growing role of ensemble learning as a foundational component in IoMT-driven healthcare systems, motivating further exploration of advanced ensemble architectures for robust and trustworthy medical image analysis.

These studies revealed promising results where the ensemble models were able to achieve top accuracy during evaluation and inference stages. However, while these studies provided strong performance when applying deep vision models and ensemble learning, to the best of our knowledge, no known study has attempted to explore the combination of base model predictions from the latest generation of CNN and ViT architectures with feature extractions to train ensemble models, which is the key focus in our study. Therefore, this distinctive approach can give a significant boost in classification performance due to our leveraging of some of the most recent deep learning models, which have been demonstrated to outperform their predecessors on benchmark datasets. Furthermore, by joining base model predictions with feature representations extracted from a dedicated pretrained CNN model, we are able to utilize both predictions and high-dimensional image feature embeddings to a classifier, also called meta model, to enhance its generalization proficiency and overall performance.

III. Methodology

This section describes our proposed feature-enhanced stacking ensemble framework for breast cancer classification. First, we present the overview of ensemble learning methods and the concept of stacked generalization. We then present our novel ensemble architecture that integrates multiple base model predictions with CNN-extracted features through a meta-learning approach. Finally, we detail the data flow in the proposed ensemble architecture.

A. Overview

Ensemble learning is a powerful and effective machine learning approach that helps to create more robust collaborative predictive models and achieve superior performance compared to any single model can do due to the combined predictive power from the group of diverse models. Our approach follows the stacked generalization methodology [36] where we combine four base models (level-0 models) including GCViT, ConvNeXt-Tiny, EfficientNetV2-B3 and ResNet-50V2 to produce diverse predictions and a custom MLP as meta model (level-1 model) to learn the patterns in these base models' predictions to determine the final classification for the input image.

In a stacked generalization framework, the level-0 models are the base learners, which are the individual models that directly process the input data and generate initial predictions, while The level-1 model, known as the meta-learner or meta model, operates on the outputs generated by the level-0 models. In this case, The level-0 dataset refers to the original input

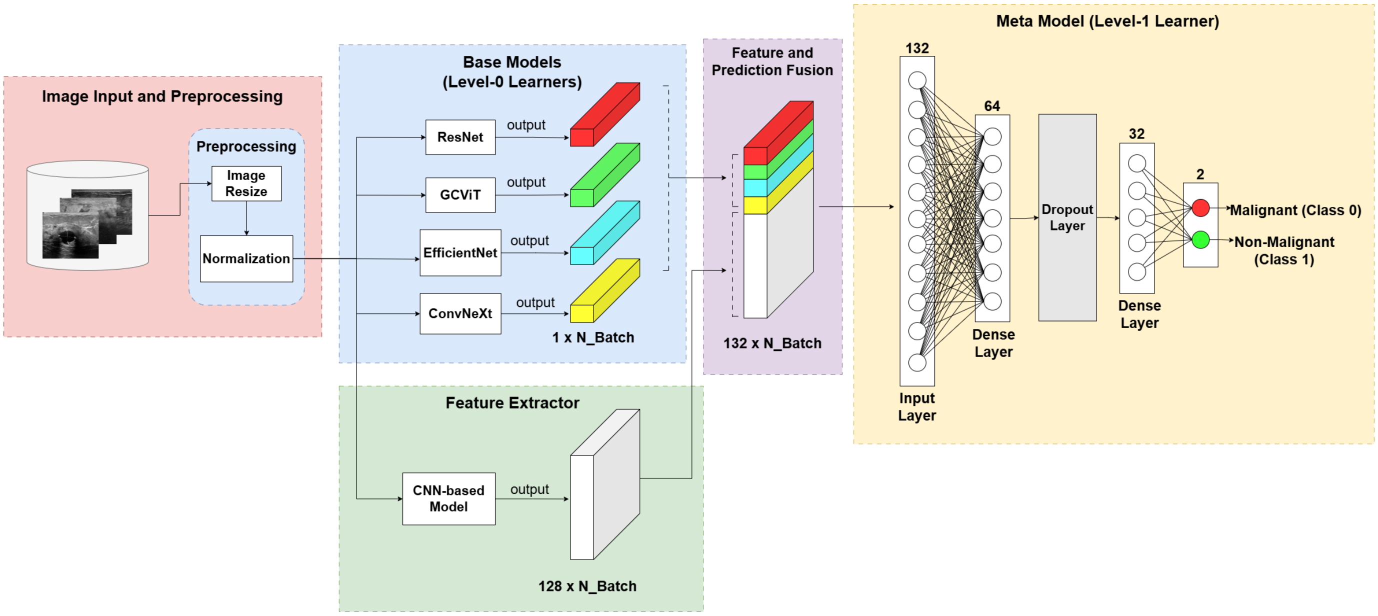

Fig. 1. Overview of the proposed ensemble architecture integrating heterogeneous base models (GCViT, ConvNeXt-Tiny, EfficientNetV2-B3, and ResNet-50V2
) and a CNN-based feature extractor, followed by a meta model for final classification.

data used to train the base (level-0) models, while the level-1 dataset is composed of the predictions made by those base models, serving as input for the meta-learner (level-1 model) in the stacked generalization framework. Our proposed ensemble learning architecture is depicted in Fig. 1, where it integrates multiple diverse models and processing strategies to enhance predictive performance in medical image classification, which follows the principle of stacked generalization. The following subsections describe the components of the proposed design in detail.

B. Base models

In the first step, raw medical images are provided as inputs to the system. These images undergo a series of preprocessing procedures as described in the Section B. Once preprocessing is complete, images are forwarded to four distinct base models:

- GCViT[22] : incorporates both local self-attention (for fine-grained features like tumor margins) and global self-attention (for contextual relationships with surrounding tissue). Unlike traditional ViTs that struggle with local feature extraction, GCViT's hierarchical design with local-global attention modules effectively addresses the dual need for detailed pathological feature detection and contextual understanding in medical diagnosis. Moreover, GCViT demonstrates superior performance on various imaging tasks compared to standard ViTs and CNNs [22], particularly excelling in capturing long-range dependencies across entire images while preserving fine-grained local details.
- ResNet-50V2 [7]: has residual connections enable training of deep networks without vanishing gradients, providing stable and reliable features. Its proven track record in medical imaging makes it a robust baseline.
- ConvNeXt-Tiny [9]: Modernizes CNN design by incorporating transformer-inspired elements (depthwise convolutions, layer normalization) while maintaining CNN efficiency. Bridges the gap between traditional CNNs and transformers. It focuses more on spatial continuity and local feature aggregation. When combined through stacked generalization, ConvNeXt-Tiny helps the meta-learner adaptively balance local texture cues with global contextual information.
- EfficientNetV2-B3[8]: Compound scaling balances depth, width, and resolution for optimal accuracy-efficiency trade-off. Particularly effective for extracting discriminative features from limited medical datasets.

Each base model has a different neural network architecture with unique feature extraction strategies, which helps capture diverse representations of image characteristics. The base models independently produce prediction vectors for the batch of images, encapsulating their unique interpretation of the input data. The diversity injected at this level forms the foundation for the ensemble's robustness and improved generalization ability.

C. Feature extractor

Parallel to the base models, a dedicated CNN-based feature extractor, implemented using a pretrained EfficientNetV2-B3 architecture, processes the input images. At this stage, the extractor outputs high-level deep representations from the last convolutional block (i.e., the penultimate feature maps before classification), which encode rich morphology and texture information. We remove the original classification head and apply global average pooling followed by a fully connected layer with 128 neurons and a rectified linear unit (ReLU) activation function to obtain a compact 128-dimensional embedding per image. These embeddings are concatenated with the base-model prediction probabilities and provide complementary feature-level evidence for the meta learner, helping explain the observed performance gain.

D. Feature and prediction fusion

The outputs of the base models (class probability vectors) and the extracted features (128-dimensional embedding vectors) are then concatenated into a unified representation. Specifically, for a mini-batch of images, the four base model predictions and the extracted features are stacked to create a composite feature vector of size 132 for each case (4 from the base models, 128 from the feature extractor). This fusion approach provides the subsequent meta learner with both the explicit predictions and the internal feature representations, thereby enriching the context available for the ensemble decision process.

E. Meta model

The final component is the meta model, or level-1 learner, which synthesizes all preceding outputs to deliver the final prediction. The meta model is implemented as an MLP consisting of an input layer of size 132, two hidden dense layers (with 64 and 32 neurons), a dropout layer for regularization, and an output head with two neurons corresponding to the malignant and non-malignant classes. The meta model learns to capture the underlying relationships between the predictions of the base models and recognize the patterns of errors made by each base model. These errors in predictions are one of the most powerful advantages of stacked generalization due to the diversity of the base models, which means that these prediction mistakes complement each other that helps the meta model identify when to trust which base models and recognize the error patterns from what it has learned from training data.

IV. Experiments

A. Dataset

We employed a publicly available dataset for breast cancer, called BreastMNIST, which is a part of the medical dataset collection MedMNIST [37]. The BreastMNIST dataset contains a total of 780 ultrasound images which were preprocessed from the BUSI dataset [38] to simplify it into binary classification task. Initially, the dataset contains 3 classes including normal, benign, and malignant. In our experiments, normal and benign images were merged into one category, thus forming 2 classes: malignant (class 0) and non-malignant (class 1). The BreastMNIST dataset is divided by default into training, validation and test sets. Based on the class distribution of the BreastMINST dataset, we can see that the dataset is imbalanced, which is a widely common challenge for medical image datasets, where the number of samples with class 1 or non-malignant label is approximately 2.7 times higher than the number of samples with class 0 or malignant label. To tackle the imbalanced problem, we apply special techniques by including focal loss [39] and class weighting during training to help the models to pay more attention to class with fewer images and focus on learning from hard image samples.

B. Preprocessing and data augmentation

The first step for model training and prediction is to apply the preprocessing technique, which is a critical step for preparing data before feeding them into models for better data quality and training efficiency. We start by ensuring that all images are of size $224 \times 224 \times 3$ by resizing them, followed by applying normalization based on the ImageNet dataset, which the CNN and GCViT models were pre-trained on, to align with the data distribution that the models have learned. After the preprocessing step, we apply data augmentation to help models increase the amount of data that they can learn from and improve their ability to generalize better on unseen data. For this approach, we include weak augmentation (basic image transformation) including only slight rotation and strong augmentation (more aggressive image transformation) including horizontal flip, gamma contrast adjustment, saturation multiplication and gaussian blur. Table I shows the augmentation techniques used during training and their corresponding parameters. We used the imgaug python library for data augmentation and were able to increase the number of training images from 546 to 3,570.

TABLE I
Data augmentation techniques and their parameters

Technique	Parameters	Rationale
Rotation	±15 degrees	Simulates natural scanning variations
Horizontal Flip	50% probability	Accounts for bilateral breast anatomy
Gamma Contrast	0.8–1.2 range	Mimics ultrasound gain variations
Saturation Adjustment	0.8–1.2 range	Handles color cast variations
Gaussian Blur	$\sigma = 5.0$	Simulates focusing variations

C. Training process

Fig. 2 illustrates the ensemble learning framework, outlining the training strategy for base models and the meta model with stacking approach, performance evaluation and explainable AI (XAI). The BreastMNIST dataset is initially split into training, validation and test sets. Every base model, pre-trained on the ImageNet dataset, undergoes model training with out-of-fold predictions. This is the recommended training strategy for stacked generalization [36] [40], where we divide the training data into 4 equal folds and then train the base model on 3 folds while leaving 1 fold for prediction. This step is repeated four times, once for each fold, to generate the final set of predictions. Each prediction comes from a model that was not trained on that specific sample. After that, we re-train another base model on the full training set and make prediction on the validation and test sets. Additionally, the CNN-based feature extractor produces feature vector embeddings for all images from the dataset. These embeddings are then combined with the unbiased predictions from the base models to form the level-1 training, validation and test sets, which are then fed into the meta model for its training, performance evaluation

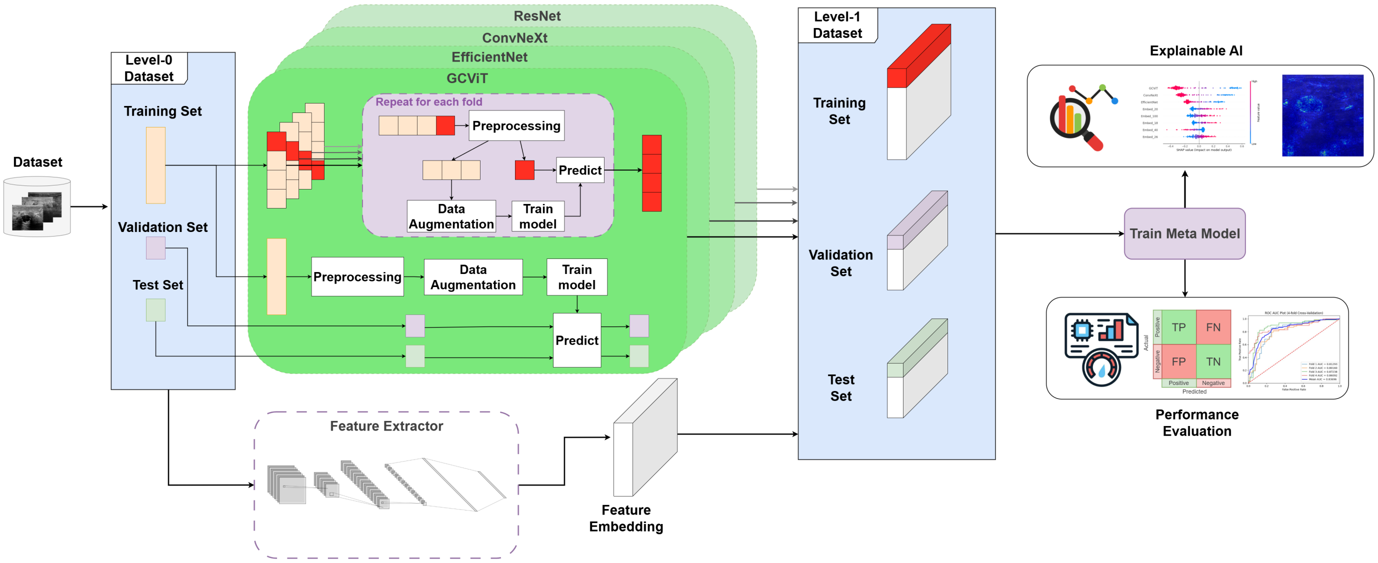

Fig. 2. Illustration of the stacked generalization (stacking) training process: base models are trained using out-of-fold predictions to prevent data leakage and their outputs, along with extracted features, are combined as input for the meta model.

and explainable AI analysis. These predictions from the base models are unbiased since they are generated from the held-out fold which is not learned during training. Thus, this helps the meta model learn to understand the patterns of how the base models make guesses on unseen data as well as prevents overfitting and data leakage. This training process allows the meta model to capture the contextual relationships between base models' predictions and the ground-truth labels, that means it learns patterns about when to trust different base models based on their historical performance on out-of-fold predictions. For instance, during training the meta model might learn that when GCViT predicts 0.75 probability of class 1 and EfficientNet predicts 0.3 probability for the same class, GCViT is usually more correct in this case. The feature vector embeddings also provide additional context, which serves as complementary information for the meta model to make a more insightful decision with features extracted from images. This makes the level-1 dataset a rich feature space that captures both high-level probabilistic information generated by the base models and detailed image characteristics extracted as feature embeddings.

$$p_t = \begin{cases} \hat{y} & \text{if } y = 0 \\ 1 - \hat{y} & \text{otherwise} \end{cases} \tag{1}$$

$$FL(p_t) = -\alpha(1 - p_t)^{\gamma} log(p_t) \tag{2}$$

We set the same hyperparameters to train across the base models and meta model. The learning rate is set to $1e - 3$ and the batch size is 32, which is recommended for good balance between computational capability and memory efficiency when training with our NVIDIA T4 graphics processing unit (GPU) for this experiment. We use adaptive moment estimation with weight decay $(AdamW)$ [41] optimization algorithm, which decouples weight decay from gradient-based optimization, providing more effective regularization than standard Adam, particularly beneficial for transformer architectures and limited medical datasets. Additionally, we apply focal loss [39] to tackle the class imbalance problem for the BreastMNIST dataset. This helps the base models and meta model to pay more attention to learning harder or misclassified samples by down-weighting the contribution of well-classified samples and emphasizing the gradient updates from difficult cases. The focal loss function can be expressed in equation (2) where α is a weight factor for the classes and γ is the focusing factor controlling how easy examples are down-weighted. p_t is defined in equation (1) where $y \in \{0, 1\}$ is the ground-truth label and $\hat{y} \in [0, 1]$ is the model's estimated probability for a sample being class 1. We set α to 0.25 and γ to 2.0, which generally tend to work best according to the original paper of focal loss. Through preliminary experiments, we trained the base models and meta model for 50 and 25 epochs, respectively, which were sufficient for the models to converge and the learning curves to stabilize.

All experiments were conducted in a Python-based training environment with an NVIDIA T4 GPU. In addition to transfer learning from ImageNet pretrained weights, we applied data augmentation using the imgaug library and class weighting using a balanced weighting strategy (computed from scikit-learn's class-weight heuristic with scaling factors) to improve robustness under limited and imbalanced medical data.

Regarding computational cost, the overall training complexity is dominated by fine-tuning the four base networks. The 4-fold out-of-fold strategy increases training cost approximately linearly with the number of folds (plus an additional fit on the full training set) to generate unbiased level-1 predictions for stacking. This extra cost is incurred offline during training, while the meta model itself is lightweight and adds negligible overhead. At inference time, the total cost is primarily the sum of forward passes through the base models, with a single additional forward pass through the CNN-based feature extractor and a small MLP head. Furthermore, we implemented class weights to address class imbalance using the approach from scikit-learn's compute_class_weight function with a little twist by adding scaling factors β_0 and β_1 to the equations (3), which assign weights inversely proportional to class frequencies. The equations for calculating class weights are given by

$$w_0 = \beta_0 \cdot \frac{n_{total}}{2 \cdot n_0}, \; w_1 = \beta_1 \cdot \frac{n_{total}}{2 \cdot n_1}, \tag{3}$$

where w_0 and w_1 denote the weight for class 0 and class 1, respectively. β_0 and β_1 represent the scaling factor for class 0 and class 1, respectively. n_0 and n_1 are number of samples with class 0 and class 1, respectively. n_{total} denotes total number of samples. This balanced weighting heuristic is inspired by methods for handling rare events in logistic regression. Specifically, for multiclass classification, the balanced class weights can be calculated as $w_j = \frac{n_{total}}{n_{classes} \cdot n_j}$, where $n_{classes}$ is the total number of classes and n_j represents the frequency of class j in the training data. For binary classification, this yields weights w_0 and w_1 with $n_{classes} = 2$ in equations (3), effectively giving higher weight to the minority class to mitigate class imbalance. Through extensive experiments, we found that setting β_0 to 1.8 and β_1 to 1 gives the best result.

D. Implementation for XAI

D1. SHAP

SHAP [23] is among the most well-known model-agnostic methods for machine learning model interpretability. It is based on game theory to understand how each feature contributes to the predictions of a black-box model. The input to SHAP consists of the complete level-1 feature vector fed into the meta model, which is formed by concatenating the prediction probabilities produced by the four base models (GCViT, ConvNeXt-Tiny, EfficientNetV2-B3, and ResNet-50V2) with the 128-dimensional feature embeddings extracted by the CNN-based feature extractor. Using this combined input representation, SHAP computes shapley values for each feature, indicating how much a feature contributes positively or negatively to the meta model's output for a given sample.

The output of SHAP, as illustrated in Fig. 5, is a set of feature-wise importance scores, visualized through a summary plot that ranks features according to their overall impact on the model's predictions. This analysis allows us to identify which base model predictions and which deep feature embeddings are most influential in driving the ensemble's final classification. Overall, SHAP is used to provide transparency into how the meta model integrates heterogeneous inputs, verify that the ensemble leverages complementary information across base models, and enhance the interpretability and clinical trustworthiness of the proposed framework.

D2. Permutation feature importance

Permutation feature importance is another common type of model-agnostic XAI approach [24] for evaluating the contributions of each feature to a machine learning model by randomly shuffling values of individual features to compute the decrease in model performance. The random permutation technique works by intentionally breaking the relationship between a specific feature and the target variable. The input to the permutation feature importance analysis is the same level-1 representation used by the meta model, consisting of the concatenated prediction probabilities from the four base models (GCViT, ConvNeXt-Tiny, EfficientNetV2-B3, and ResNet-50V2) and the 128-dimensional feature embeddings extracted by the CNN-based feature extractor. For each feature or predefined feature group, values are randomly shuffled across samples while keeping all other features unchanged.

The output of the permutation feature importance method, as illustrated in Fig. 6, is a feature importance score computed as the decrease in area under the receiver operating characteristic curve (AUC-ROC, used interchangeably with AUC) of the meta model after permutation, with larger decreases indicating greater importance. This analysis reveals how strongly the meta model depends on each base model and their interactions, as well as the relative contribution of CNN-extracted feature embeddings. Permutation feature importance is used to complement SHAP by providing a global, performance-based explanation of the ensemble's behavior. While SHAP quantifies directional contributions to individual predictions, permutation importance highlights which features or feature groups are critical for maintaining overall classification performance. Together, these results validate that the meta model primarily relies on synergistic combinations of base model predictions, with feature embeddings serving as supplementary contextual information to refine final decisions.

D3. Saliency maps

We also produce saliency maps [25] for both the proposed ensemble model and each individual base model to visualize the spatial regions that most influence classification decisions. For each test image, the input to the saliency method is the original ultrasound image, and gradients are computed with respect to the predicted class score. For the base models, saliency maps are obtained directly from the gradient of the model output with respect to the input image. For the proposed ensemble, gradients are propagated through the entire end-to-end pipeline, including the base models, feature extractor, and meta model, allowing the saliency map to reflect how the ensemble jointly leverages all components to produce the final prediction.

The output of this process, as illustrated in Fig. 7, is a heatmap overlaid on the input image, where higher-intensity regions indicate pixels that contribute more strongly to the predicted class. Saliency maps are used to qualitatively assess the interpretability of the proposed framework and to verify that the ensemble focuses on meaningful anatomical structures rather than irrelevant background regions. This visualization demonstrates how stacked generalization enables the ensemble to combine complementary spatial cues from heterogeneous base models, resulting in more comprehensive and reliable feature localization for breast cancer classification.

E. Evaluation metrics

To evaluate the ensemble model, we use various performance metrics to assess its effectiveness for the binary classification task including accuracy, AUC-ROC, recall, precision and F1 score. Given $TP = True\ Positive$, $TN = True\ Negative$, $FP = False\ Positive$, $FN = False\ Negative$ as

- **Accuracy** measures the overall proportion of correct predictions across all classes

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

- **AUC-ROC** measures the area under the receiver operating characteristic curve and ranges from 0 to 1.

$$TPR = \frac{TP}{TP + FN},\ FPR = \frac{FP}{FP + TN} \tag{5}$$

$$AUC - ROC = \int_0^1 TPR(FPR^{-1}(x))\, dx \tag{6}$$

- **Recall** (also known as sensitivity or true positive rate) measures the proportion of actual positives that were correctly identified.

$$Recall = \frac{TP}{TP + FN} \tag{7}$$

- **Precision** (also known as positive predictive value) measures the proportion of true positive predictions among all positive predictions.

$$Precision = \frac{TP}{TP + FP} \tag{8}$$

- **F1 Score** is the harmonic mean of precision and recall, providing a single metric that balances both measures.

$$F1 = \frac{2 \cdot TP}{2 \cdot TP + FP + FN} \tag{9}$$

These mathematical formulations provide the basis for evaluating the ensemble model across multiple perspectives, thus allowing us to conduct a comprehensive overall assessment of the model performance beyond simple accuracy measures and consider the trade-offs between different metrics in the context of medical diagnosis and healthcare applications.

V. Results & Analysis

A. *Experiment results of GCViT and our proposed ensemble model for BreastMNIST*

In this section, we will present and discuss the training results of GCViT and our proposed ensemble models evaluated on the test set to assess their predictive performance on unseen data and generalization capability. Fig. 3 shows the confusion matrix for our proposed stacking ensemble model trained with EfficientNetV2-B3 as the feature extractor. It demonstrates the model's effectiveness and ability to handle the innate challenge of the class imbalance in the BreastMNIST dataset.

Fig. 3 indicates 9 incorrect predictions on the test set, consisting of 3 false negatives (malignant, class 0, predicted as non-malignant, class 1) and 6 false positives (non-malignant, class 1, predicted as malignant, class 0). The false negatives are mainly associated with malignant lesions that visually resemble benign patterns in ultrasound (e.g., round/oval appearance, smoother margins, and relatively homogeneous internal echoes), which can reduce the confidence of malignancy cues. Conversely, the false positives are mainly associated with benign lesions exhibiting malignant-like characteristics such as irregular margins or shape, heterogeneous internal texture, and darker regions consistent with posterior acoustic shadowing. These failure cases reflect the inherent overlap of sonographic appearance between benign and malignant lesions, especially in a small and imbalanced dataset.

The model achieved good sensitivity for classifying images with malignant tumors by successfully identifying 39 out of 42 malignant cases, while only 3 malignant cases are misclassified. This high sensitivity is clinically crucial, since it directly impacts patient outcomes by minimizing the risk of missed cancer diagnoses. Additionally, the ensemble model correctly classified 108 out of 114 non-malignant cases, showing good precision for diagnosing patients with minimal chance of false alarms of malignancy.

We monitor training and validation loss/accuracy curves throughout the training process to detect overfitting (diverging curves), underfitting (premature plateau), and ensure optimal convergence. Fig. 4 shows the loss and accuracy curves of both training and validation over 25 epochs for the proposed ensemble model with EfficientNetV2-B3 as feature extractor. The training loss drops sharply from above 0.05 during the first 5 epochs, then continues to decrease gradually and reaches approximately below 0.01 by epoch 25. The validation loss also reduces quickly for the first 5 epochs and then maintains a stable line with small fluctuations in the range 0.01-0.02. Similarly, the training accuracy starts around 50% and increases rapidly for the first 5 epochs, reaching around 90% and then rises more slowly to approximately 95% by epoch 25. On the other hand, the validation accuracy starts higher at around 70% and increases quickly to around 93% by epoch 6, then fluctuates slightly but remains above 90% at the end. From epoch 13 onward, the training accuracy increases slowly while the validation accuracy stays a little

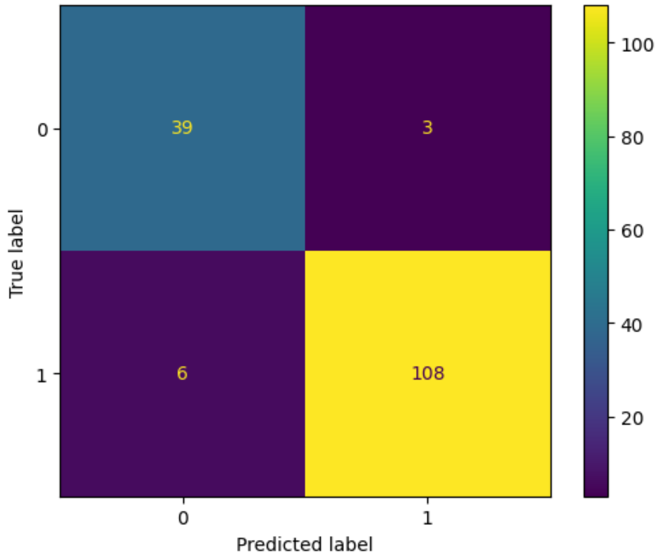

Fig. 3. Confusion Matrix of the proposed ensemble model with 128-dimensional features extracted from EfficientNetV2-B3 as the Feature extractor for BreastMNIST.

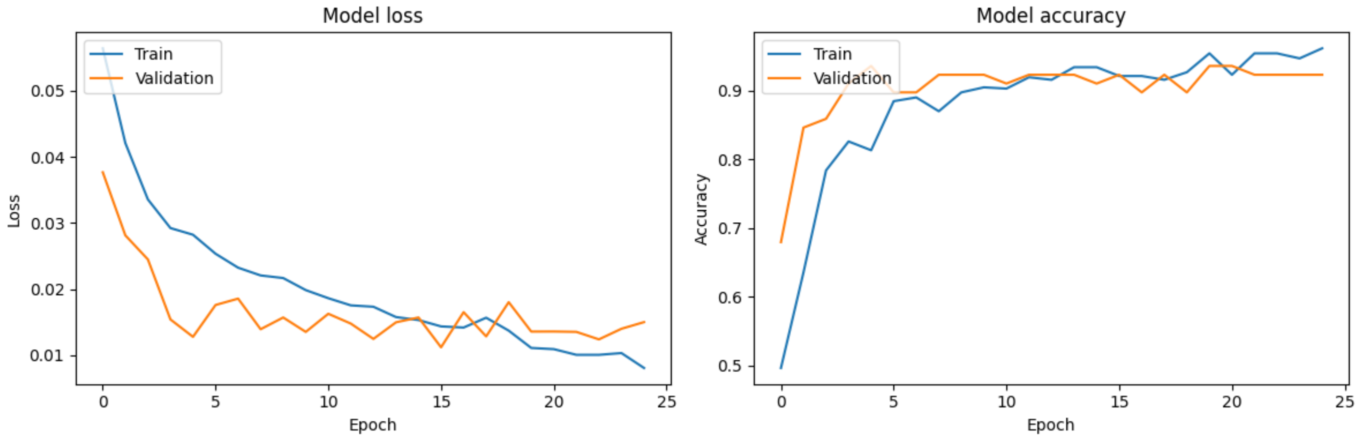

Fig. 4. Learning Curves of the Proposed Ensemble Model with EfficientNetV2-B3 as the Feature Extractor for BreastMNIST.

TABLE II
Evaluation metrics of our proposed ensemble model with 128-dimensional features extracted from EfficientNetV2-B3, GCViT, baseline ViT and other CNN-based models for BreastMNIST Dataset

Method	**ACC**	**AUC**	**Recall**			**Precision**			**F1 Score**		
			Malignant	Non-Malignant	Mean	Malignant	Non-Malignant	Mean	Malignant	Non-Malignant	Mean
DenseNet-121	0.8718	0.9206	0.81	0.89	0.85	0.74	0.93	0.835	0.77	0.91	0.84
Xception	0.8846	0.9127	0.69	**0.96**	0.825	0.85	0.89	0.87	0.76	0.92	0.84
InceptionV3	0.8782	0.9121	0.79	0.91	0.85	0.77	0.92	0.845	0.78	0.92	0.85
VGG19	0.7372	0.8517	0.81	0.71	0.76	0.51	0.91	0.71	0.62	0.8	0.71
MobileNetV3-Large	0.8589	0.9077	0.64	0.94	0.79	0.79	0.88	0.835	0.71	0.91	0.81
NASNet-Large	0.8013	0.8083	0.74	0.82	0.78	0.61	0.9	0.755	0.67	0.86	0.765
InceptionResNetV2	0.8846	0.9267	0.81	0.91	0.86	0.77	0.93	0.85	0.79	0.92	0.855
ResNet-50V2	0.8269	0.8814	0.62	0.9	0.76	0.7	0.87	0.785	0.66	0.88	0.77
EfficientNetV2-B3	0.8654	0.8924	0.76	0.9	0.83	0.74	0.91	0.825	0.75	0.91	0.83
ConvNeXt-Tiny	0.8974	0.9236	0.79	0.94	0.865	0.82	0.92	0.87	0.8	0.93	0.865
ViT [42]	0.9038	0.895	0.79	0.95	0.87	0.85	0.92	0.885	0.81	0.94	0.875
MedVIT-S [43]	0.897	0.938	-	-	-	-	-	-	-	-	-
GCViT	0.9103	0.9457	0.83	0.94	0.885	0.83	0.94	0.885	0.83	0.94	0.885
Proposed Ensemble Method	**0.9423**	**0.9547**	**0.93**	0.95	**0.94**	**0.87**	**0.97**	**0.92**	**0.9**	**0.96**	**0.93**

lower with a small gap between the two lines, indicating that training for 25 epochs or fewer is sufficient to achieve a good balance between model's performance on the training data and its generalization to unseen data. These learning curves show that our stacking ensemble approach can help the meta model learn effectively with stable learning curves and is capable of preventing overfitting due to combining different techniques, including out-of-fold prediction strategy, training on predictions of diverse base models with different error rates, data augmentation, transfer learning and $AdamW$.

Table II presents the performance results of our proposed ensemble model against individual base models (GCViT, ResNet-50V2, ConvNeXt-Tiny, and EfficientNetV2-B3) and other CNN-based models across different evaluation metrics on the test set. The proposed ensemble model notably outperforms both traditional CNN models and state-of-the-art vision transformers in most of performance metrics. It demonstrates balanced and robust performance across both benign and malignant classifications. For malignant cases, the ensemble model achieves 93% recall, 87% precision and 90% F1 score while for non-malignant cases, it accomplishes 95% recall, 97% precision and 96% F1 score. In terms of mean recall, precision and F1 score, The proposed ensemble model achieves the highest results among other models. This balanced performance is specifically vital in medical applications, since it indicates the model's reliability in both correctly identifying malignant cases (high sensitivity) and classifying benign cases (high specificity). Thus, the recall of malignant cases is crucial to determine the percentage of actual patients with malignant tumors that a model can correctly identify. The proposed ensemble model is able to achieve the highest score with 93% recall for malignant cases while GCViT achieves the second-highest result with 83%, which demonstrates a 10% improvement in malignancy recall compared to GCViT. The proposed ensemble also scores the highest in accuracy and AUC with 94.23% and 95.47%, respectively, while GCViT scores the second-highest with 91.03% accuracy and 94.57% AUC, which is an improvement of 3.2% for accuracy and 0.9% for AUC over the second-best performing model.

GCViT also shows really strong competence in predicting unseen data. It is able to outperform traditional CNN models and original ViT [42] in various metrics. This shows that the architecture of GCViT effectively leverages the self-attention mechanism enhanced with both local and global contexts to achieve better predictive performance and generalization capabilities. For origninal ViT model, GCViT scores moderately higher in some metrics, indicating an improvement of 0.65% accuracy, 5.07% AUC, 1.5% mean recall and 1% mean F1 score. This shows that the recent variant of ViT architecture successfully addresses some of the limitations inherent in the original ViT. In terms of overall performance, individual assessment of the four base models reveals the following performance ranks in descending order: GCViT achieved the highest overall performance, followed by ConvNeXt-Tiny, EfficientNetV2-B3 and ResNet-50V2. Thus, this ranking order of performance is important to evaluate how the meta model learn to recognize each base model's competence and combine their guesses to obtain the best prediction results. We will discuss this performance ranking phenomenon further in the explainable AI analysis section to unveil how the meta model weights each base model's predictions on unseen data.

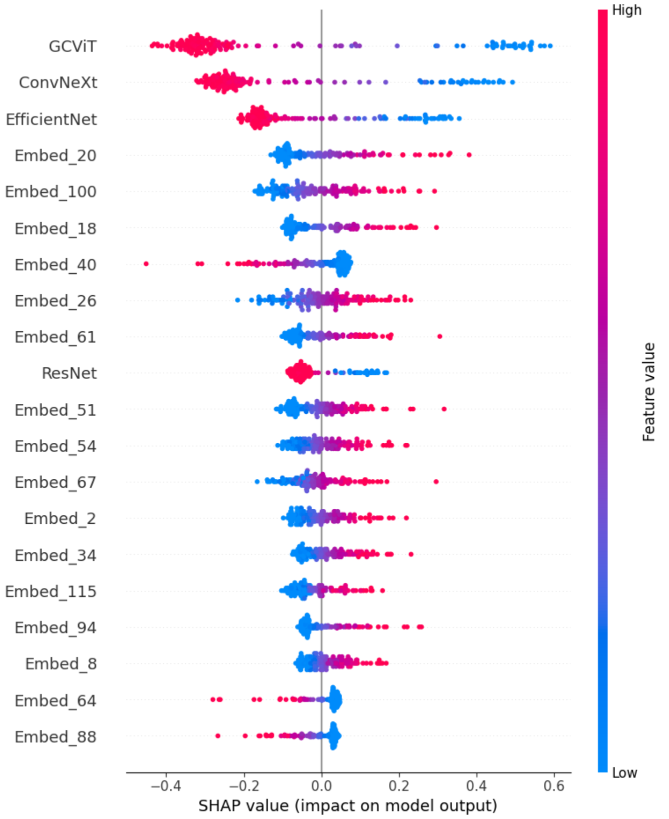

Fig. 5. Shapley values summary plot of BreastMNIST testing dataset for input features of the meta model with EfficientNetV2-B3 as the Feature Extractor where points residing further to the right means positive contribution and further to the left means negative contribution to malignant predictions.

B. *Explainable AI analysis for meta model*

The deployment of AI systems in medical applications demands not only high performance but also transparency and interpretability to gain trust from healthcare professionals and ensure safe clinical adoption. As our proposed ensemble framework combines multiple complex deep learning models, understanding how each component contributes to the final prediction becomes crucial for clinical acceptance and regulatory approval. XAI methods provide insights into model behavior by generating visual and statistical explanations that illuminate the decision-making process. Using multiple XAI methods is essential in medical imaging because each technique provides unique insights into model behavior and decision-making. Saliency Maps, local interpretable model-agnostic explanations (LIME) and SHAP highlight different aspects such as spatial attention, feature relevance, and global contributions. Combining these approaches helps overcome the limitations of single XAI methods, ensures more robust and reliable explanations, and supports diverse clinical needs. Therefore, this can ultimately help to increase transparency and trust in AI-driven diagnostic tools for both practitioners and patients. In the following analysis, we apply multiple XAI techniques including Saliency Maps, LIME and SHAP to provide comprehensive interpretability of our ensemble model.

B1. *SHAP*

Fig. 5 shows the summary plot for shapley values of the top 20 most important features as inputs to the meta model. The plot demonstrates the most influential features to the least ones from top to bottom. The x-axis represents the scale of shapley value where the higher positive values and the lower negative values indicate more contributions. Each data point represents a sample image from the test set that the ensemble model is evaluated on. Red color of a point indicates high feature value while blue color means low feature value and this red-blue color scale represents the magnitude of feature values of data points in the visualization. Points that stay further to the right positively contribute or push the meta model's prediction towards malignant outcome, and vice versa, points further to the left negatively contribute or push the prediction towards non-malignant outcome.

As we can observe from the SHAP values summary plot, the estimated prediction probabilities from GCViT have the highest contributions to the final classification decisions of the meta model. It shows that GCViT model has really high confidence in its predictions for most of the data points since there are two major clusters that are further away to the left and right sides while only a few data points are around the zero contribution line, reflecting higher uncertainty. This can be interpreted as the indication that the more confident the GCViT model is, the greater its influence on the final classification outcome of the meta model and its two major clusters of data points have the highest contributions on average compared to other features' clusters. The ConvNeXt-Tiny model demonstrates similar pattern but most of its predictions have lower contributions compared to GCViT. Most data points of EfficientNetV2-B3 have lower shapley values compared to GCViT and ConvNeXt-Tiny, however the model still demonstrates relatively high positive contributions for samples that are predicted as malignant with high confidence. Among the 4 base models, ResNet-50V2 model's predictions has the least influence to final decision of the meta model as most of its data points are clustered near the zero contribution line. Out of 128 feature embeddings, there are 6 features that have

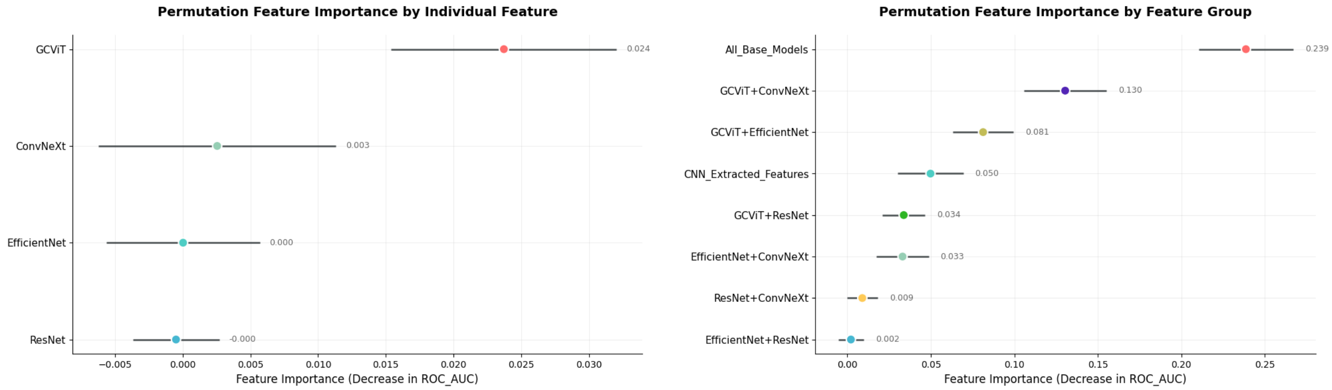

Fig. 6. Permutation feature importance of the proposed ensemble model with EfficientNetV2-B3 as the Feature extractor for BreastMNIST.

higher contributions than ResNet-50V2's predictions. Most of the embedding features, which have high feature values, show positive contributions towards malignant outcomes. In contrast, some embedding features such as embed_40 with high feature values contributes negatively or pushes predictions of the meta model away from malignant outcomes. Notably, the summary plot shows that the meta model has learned to accurately identify to rank the 4 base models based on their historical performance from the training set, consistent with the individual performance results we have mentioned in the previous section, reflecting the effectiveness of our ensemble approach.

In addition, the SHAP summary plot shows that the meta model does not rely on a single base model but instead learns to combine complementary information from multiple predictors. While GCViT exhibits the largest individual contribution, predictions from ConvNeXt-Tiny and EfficientNetV2-B3 also have substantial shapley values, and several CNN-extracted feature embeddings contribute more than the ResNet-50V2 output alone. This indicates that the ensemble leverages diverse and complementary signals rather than over-trusting any single model. By adaptively weighting these contributions, the meta model can compensate for individual model weaknesses and reduce misclassification risk, which explains why the proposed ensemble achieves more robust and consistently higher performance than single-model predictions.

B2. Permutation feature importance

Fig. 6 shows two feature importance plots. The left and right plots present feature importance by individual base models and feature group, respectively. The feature importance for each feature is calculated by measuring the decrease in AUC-ROC of the meta model.

The permutation feature importance analysis reveals several key insights into the ensemble's decision-making behavior. For ResNet-50V2, EfficientNetV2-B3 and ConvNeXt-Tiny, they show little to no impact on the decline of AUC-ROC, thus they have low feature importance. This can be explained by the fact that the meta model can still depend on other base models' predictions and feature embeddings to make accurate decisions when predictions of a base model are shuffled randomly. On the other hand, the GCViT model has relatively high feature importance compared to other base models, demonstrating that the model's predictions on some samples can help the meta model to make more accurate predictions, while other base models and feature embeddings are not helpful for these cases. However, the right plot shows that when randomly shuffling predictions of both EfficientNetV2-B3 and ConvNeXt-Tiny, the performance of the meta model shows a sign of decline in AUC-ROC, demonstrating that the meta model not only learns to look at predictions of each base model separately, it can also capture relationships between features by identifying patterns in combinations of predictions from multiple base models to make more accurate classification and this also happens similarly when shuffling any two base models' predictions most of the time. The group feature of all 4 base models has the highest value of feature importance of 0.239 while the 128-dimensional feature embeddings show a relatively lower feature importance of 0.05. Evidently, this disparity reveals that the meta model primarily relies on the collective predictions of the base models for its decision-making, while the feature embeddings serve a supplementary role by providing contextual information to refine these predictions.

The collective predictions of all four base models achieve the highest feature importance score, demonstrating the meta model's sophisticated capability to identify and exploit synergistic relationships between multiple model predictions rather than relying on individual outputs alone. Notably, the combination of GCViT+ConvNeXt-Tiny shows the second-highest group importance (0.130), suggesting these two architecturally diverse models (transformer and CNN) provide particularly complementary information when combined. The 128-dimensional CNN extracted features exhibit relatively low importance (0.050), confirming their supplementary role in providing contextual refinement to the primary decision-making process driven by base model predictions. Interestingly, most pairwise combinations of base models show minimal importance (0.002-0.034), except for GCViT combinations, reinforcing that GCViT's predictions are particularly valuable for the ensemble's performance. This hierarchical importance structure demonstrates that while individual CNN models may have limited standalone contribution, their collective integration with the transformer model creates a powerful synergistic effect that drives the ensemble's superior performance.

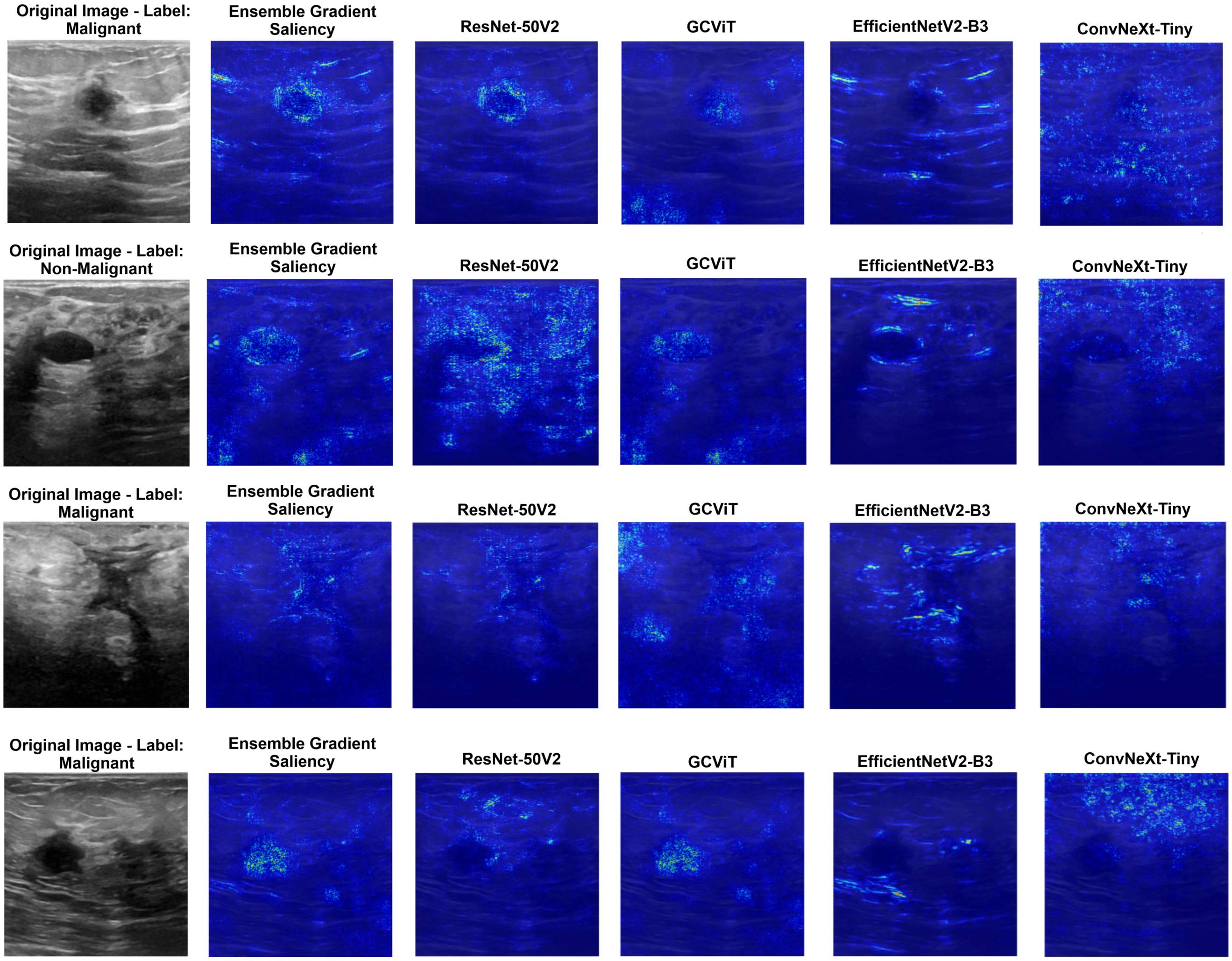
Original Image - Label: Malignant
Ensemble Gradient Saliency
ResNet-50V2
GCViT
EfficientNetV2-B3
ConvNeXt-Tiny
Original Image - Label: Non-Malignant
Ensemble Gradient Saliency
ResNet-50V2
GCViT
EfficientNetV2-B3
ConvNeXt-Tiny
Original Image - Label: Malignant
Ensemble Gradient Saliency
ResNet-50V2
GCViT
EfficientNetV2-B3
ConvNeXt-Tiny
Original Image - Label: Malignant
Ensemble Gradient Saliency
ResNet-50V2
GCViT
EfficientNetV2-B3
ConvNeXt-Tiny

Fig. 7. Saliency maps from the proposed Ensemble model and the base models for BreastMNIST samples.

TABLE III
EVALUATION METRICS OF ENSEMBLE MODELS WITH DIFFERENT FEATURE EXTRACTORS AND WITHOUT FEATURE EXTRACTION FOR BREASTMNIST DATASET

Feature Extractor	**ACC**	**AUC**	**Recall**			**Precision**			**F1 Score**		
			Malignant	Non-Malignant	Mean	Malignant	Non-Malignant	Mean	Malignant	Non-Malignant	Mean
N/A	0.8717	0.9371	0.9	0.86	0.88	0.7	0.96	0.83	0.79	0.91	0.85
EfficientNetV2-B3	**0.9423**	**0.9547**	**0.93**	**0.95**	**0.94**	**0.87**	**0.97**	**0.92**	**0.9**	**0.96**	**0.93**
ResNet-50V2	0.9038	0.9204	0.79	0.94	0.87	0.84	0.92	0.88	0.81	0.93	0.87
ConvNeXt-Tiny	0.9038	0.9252	0.83	0.93	0.88	0.82	0.94	0.88	0.83	0.94	0.89

TABLE IV
EVALUATION METRICS OF ENSEMBLE MODELS WITH DIFFERENT FEATURE DIMENSIONS EXTRACTED FROM EFFICIENTNETV2-B3 FOR BREASTMNIST DATASET

Feature Dimension	**ACC**	**AUC**	**Recall**			**Precision**			**F1 Score**		
			Malignant	Non-Malignant	Mean	Malignant	Non-Malignant	Mean	Malignant	Non-Malignant	Mean
32	0.936	0.941	0.88	0.96	0.92	0.88	0.96	0.92	0.88	0.96	0.92
64	0.9231	0.9327	0.83	0.96	0.9	0.88	0.94	0.91	0.85	0.95	0.9
128	0.9423	**0.9547**	**0.93**	0.95	**0.94**	0.87	**0.97**	0.92	0.9	0.96	0.93
256	**0.9487**	0.9317	0.86	**0.98**	0.92	**0.95**	0.95	**0.95**	0.9	**0.97**	**0.935**
512	0.9423	0.948	0.88	0.96	0.92	0.9	0.96	0.93	0.89	0.96	0.925

TABLE V
EVALUATION METRICS OF OUR PROPOSED ENSEMBLE MODEL WITH 128-DIMENSIONAL FEATURES EXTRACTED FROM EFFICIENTNETV2-B3, GCVIT, BASELINE VITS AND OTHER CNN-BASED MODELS FOR MELANOMA DATASET

Method	**ACC**	**AUC**	**Recall**			**Precision**			**F1 Score**		
			Benign	Malignant	Mean	Benign	Malignant	Mean	Benign	Malignant	Mean
DenseNet-121	0.938	0.9817	0.95	0.93	0.94	0.93	0.95	0.94	0.94	0.94	0.94
Xception	0.934	0.9842	0.97	0.9	0.935	0.9	0.97	0.935	0.94	0.93	0.935
InceptionV3	0.908	0.9684	0.89	0.93	0.91	0.93	0.89	0.91	0.91	0.91	0.91
VGG19	0.817	0.9777	**0.98**	0.65	0.815	0.74	**0.97**	0.855	0.84	0.78	0.81
MobileNetV3-Large	0.8	0.9496	0.97	0.63	0.8	0.73	0.95	0.84	0.83	0.76	0.795
NASNet-Large	0.886	0.9478	0.87	0.91	0.89	0.9	0.87	0.885	0.88	0.89	0.885
InceptionResNetV2	0.948	0.9853	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95
ResNet-50V2	0.924	0.9799	0.9	0.95	0.925	0.95	0.9	0.925	0.92	0.93	0.925
EfficientNetV2-B3	0.877	0.9455	0.83	0.92	0.875	0.91	0.85	0.88	0.87	0.88	0.875
ViT-Base-16 [44]	0.92	-	0.95	0.89	0.92	0.89	0.94	0.915	0.92	0.91	0.915
ConvNeXt-Base [44]	0.915	-	0.92	0.9	0.91	0.9	0.92	0.91	0.91	0.91	0.91
Swin-v2-s [44]	0.91	-	0.85	0.97	0.91	0.96	0.86	0.91	0.9	0.91	0.905
GCViT	0.953	**0.9891**	0.94	0.96	0.95	0.96	0.94	0.95	0.95	0.95	0.95
Proposed Ensemble Method	**0.954**	0.9871	0.93	**0.98**	**0.955**	**0.98**	0.93	**0.955**	**0.95**	**0.96**	**0.955**

B3. *Saliency maps*

Fig. 7 shows the saliency maps for the end-to-end ensemble model and each individual base model. While every base model highlights different parts of an images, the ensemble is able to integrate these diverse attention patterns and focus on the most discriminative features, resulting in more comprehensive and robust feature localization for accurate classification. This finding emphasizes how our stacking ensemble method can leverage the complementary strengths of different base models by combining their diverse attention mechanisms along with additional insights from feature embeddings to achieve a richer and more precise understanding of the underlying visual patterns.

In the first row (malignant case), GCViT demonstrates strong focus on the lesion region with relatively coherent attention, reflecting its superior ability to capture both local abnormalities and broader contextual information. ResNet-50V2 emphasizes fine-grained texture patterns and boundary details, yet its attention appears scattered and less concentrated on the lesion core. EfficientNetV2-B3 predominantly focuses on lesion edges and boundary transitions. On the other hand, ConvNeXt-Tiny exhibits more spatially distributed activations, reflecting strong sensitivity to fine-grained texture patterns across the image rather than highly localized lesion regions. The ensemble saliency map effectively integrates these behaviors, yielding a more compact and well-localized activation that clearly highlights the malignant mass and its irregular margins. In the second row (non-malignant case), GCViT again exhibits focused attention on the central lesion structure while maintaining awareness of surrounding context, consistent with its strong standalone performance. ResNet-50V2 and ConvNeXt-Tiny produce widespread activation across the image, indicating reduced specificity, whereas EfficientNetV2-B3 concentrates on localized edges and textures. By combining the behaviors of these four base models, the ensemble produces a clean, lesion-centered saliency map that aligns with the benign appearance. Similar trends also occur in the third and fourth rows.

By fusing GCViT's robust global-context understanding with EfficientNetV2-B3's strong sensitivity to lesion boundaries, ConvNeXt-Tiny's rich texture-oriented responses, and ResNet-50V2's stable structural representations, the ensemble produces saliency maps that are more spatially coherent and clinically interpretable than those of any individual model. These observations support the quantitative results and demonstrate that the proposed ensemble not only outperforms individual models in accuracy but also provides more consistent and reliable visual explanations for breast cancer classification.

C. *Performance comparison by feature extractors*

We also conduct additional experiments to train the ensemble model with and without feature embeddings extracted from different variants of CNN-based model and the results are shown in Table III. This analysis was conducted to evaluate how much additional value, which different CNN-based feature extractors contribute to the ensemble's performance, and to determine whether combining learned image embeddings with base model predictions leads to more accurate and robust classification than relying on model outputs alone. In Table III, the N/A row corresponds to a simpler prediction-only stacking baseline (i.e., using only base-model outputs without feature embeddings). When training the meta model without feature embeddings, it achieves 87.17% accuracy, 93.71% AUC, 88% mean recall, 83% mean precision, and 85% mean F1 score, which is a decent result compared to most of traditional CNN models but lower than the meta models trained with feature embeddings. This demonstrates that feature extractors provide additional context that enhances the meta model's ability to make accurate predictions rather than identifying the class solely based on the guesses from the base models. Furthermore, we found that using EfficientNetV2-B3 as the feature extractor yields the best performance result compared to ResNet-50V2 and ConvNeXt-Tiny.

D. *Performance comparison by feature embedding dimensions*

Understanding the optimal dimensionality for feature representations is crucial in ensemble learning systems, as it directly impacts both model performance and computational efficiency. To investigate the effect of feature embedding dimensions on our ensemble framework, we conducted systematic experiments using different dimensional outputs (32, 64, 128, 256 and 512 dimensions) from the EfficientNetV2-B3 feature extraction network on the BreastMNIST dataset. Table IV presents the comprehensive evaluation results across all tested dimensions. The results reveal that 256-dimensional features achieve the highest overall accuracy of 94.87%, closely followed by 128-dimensional features at 94.23%. This suggests that the optimal feature dimension for our ensemble framework lies within the 128-256 range, where the feature extractor captures sufficient discriminative information without introducing excessive noise or redundancy. The 32-dimensional and 64-dimensional feature representations show decreased performance (93.6% and 92.31% accuracy respectively), indicating that insufficient feature capacity limits the ensemble's ability to capture complex patterns in breast cancer ultrasound images. Interestingly, the 512-dimensional features (94.23% accuracy) do not outperform the 256-dimensional variant, suggesting the presence of diminishing returns and potentially introducing noise that affects the meta-learner's decision-making capability. This phenomenon is consistent with the curse of dimensionality [45], where excessively high-dimensional features can lead to sparse data representation and reduced generalization.

E. Generalization on an additional dataset

To further assess the robustness, adaptability, and generalizability of our framework, we also evaluated it on the melanoma cancer image dataset [46]. This publicly available dataset includes thousands of high-quality, labeled dermoscopic images of skin lesions used to differentiate between benign and malignant melanoma cases. These images were collected from diverse sources, representing a wide range of lesion appearances, skin types, and image acquisition conditions. Each image is accompanied by a binary label indicating the presence or absence of melanoma. By including this dataset, we aim to assess the effectiveness of our ensemble model across different cancer types and imaging modalities, ultimately demonstrating its potential adaptability for broader medical image classification tasks. The training procedure for both the base models and the meta-model follows the setup used for the BreastMNIST dataset.

Table V shows the performance comparison of our proposed ensemble model against GCViT, baseline vision transformers and various CNN-based architectures on the melanoma dataset. The results reveal several notable findings that validate the effectiveness of our ensemble approach in different medical imaging domains. Our proposed ensemble method achieves exceptional performance with 95.4% accuracy and 98.71% AUC-ROC, demonstrating its strong capability for melanoma classification. Particularly impressive is the model's balanced performance across both benign and malignant classes, with mean precision, recall, and F1-score all reaching 95.5%. This balanced performance is crucial in medical applications where both false positives and false negatives carry significant clinical consequences. In addition, Our ensemble model achieves 98% malignant recall, which is clinically critical as it indicates the model's ability to correctly identify 98% of actual melanoma cases.

Among individual models, GCViT demonstrates superior performance with 95.3% accuracy and the highest AUC-ROC of 98.91%, confirming its effectiveness as a state-of-the-art vision transformer for medical image analysis. The strong performance of GCViT validates our choice of including this advanced transformer architecture as a key base model in our ensemble approach. Traditional CNN models show varied performance levels, with InceptionResNetV2 achieving the best results among CNNs (94.8% accuracy, 98.53% AUC), followed closely by DenseNet-121 (93.8% accuracy, 98.17% AUC). Interestingly, some models like VGG19 and MobileNetV3-Large exhibit significant performance class imbalance issues, achieving high recall for benign cases (98% and 97% respectively) but poor malignant recall (65% and 63% respectively).

VI. Conclusion

In this study, we proposed a feature-enhanced stacking ensemble framework for breast cancer classification in ultrasound images. The framework integrates GCViT with complementary CNN architectures (ResNet-50V2, ConvNeXt-Tiny, and EfficientNetV2-B3). By exploiting architectural diversity and combining prediction-level outputs with deep feature embeddings through a trainable meta model, the proposed ensemble achieves 94.23% accuracy and 95.47% AUC-ROC on BreastMNIST, with cross-validation on melanoma dataset (95.4% accuracy) demonstrating robustness across cancer types. It consistently performs better individual models and several other CNNs across multiple evaluation metrics. Key contributions of this work include the effective integration of GCViT's global-context attention with CNN-based models that capture boundary, texture, and structural patterns, as well as the incorporation of feature-level representations extracted from a dedicated CNN model to enhance decision-making at the ensemble level. The proposed approach demonstrates superior performance on both breast ultrasound and melanoma datasets, highlighting its robustness and cross-dataset generalizability. Furthermore, comprehensive explainability analyses using SHAP, permutation feature importance, and saliency maps provide transparent insights into the ensemble's behavior and confirm that the meta model adaptively balances contributions from all base models and feature embeddings. It reveals that GCViT carries the highest decision weight, validating transformer effectiveness in medical imaging; feature embeddings contribute to base predictions; and attention maps show how the model makes its prediction to support radiological diagnostics. By maintaining balanced sensitivity and specificity, this framework provides a practical and interpretable solution for clinical diagnostic support across a range of medical imaging tasks.

References

[1] M. Arnold *et al.*, "Current and future burden of breast cancer: Global statistics for 2020 and 2040," *The Breast*, vol. 66, pp. 15–23, Dec. 2022.

[2] L. F. Ellison and N. Saint-Jacques, "Five-year cancer survival by stage at diagnosis in Canada," *Health Rep.*, vol. 34, no. 1, pp. 3–15, Jan. 2023.

[3] R. Sood *et al.*, "Ultrasound for breast cancer detection globally: A systematic review and meta-analysis," *J. Glob. Oncol.*, vol. 5, no. 5, pp. 1–17, Aug. 2019.

[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, Lake Tahoe, Nevada, USA, Dec. 2012, pp. 1097–1105.

[5] W. Tang, J. Sun, S. Wang, and Y. Zhang, "Review of AlexNet for medical image classification," *EAI Endorsed Trans. e-Learn.*, vol. 9, no. 1, Dec. 2023.

[6] M. Liu, L. Hu, Y. Tang, C. Wang, Y. He, C. Zeng, K. Lin, Z. He, and W. Huo, "A deep learning method for breast cancer classification in the pathology images," *IEEE J. Biomed. Health Inform.*, vol. 26, no. 10, pp. 5025–5032, Oct. 2022.

[7] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Las Vegas, NV, USA, Jun. 2016, pp. 770–778.

[8] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Machine Learning (ICML)*, Long Beach, CA, USA, Jun. 2019, pp. 6105–6114.

[9] Z. Liu *et al.*, "A ConvNet for the 2020s," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, New Orleans, LA, USA, Jun. 2022, pp. 11 966–11 976.

[10] A. Dosovitskiy *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learning Representations (ICLR)*, May 2021.

[11] C. Meng, W. Lin, B. Liu, H. Zhang, Z. Gan, and C. Ouyang, "RTS-ViT: Real-time share vision transformer for image classification," *IEEE J. Biomed. Health Inform.*, vol. 29, no. 5, pp. 3576–3586, May 2025.

[12] G. Wang, Q. Zhu, C. Song, B. Wei, and S. Li, "Medkaformer: When Kolmogorov–Arnold theorem meets vision transformer for medical image representation," *IEEE J. Biomed. Health Inform.*, vol. 29, no. 6, pp. 4303–4313, Jun. 2025.

[13] H. Q. Vo *et al.*, "Frozen large-scale pretrained vision-language models are the effective foundational backbone for multimodal breast cancer prediction," *IEEE J. Biomed. Health Inform.*, vol. 29, no. 5, pp. 3234–3246, May 2025.

[14] L. Liu, Y. Wang, P. Zhang, H. Qiao, T. Sun, H. Zhang, X. Xu, and H. Shang, "Collaborative transfer network for multi-classification of breast cancer histopathological images," *IEEE J. Biomed. Health Inform.*, vol. 28, no. 1, pp. 110–121, Jan. 2024.

[15] N. Tsiknakis *et al.*, "Unveiling the power of model-agnostic multiscale analysis for enhancing artificial intelligence models in breast cancer histopathology images," *IEEE J. Biomed. Health Inform.*, vol. 28, no. 9, pp. 5312–5322, Sept. 2024.

[16] V. Tsafas, I. Oikonomidis, E. Gavgiotaki, E. Tzamali, G. Tzedakis, C. Fotakis, I. Athanassakis, and G. Filippidis, "Application of a deep-learning technique to non-linear images from human tissue biopsies for shedding new light on breast cancer diagnosis," *IEEE J. Biomed. Health Inform.*, vol. 26, no. 3, pp. 1188–1195, March 2022.

[17] B. Wang, X. Hu, J. Zhang, C. Xu, and Z. Gao, "Intelligent internet of things in mammography screening using multicenter transformation between unified capsules," *IEEE Internet Things J.*, vol. 10, no. 2, pp. 1536–1545, Jan. 2023.

[18] A. Kumar, A. Sharma, V. Bharti, A. K. Singh, S. K. Singh, and S. Saxena, "Mobihisnet: A lightweight cnn in mobile edge computing for histopathological image classification," *IEEE Internet Things J.*, vol. 8, no. 24, pp. 17 778–17 789, Dec. 2021.

[19] M. Bohra, K. U. Singh, I. Kumar, and S. Mishra, "Multiresolution wavelet packet-driven dual path cnn for breast lesion classification," *IEEE Internet Things J.*, vol. 12, no. 21, pp. 44 750–44 762, Nov. 2025.

[20] A. Alnemer and J. Rasheed, "An efficient transfer learning-based model for classification of brain tumor," in *Proc. Int. Symp. Multidisciplinary Studies and Innovative Technologies (ISMSIT)*, Ankara, Turkey, Oct. 2021, pp. 478–482.

[21] O. S. Khedr, M. E. Wahed, A.-S. R. Al-Attar, and E. A. Abdel-Rehim, "The classification of the bladder cancer based on vision transformers (ViT)," *Sci. Rep.*, vol. 13, no. 1, p. 20639, Nov. 2023.

[22] A. Hatamizadeh, H. Yin, G. Heinrich, J. Kautz, and P. Molchanov, "Global context vision transformers," in *Proc. Int. Conf. Machine Learning (ICML)*, Honolulu, Hawaii, USA, Jul. 2023, pp. 12 633–12 646.

[23] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. Int. Conf. Neural Information Processing Systems (NeurIPS)*, Long Beach, California, USA, December 2017, p. 4768–4777.

[24] A. Fisher, C. Rudin, and F. Dominici, "All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously," *J. Mach. Learn. Res.*, vol. 20, no. 177, p. 181, 2019.

[25] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," *arXiv preprint arXiv:1312.6034*, Dec. 2013.

[26] Y. K. Aluri, A. Nellore, N. S. Kaza, N. B. M. Pajjuru, and V. N. S. A. Palnati, "Histopathological image-based breast cancer detection using deep learning models," in *Proc. Int. Conf. Trends in Electronics and Informatics (ICOEI)*, Tirunelveli, India, Apr. 2025, pp. 1498–1504.

[27] V. R. Patheda, G. Laxmisai, B. V. Gokulnath, S. P. Siddique Ibrahim, and S. Selva Kumar, "A robust hybrid CNN+ViT framework for breast cancer classification using mammogram images," *IEEE Access*, vol. 13, pp. 77 187–77 195, Apr. 2025.

[28] S. Garg and P. Singh, "Transfer learning based lightweight ensemble model for imbalanced breast cancer classification," *IEEE/ACM Trans. Comput. Biol. Bioinform.*, vol. 20, no. 2, p. 1529–1539, May 2022.

[29] S. Sasidharan Nair and M. Subaji, "Automated identification of breast cancer type using novel multipath transfer learning and ensemble of classifier," *IEEE Access*, vol. 12, pp. 87 560–87 578, Jun. 2024.

[30] A. Manna, R. Kundu, D. Kaplun, A. Sinitca, and R. Sarkar, "A fuzzy rank-based ensemble of cnn models for classification of cervical cytology," *Sci. Rep.*, vol. 11, no. 1, p. 14538, Jul. 2021.

[31] W. Zhang *et al.*, "Dynamic-fusion-based federated learning for covid-19 detection," *IEEE Internet Things J.*, vol. 8, no. 21, pp. 15 884–15 891, Nov. 2021.

[32] N. Rashid, T. Mortlock, and M. A. A. Faruque, "Stress detection using context-aware sensor fusion from wearable devices," *IEEE Internet Things J.*, vol. 10, no. 16, pp. 14 114–14 127, Aug. 2023.

[33] S. Kumar, A. V. Shvetsov, and S. H. Alsamhi, "Empowering remote healthcare with federated learning for early diagnosis of pulmonary disease," *IEEE Internet Things J.*, vol. 12, no. 13, pp. 23 288–23 296, Jul. 2025.

[34] ——, "Fuzzyguard: A novel multimodal neuro-fuzzy framework for copd early diagnosis," *IEEE Internet Things J.*, vol. 12, no. 8, pp. 9627–9637, Apr. 2025.

[35] S. Abdelfattah, M. M. Badr, M. M. E. A. Mahmoud, K. Abualsaud, E. Yaacoub, and M. Guizani, "Efficient and privacy-preserving cloud-based medical diagnosis using an ensemble classifier with inherent access control and micro-payment," *IEEE Internet Things J.*, vol. 10, no. 24, pp. 22 096–22 110, Dec. 2023.

[36] D. H. Wolpert, "Stacked generalization," *Neural Netw.*, vol. 5, no. 2, pp. 241–259, 1992.

[37] J. Yang *et al.*, "MedMNIST v2 - a large-scale lightweight benchmark for 2D and 3D biomedical image classification," *Sci. Data*, vol. 10, no. 1, p. 41, Jan. 2023.

[38] W. Al-Dhabyani, M. Gomaa, H. Khaled, and A. Fahmy, "Dataset of breast ultrasound images," *Data Br.*, vol. 28, p. 104863, Feb. 2020.

[39] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 318–327, Feb. 2020.

[40] L. Breiman, "Stacked regressions," *Mach. Learn.*, vol. 24, no. 1, pp. 49–64, Jul. 1996.

[41] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. Int. Conf. Learning Representations (ICLR)*, New Orleans, LA, USA, May 2019.

[42] A. Halder, S. Gharami, P. Sadhu, P. K. Singh, M. Woźniak, and M. F. Ijaz, "Implementing vision transformer for classifying 2D biomedical images," *Sci. Rep.*, vol. 14, no. 1, p. 12567, May 2024.

[43] O. N. Manzari, H. Ahmadabadi, H. Kashiani, S. B. Shokouhi, and A. Ayatollahi, "Medvit: A robust vision transformer for generalized medical image classification," *Comput. Biol. Med.*, vol. 157, p. 106791, May 2023.

[44] S. Aksoy, P. Demircioglu, and I. Bogrekci, "Enhancing melanoma diagnosis with advanced deep learning models focusing on vision transformer, swin transformer, and ConvNeXt," *Dermatopathology (Basel)*, vol. 11, no. 3, pp. 239–252, Aug. 2024.

[45] E. Keogh and A. Mueen, *Curse of Dimensionality*. Boston, MA, USA: Springer US, 2017.

[46] B. Mittal, "Melanoma Cancer Image Dataset," Kaggle, Feb. 2024. [Online]. Available: https://www.kaggle.com/datasets/bhaveshmittal/melanoma-cancer-dataset