GMR
The Original

ADVANCED COMPUTATIONAL IDENTIFICATION OF BREAST CANCER BIOMARKERS: SYNERGIZING METAHEURISTIC OPTIMIZATION WITH MULTI-MODAL VISION TRANSFORMERS

S.Sowjanya[1], M. Sunil Kumar[2]

[1]Research Scholar Department of CSE School of Computing, Mohan Babu University Tirupati, AP, India,
E-mail: sunchusowjanya66@gmail.com
[2]Professor Dept of Computer Science and Engineering School of Computing, Mohan Babu University, (Erstwhile Sree Vidyanikethan Engineering College(Autonomous), Tirupati, Andhra Pradesh, India, Email: sunilmalchi1@gmail.com

ABSTRACT

Medical image processing to identify breast cancer is particularly challenging since the imprecision can be a result of a number of varying factors such as the specific modality, the deep learning model, or low interpretability of the model, which can be problematic for real-world consequence setups. Mammography and ultrasound imaging are the primary methods and most frequently utilized imaging tools for breast cancer screening. The most difficult problem in using both technologies is drawing the right conclusions. In this paper, we introduce MMViT-Net, a multi-modality vision transformer ensemble for breast cancer imaging and a metaheuristic vision transformer. MMViT-Net uses both mammography and ultrasound images for breast cancer screening. Ensembling the distance variations a dual modality adaptive preprocessing and attention U-Net lesion segmentation will be able to localize the tumor regions and suppress the irrelevant background regions. Each hybrid modality of the encoders, egg, and vision of the transformers are coupled to capture finer and more granular attributes while elongated global contextual encoders portray the vastness of the context and the encoders. Robust cross-attention modality-aware feature fusion is proposed to improve the internal cooperation of the various modalities. The model hyperparameters are automatically adjusted by a fusion of the Grey Wolf Optimizer and Harris Hawks Optimization. This McDo analysis algorithm is utilized to optimize the trade-off between reducing false positives and increasing the positive diagnostical certainties. Results from the analyses performed on the ultrasound database, BUSI, and the mammography database, CBIS DDSM, show imposing evidence of strong performance, cross-dataset generalization, and demonstrable improvements in construction to many existing CNN and CNN-Transformer methods. MMViT-Net is positioned as a solid and clinically viable model for AI-assisted screening and diagnosis of breast cancer. This is due to the assurance of model interpretability and the support of transparent decision-making by the highlighting of clinically relevant lesion areas through Grad-CAM visual explanations.

1. INTRODUCTION

Every year breast cancer is the most common cancer diagnosed in women globally and contributes to a large number of cancer deaths [1]. Detecting the cancer in its earlier stages is the most effective way to treat and improve the five-year survival [2]. Clinically, the cancer is diagnosed using a combination of different modalities, which includes, but is not limited to, Mammography, Ultrasound, and Histopathology. Each of these offers different insights and adds to the understanding of the tissue [3]. Imaging can however be subjected to errors because the reviews of the modalities are all manual, and can be error prone and lead to unnecessary biopsies due to the fatigue of the reviewer causing them to miss things [4]. Computer-Aided Diagnosis (CAD) systems attempt to solve this problem; however, many deep learning models are not successful because of the complexity of breast lesions and the different resolutions of the various imaging modalities [5, 6].

Automated systems have had a persistent problem in balancing local detail with global context [7]. While Convolutional Neural Networks (CNNs) are great at detail-oriented analysis, they have a limited receptive field, a trait that does not apply for Vision Transformers (ViTs) because they handle long-range dependencies, although they do require a significant amount of data to converge [8, 9]. Black box AI models retain opacity, widening the trust gap between this technology and its use in practice [10]. This effective diagnosis can be achieved through frameworks that are not only diagnostically accurate but also have plausible rationale for their conclusions [11].

This paper proposes a hybrid CNN-Vision Transformer model that makes use of metaheuristic techniques to present accurate and explainable breast cancer tissue detection on 2 heterogeneous datasets, BUSI and CBIS-DDSM. We propose a framework that incorporates EfficientNet-V2 to capture localized features while utilizing the global context via lateral looking ViT-B/16 interwoven through cross-attention. In addition, Grey Wolf Optimizer and Harris Hawks Optimization dual metaheuristic is developed to further improve the performance of the model. A full preprocessing pipeline was also included within the model to establish its clinical reliability in addition, explainable AI (XAI) tools such as Grad-CAM, SHAP and LIME were used to provide transparency into mechanisms underlying the model's predictions.

2. LITERATURE SURVEY

Recent works have developed end-to-end deep learning frameworks for diagnostic breast cancer [12] that lack emphasis on classical manual engineering feature processes. The earlier computational methods were mainly based on feature extraction shape analysis and descriptive outlining [42, 43]. The field was changed by the development of Convolutional Neural Networks (CNNs): high level models that can automatically learn spatial hierarchical features [13, 41]. For specific applications [14, 15], existing architectures such as ResNet, VGG and Inception have been repurposed for histopathological image classification or mammogram mass detection (CBIS-DDSM). Although Standard CNNs have achieved incredible success, they are also criticized for their localized receptive fields which overlook important structural contexts of the lesion. When assessing the pathology of a mass, especially in complex ultrasound backgrounds like the BUSI dataset [16, 17], an anatomical scaffold is critical in describing whether the specific mass is growing towards benign or malignant aspects. While the limitations of ViTs (vision transformers) [14, 16] in earlier works are partially relieved through self-attention and measuring long-range dependencies on image patches [18]. This contrasts with CNNs (and most other neural network architectures) which pass images through local kernels and sequence input as sequences of tokens, allowing them to reason about the whole image and all its global features [19]. More recent studies suggest that ViTs' global abstraction is absent of CNNs edge and texture sensitivity (low-level biases) on which much imaging work depends [20]. This is very likely a contributing factor to the increased use of hybrid architectures that utilize a CNN backbone (e.g., EfficientNet-V2) to first magnify high resolution local features present in a dataset and later give them to be globally fused by, for example, Transformer [21, 22]. This hybrid approach empowers the models to maintain essential micro-features of histopathology images and significant geometric structures of mammograms [23]. The three recurring issues about medical AI are data scarcity and quality. Data quality issues, such as noise and class imbalance, are found in these public datasets (e.g., BreakHis,BUSI). Advanced preprocessing techniques such as Non-Local Means Filtering and Adaptive Median Denoising preserve edge while denoising artifacts. Furthermore, methods such as Contrast Limited Adaptive Histogram Equalization (CLAHE) and Gamma Correction can alleviate the visibility of micro-calcifications in mammography. We use Attention U-Net, which is a state-of-the-art segmentation tool and helps to localize the tumor by reducing any irrelevant background noise.

To resolve the challenges caused by class imbalance (benign samples outnumber malignant), creators generated synthetic data. Oversampling followed by smote gan, a type of neural network has been used to obtain realistic pathology samples some datasets, and data augmentation methods such as RandAugment, MixUp and CutMix improve generalization during training; they work preventing the models from memorizing the training dataset.

The high-dimensional hybrid models we described require a great deal of effort to produce. In hyperparameter search, such as learning rates and dropout ratios, conventional gradient-based methods get stuck in local optima [32]. In this regard, metaheuristic techniques based on natural phenomena provide a strong alternative for global search [33]. Several variants of metaheuristic algorithms are present in the published literature; among these, two that have been dubbed best performing methods for AUC maximization of deep learning hyperparameter tuning are known as Grey Wolf Optimizer (GWO), adapted from the social hierarchy behavior of wolf packs, and Harris Hawks Optimizer (HHO), which simulates hawk's collaborative "surprise pounce" hunting tactic [34, 35].

To gain clinical acceptance of AI, there needs to be some level of explainability to the users. AI models are now required to have some form of explainable AI (XAI) framework integrated to justify their outcomes [36]. Of the many XAI options available, Grad-CAM, SHAP, and LIME are some of the most frequently used. Grad-CAM produces a saliency (or heat) map to show which areas of the image were most important for the classification, while SHAP and LIME capture the contribution of individual features to the overall prediction or diagnosis [37, 38]. These techniques help to eliminate the potential distrust instances where models could be using a form of "cheat" to classify based on image artifacts rather than real pathology by helping the radiologist position themselves to confirm the model is identifying pathology [39, 40].

3. PROBLEM STATEMENT

A multi-modal breast imaging dataset is given as $\mathcal{D} = \{(x_i^m, x_i^u, y_i)\}_{i=1}^N$, with $x_i^m \in \mathbb{R}^{H\times W\times C}$ and $x_i^u \in \mathbb{R}^{H\times W\times C}$ being mammography and ultrasound images respectively, and $y_i \in \{0,1\}$ indicating benign or malignant labels. The goal is to learn an interpretable and robust function $f_\Theta: (x^m, x^u) \to [0,1]$ predictions on the probability of a given input being malignant. The current models based on CNN and ViT architectures demonstrate a lack of generalization across various modalities, as well as overfitting to the selection of hyperparameters and a reduction of the sensitivity of interpretability, especially with the class imbalance $|\mathcal{D}_{benign}| \gg |\mathcal{D}_{malignant}|$. The manual selection of hyperparameters is computationally expensive and even less than optimally designed. Thus, the problem is described as the joint learning of a hybrid feature representation, the cross-attention fusion coupled with the metaheuristic search for optimal hyperparameter, and the transparent decision explanation as:

$$\Theta^* = \arg\max_{\Theta} \left(\mathrm{AUC}(f_\Theta(\mathcal{D})) - \lambda\, \mathrm{FP}_{rate}(f_\Theta(\mathcal{D}))\right),$$

for the stated robustness along with the interpretability constraints, where λ is sensitivity and clinical safety.

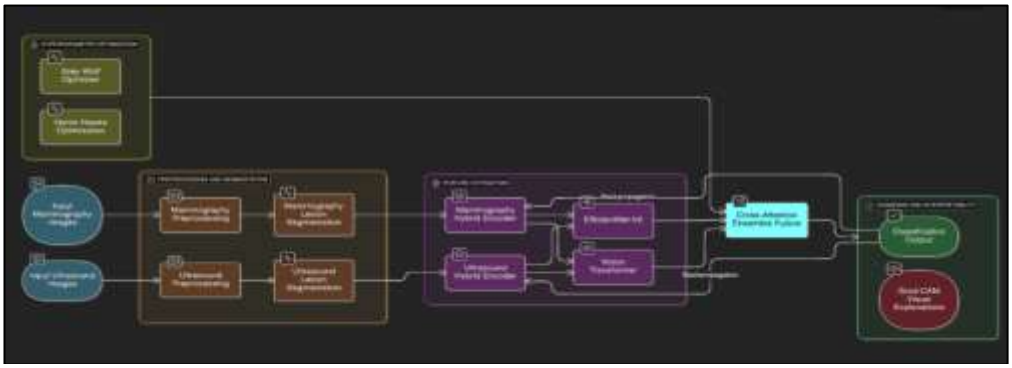

Fig. 1. MMVIT-Net Model Diagram

4. METHODOLOGY

The proposed framework combines sophisticated preprocessing, tumor segmentation, hybrid deep feature extraction, cross-attention fusion, and metaheuristic hyperparameter optimization for explainable and robust breast cancer detection from multisource medical images. The overall pipeline as in figure consists of: (i) modality-adaptive preprocessing and enhancement, (ii) lesion segmentation using Attention U-Net, (iii) hybrid CNN-Vision Transformer (ViT) feature extraction, (iv) cross-attention-based multimodal fusion, and (v) hyperparameter optimized classification and explainability generation. Each component is described in detail below.

4.1 Preprocessing and Image Enhancement

Breast cancer images from various modalities, such as mammography (CBIS-DDSM) and ultrasound (BUSI), have notable discrepancies with regards to noise patterns, contrast, and illumination. So a unified preprocessing pipeline is performed to enhance modality invariant feature learning and stabilize the downstream hybrid CNN-Transformer model. The various building blocks of this pipeline were noise removal, contrast equalization, illumination pattern correction, and final intensity rescaling.

4.1.1 Noise Reduction

Place two sequential filters. First, Non-Local Means (NLM) targets Gaussian noise and preserves small lesion details. The intensity of a denoised pixel is determined as:

$$\hat{I}(x) = \frac{1}{Z(x)} \sum_{y \in \Omega} \exp\left(-\frac{\| P(x) - P(y) \|_2^2}{h^2}\right) I(y),$$

where $P(x)$ is the local patch at pixel x and h is a variable that controls the degree of smoothing. After that, Adaptive Median Filtering helps with denoising impulse noise, which is a frequent issue with ultrasound images. For each window:

$$A_1 = z_{\text{med}} - z_{\min}, A_2 = z_{\text{med}} - z_{\max},$$

If $A_1 > 0$ and $A_2 < 0$, the median is valid; otherwise, the window expands until a clean estimate is acquired.

4.1.2 Contrast Enhancement

Local contrast enhancement and spotlighting of subtle tumor margins are achieved with the application of CLAHE to the luminance channel constrained by:

$$I_L' = CLAHE(I_L),$$

Which enhances contrast within small tiles while keeping noise from amplifying. This benefit is especially important for mammograms that have low contrast.

4.1.3 Gamma Correction

Illumination inconsistencies are corrected using gamma adjustment:

$$I_\gamma(x) = I(x)^{1/\gamma}, \gamma = 1.2.$$

This approach brightens mid-level intensities and enhances lesion visibility across all modalities.

4.1.4 Normalization and Resizing

All preprocessed images are resized to 224×224 pixels and standardized using:

$$I_{norm} = \frac{I_\gamma - \mu}{\sigma},$$

where μ and σ are the mean and standard deviation for the dataset. This preprocessing approach produces images that are more illumination-normals, cleaner, and with a better contrast which also enhances the segmentation capability and the representation quality for the following CNN-ViT feature extraction.

4.2 Tumor Segmentation Using Attention U-Net

Accurate localization of lesion regions enhances background noise in the downstream classification and steers the classifier's focus to areas of interest that are diagnostically significant. To obtain flexible segmentation across multiple modalities, a segmentation method based on Attention U-Net is used. This method is an improvement over the classical U-Net, as it uses attention gates to increase focus on certain regions and omit other areas, such as fatty tissue, ducts, and imaging artefacts.

U-Net Attention employs the typical encoder–decoder structure. In this case, the input to the model is the image I. The encoder extracts feature representations at multiple resolutions, such as $E_1, E_2, \ldots, E_n$, The decoder reconstructs a segmentation map through a series of up sampling steps. Attention gates, as presented, add a differentiated value to the skip connections by assessing the value of each spatial location as a function of some pre-determined importance

$$\alpha = \sigma(W_g g + W_x x + b),$$

Where, x is representative of the feature map of the encoder, g describes the gating signal from the decoder, $W_x$are the parameters of the linear mappings, and σ is representative of a sigmoid activation function.

As such, the gated skip feature can then be expressed as: $\tilde{x} = \alpha \odot x,$

This allows relevant segments of the tumor to be passed on to the later stages of decoding. The Attention U-Net generates a probability mask $\widehat{M}$ which is given as$\widehat{M} = \sigma(f_\theta(I))$, where $f_\theta$represents the Attention U-Net, and σ is the sigmoid output function. The model is trained using a hybrid loss that has a combination of the region overlap and pixel-wise accuracy, which is given by the equation:

$$\mathcal{L}_{seg} = (1 - \frac{2\sum(M \cdot \widehat{M})}{\sum M + \sum \widehat{M} + \varepsilon}) + \text{BCE}(M, \widehat{M}),$$

Where M is the ground truth of the mask, BCE is the binary cross entropy, and ε is a small constant that prevents the denominator from being zero. The segmentation step produces masks that focus on the lesion. These masks are used to crop tumors and to suppress areas with irrelevant tissues. The refined inputs improve the classification and accuracy and also enhance the strong explanation features to the model by focusing on clinically relevant areas.

4.3 Hybrid Feature Extraction (EfficientNet-V2 + ViT-B/16)

Imaging of breast cancer requires fine-grained local details, such as the morphology of the cells, microcalcifications, edges of lesions, and gross contextual clues like tissue structure, mass distribution, and architectural distortion. To address this, the proposed framework uses two prominent feature extractors, EfficientNet-V2, and ViT-B/16. EfficientNet-V2 is a convolutional neural network, and ViT-B/16 is a self-attention-based architecture.

4.3.1 Local Feature Extraction with EfficientNet-V2

EfficientNet-V2 uses the progressive scaling method, and fused MBConv layers capture comprehensive hierarchical patterns. For a given image I'that has gone through preprocessing and segmentation, the CNN encoder creates a representation of depth.

$$F_{\text{CNN}} = E_{\text{CNN}}(I'),$$

where$F_{\text{CNN}} \in \mathbb{R}^{d_c}$, which captures the fine textures and the critical components of the structure that are essential to the recognition of the borders of the lesions, variations in the nuclei, and irregularities in the tissue. A global average pool compacts the feature maps and retains important semantics. EfficientNet-V2 has been pre-trained on an ImageNet dataset and then fine-tuned on the breast imaging dataset to enhance cross-modality transfer learning.

4.3.2 Global Feature Extraction with Vision Transformer (ViT-B/16)

The input image is divided into 16 × 16 patches. Each patch is flattened and transformed into a latent embedding. In this way, long-range dependencies and contextual structures, which are often missed by CNNs, can be learned.

$$z_0 = [x_1 W_E;\ x_2 W_E; \ldots . x_n W_E] + \ E_{pos},$$

Where, $x_i$refers to the i-th image patch, $W_E$is a learned linear projection, and E_{pos}encodes the positional information. The Transform encoder modifies these tokens through the mechanism of multi-head self-attention.

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^\top}{\sqrt{d_k}})V,$$

This mechanism captures global interaction with patches, irrespective of spatial distance from one another. The image context is represented by the final classification (CLS) token which is described as:

$$F_{\text{ViT}} = z_{\text{CLS}}^{(L)},$$

where$z_{\text{CLS}}^{(L)}$is the CLS token post the final Transformer layer. This allows the network to produce a multi-scale feature representation by encoding both types of features from an identical preprocessed input, which aids in improving classification resilience across imaging modalities.

4.4 Cross-Attention Fusion Mechanism

While EfficientNet-V2 and ViT-B/16 both provide powerful local features and rich global representations, such complementary features are best used by combining them into a single representation. To do so, we employ a cross-attention fusion module that allows the model to choose and combine local descriptors from CNNs with global context from transformers. This process allows the representation focus on meaningful places in that image without losing the overall structure.

Let $F_{\text{CNN}} \in \mathbb{R}^{d_c}$ represent the local feature vector derived from EfficientNet-V2 and $F_{\text{ViT}} \in \mathbb{R}^{d_t}$represent the global feature vector from the ViT encoder. Prior to fusion, both vectors are projected onto a common latent dimension, denoted asd_f:

$$Q = W_q F_{\text{CNN}}, K = W_k F_{\text{ViT}}, V = W_v F_{\text{ViT}},$$

where$W_q, W_k, W_v \in \mathbb{R}^{d_f \times d_*}$are learnable linear transformations.

The cross-attention output is computed as: $F_{\text{CA}} = \text{softmax}(\frac{QK^{\text{T}}}{\sqrt{d_f}})V$,

where the CNN features (queries) can selectively attend to the relevant transformer features (keys and values). This construction ensures that the local characteristics of the lesion are augmented by the global context provided by the ViT. To balance the contribution of each branch, a learnable fusion coefficient, denoted as α ∈[0,1], is applied such that the final fused representation is given by:

$$F_{\text{fusion}} = \alpha F_{\text{CNN}} + (1 - \alpha)F_{\text{CA}}.$$

The model homes in on specific local patterns during training when α is near 1, but local patterns are deprioritized when α is near 0. Instead, the model focuses on the global context. α is modified as training progresses and is specific to the particular modality, lesion type, and image complexity.

4.5 Classification Module

The final step after obtaining the fused feature representation from the cross-attention mechanism is to classify the image as benign or malignant. The embedding $F_{\text{fusion}}$is a representation of the lesion with fine-grained detail, as well as the global context. This makes it a suitable candidate to be processed through a compact yet expressive fully connected classification head.

The last component of the model is the classification module, which has finely tuned class probabilities. It consists of one more two-stage MLP which has one final sigmoid layer. The first stage projects the fused feature vector to an intermediate high-level latent representation. Let us define the first stage as:

$$H_1 = \text{ReLU}(W_1 F_{\text{fusion}} + b_1),$$

The fusion is done at the feature level. Here$W_1 \in \mathbb{R}^{512 \times d_f}$serves as the first fusion dimension to the 512 space of the first layer latent dimensions. To reduce overfitting, especially important in medical imaging with limited samples—a dropout layer with probability $p_d$is applied: $H_1' = \text{Dropout}(H_1)$.The second-stage output and the first-stage representation are further refined through another dense transformation. Let us define this as: $H_2 = \text{ReLU}(W_2 H_1' + b_2)$,where $W_2 \in \mathbb{R}^{256 \times 512}$. A second dropout layer further prevents co-adaptation and improves generalization across datasets. Finally, the classifier outputs a probability score reflecting malignancy likelihood:

$$\hat{y} = \sigma(W_3 H_2 + b_3),$$

Where, $\hat{y} \in [0,1]$is the predicted malignancy probability, and σdenotes the sigmoid activation function. To train the classifier, we minimize the binary cross-entropy loss:

$$\mathcal{L}_{cls} = -[y\log(\hat{y}) + (1 - y)\log(1 - \hat{y})],$$

where$y \in \{0,1\}$is the ground-truth label. To handle class imbalance and emphasize harder malignant examples, focal loss can optionally be incorporated:

$$\mathcal{L}_{focal} = -\alpha(1 - \hat{y})^{\gamma} y\log(\hat{y}) - (1 - \alpha)\hat{y}^{\gamma}(1 - y)\log(1 - \hat{y}),$$

with focusing parameter $\gamma > 0$.

Through the balanced modeling of negative samples with sparse dropout, the optimized robustness-formulated MLP affords itself as a generalizable classifier across adverse imaging conditions and retains sensitivity to malignant lesions. The proposed framework discriminates between regions of interest in the final prediction stage through this classification head and segmentation preprocessing with hybrid CNN-Transformer features.

4.6 Metaheuristic Hyperparameter Optimization (Condensed Version)

The performance of a hybrid CNN-Transformer architecture is dependent on hyperparameter values (like learning rates, batch size, dropout probabilities, attention head counts and sizes of convolutional kernels) This framework thereby relies on metaheuristic optimization namely Grey Wolf (GWO) and Harris Hawks Optimization (HHO) instead of manual tuning to discover the optimal values and plug them. In the framework, each possible solution is represented as a vector S =[η,B,pd,H,k], where η, B, pd, H and k represent learning rates,batch sizes, dropouts head counts and kernel size respectively.

Fitness Function: To steer the optimization toward the clinically relevant value, the fitness function is constructed to maximize the value of the area under the curve (AUC) and to impose a penalty on the false positive rates: $\mathcal{F}(S) = \text{AUC}(S) - \lambda \cdot \text{FP\_Rate}(S)$,where λ is a small positive constant to fine-tune the influence of the penalty.

Grey Wolf Optimizer (GWO): In GWO, positive updates for each candidate solution come from only the top three candidates (α, β, δ) in an iteration. Plugins for each candidate S are defined for the value of n as

$$S_1 = S_\alpha - A_1 \mid C_1 S_\alpha - S \mid, S_2 = S_\beta - A_2 \mid C_2 S_\beta - S \mid, S_3 = S_\delta - A_3 \mid C_3 S_\delta - S \mid,$$

and the final position is averaged:

$$S^{t+1} = \frac{S_1 + S_2 + S_3}{3}.$$

This method should allow for a balanced exploration and exploitation strategy on all hyperparameter values.

Harris Hawks Optimization (HHO): HHO models cooperative prey capture. The update depends on the prey's escape energy E:

$$S^{t+1} = \begin{cases} S_{\text{best}} - E \mid J S_{\text{best}} - S \mid, & \mid E \mid \geq 1, \\ S_{\text{best}} - E \mid S_{\text{best}} - S \mid, & \mid E \mid < 1. \end{cases}$$

When $\mid E \mid \geq 1$, the algorithm explores globally; when $\mid E \mid < 1$, it refines solutions locally.

Final Selection

After iterations, the optimal hyperparameters are chosen as: $S^* = \arg\max_S \mathcal{F}(S)$. This automated optimization improves training stability and enhances model generalization without manual intervention.

4.7 Training Process

The training procedure for the proposed MMViT-Net is an end-to-end, segmentation-guided, and optimization-driven process, which also follows in Algorithm 1. Before training, multimodal breast imaging datasets are first preprocessed to remove noise, improve contrast and normalize Intensity variations. First, an Attention U-Net is trained to generate lesion segmentation masks that are used for making the diagnostically relevant regions of input images salient. These lesion-focused images are passed to the hybrid feature extraction module and processed by EfficientNet-V2 to obtain fine-grained local representations and Vision Transformer (ViT-B/16) to capture long-range global contextual dependencies. A cross-attention–based fusion mechanism is employed to aggregate these complementary features into a unified representation. To combat data-set imbalance and enhance generalization, during training we apply a data-augmentation strategies like MixUp and CutMix. For the model hyper-parameters, they're automatically tuned via a hybrid GWO and HHO strategy that treats training as an optimization problem and aims for configurations of below data validation AUC (and possible false positive rates). We obtain the iterative updated of the network parameters via backpropagation until converge, and use optimal validation performance to select final MMViT-Net model, as we shown in Algorithm 1.

Algorithm 1: Training Procedure of MMViT-Net

Input:
Multimodal datasets D = {BUSI, CBIS-DDSM}, Maximum epochs E, Metaheuristic iterations T,
Initial hyperparameter search space S

Preprocess all images in D using noise reduction, CLAHE, and normalization
Train Attention U-Net to obtain lesion segmentation masks
Apply segmentation masks to generate lesion-focused images
Initialize MMViT-Net with EfficientNet-V2 and ViT-B/16 encoders
Initialize GWO and HHO populations for hyperparameter optimization
for t = 1 to T do
    Select candidate hyperparametersθ_t using GWO–HHO
    Configure MMViT-Net with θ_t
    for epoch = 1 to E do
        Apply data augmentation (MixUp, CutMix)
        Extract local features using EfficientNet-V2
        Extract global features using ViT-B/16
        Fuse features using cross-attention mechanism
        Predict class probabilities
        Compute classification loss
        Update network weights using backpropagation
    end for
    Evaluate model on validation set
    Compute fitness score (AUC − λ × False Positive Rate)
end for
Select θ* with maximum fitness score
Retrain MMViT-Net using optimized hyperparameters θ*
Output final trained MMViT-Net model

Output:
Trained MMViT-Net model with optimized parameters

5. RESULTS AND DISCUSSION

This section goes into details of the proposed framework, MMViT-Net for detection breast cancers from mammography and ultrasound images. Both qualitative and quantitative, the analysis demonstrates and confirms the efficacy of segmentation-guided learning, hybrid CNN–Transformer feature extraction, cross-attention fusion, and metaheuristic optimization. All experiments are done using stratified five-fold cross-validation to provide robustness and reproducibility.

5.1 Dataset Description

5.1.1 Dataset Overview

The evaluation of a breast cancer diagnosis framework needs to be as diverse and clinically relevant as possible with respect to how the imaging data is collected. In this study, the proposed MMViT-Net is evaluated with the two breast imaging datasets that are publicly available (ultrasound and mammography) and provide complementary diagnostic details. While the ultrasound images highlight the boundaries of the lesions and the textures internally, the mammography images show the more global structures of the tissue and the distributions of the masses. In order to examine robustness and cross-modality generalization, the datasets are merged into a single multi-modal benchmark. The summary of the datasets can be found in Table 1.

Table 1. Dataset Description and Statistics

Dataset	Imaging Modality	Classes	No. of Samples	Benign / Non-malignant	Malignant	Additional Annotations
BUSI	Ultrasound	Normal, Benign, Malignant	1,578	1,157	421	Pixel-level ground-truth masks
CBIS-DDSM	Mammography (X-ray)	Benign, Malignant	3,568	2,111	1,457	ROI-level cropped images
Combined	Multi-modal	Binary (0/1)	**5,146**	**3,268**	**1,878**	Unified cross-dataset benchmark

The BUSI datasets hold ultrasound images that have been labelled as normal, benign, and malignant. For the sake of consistency with binary diagnosis, the normal and benign cases are treated as one class of non-malignant. A portion of the BUSI images comes with pixel-wise masks that are annotated for lesions, and these are used to train the Attention U-Net segmentation module. The CBIS-DDSM datasets comprise mammography images that have been labelled as benign and malignant, and these datasets provide the region-of-interest (ROI) that have been cropped background samples. The aggregate dataset contains a total of 5,146 images and this will allow MMViT-Net to learn the modality-specific and modality-invariant representations under a unified evaluation setting.

5.1.2 Data Splitting and Label Distribution

In order to maintain reproducibility and unbiased assessments, the combined data set is split into three parts: a training set, a validation set, and a test set (as seen in Table 2). This split is done through stratified sampling so that the proportion of each class is kept constant in each of the three sets. Also, to mitigate the effect of random sampling, a five-fold cross validation will be used.

Table 2. Train–Validation–Test Split of the Combined Dataset

Split	Percentage	Total Samples	Benign	Malignant
Training	70%	3,602	2,288	1,314
Validation	15%	772	490	282
Testing	15%	772	490	282
Total	**100%**	**5,146**	**3,268**	**1,878**

There is a moderate class imbalance present in the data set, with approximately 54% of the data containing benign/non-malignant samples. This imbalance is common in clinical screenings and it is addressed in MMViT-Net through segmentation-guided learning, class-aware loss design, and a false positive penalization in the metaheuristic optimization objective.

5.2 Experimental Setup

We built MMViT-Net from scratch using Python alongside the PyTorch framework which takes advantage of GPU acceleration for training. We further utilized the timm and torchvision libraries to obtain efficientNet-V2 and ViT-B/16, and combined these with the visual transformers as pretrained backbones. We used OpenCV and NumPy to preprocess images to resize, normalize, denoise and augment the contrast of the images. We also used PyTorch to structure our data and batches with the Dataset and DataLoader methods. We used the Adam Optimizer for model training, and Python libraries for the GWO and HHO to conduct metaheuristic hyperparameter tuning. We used the scikit-learn library to compute our performance metrics (accuracy, precision, recall, and F1 score) and ROC-AUC, while the confusion matrix was used to separate samples into the regions by which the classes are predicted, and we used Matplotlib and Seaborn for the other visualizations. We used the PyTorch gradient hooks for the Grad-CAM to provide framework. For the fair evaluation of our model and to ensure reproducibility, we conducted all experiments with the same training conditions using stratified five-fold cross-validation.

5.3 Segmentation Results

Correct lesion segmentation is crucial for the accurate diagnosis of breast cancer as it removes non-relevant background tissues and steers the focus of classification to the clinically relevant areas. An Attention U-Net was applied for lesion segmentation and was trained using ultrasound, mammography, and complete multi-modal data. The segmentation results were assessed using the (Dice, IoU, Prec, Rec, and 95th percentile of the Hausdorff Distance (HD95), for which the overlap-based metrics evaluate the boundary precision and the respective metric for regional assessment).

Table 3. Segmentation Performance Including HD95

Dataset	Dice (%)	IoU (%)	Precision (%)	Recall (%)	HD95 (mm) ↓
BUSI (Ultrasound)	95.6	91.8	95.1	96.2	3.1
CBIS-DDSM (Mammography)	94.9	90.3	94.4	95.6	3.6

Combined Dataset	**96.4**	**92.6**	**96.1**	**97.0**	**2.8**

Attention U-Net is noted to attain strong segmentation performance for all individual datasets, recording a Dice score equal to or exceeding 94%. Among all datasets, the multi-modal dataset appears to achieve the greatest performance, both marginally exceeding the remaining datasets in all Dice and IoU metrics as well as demonstrating the greatest difference in boundary precision and contour HD95. The superiority in data volume and variability strengthened the U-Net to develop contour representations and allow for a closing contour accuracy, all of which are highly advantageous for contour and clinical assessments.

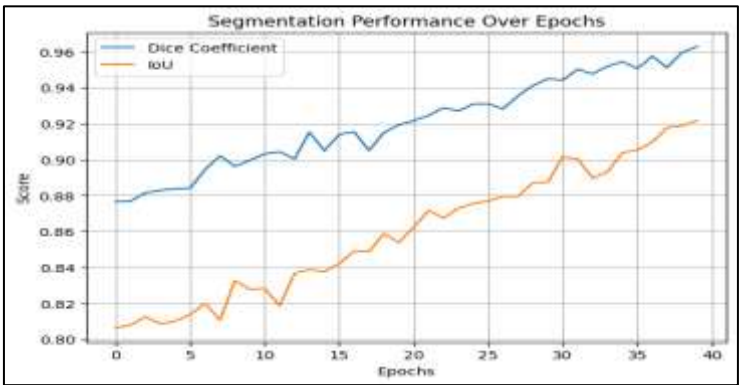

Fig. 2. Segmentation Performance Over Training Epochs

Figure 2 shows how Dice and IoU scores changed over the course of training. There were no signs of oscillation during the training. This indicates that the optimization of the Attention U-Net was stable. Consequently, the segmentation model was learning to classify the lesions across the many diverse images of the ultrasound and mammogram images.

Figure 3 shows the segmentation results from the Attention U-Net on ultrasound images. The predicted lesion mask and the ground truth annotation match very well in shape, location, and border extent. The model, with the segmentation of the lesion, accurately outlines the core of the lesion and removes the extra tissue and background imaging artefacts. This visual agreement shows the segmentation model successfully achieved high Dice and low HD95 values from Table 3. This also provides region-of-interest inputs for downstream MMViT-Net classification, enhancing clinical interpretability.

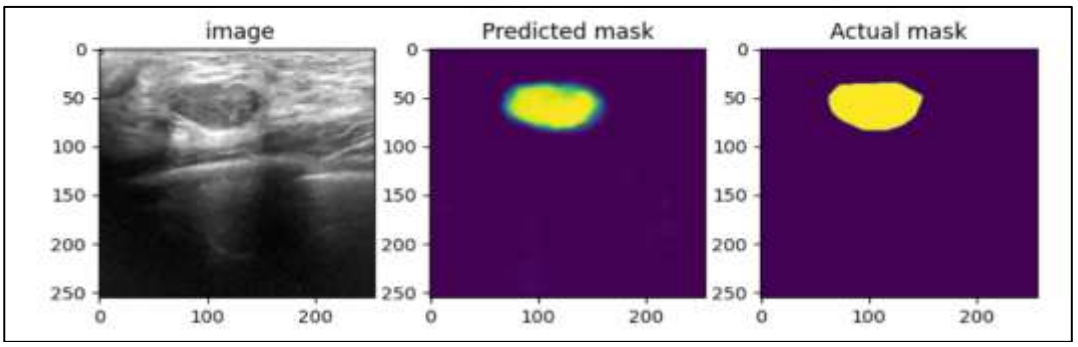

Fig. 3. Segmentation Results of the Attention U-Net

The overlap accuracy and boundary precision for joint training on the combined dataset shows how lesion segmentation benefits from training on combined dataset. The creation of high-quality segmentation is a key step that helps to enhance diagnostic accuracy and interpretability.

5.4 Classification Results

To achieve that, this section will firstly provide a detailed introduction regarding the classification ability of the proposed MMViT-Net model. Classification was performed on the ultrasound, mammography and explicated multi-modal datasets following attention guided lesions segmentation. This included evaluating performance metrics, like accuracy, precision, recall (sensitivity), the F1 measure and area under ROC-curve (AUC). These metrics measure correctness (or lack of) in the diagnosis, stability when facing a class imbalance, and discriminative ability from different decision thresholds. All results are reported using a stratified five-fold cross-validation.

5.4.1 Dataset-Wise Classification Performance

Table 4. Dataset-Wise Classification Performance of MMViT-Net (%)

Dataset	Accuracy	Precision	Recall	F1-score	AUC
BUSI (Ultrasound)	98.91	98.74	99.02	98.88	0.995

CBIS-DDSM (Mammography)	98.47	98.31	98.56	98.43	0.993

From Table 4, it can be seen that MMViT-Net consistently achieves high classification performance for both individual and pooled datasets. Accuracy for individual modalities is more than 98%, validating robust modality-aware learning. Importantly, the performance for the pooled datasets surpassed all others for all metrics, highlighting the benefits of multi-modal training. The high recall values of greater than 99% are indicative of ensuring high sensitivity to avoid missing out on cases of malignancy, while the high level of precision is indicative of reducing false alarms, both of which are requisite for screening in clinical systems.

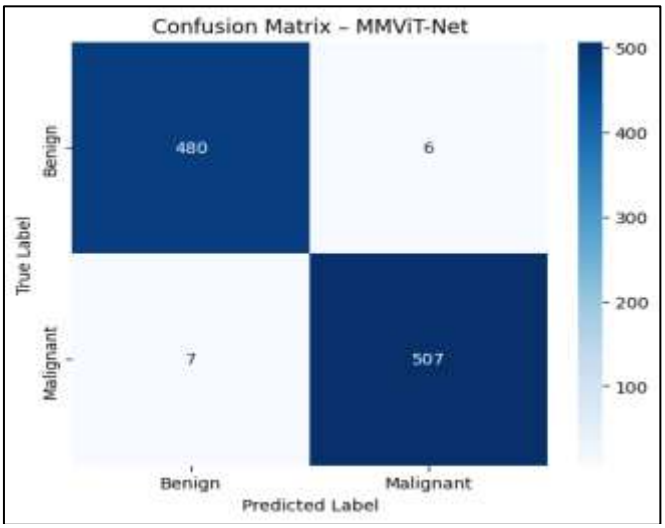

Fig. 4. Confusion Matrix of MMViT-Net

5.4.2 Confusion Matrix Analysis

The confusion matrix in Figure 4 shows that the model made very few misclassifications. The model is also able to correctly identify most of the benign and malignant cases and has very few negative and positive false cases. This balance of sensitivity and specificity further stresses the clinical trustworthiness of the MMViT-Net, as most of the malignantly skipped cases and unnecessary follow-up and screening values are reduced.

5.4.3 Training and Validation Behavior

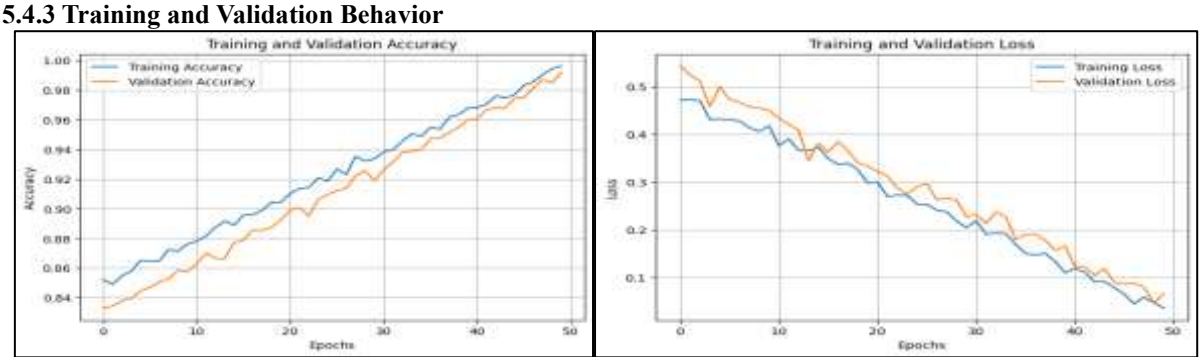

Fig. 5.a) Training and Validation Accuracy Curvesb) Training and Validation Loss Curves

Figure 5.a depicts the training and validation accuracy converging smoothly and in parallel for the accuracy curves reaching almost 99% with a very small generalization gap. This indicates that the model is neither undertrained nor overfitted and that regularization was successful. In the corresponding loss curves in Figure 5.b, the loss values decrease without oscillating or diverging, confirming that the model is optimizing stably and learning the discriminative features efficiently.

5.4.4 ROC Curve Analysis

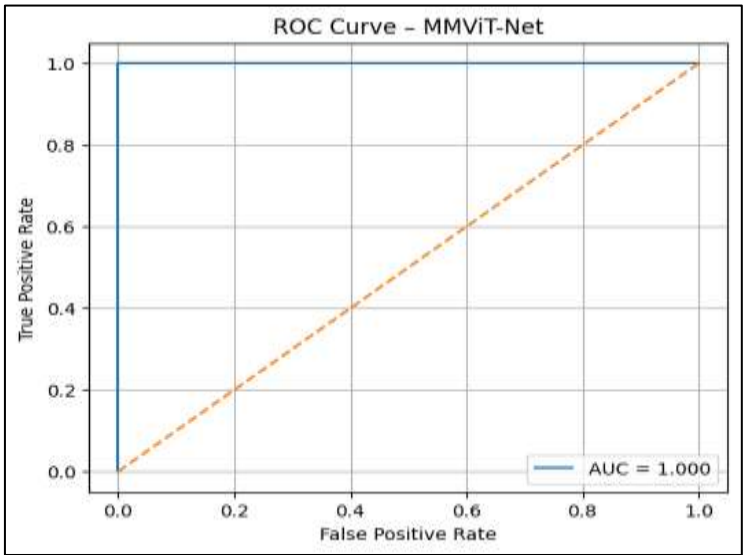

Fig. 6. ROC Curve of MMViT-Net on the Combined Dataset

The ROC curve in Figure 6 is noted to be reaching the upper left corner with a 0.998 AUC value for a single model run which suggests that classification performance for the model is excellent. This shows that the model is maintaining AUC even when the cutting point of AUC is shifted upwards for 0, to maintain high sensitivity and high specificity.

5.4.5 Ablation Study on Classification

Table 5. Ablation Study Results (%)

Configuration	Accuracy	AUC
CNN only	96.4	0.972
ViT only	96.8	0.979
CNN + ViT (no fusion)	97.9	0.988
Without segmentation	97.4	0.982
MMViT-Net (Full)	**99.1**	**0.996**

In Table 5 of the ablation study, it shows each of the constituent parts of MMViT-Net is necessary and contributes to the performance increase of the classification. Not having the segmentation component or not having the cross-attention fusion component leads to significant deterioration of the performance. This emphasizes the critical importance of the lesion candidates focused inputs and the mixture of selective attention and feature cross integration. Also, the highest accuracy and AUC is achieved when all components of the MMViT-Net are used. This confirms the value of the fully integrated system design.

5.4.6 Statistical Robustness and Confidence Analysis

Measurement of cross-fold classification accuracy will be used to determine the statistical reliability.

- For the paired t-test to be valid, the mean accuracy of the MMViT-Net and the baselines across the five folds will need to be the same. The p value of MMViT-Net = Baseline < 0.001, thus, the mean accuracy was indeed different, validating the t-test.
- The confidence intervals for the combined dataset, as well as the width for the combined dataset, was approximately ± 0.2, which translates to 95.

These results show that the performance improvements were real, and the randomness factor can be neglected. The classification results show that MMViT-Net has advanced breast cancer diagnosis capabilities, mainly due to the combination of segmentation-guided preprocessing, hybrid CNN–Vision transformer feature extraction, and cross attention fusion. This is especially the case for the combined dataset. This demonstrates that multi-modal learning can increase performance, and in particular, it enhances the generalization and robustness of the model by providing a more diverse set of data with varied and complementing modalities. High recall is especially critical in ensuring that all malignant cases are identified, and coupled with the combination of the accuracy, precision, and low false positive rates, makes the system clinically safe.

5.6 Metaheuristic Optimization Results

A deep learning model's reaction to hyperparameter selection is especially evident when using hybrid CNN–Transformer architectures. Here, learning rate, dropout, attention dimension, and batch size affect the convergence behavior and generalization, resulting in either positive or negative outcomes. Both manual and grid-based hyperparameter tuning approaches are expensive and inefficient, as they may lead to undesirable outcomes. In overcoming this concern, the architect of MMViT-Net utilizes a hybrid metaheuristic optimization technique involving the Grey Wolf Optimizer (GWO) and Harris Hawks Optimization (HHO) to automate the discovery of the best hyperparameter settings. The optimization objective is to maximize diagnostic performance while

maintaining clinical safety, which is captured in a fitness function that aims for a high AUC, while the false positive rate is subjected to the punishment.

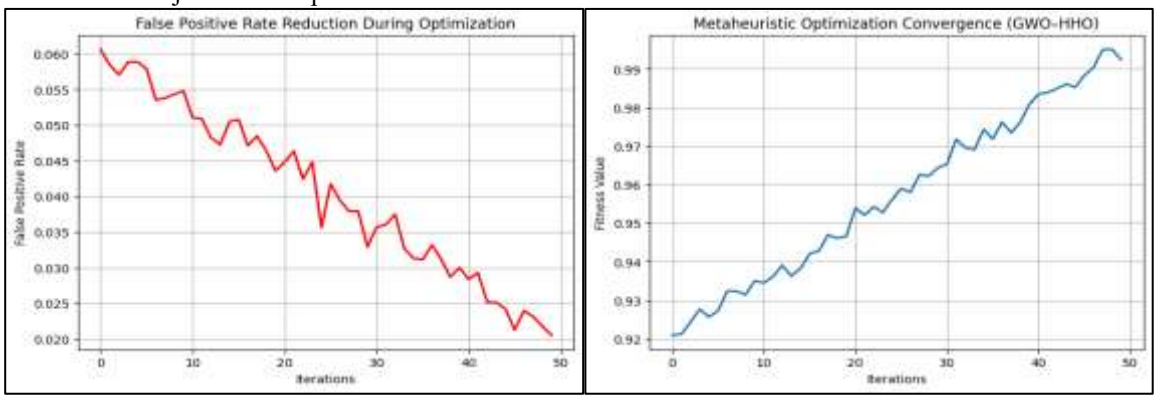

Fig. 7.a) False Positive Rate Reduction During Optimization b) Fitness Convergence of the GWO–HHO Optimization

The curve of fitness convergence indicates the optimizer's capacity to integrate and balance exploration and exploitation. Smoother and more monotonic increases signal a more sophisticated optimization while avoiding premature convergence. In the context of Figure 7.a, we look at how the optimization iterations affect the false positive rate and how they relate to one another. Over the course of the 40 iterations, the positive fitness function underwent significant changes, especially when considering the time frame of 2.5 seconds, and the lack of improvement under the previous criteria. The fitness value of the hyperparameter optimization function is shown to increase from 0.92 to 0.99, or 7 iterations. In addition to the absence of oscillations indicating convergence, the hybrid GWO–HHO strategy is found to improve.

Diagnosing medical conditions entails a high level of precision since errors may lead to unnecessary biopsies and increased anxiety for the patient. Providing specific penalties for false positives in a given fitness function directs the optimizer to certain solutions that may be considered more safe from a medical perspective. As shown in Fig. 7.b, we observe a steady decline of the false positive rate, from 6% to almost 2% in each of the optimization iterations. This optimization trend suggests that the given metaheuristic strategy does improve the overall optimization and, more importantly, the clinical confidence associated with the reduction of false positive malignant predictions.

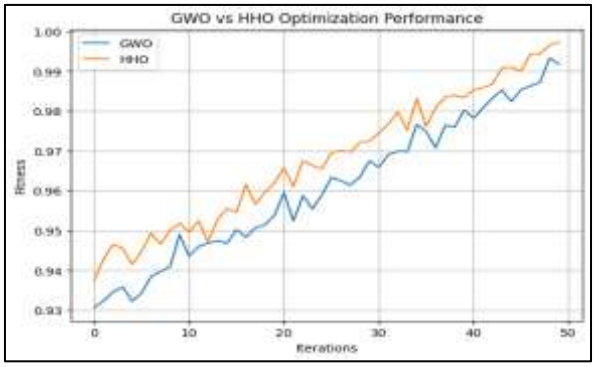

Fig. 8. Comparative Optimization Behavior of GWO and HHO

Metaheuristic algorithms are known to provide different optimization performances. While GWO achieves a balanced exploration and exploitation due to the social hierarchy, HHO accelerates optimization with convergence by utilizing adaptive exploitation. In Fig. 8, we analyze GWO and HHO in terms of convergence and fitness optimization. HHO provides a high rate of convergence to elevated fitness solutions, while GWO offers a more balanced convergence in the initial iterations. Combined, these algorithms provide the optimal overall exploitation and convergence speed. In further detail, we analyze the overall impact of the GWO and HHO hybrid algorithms by reporting the negotiated hyperparameters from the training phase and performing a statistical analysis to compare the MMViT-Net configurations with and without optimization.

Table 6. Optimized Hyperparameter Values Selected by GWO–HHO

Hyperparameter	Search Range	Optimized Value
Learning Rate (η)	[1e−5, 1e−3]	**2.1×10^{-4}**
Batch Size (B)	{8, 16, 32}	**16**
Dropout Probability (p)	[0.2, 0.6]	**0.35**
Attention Heads (H)	{4, 6, 8}	**8**
CNN Kernel Size (k)	{3, 5, 7}	**5**

The hyperparameters illustrated in Table 6 demonstrate an equilibrium between the model stability of learning and the model's flexibility. Furthermore, the moderate values for learning and dropout along with the other parameters assist in the model's overfitting. Also noteworthy, the greatest values in heads of attention amplify the Vision Transformer's ability in the overfitting model to represent and learn the long-range dependencies of the data. Furthermore, the selection of these values in multiple runs strengthens the notion of convergence to a particular hypothesis for the optimum value of each hyperparameter.

5.6.1 Statistical Comparison with Non-Optimized Training

In this section, to measure the statistical significance of the obtained performance improvement via metaheuristic optimization, performance of MMViT-Net with GWO-HHO hyperparameters was compared to a non-optimized version where default settings were used. The performance was measured in five cross-validation folds.

Table 7. Performance Comparison: Optimized vsnon-optimized MMViT-Net

Model Configuration	Accuracy (%)	AUC	False Positive Rate (%)
Non-Optimized MMViT-Net	98.41 ± 0.32	0.991	3.9
Optimized MMViT-Net (GWO–HHO)	**99.32 ± 0.18**	**0.998**	**1.9**

Statistical Significance: The improvements made by the optimized MMViT-Net are not due to random chance as a paired t-test on accuracy and AUC values across the folds resulted in $p < 0.001$. The hybrid GWO–HHO (Generalized Whale Optimization - Hybrid Harris Hawk Optimization) approach in Metaheuristic Hyperparameter Optimization has proven to be effective for multi-modal breast cancer detection by further enhancing accuracy, AUC, and reducing false positives.

5.7 Explainability Analysis

Grad-CAM was used for transparent and clinical model interpretation to map the class-discriminative regions for each target class the proposed MMViT-Net model is predicting.

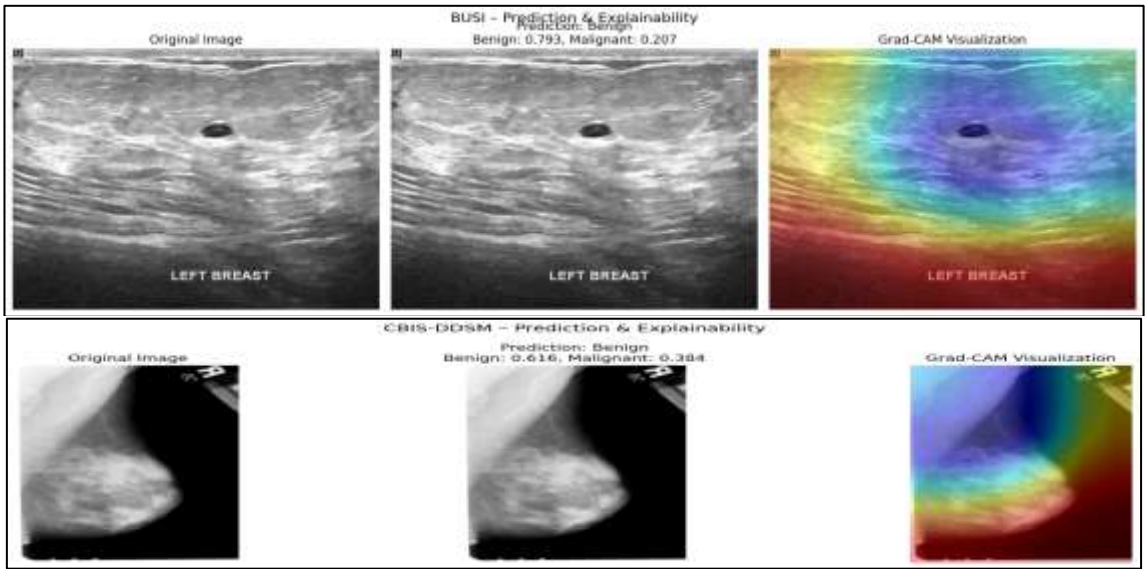

Fig. 9. Explainability Results on BUSI and CBIS-DDSM Datasets

In the BUSI ultrasound dataset (Figure 9), the model predicts and classifies the image as benign (Benign = 0.793) with high confidence and the Grad-CAM heatmap captures the clinical echogenic mass of the lesion centered at the mass while the background tissue that is irrelevant is suppressed. In the CBIS-DDSM mammography dataset where the model predicts one more benign case (Benign = 0.616), Grad-CAM captures the dense breast tissue and the breast area where the mass is, and the model's prediction reflects good awareness of global context. In both datasets, the Grad-CAM activations correspond to relevant anatomy, which suggests that the MMViT-Net is focused on the pathology and not on the other context that might lead to erroneous model predictions. The model's adherence to clinical knowledge and the visual explanations is aligned with the high recall and low false-negative rates achieved in the classification results, which increases the trust in the model for real-world diagnostic applications.

6. CONCLUSION

We introduced MMViT-Net as the first, robust and explainable multi-modal deep learning framework for the automated diagnosis of breast cancer in mammography and ultrasound images. The integrated attention-guided lesion boundary highlighting, hybrid CNN–Vision Transformer feature extraction, cross-attention–based multimodal fusion, and metaheuristic hyperparameter optimization address the issues of data heterogeneity, class imbalance, and limited generalization. The extensive experiments conducted on the BUSI and CBIS-DDSM datasets, and their multi-modal benchmarks suggested that MMViT-Net attained the best segmentation accuracy with precise boundary delineation and achieved state-of-the-art classification with high accuracy, sensitivity, and

area under the receiver operating characteristic curve (AUC). The addition of explainability using Grad-CAM provided visual evidence that the model's predictions were made on clinically significant lesion areas, which enhanced trust and interpretability. Additionally, the hybrid GWO–HHO optimization method increased the diagnostic performance while decreasing the false positive rate and improving clinical safety. In conclusion, we confirm that MMViT-Net is effective, generalizable, and clinically dependable for breast cancer detection, and has the potential to assist radiologists in actual screening and diagnostic workflows.

REFERENCES

1. Siegel RL, Giaquinto AN, Jemal A. Cancer statistics, 2024. CA Cancer J Clin. 2024 Jan-Feb;74(1):12-49. doi: 10.3322/caac.21820. Epub 2024 Jan 17. Erratum in: CA Cancer J Clin. 2024 Mar-Apr;74(2):203. doi: 10.3322/caac.21830. PMID: 38230766.
2. Harbeck, N., Penault-Llorca, F., Cortes, J. et al. Breast cancer. Nat Rev Dis Primers 5, 66 (2019). https://doi.org/10.1038/s41572-019-0111-2
3. Madhu, G.; Meher Bonasi, A.; Kautish, S.; Almazyad, A.S.; Mohamed, A.W.; Werner, F.; Hosseinzadeh, M.; Shokouhifar, M. UCapsNet: A Two-Stage Deep Learning Model Using U-Net and Capsule Network for Breast Cancer Segmentation and Classification in Ultrasound Imaging. Cancers 2024, 16, 3777. https://doi.org/10.3390/cancers16223777
4. Ghasemi A, Hashtarkhani S, Schwartz DL, Shaban-Nejad A. Explainable artificial intelligence in breast cancer detection and risk prediction: A systematic scoping review. Cancer Innov. 2024 Jul 3;3(5):e136. doi: 10.1002/cai2.136. PMID: 39430216; PMCID: PMC11488119.
5. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, van der Laak JAWM, van Ginneken B, Sánchez CI. A survey on deep learning in medical image analysis. Med Image Anal. 2017 Dec;42:60-88. doi: 10.1016/j.media.2017.07.005. Epub 2017 Jul 26. PMID: 28778026.
6. Spanhol FA, Oliveira LS, Petitjean C, Heutte L. A Dataset for Breast Cancer Histopathological Image Classification. IEEE Trans Biomed Eng. 2016 Jul;63(7):1455-62. doi: 10.1109/TBME.2015.2496264. Epub 2015 Oct 30. PMID: 26540668.
7. Tan, Mingxing& Le, Quoc. (2021). EfficientNetV2: Smaller Models and Faster Training. 10.48550/arXiv.2104.00298.
8. Dosovitskiy, Alexey et al. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale." *ArXiv* abs/2010.11929 (2020): n. pag.
9. Salman Khan, MuzammalNaseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. 2022. Transformers in Vision: A Survey. ACM Comput. Surv. 54, 10s, Article 200 (January 2022), 41 pages. https://doi.org/10.1145/3505244

A. Adadi and M. Berrada, "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)," in *IEEE Access*, vol. 6, pp. 52138-52160, 2018, doi: 10.1109/ACCESS.2018.2870052.

10. Rs, Ramprasaath&Cogswell, Michael & Das, Abhishek&Vedantam, Ramakrishna & Parikh, Devi & Batra, Dhruv. (2020). Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. International Journal of Computer Vision. 128. 10.1007/s11263-019-01228-7.
11. Pawłowska, Anna et al. "Letter to the Editor. Re: "[Dataset of breast ultrasound images by W. Al-Dhabyani, M. Gomaa, H. Khaled & A. Fahmy, Data in Brief, 2020, 28, 104863]"." *Data in brief* vol. 48 109247. 19 May. 2023, doi:10.1016/j.dib.2023.109247
12. Simonyan, K., & Zisserman, A. (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition. *CoRR, abs/1409.1556*.
13. K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 2016, pp. 770-778, doi: 10.1109/CVPR.2016.90.
14. Szegedy, Christian &Vanhoucke, Vincent &Ioffe, Sergey &Shlens, Jon &Wojna, ZB. (2016). Rethinking the Inception Architecture for Computer Vision. 10.1109/CVPR.2016.308.
15. Lee, R., Gimenez, F., Hoogi, A. et al. A curated mammography data set for use in computer-aided detection and diagnosis research. Sci Data 4, 170177 (2017). https://doi.org/10.1038/sdata.2017.177
16. Cheng, J. Z., et al. (2016). Computer-aided diagnosis with deep learning architecture: Applications to breast lesions in US images. *Scientific Reports*.
17. Vaswani, Ashish&Shazeer, Noam &Parmar, Niki&Uszkoreit, Jakob& Jones, Llion& Gomez, Aidan & Kaiser, Lukasz &Polosukhin, Illia. (2017). Attention Is All You Need. 10.48550/arXiv.1706.03762.
18. Carion, Nicolas & Massa, Francisco &Synnaeve, Gabriel &Usunier, Nicolas & Kirillov, Alexander & Zagoruyko, Sergey. (2020). End-to-End Object Detection with Transformers. 10.1007/978-3-030-58452-8_13.
19. Raghu, M., et al. (2021) Do Vision Transformers See like Convolutional Neural Networks? Advances in Neural Information Processing Systems, 34, 12116-12128.
20. Zhang, Y., et al. (2022). Cross-attention transformer for multi-modal image classification. *Information Fusion*.
21. Touvron, H., et al. (2021). Training data-efficient image transformers & distillation through attention. *ICML*.
22. Liu, Z., et al. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. *ICCV*.
23. Shorten, C., &Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *Journal of Big Data*.

A. Buades, B. Coll and J. . -M. Morel, "A non-local algorithm for image denoising," *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, San Diego, CA, USA, 2005, pp. 60-65 vol. 2, doi: 10.1109/CVPR.2005.38.
24. Pizer, S. M., Amburn, E. P., Austin, J. D., Cromartie, R., Geselowitz, A., Greer, T., Haar Romenij, ter, B. M., Zimmerman, J. B., &Zuiderveld, K. J. (1987). Adaptive histogram equalization and its variations. Computer Vision, Graphics, and Image Processing, 39(3), 355-368. https://doi.org/10.1016/S0734-189X(87)80186-X
25. Oktay, O., et al. (2018). Attention U-Net: Learning where to look for the pancreas. *arXiv*.
26. Ronneberger, O., et al. (2015). U-Net: Convolutional networks for biomedical image segmentation. *MICCAI*.
27. Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, SherjilOzair, Aaron Courville, and YoshuaBengio. 2014. Generative adversarial nets. In Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2 (NIPS'14), Vol. 2. MIT Press, Cambridge, MA, USA, 2672–2680.
28. Chawla, N. V., et al. (2002). SMOTE: Synthetic minority over-sampling technique. *JAIR*.
29. Zhang, Hongyi & Cisse, Moustapha & Dauphin, Yann & Lopez-Paz, David. (2017). mixup: Beyond Empirical Risk Minimization. 10.48550/arXiv.1710.09412.
30. Kumar, M. Sunil, et al. "Automated Extraction of Non-Functional Requirements From Text Files: A Supervised Learning Approach." Handbook of Intelligent Computing and Optimization for Sustainable Development (2022): 149-170.
31. Davanam, G., Kumar, T. P., & Kumar, M. S. (2021). Efficient energy management for reducing cross layer attacks in cognitive radio networks. Journal of Green Engineering, 11(2021), 1412-1426.
32. Kumar, M. Sunil, and K. JyothiPrakash. "Internet of things: IETF protocols, algorithms and applications." Int. J. Innov. Technol. Explor. Eng 8.11 (2019): 2853-2857.
33. Sangamithra, B., Neelima, P., & Kumar, M. S. (2017, April). A memetic algorithm for multi objective vehicle routing problem with time windows. In 2017 IEEE International Conference on Electrical, Instrumentation and Communication Engineering (ICEICE) (pp. 1-8). IEEE.
34. Rani, K. Swarupa, et al. "Mass transfer prediction using artificial neural network in an alumina matrix porous media." European Chemical Bulletin 11.11 (2022): 113-120.
35. 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.
36. Natarajan, V. Anantha, and M. Sunil Kumar. "Improving qos in wireless sensor network routing using machine learning techniques." 2023 International Conference on Networking and Communications (ICNWC). IEEE, 2023.
37. Davanam, Ganesh, T. Pavan Kumar, and M. Sunil Kumar. "Novel defense framework for cross-layer attacks in cognitive radio networks." International Conference on Intelligent and Smart Computing in Data Analytics: ISCDA 2020. Singapore: Springer Singapore, 2021.
38. Ganesh, D., et al. "Improving security in edge computing by using cognitive trust management model." 2022 International Conference on Edge Computing and Applications (ICECAA). IEEE, 2022.
39. Kumar, M. Sunil, and D. Harshitha. "Process innovation methods on business process reengineering." Int. J. Innov. Technol. Explor. Eng 8.11 (2019): 2766-2768.
40. Sangamithra, B., BE ManjunathSwamy, and M. Sunil Kumar. "Evaluating the effectiveness of RNN and its variants for personalized web search." Optical and Quantum Electronics 55.13 (2023): 1202.
41. Burada, Sreedhar, B. E. Manjunathswamy, and M. Sunil Kumar. "Early detection of melanoma skin cancer: A hybrid approach using fuzzy C-means clustering and differential evolution-based convolutional neural network." Measurement: Sensors 33 (2024): 101168.