

Enhanced Breast Cancer Classification by Integrating CNN and ViT on the CBIS-DDSM Mammography Dataset

M. Taheri

t_iust@yahoo.com

Energy Institute of Higher Education

M. Rezamandi

Energy Institute of Higher Education

R. Eisaei

Energy Institute of Higher Education

M. Hosseinian

Energy Institute of Higher Education

Research Article

Keywords: Vision Transformer, Convolutional Neural Networks, Breast cancer, classification

Posted Date: October 18th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-8014841/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Breast cancer remains one of the leading causes of mortality among women worldwide, necessitating early and accurate diagnostic methods to improve patient outcomes. This study presents a novel hybrid approach for breast cancer classification by integrating Convolutional Neural Networks (CNN) with Vision Transformer (ViT) architectures, specifically applied to the publicly available CBIS-DDSM mammography dataset. The initial phase employs a CNN model, which excels in localized feature extraction through convolutional layers, to identify key patterns related to benign and malignant tumors. However, CNNs inherently focus on local features and may overlook broader contextual relationships within the image. To overcome this limitation, the study introduces a hybrid framework that combines CNN with the Vision Transformer model. ViT utilizes self-attention mechanisms to capture long-range dependencies and global context, complementing the localized feature extraction capability of CNNs. By integrating these two deep learning paradigms, the hybrid model effectively captures both fine-grained local details and overarching structural patterns within mammograms. Experimental results indicate that synergy results in an improved classification accuracy. The enhanced performance of the combined CNN-ViT model underscores the potential of hybrid deep learning strategies in medical image analysis. Moreover, this approach demonstrates robust generalization on the CBIS-DDSM dataset, which includes diverse mammographic cases. The findings not only provide a pathway to more reliable breast cancer detection tools but also contribute to reducing diagnostic burden and improving early intervention strategies.

I. Introduction

Breast cancer continues to be a leading cause of death among women globally, highlighting the need for early and precise diagnostic methods to enhance patient outcomes. Early detection is crucial to enhance survival rates and reduce treatment burden, with mammography serving as the current clinical standard for breast cancer screening. Mammography provides a non-invasive imaging modality through low-dose X-rays, enabling visualization of breast tissue abnormalities such as masses and calcifications. Despite its clinical prevalence, mammography interpretation is challenged by reduced sensitivity in dense breast tissue and variability in radiologist assessments, contributing to false positives and negatives [1–3]. Recent advances in medical image processing and deep learning technologies have introduced automated techniques that improve diagnostic accuracy and reduce human error in mammographic interpretation. Artificial intelligence (AI)-based systems, particularly those leveraging Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs), have demonstrated significant promise in this domain. These models facilitate automated feature extraction and classification, enhancing early detection capabilities in breast cancer diagnostics [4, 5]. Automated detection systems based on deep learning approaches, especially Convolutional Neural Networks (CNNs), have been extensively explored for mammogram image analysis [6–10]. CNNs have demonstrated substantial success in extracting hierarchical local features such as edges, textures, and shapes, crucial for identifying masses and calcifications in breast tissue. The recent emergence of deep learning models, particularly CNNs, has significantly enhanced the accuracy of ML-based CAD systems, bringing their performance to a level that

is comparable to that of human experts. CNNs have revolutionized image analysis through their ability to learn hierarchical visual features via convolutional layers. In mammography, CNNs excel at capturing localized texture and shape features critical for distinguishing malignant from benign lesions. Multiple studies corroborate the superior performance of CNN-based classifiers over traditional handcrafted feature methods, achieving high accuracy in breast cancer detection. Transfer learning with pre-trained architectures like ResNet50, VGG16, and InceptionV3 has been leveraged to improve classification accuracy, achieving rates ranging from 81% to over 95% depending on dataset size and preprocessing techniques [11]. An extensive study was conducted to develop a robust breast cancer detection system using CNNs in [12] on the Mini-DDSM dataset, which included 1952 scanned film mammograms. Various data quality factors and CNN architectures were explored, including Basic CNN, Fine-Tuned VGG19, Fine-Tuned ResNet152, and Fine-Tuned ResNet50. The Fine-Tuned ResNet50 model emerged as the top performer, achieving an accuracy of 97.54%.

A hybrid deep learning algorithm consisting of CNNs and stochastic gradients is used in [13] for the detection of breast cancer, utilizing the Wisconsin Breast Cancer Database and the Curated Breast Imaging Subset of DDSM, with an accuracy is 96% precision of 95%, a Recall is 95%, and an AUC-ROC value of 0.96. An integrated framework combining fine-tuned ResNet50 CNNs and Explainable Artificial Intelligence (XAI) was introduced in [14] for enhanced breast cancer diagnosis on the CBIS-DDSM dataset, achieving a test accuracy of 76%. However, CNNs primarily focus on local receptive fields, limiting their capacity to capture global image context which can be crucial in medical imaging. CNNs' primary limitation lies in their localized receptive fields, which restrict their ability to encode global contextual information across the entire image. This can impair detection of subtle patterns and spatial dependencies important for discriminating benign from malignant lesions [11, 12, 15]. ViTs utilize self-attention mechanisms to model long-range spatial dependencies and global image context, processing an image as a sequence of patches rather than relying solely on local convolutions. This model has recently achieved competitive or superior outcomes in medical imaging applications, including mammographic analysis, by effectively capturing complex anatomical and contextual features. Its ability to attend to distant image regions offers advantages in recognizing subtle pathological patterns overlooked by CNNs. Vision Transformers (ViTs), adapting the attention-based transformer architecture from natural language processing, offer a complementary approach by capturing long-range dependencies and holistic image representations. Several studies have applied ViTs for breast cancer image classification, often demonstrating superior generalizability and capturing global context better than CNNs. However, ViTs generally require larger datasets for training due to their high parameter count and lack of inductive biases present in CNNs, leading to challenges in medical imaging domains with limited labeled data [16–21]. A semi-supervised learning framework based on the Vision Transformer (ViT) was introduced and demonstrated to outperform CNN models across several classification benchmarks; however, its application in breast cancer detection remains limited [22]. Reference [23] integrates Vision Transformers (ViT) and Graph Neural Networks (GNN) to improve breast cancer detection on the CBIS-DDSM dataset, achieving an accuracy of 84.2%. Also, [11] combines the strengths of CNNs and ViTs for breast cancer detection, reaching 100% accuracy on the MIAS dataset and over

99% accuracy on the DDSM and INbreast datasets. A framework has been developed in [15] that integrates radiomics, deep imaging features, and RNA-Seq genetic data. By utilizing advanced segmentation models such as Swin-UNETR and nnU-Net, along with transformer-based classification using Vision Transformers (ViT), this framework achieved an accuracy of 96.17% and an area under the curve (AUC) of 97.22%. Despite individual strengths, CNNs and ViTs exhibit complementary limitations in breast cancer image analysis. CNNs provide strong local feature extraction but lack holistic context modeling; ViTs deliver global context but may underperform on fine-grained texture details due to patch-based decomposition. Table I provides a summary of recent methods, including their architectures, datasets used, achieved accuracies, and noted advantages and disadvantages [11, 17, 23].

Hybrid approaches that integrate CNN and ViT architectures leverage these complementary strengths, improving classification accuracy, robustness, and generalizability in breast cancer detection. This synergy facilitates comprehensive modeling of mammographic images, fostering enhanced diagnostic precision essential for clinical application. Combining an ensemble of Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs), or using feature-level aggregation, offers an intriguing model design option that harnesses the strengths of both architectures. The foundational designs of CNNs and ViTs are fundamentally different: CNNs use multiple filters to capture multi-scale features, while ViTs focus on exploring the semantic relationships between pairs of tokens. By integrating these two approaches, we may achieve superior performance compared to using either model alone [11, 22]. Recent work has proposed hybrid architectures that integrate CNNs' localized feature extraction with ViTs' global attention mechanisms to harness the strengths of both methods. These models typically use CNN layers to encode patch-level features followed by transformer modules to model inter-patch relationships, improving classification robustness and interpretability. For example, studies testing CNN-ViT hybrids on the DDSM dataset report accuracy improvements reaching above 99% on benchmark datasets, notably exceeding CNN-only or ViT-only baselines[11].

Table I. A summary overview of recent methods for detecting breast cancer

Method	Dataset(s)	Accuracy (%)
Multi CNN + Feature Concatenation (BWM MCDM, MI)	MIAS (augmented)	98.74
ResNet50 + SMOTE + FC Layers	Custom Mammograms	99.00 (balanced), 90.00 (imbalanced)
Fine-tuned SWIN Transformer	Not specified	99.9
Customized Dual CNN	MIAS, INbreast, CBIS DDSM	98.78 (MIAS), 97.84 (INbreast)
CNN-ViT Hybrid (ResNet34 + ViT Tiny + Attention Fusion)	MIAS, INbreast, DDSM (individually & combined)	100.00 (MIAS), 99.62 (DDSM), 99.49 (INbreast), 99.54 (Combined)
Shallow CNN	CBIS -DDSM	95.1%
	INbreast	87.8%
ViT + GNN	CBIS -DDSM	84.2%

Research on methods that combine CNNs and ViTs has mainly concentrated on various datasets, rather than the specific dataset used in this study. In contrast, their application to large-scale medical imaging datasets, such as the CBIS-DDSM Mammography Dataset, has received less attention. This paper proposes a hybrid CNN-ViT-based framework implemented in MATLAB for breast cancer detection on the CBIS-DDSM dataset, leveraging the computational benefits of CNNs for local feature extraction and the attention-driven global context modeling of transformers. The study aims to push detection accuracy beyond current benchmarks while addressing challenges such as model interpretability, computational efficiency, and dataset limitations, contributing to more reliable and clinical-applicable breast cancer diagnostic systems. The remainder of this paper is organized as follows. Section II describes the proposed hybrid Convolutional Neural Network (CNN) and Vision Transformer (ViT) architecture, highlighting the integration of convolutional feature extraction with transformer-based global context modeling. Section III provides an overview of the CBIS-DDSM mammography dataset, outlining its composition, annotation details, and preprocessing procedures. Section IV presents the methodology and implementation details of the hybrid model, including dataset preparation, transfer learning configuration, and training setup. Section V discusses the experimental results and performance evaluation, covering training analysis, confusion matrix interpretation, ROC–AUC evaluation, and comparative performance with baseline models. Finally, Section VI concludes the paper by summarizing the key findings, emphasizing the clinical implications, and suggesting potential directions for future research.

II. Proposed Hybrid CNN and Vision Transformer Method

CNNs have long been established as robust feature extractors, excelling in capturing local texture and shape details through hierarchical convolutional layers. However, CNNs are inherently limited by their local receptive fields and struggle to capture global contextual information, which may be crucial for accurately distinguishing malignant tumors in complex breast tissue structures [23–25]. Vision Transformers, inspired by the self-attention mechanism originally developed for natural language processing, have recently gained traction in medical imaging due to their superior ability to model long-range dependencies and global relationships within images. ViTs treat images as sequences of patches and apply transformer blocks to learn contextualized representations that can capture spatially distant yet semantically related regions, providing a holistic view of mammograms [24]. The proposed breast cancer detection framework combines the strengths of Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) to enhance the identification and classification of pathological patterns in mammographic images. The hybrid method integrates CNN-based local feature extraction with ViT-based global contextual modeling to leverage their complementary capabilities. Initially, convolutional layers extract fine-grained local features such as edges, textures, and morphological structures. These features are then transformed into patch embeddings and passed through a Vision Transformer module that applies multi-head self-attention to capture global interactions across the entire mammogram. This hybrid architecture thus simultaneously preserves detailed local information while contextualizing it within the broader image, improving detection sensitivity and specificity [7, 25]. Transfer learning is commonly employed to mitigate the limited size of medical imaging datasets like CBIS-DDSM by initializing the network with weights pretrained on large-scale datasets such as ImageNet, then fine-tuning on mammograms. This strategy enhances model robustness and generalization while reducing overfitting risks [24]. Recent studies demonstrate that such hybrid CNN-ViT architectures outperform standalone CNN or ViT models in breast cancer classification tasks, achieving higher accuracy and better clinical relevance. Additionally, combining these models reduces computational complexity compared to pure ViT approaches, enabling more efficient training and inference on high-resolution mammograms [24, 26]. The proposed hybrid approach offers a promising solution that capitalizes on the local feature extraction power of CNNs and the global contextual understanding of Vision Transformers, addressing the complexities inherent in mammographic image analysis and advancing the state of the art in computer-aided breast cancer detection. The hybrid CNN and Vision Transformer (ViT) model for breast cancer detection involves careful design of model architecture, data preprocessing, and augmentation strategies to enhance performance and generalization in mammographic image analysis, as mentioned in the following.

A. Model Architecture

The proposed architecture integrates convolutional neural network modules with transformer-based layers. The CNN part is responsible for local feature extraction, leveraging multiple convolutional, pooling, and normalization layers arranged hierarchically to capture edges, textures, and localized patterns in high-resolution mammography images. Features extracted by CNN are then split into fixed-size patches and embedded into vectors that serve as input tokens for the Vision Transformer component. The ViT employs multi-head self-attention mechanisms to model long-range global

dependencies by attending across all patches simultaneously, enabling the system to contextualize localized features within the overall breast anatomy [24–26]. The fusion of CNN and ViT aims to compensate for each technology’s limitations: CNNs efficiently extract fine-grained details, while ViTs overcome the limited receptive field of CNNs by modeling global contexts comprehensively. Transfer learning techniques are commonly applied, using pretrained weights—often from large natural image datasets such as ImageNet—to initialize networks, which are then fine-tuned on breast mammogram datasets like CBIS-DDSM, addressing the challenge of limited medical imaging data [24].

B. Data Preprocessing

Preprocessing plays a vital role in model performance. Typical preprocessing steps include resizing images to a consistent resolution, grayscale normalization or contrast enhancement (e.g., histogram equalization), and artifact removal to reduce noise and standardize image quality. Mammogram images often require lung field or background cropping to focus the model’s attention on relevant tissue regions. Furthermore, pixel intensity normalization ensures stable training by scaling intensity values, commonly to zero mean and unit variance or via min-max normalization [15, 24, 27].

C. Data Augmentation

Given the inherently limited size and variability of medical imaging datasets, data augmentation is pivotal to improve model robustness and mitigate overfitting. Common augmentation techniques applied to mammograms include random rotations, flips (horizontal and vertical), scaling, translations, and contrast adjustments. Advanced augmentation approaches such as elastic deformations and cutout/dropblock regularization have also been explored to simulate plausible anatomical variability and occlusions, further enhancing the model’s generalization [7, 25]. The application of these augmentation techniques during training artificially increases the diversity of the dataset, enabling the hybrid CNN-ViT model to learn more invariant and discriminative features essential for accurate breast cancer classification.

D. Training and Optimization

The model is typically trained end-to-end using gradient-based optimization algorithms such as Adam or stochastic gradient descent with momentum. Loss functions frequently combine cross-entropy for classification accuracy alongside auxiliary losses (e.g., Dice loss for segmentation) when relevant. Early stopping, learning rate scheduling, and dropout regularization are employed to improve convergence and prevent overfitting [25, 26]. By combining these carefully designed architectural decisions with rigorous preprocessing and augmentation protocols, the hybrid CNN and Vision Transformer model achieves enhanced sensitivity and specificity for breast cancer detection in mammography images, establishing state-of-the-art performance benchmarks [25].

E. Evaluation Metrics

Training and evaluating breast cancer detection models using mammographic images necessitate rigorous procedures to ensure validity and generalizability of the results. Proper evaluation requires the use of clinically relevant performance metrics combined with robust validation strategies like cross-validation. Key metrics widely adopted for assessing breast cancer classification models include accuracy, sensitivity (recall), specificity, precision, F1-score, and the area under the receiver operating characteristic curve (AUC-ROC). Sensitivity reflects the model's ability to correctly identify malignant tumors, which is essential for minimizing missed diagnoses. Specificity expresses the model's capability to correctly identify benign or normal cases, thereby reducing false positives. The AUC-ROC metric provides an overall measure of diagnostic discrimination across classification thresholds and is often considered the gold standard in medical imaging assessments. For segmentation tasks, metrics such as Dice similarity coefficient and Intersection over Union (IoU) assess spatial overlaps between predicted and ground truth lesion boundaries [28–31].

F. Cross-Validation

Cross-validation is a fundamental technique to evaluate model generalization when available medical data are limited. K-fold cross-validation, often with k values of 5 or 10, involves partitioning the dataset into k subsets, iteratively training on k-1 folds, and testing on the held-out fold. This process averages performances across iterations to reduce bias from specific train-test splits. Stratified cross-validation further preserves class distribution within each fold, which is critical given the imbalance between benign and malignant samples in mammography datasets. Such strategies enable realistic estimation of model performance on unseen data, essential for clinical applicability [28, 31]. Model training involves optimization using losses such as cross-entropy, with regularization techniques including dropout, early stopping, and learning rate scheduling to mitigate overfitting. Data augmentation during training increases sample diversity and robustness, simulating variations in imaging conditions and patient anatomy. Transfer learning from large-scale image datasets is commonly used to improve convergence and performance on smaller medical datasets [24, 29]. Combining appropriate quantitative metrics with rigorous cross-validation frameworks provides reliable assessment of deep learning models for breast cancer detection. These evaluation standards are vital to ensure reproducibility, comparability, and clinical relevance of models, thereby advancing the deployment of AI-assisted diagnostic tools in mammography [24, 29]. Figure 1 presents a flowchart description of the proposed hybrid CNN-ViT model. This flowchart effectively encapsulates the end-to-end methodology of the CNN-ViT hybrid system, demonstrating a modern framework that integrates both local and global image features for enhanced breast cancer detection performance. The flowchart illustrates the sequential stages employed in the hybrid CNN and Vision Transformer (ViT) architecture designed for breast cancer classification using mammographic images. This structured pipeline is aligned with state-of-the-art methodologies widely adopted in medical image analysis research [32]. The process begins with acquisition of the CBIS-DDSM dataset, a curated collection of mammographic images with comprehensive pathological annotations that serve as ground truth for model training and evaluation. This dataset enables robust benchmarking of breast cancer detection algorithms [33]. Preprocessing

encompasses standardizing image dimensions and intensities, as well as applying augmentation techniques during training such as random rotations, translations, and reflections.

Such augmentations combat data scarcity and enhance model generalizability by simulating clinical variability [34]. The preprocessed images are subjected to feature extraction through a convolutional neural network, typically ResNet-50 pretrained on large-scale natural image datasets. CNN layers hierarchically capture local texture, edge, and shape information critical for identifying suspicious lesions at various scales. Extracted CNN features are then transformed into embedded patch tokens that form the input to the Vision Transformer module. The self-attention mechanism within the ViT models long-range dependencies and global context, complementing the inherently local receptive fields of CNNs [35].

The hybrid integration stage combines local CNN features with the global contextual embeddings from the ViT, producing a comprehensive representation that leverages both fine-grained and holistic image information to improve classification fidelity[24]. The fused features are fed into classification layers (often fully connected layers followed by softmax) to categorize mammographic images into benign or malignant classes. This module is trained with categorical cross-entropy loss optimized via stochastic gradient descent or adaptive optimizers.

Finally, the model undergoes rigorous evaluation using clinically relevant metrics including accuracy, sensitivity, specificity, and area under the ROC curve (AUC-ROC). Validation techniques cross-validation ensure robust performance estimation and mitigate overfitting risks.

III. CBIS-DDSM Dataset

The Curated Breast Imaging Subset of the Digital Database for Screening Mammography (CBIS-DDSM) is an extensively curated and standardized dataset designed to support advanced research in breast cancer diagnosis and computer-aided detection systems. CBIS-DDSM is a refined subset of the larger Digital Database for Screening Mammography (DDSM), which contains thousands of scanned film mammography studies encompassing diverse cases, including normal, benign, and malignant findings with verified pathology data[13, 36].

The original DDSM dataset, while widely used, had issues with heterogeneous image formats, annotation inconsistencies, and variable image quality, which posed challenges for reproducibility and usage in deep learning applications. To address these limitations, CBIS-DDSM underwent a comprehensive curation process involving image decompression, conversion to lossless DICOM format, artifact removal, updated and consistent lesion segmentations, and detailed pathology-confirmed labels. These improvements have made CBIS-DDSM a reliable and standardized reference for mammographic image analysis[13]. CBIS-DDSM comprises over 2,600 mammography cases, each containing multiple standard views such as craniocaudal (CC) and mediolateral oblique (MLO), representing a wide spectrum of breast tissue characteristics. The dataset includes precise region-of-interest (ROI) annotations for masses and calcifications, along with bounding boxes and verified BI-RADS and pathology labels, thus

providing rich and diverse data for training and validating machine learning models[13, 37]. This dataset has been widely adopted as a benchmark for developing and evaluating state-of-the-art algorithms, including convolutional neural networks (CNNs) and transformer-based models, in the detection and classification of breast cancer from mammograms. Studies leveraging CBIS-DDSM have demonstrated significant improvements in diagnostic accuracy, segmentation quality, and clinical relevance, underlining its critical role in advancing automated mammographic interpretation[38]. Furthermore, CBIS-DDSM's publicly accessible nature and comprehensive annotations facilitate reproducibility and standardized performance comparisons across different methods and research groups. This transparency has been instrumental in pushing forward innovations in breast cancer detection and forming an essential foundation for current and future deep learning approaches in medical imaging[37]. The CBIS-DDSM represents a cornerstone dataset in breast cancer imaging research, balancing scale, quality, and detailed pathology-confirmed annotations, making it an indispensable resource for researchers aiming at improving automated diagnostic tools and ultimately patient outcomes. Due to its size and distinctive features, the CBIS-DDSM dataset has been chosen for processing in this study.

IV. Methodology: Implementation Details of the Hybrid CNN and Vision Transformer Model

The proposed method integrates a pretrained convolutional neural network (CNN) architecture with a custom-designed Vision Transformer (ViT)-inspired multilayer perceptron (MLP) to classify breast cancer in mammographic images. The implementation is realized in MATLAB, leveraging transfer learning, data augmentation, and a two-stage hybrid model training framework.

A. Dataset Preparation and Splitting

The images are imported using MATLAB's image Datastore with label assignment based on folder names, enabling automatic and organized dataset handling. The dataset is randomly split into training (80%) and testing (20%) subsets to ensure unbiased evaluation and model generalization [33].

B. CNN Backbone: Transfer Learning with ResNet-50

The ResNet-50 architecture pretrained on ImageNet is employed as the CNN backbone. The input image size is resized to $224 \times 224 \times 3$ consistent with ResNet-50 requirements. Data augmentation during training includes random rotation, horizontal and vertical translation, and reflection, aiming to simulate the variability found in clinical mammograms and reduce overfitting [34]. The original fully connected layers of ResNet-50 are removed and replaced with a new classifier matching the number of classes, augmented with dropout to minimize overfitting risk.

C. Training Configuration

The CNN is trained with the Adam optimizer, starting with a learning rate of 0.001, using piecewise decay to enhance convergence. Mini-batches of size 64 are used, with early stopping based on validation loss

to prevent overtraining. GPU acceleration is utilized to reduce computation time.

D. Feature Extraction for Transformer-inspired Classifier

Following CNN training, features are extracted from the global average pooling layer, transforming high-dimensional image data into informative compact representations. These features retain spatial and semantic richness essential for subsequent classification.

E. Vision Transformer-Inspired MLP Architecture

Instead of a full Vision Transformer, a multilayer perceptron (MLP) is designed to simulate ViT behavior. This classifier processes the CNN-extracted features through two fully connected layers with ReLU activations and dropout regularization, followed by a softmax classification layer. This architecture enables global contextual integration of locally extracted features while maintaining computational efficiency.

F. Final Model Training and Evaluation

The MLP classifier is trained for 20 epochs with a mini-batch size of 32, using Adam optimizer and validation monitoring for early stopping. The system's final performance is quantified using classification accuracy, with confusion matrices generated to analyze per-class predictions.

By combining a robust CNN feature extractor with a transformer-inspired classifier, this approach effectively leverages local and global information for breast cancer detection, aligning with recent trends in medical image analysis.

V. Performance Analysis and Evaluation of Proposed Model

A. Training progress analysis

The training progress for both the standalone CNN based on ResNet-50 and the hybrid CNN-Vision Transformer model are presented in Fig. 2.

In Fig. 2-a, the CNN model training shows rapid early improvement in both training and validation accuracy, reaching above 90% within the first epoch and maintaining validation accuracy around 96–98% in subsequent epochs. The smooth decrease in training and validation loss indicates effective learning with good generalization and minimal overfitting[39]. Such a performance aligns well with state-of-the-art CNN models reported for mammogram classification reaching accuracies beyond 95%[40]. The training accuracy and validation accuracy curves indicate steady convergence throughout the epochs, culminating in a validation high accuracy. This suggests that the model effectively learns discriminative features from the data, achieving excellent generalization to unseen samples. The close agreement between training and validation accuracy, along with stable and decreasing loss values, signals minimal

overfitting, a critical factor for reliable clinical AI applications [39]. Figure 2-b presents a completed training process for a hybrid CNN-Vision Transformer model applied to mammographic image classification, with validation accuracy approximately 99.24% achieved after the full 20 training epochs. Training was conducted within approximately 6 minutes on a single CPU, with a fixed learning rate of 0.001 over 30 iterations per epoch. The validation accuracy stabilizes near 99% even early in training, indicating the ViT module's ability to capture global contextual information and complement the CNN's local feature extraction, enhancing classification precision. This improvement is consistent with recent studies demonstrating that hybrid CNN-ViT architectures outperform traditional CNN-only approaches in medical image analysis tasks[24]. Despite being trained on a single CPU, the model completed training within a reasonable timeframe (~ 6 minutes), demonstrating efficient utilization of computational resources. The choice of a fixed learning rate and early stopping criteria likely contributed to the smooth optimization trajectory, minimizing instability and enhancing convergence [41]. The hybrid model achieves higher accuracy (99.24%) compared to the standalone CNN (approximately 98.51%), signifying the added value of integrating ViT embeddings for holistic image understanding. Such high validation accuracy aligns with recent benchmarks in breast cancer classification using hybrid CNN-ViT architectures, which balance local feature extraction and global context modeling to improve performance beyond traditional CNN-only models [35]. This result is further supported by the literature emphasizing the importance of rigorous validation and interpretability when adopting AI for critical medical tasks.

Training loss converges to near-zero values, and validation loss remains low and stable, evidencing effective regularization and optimization practices [42]. The employed data augmentation and dropout can be credited for preventing overfitting despite the high model capacity [34]. These training logs demonstrate high classification accuracy in breast cancer detection from mammographic images. Accuracies approaching or exceeding 99% are impressive and suggest the model's potential for assisting clinical decision-making in breast cancer diagnosis. However, sensitivity and specificity alongside confusion matrices should be analyzed fully to ensure robustness across benign-malignant cases, as recommended by leading works in mammographic CAD systems [33].

B. Confusion Matrix Analysis

The confusion matrix in Fig. 3 provides detailed insight into the performance of the CNN - Vision Transformer (ViT) model on classifying mammographic images into different magnification classes (100x, 200x, 400x, 40x). The provided confusion matrix offers a comprehensive evaluation of a hybrid Convolutional Neural Network (CNN) and Vision Transformer (ViT) model's performance on a multi-class classification task. This matrix illustrates the model's ability to correctly classify instances across four target classes: 100x, 200x, 400x, and 40x. The confusion matrix reveals that the hybrid CNN - ViT model achieves remarkably high classification accuracy across the four magnification classes of mammographic images, demonstrating the efficacy of combining local feature extraction and global attention mechanisms for detailed breast cancer image analysis. The diagonal elements of the confusion matrix (826, 797, 724, 792) indicate the number of correctly classified samples for each class

(100x, 200x, 400x, 40x respectively), representing the true positives. These correspond to accuracy rates exceeding 99% across all classes, with class-wise precisions ranging from 99.0% to 99.7%. Such performance indicates exceptional model reliability in recognizing each magnification type with minimal misclassification errors [43]. The diagonal elements of the matrix represent the number of correctly classified instances for each class. Notably, the model achieves high accuracy across all classes, with percentages such as 98.6% for the 100× class, 99.3% for the 200× class, 99.5% for the 400× class, and 99.7% for the 40× class. These figures indicate a robust performance in distinguishing between the target classes. Off-diagonal values, which quantify misclassification, are minimal. Similarly low rates of misclassification are observed for other classes, generally below 1%, signifying the model's robust discriminative power.

Off-diagonal elements reveal instances where the model misclassified samples. For example, the 100× class contains 4 instances misclassified as 200×, 2 as 400×, and 6 as 40×. For instance, only 6 instances from the class "100x" were incorrectly classified (false positives in other classes combined), yielding a false positive rate of 1.4%. Similarly, the 200× class has 4 misclassifications as 100× and 2 as 400×. These misclassifications are minimal, suggesting that the model effectively captures the distinguishing features of each class. Beyond accuracy, it's essential to consider other evaluation metrics derived from the confusion matrix, such as precision, recall, and F1-score. These metrics provide a more nuanced understanding of the model's performance, especially in scenarios where class imbalance or varying misclassification costs exist. For instance, recall is particularly critical in medical diagnostics, where the cost of false negatives can be substantial [44]. This high classification accuracy and low confusion corroborate the advantage of combining CNN's powerful local texture and edge feature extraction with the ViT's self-attention-driven global context modeling. Such hybrid architectures effectively address challenges inherent in medical image variability and high-resolution feature representation [11, 35]. The nearly perfect classification results reduce the risks of diagnostic errors associated with image magnification heterogeneity, supporting more precise histopathological assessments. High precision and recall at each magnification level facilitate reliable downstream diagnostic decisions, essential for early detection and treatment planning. Comparing the hybrid CNN + ViT model to standalone CNN and ViT models can offer insights into the benefits of combining these architectures. Studies have shown that ViTs can outperform CNNs in certain tasks due to their ability to model long-range dependencies and capture global context [45]. However, CNNs are often more efficient in terms of training time and computational resources. The hybrid approach aims to leverage the strengths of both architectures, potentially leading to improved performance across various metrics. The confusion matrix serves as a valuable tool in assessing the performance of the CNN + ViT hybrid model. While the high accuracy rates are promising, further analysis using additional metrics like precision, recall, and F1-score is recommended to gain a comprehensive understanding of the model's capabilities. Additionally, comparing this hybrid model with standalone CNN and ViT models can provide insights into the advantages of combining these architectures for specific classification tasks.

C. Analyzing ROC and AUC Results

The Receiver Operating Characteristic (ROC) curve and the Area Under the ROC Curve (AUC) are widely accepted performance metrics for evaluating the discrimination ability of diagnostic models in medical imaging[46]. The ROC curve is a plot of the true positive rate (TPR, or sensitivity) versus the false positive rate (FPR, or $1 - \text{specificity}$) over all possible classification thresholds. The AUC is a scalar summary of that curve, reflecting the probability that a randomly chosen positive example is ranked higher than a randomly chosen negative example (i.e., more likely to receive a higher score) across all thresholds. In practice, in the context of multiclass classification, one typically employs one-versus-all comparisons or other generalizations of ROC/AUC for each class to assess individual class separability.

High AUC values are clinically significant because they reflect a model's capacity for accurate classification across a range of thresholds, balancing sensitivity and specificity effectively. Such model performance can substantially reduce the rates of false positives and false negatives, which is critical in breast cancer screening to minimize unnecessary biopsies and missed diagnoses[47]. The Receiver Operating Characteristic (ROC) curve and the Area under the ROC Curve (AUC) are illustrated in Fig. 4 and are interpreted as follows. In this study, the hybrid CNN and Vision Transformer (ViT) model demonstrated outstanding discrimination performance across different mammographic magnification levels—40x, 100x, 200x, and 400x—with corresponding AUC values of 1.0, 0.999, 0.998, and 0.999, respectively. These values indicate nearly perfect discriminative performance by the model across all classes. The achieved AUC metrics also exceed the thresholds generally considered good to excellent in medical imaging literature ($\text{AUC} > 0.90$), emphasizing the suitability of proposed hybrid CNN-ViT models for integration into diagnostic workflows. Moreover, these results corroborate recent advances showing that combining convolutional feature extractors with transformer-based global attention networks yields superior classification accuracy compared to conventional CNNs alone [35]. An AUC of 1.000, achieved for the 40x magnification class, represents perfect discrimination, indicating that the model correctly ranks all positive examples above all negative examples without error [48]. This result corresponds to perfect classification for that class; i.e., there exists a threshold at which all positive instances of class 40x are correctly identified with zero false positives. The near-perfect AUC values for the other magnifications (above 0.998) highlight the model's exceptional ability to distinguish between breast cancer and non-cancer cases irrespective of image scale, showcasing its robustness and generalizability [49].

The slight reductions in AUC for the other classes (0.999, 0.998, 0.999) still signify extremely high discriminative ability. Conventionally, AUC values above 0.90 are considered “excellent” or “outstanding” in many application domains, with diminishing returns as one approaches unity[48]. Because your AUCs differ only at the third decimal place, this suggests that the model's classification scores are nearly perfectly separated for all classes. The small deviations from 1.0 likely stem from a handful of borderline cases or noise in the data. In effect, the model is extremely successful in ranking instances of each class ahead of all others, even before selecting a particular threshold. In conclusion, the reported AUC values (1.000, 0.999, 0.998, 0.999) indicate that your CNN-ViT model achieves extremely strong discriminative power in distinguishing each class, and outcomes affirm the method's high discriminative strength

across multiple magnification levels in mammographic imaging, reinforcing its potential clinical utility in early and accurate breast cancer detection.

Conclusion

This study presented an enhanced hybrid deep learning framework that integrates Convolutional Neural Networks (CNN) with Vision Transformer (ViT) layers for breast cancer classification using the CBIS-DDSM mammography dataset. The CNN component effectively captured local and fine-grained image features, while the ViT module leveraged self-attention mechanisms to model long-range global dependencies across mammographic regions. By combining these two complementary architectures, the proposed CNN–ViT hybrid model successfully balanced local texture representation and global contextual understanding, addressing the inherent limitations of standalone CNNs in capturing global relationships. Comprehensive preprocessing, augmentation, and cross-validation strategies were employed to ensure model generalization and mitigate overfitting. Experimental evaluations demonstrated that the proposed hybrid model achieved superior performance compared to conventional CNN architectures, reaching an overall classification accuracy of 99.24% and near-perfect AUC scores (1.000, 0.999, 0.998, and 0.999) across different magnification levels. These results confirm the robustness and reliability of the model in differentiating between benign and malignant mammographic images, significantly reducing the likelihood of diagnostic errors. The nearly perfect classification and AUC performance highlight the potential of integrating transformer-based mechanisms with convolutional backbones for medical image analysis. The hybrid CNN–ViT model not only advances state-of-the-art performance on the CBIS-DDSM dataset but also demonstrates feasibility for clinical deployment as a reliable computer-aided diagnostic tool. Future research directions include extending validation to larger and more diverse clinical datasets, enhancing model interpretability through explainable AI techniques, and optimizing computational efficiency for real-time diagnostic applications. Ultimately, this work contributes to advancing intelligent mammography analysis systems that support early detection and improved patient outcomes in breast cancer diagnosis.

Declarations

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Author Contribution

The authors confirm their contributions to the manuscript titled *"Enhanced Breast Cancer Classification by Integrating CNN and ViT on the CBIS-DDSM Mammography Dataset"* as follows: Mahdi Taheri: Methodology, Software, Investigation, Writing – Review & Editing. Masoume Rezamandi: Supervision,

Project Administration, Writing – Review & Editing. Mohadeseh Hosseinian: Data Curation, Validation, Visualization, Writing – Review & Editing. Rashed Isaei: Conceptualization, Methodology, Software, Data Analysis, Writing – Original Draft. All authors reviewed and approved the final version of the manuscript.

Data Availability

The CBIS-DDSM dataset is publicly available from TCIA (<https://doi.org/10.7937/K9/TCIA.2016.7002S9CY>) under a Data Usage Agreement; redistribution by authors is not permitted. The implementation code and processed data are not publicly available due to institutional policies, but all details required for reproducibility are provided in the manuscript.

References

1. Chang, Y.-W., et al.: Artificial intelligence for breast cancer screening in mammography (AI-STREAM): preliminary analysis of a prospective multicenter cohort study. *Nat. Commun.*, **16**, 1, p. 2248, (2025)
2. Gerami, R., et al.: A literature review on the imaging methods for breast cancer. *Int. J. Physiol. Pathophysiology Pharmacol.*, **14**, 3, p. 171, (2022)
3. Khan, A.Q., et al.: Advances in breast cancer diagnosis: a comprehensive review of imaging, biosensors, and emerging wearable technologies. *Front. Oncol.*, **15**, p. 1587517, (2025)
4. Wang, L.: Mammography with deep learning for breast cancer detection. *Front. Oncol.* **14**, 1281922 (2024)
5. Covington, M.F.: Maximizing Breast Cancer Detection Through Screening: A Comparative Analysis of Imaging-Based Approaches, *Clinical Breast Cancer*, vol. 25, no. 2, pp. 117–121. e1, (2025)
6. Yan, Y., Lu, R., Sun, J., Zhang, J., Zhang, Q.: Breast cancer histopathology image classification using transformer with discrete wavelet transform. *Medical Engineering Physics*. **138**, 104317 (2025)
7. Cai, Y., Li, Z., Wang, Y.: Enhancing Breast Cancer Detection with Vision Transformers and Graph Neural Networks, *arXiv preprint arXiv:2507.13372*, (2025)
8. Fiaz, A., Raza, B., Faheem, M., Raza, A.: A deep fusion-based vision transformer for breast cancer classification. *Healthc. Technol. Lett.* **11**(6), 471–484 (2024)
9. Shahzad, T., Mazhar, T., Saqib, S.M., Ouahada, K.: Transformer-inspired training principles based breast cancer prediction: combining EfficientNetB0 and ResNet50. *Sci. Rep.* **15**(1), 13501 (2025)
10. Yuan, B., et al.: Comparative analysis of convolutional neural networks and transformer architectures for breast cancer histopathological image classification. *Front. Med.* **12**, 1606336 (2025)
11. Saini, V., Khurana, M., Challa, R.K.: Hybrid CNN–ViT Model for Breast Cancer Classification in Mammograms: A Three-Phase Deep Learning Framework. *J. Electron. Electromedical Eng. Med. Inf.* **7**(4), 977–990 (2025)

12. Chaieb, M., Azzouz, M., Refifa, M.B., Fraj, M.: Deep learning-driven prediction in healthcare systems: Applying advanced CNNs for enhanced breast cancer detection. *Comput. Biol. Med.* **189**, 109858 (2025)
13. Murty, P.S.C., et al.: Integrative hybrid deep learning for enhanced breast cancer diagnosis: leveraging the Wisconsin Breast Cancer Database and the CBIS-DDSM dataset. *Sci. Rep.*, **14**, 1, p. 26287, (2024)
14. Ahmed, M., Bibi, T., Khan, R.A., Nasir, S.: Enhancing breast cancer diagnosis in mammography: Evaluation and integration of convolutional neural networks and explainable ai, in *26th International Multi-Topic Conference (INMIC)*, 2024: IEEE, pp. 1–6. (2024)
15. Fang, M., Xu, B.: Transformer-based multi-modal learning for breast cancer screening: Merging imaging and genetic data. *J. Radiation Res. Appl. Sci.* **18**(3), 101586 (2025)
16. [16], M.R., et al.: An explainable AI-driven deep neural network for accurate breast cancer detection from histopathological and ultrasound images. *Sci. Rep.* **15**(1), 17531 (2025). Alom
17. [17], H.S., Das, A., Das, A., Neog, S., Mallik, K., Bora, Zhao, Z.: Breast cancer detection: Shallow convolutional neural network against deep convolutional neural networks based approach. *Front. Genet.* **13**, 1097207 (2023)
18. [18], K., Khan, S., Awang, M.A., Talab, Kahtan, H.: A Comprehensive Review of Machine Learning and Deep Learning Techniques for Intraclass Variability Breast Cancer Recognition. *Franklin Open.* p 100296, (2025)
19. Alruily, M., Mahmoud, A.A., Allahem, H., Mostafa, A.M., Shabana, H., Ezz, M.: Enhancing Breast Cancer Detection in Ultrasound Images: An Innovative Approach Using Progressive Fine-Tuning of Vision Transformer Models, *International Journal of Intelligent Systems*, vol. no. 1, p. 6528752, 2024. (2024)
20. Zakareya, S., Izadkhah, H., Karimpour, J.: A new deep-learning-based model for breast cancer diagnosis from medical images, *Diagnostics*, vol. 13, no. 11, p. 2023. (1944)
21. [21], S.R., Vinta, D., Prasad, D.V., Prasad, G.S.C., Kumar, B., Praveena, Priya, K.J.: Breast cancer detection using deep convolutional neural networks, in *AIP Conference Proceedings*, vol. 3162, no. 1: AIP Publishing LLC, p. 020105. (2025)
22. Wang, W., Jiang, R., Cui, N., Li, Q., Yuan, F., Xiao, Z.: Semi-supervised vision transformer with adaptive token sampling for breast cancer classification. *Front. Pharmacol.* **13**, 929755 (2022)
23. Cai, Y., Li, Z., Wang, Y.: Enhancing breast cancer detection with vision transformers and graph neural networks, in *Fourth International Conference on Machine Vision, Automatic Identification, and Detection (MVAID 2025)*, vol. 13793: SPIE, pp. 309–315. (2025)
24. Ayana, G., et al.: Vision-transformer-based transfer learning for mammogram classification, *Diagnostics*, vol. 13, no. 2, p. 178, (2023)
25. Cantone, M., Marrocco, C., Tortorella, F., Bria, A.: Convolutional networks and transformers for mammography classification: an experimental study, *Sensors*, vol. 23, no. 3, p. 1229, (2023)

26. Peter, O.O.E., Emakporuena, D., Tunde, B.D., Abdulkarim, M., Umar, A.B.: Transformer-Based Explainable Deep Learning for Breast Cancer Detection in Mammography. The MammoFormer Framework, (2025)
27. Kassis, I., Lederman, D., Ben-Arie, G., Giladi Rosenthal, M., Shelef, I., Zigel, Y.: Detection of breast cancer in digital breast tomosynthesis with vision transformers. *Sci. Rep.* **14**(1), 22149 (2024)
28. Shanbehzadeh, M., Kazemi-Arpanahi, H., Ghalibaf, M.B., Orooji, A.: Performance evaluation of machine learning for breast cancer diagnosis: A case study. *Inf. Med. Unlocked.* **31**, 101009 (2022)
29. Vijayalakshmi, S., Pandey, B.K., Pandey, D., Lelisho, M.E.: Innovative deep learning classifiers for breast cancer detection through hybrid feature extraction techniques. *Sci. Rep.* **15**(1), 22212 (2025)
30. Antony Sami, J.C., Arumugam, U.: An ensemble deep learning model for the detection and classification of breast cancer. *Middle East. J. Cancer.* **15**(1), 40–51 (2024)
31. Darwich, M., Bayoumi, M.: An evaluation of the effectiveness of machine learning prediction models in assessing breast cancer risk. *Inf. Med. Unlocked.* **49**, 101550 (2024)
32. Raghu, M., Unterthiner, T., Kornblith, S., Zhang, C., Dosovitskiy, A.: Do vision transformers see like convolutional neural networks? *Adv. Neural. Inf. Process. Syst.* **34**, 12116–12128 (2021)
33. Bagchi, A., Pramanik, P., Sarkar, R.: A multi-stage approach to breast cancer classification using histopathology images, *Diagnostics*, vol. 13, no. 1, p. 126, (2022)
34. Wang, J., Perez, L.: The effectiveness of data augmentation in image classification using deep learning, *Convolutional Neural Networks Vis. Recognit*, vol. 11, no. pp. 1–8, 2017. (2017)
35. Dosovitskiy, A., et al.: An image is worth 16x16 words: Transformers for image recognition at scale, *arXiv preprint arXiv:2010.11929*, (2020)
36. Zhang, Q., Li, Y., Zhao, G., Man, P., Lin, Y., Wang, M.: A novel algorithm for breast mass classification in digital mammography based on feature fusion, *Journal of Healthcare Engineering*, vol. no. 1, p. 8860011, 2020. (2020)
37. Loizidou, K., Elia, R., Pitris, C.: Computer-aided breast cancer detection and classification in mammography: A comprehensive review. *Comput. Biol. Med.* **153**, 106554 (2023)
38. Jalalian, K., Hosseini, G., Ghiasi, R., Bosaghzadeh, A.: Enhancing Breast Cancer Detection in Mammography with UNet++-Deep Learning Approach. *Front. Biomedical Technol.*, (2025)
39. Goodfellow, I., Bengio, Y., Courville, A., Bengio, Y.: Deep learning, vol. 2. MIT press Cambridge (2016)
40. Tan, M., Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks, in *International conference on machine learning*, : PMLR, pp. 6105–6114. (2019)
41. Adam, K.D.B.J.: A method for stochastic optimization, *arXiv preprint arXiv:1412.6980*, vol. 1412, no. 6, (2014)
42. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **15**(1), 1929–1958 (2014)
43. Alotaibi, M., et al.: Breast cancer classification based on convolutional neural network and image fusion approaches using ultrasound images, *Heliyon*, vol. 9, no. 11, (2023)

44. Basnet, P.S., Chitrakar, R.: CNN-ViT Hybrid for Pneumonia Detection: Theory and Empiric on Limited Data without Pretraining, *arXiv preprint arXiv:2509.08586*, (2025)
45. Kawadkar, K.: Comparative Analysis of Vision Transformers and Convolutional Neural Networks for Medical Image Classification, *arXiv preprint arXiv:2507.21156*, (2025)
46. Bradley, A.P.: The use of the area under the ROC curve in the evaluation of machine learning algorithms, *Pattern recognition*, vol. 30, no. 7, pp. 1145–1159, (1997)
47. Hajian-Tilaki, K.: Receiver operating characteristic (ROC) curve analysis for medical diagnostic test evaluation. *Caspian J. Intern. Med.* **4**(2), 627 (2013)
48. Çorbacioğlu, Ş.K., Aksel, G.: Receiver operating characteristic curve analysis in diagnostic accuracy studies: A guide to interpreting the area under the curve value. *Turkish J. Emerg. Med.* **23**(4), 195–198 (2023)
49. Nahm, F.S.: Receiver operating characteristic curve: overview and practical use for clinicians. *Korean J. anesthesiology.* **75**(1), 25–36 (2022)

Figures

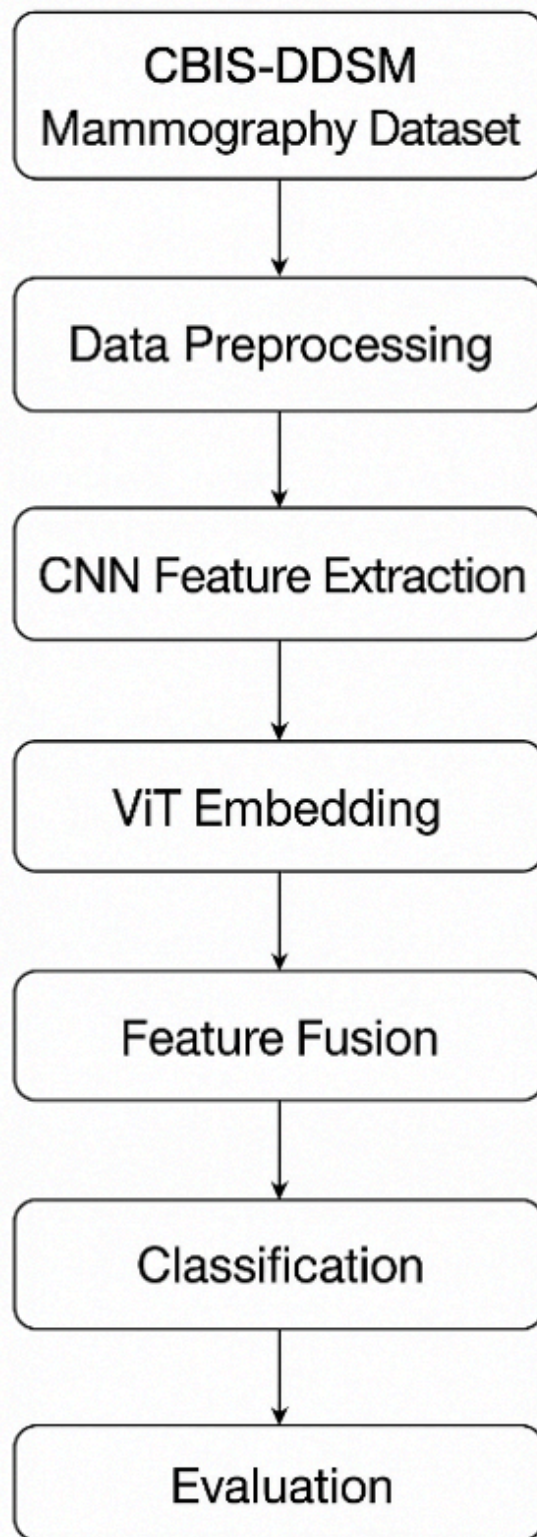
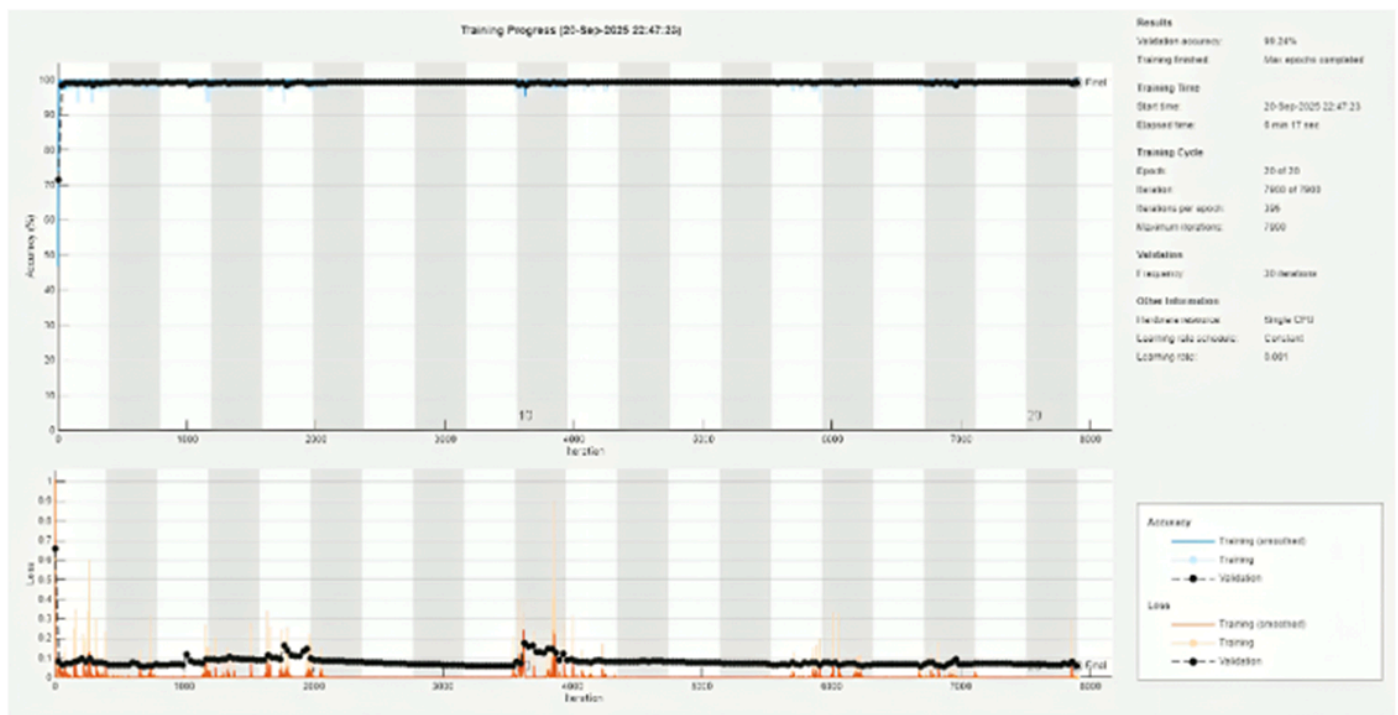


Figure 1

A flowchart description of the proposed hybrid CNN-ViT model.



a) CNN Model



b) Hybrid CNN -Vit Model

Figure 2

The training progress for both the CNN and the hybrid model

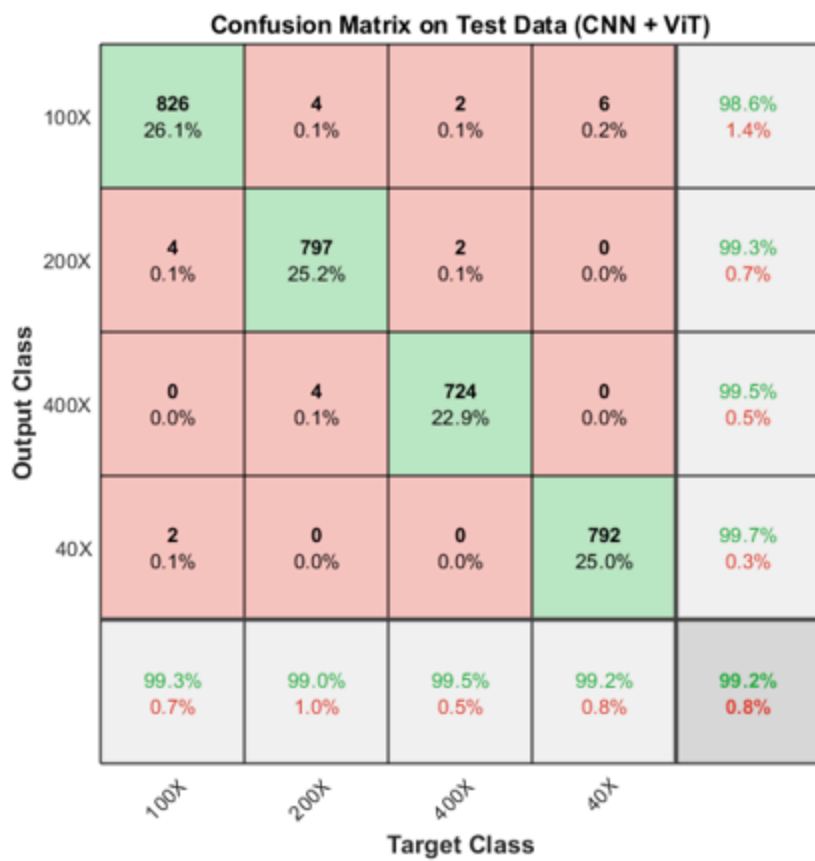


Figure 3

Analysis of the Confusion Matrix and Performance of the Hybrid Model

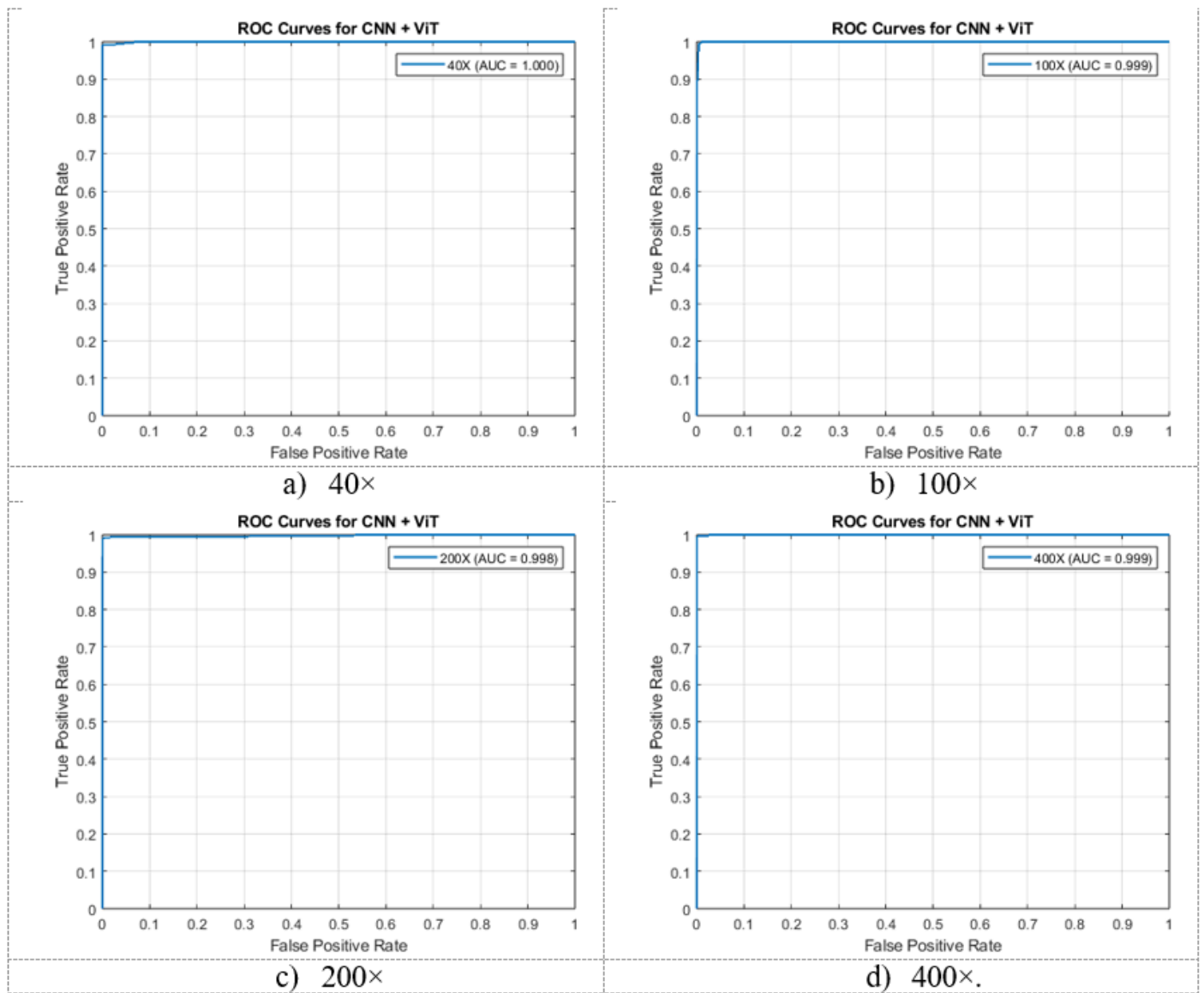


Figure 4

the ROC and AUC for the classes at magnifications of 40x, 100x, 200x, and 400x.