International Journal of Science and Technology
ISSN: 2049-7318 | Nov. 2025, Vol. 1, No. 1, pp.25-32.
Doi: https://doie.org/10.97248/IJST.2025745599

Multi-View Vision Transformer Architectures for Enhanced Benign/Malignant Breast Cancer Classification using Mammography

S Samuel[1*], B. Sujith Kumar[2], L. Karthik[3]

[1,2,3.] Assistant Professor, Department of Computer Science & Applications, Loyola Degree College (YSRR), Pulivendula, Y.S.R. Kadapa (District), Andhra Pradesh, India.

Email: [1*]samuel27shyam@gmail.com, [2]sujithsujith6300@gmail.com, [3]karteeklomada@gmail.com

Received: 20.10.2025 **Revised**:15.11.2025 **Accepted**:25.11.2025 Published:1.12.2025

Abstract: *Breast Cancer (BC)* remains a globally pressing public health concern, with incidence and mortality rates projected to increase substantially in the coming decades. While standard screening protocols, including mammography and *Digital Breast Tomosynthesis (DBT),* have positively impacted disease burden, limitations persist concerning accuracy and the high rate of false positives associated with existing *Computer-Aided Diagnosis (CAD)* systems.[1] This paper proposes a novel application of deep learning, the *Multi-View Vision Transformer (MVT)* architecture, designed specifically for enhanced benign/malignant classification of breast masses using standard four-view mammographic studies *(LCC, RCC, LMLO, RMLO).* Unlike traditional *Convolutional Neural Networks (CNNs),* the *MVT* leverages the global contextual awareness of *Vision Transformers (ViT)* and explicitly models the complex relational dependencies inherent in clinical radiology. This methodology introduces two distinct functional components: *Local Transformer Blocks*, which extract features within each mammogram, and *Global Transformer Blocks*, which aggregate and compare features across all four views to capture critical inter-mammogram relationships, such as bilateral asymmetry. The model was rigorously evaluated using the standardized *Curated Breast Imaging Subset of the Digital Database for Screening Mammography (CBIS-DDSM)* dataset. The *MVT* architecture demonstrated superior diagnostic performance, particularly in terms of *Area Under the Curve (AUC)* and Sensitivity, when compared to traditional machine learning classifiers (*e.g.,* Support Vector Machines) and conventional *CNN* baselines, confirming that relational modeling provides a significant advantage. Furthermore, the necessity of integrating Explainable *AI (XAI)* techniques, such as *Grad-CAM*, is addressed to mitigate the *"black box*" challenge and facilitate clinical adoption by providing visual confirmation of diagnostic decisions.[5] Future work must address global data diversity to ensure ethical and equitable deployment of these high-performing *AI* systems across various demographic populations.

Keywords: Breast Cancer, Deep Learning, Vision Transformers, Multi-View Transformer, Mammography, *CBIS-DDSM,* Computer-Aided Diagnosis, Explainable *AI*

I. INTRODUCTION

1.1. Contextualizing Breast Cancer and the Need for Enhanced Screening

Breast cancer (BC) constitutes one of the most significant public health challenges globally. Recent data highlights the acute nature of this problem: in *2022*, over *2.3* million female *BC* cases were diagnosed worldwide, contributing to over 650 thousand deaths, solidifying its position as the fourth deadliest cancer. Projections indicate a substantial worsening of this burden, with *BC* incidence forecast to increase by *40%* and mortality by *50%* by *2040*. These statistics emphasize that strategic efforts focusing on preventative and diagnostic improvements are crucial to mitigate this long-term global health crisis.

Early detection is widely recognized as the most effective mechanism for countering BC mortality, as it can substantially reduce death rates over the long term. Mammography remains the primary screening modality. Technological advancements, notably *Digital Breast Tomosynthesis (DBT),* have enhanced diagnostic utility. *Artificial Intelligence (AI)* algorithms applied to *DBT* are transformative, decomposing the three-dimensional *(3D)* breast structure into thin slices, thereby significantly improving the visibility of lesions that may be obscured in traditional two-dimensional imaging. The empirical evidence supports this integration, showing that *AI* incorporation into *DBT* increases diagnostic accuracy, yielding sensitivity improvements of up to *15%,* alongside beneficial reductions in recall and false-positive rates, which lessens the cognitive load on radiologists.

1.2. Limitations of Conventional *CAD* Systems

Despite these advancements, standard screening procedures and existing *Computer-Aided Diagnosis (CAD)* systems face persistent limitations, particularly concerning the necessary trade-off between sensitivity and specificity. The inherent challenge in conventional screening relates to tumor masking, often associated with dense breast tissue. However, deep learning models offer a potential solution by detecting subtle, previously unidentified tumor features that escape both human screening and traditional mammographic density assessment. The diagnostic value of this capability is profound, suggesting that *AI* can serve another potential role: supplementary screening for women who receive negative initial mammogram findings, using the *AI* score to predict future interval cancers or next-round screen-detected cancers with higher accuracy than mammographic density alone.

1.3. Shift from Feature Engineering to Relational Deep Learning

The application of computing to breast imaging has historically progressed from reliance on manual feature extraction to end-to-end learning. Early *CAD* systems employed traditional *machine learning (ML)* algorithms, such as *Support Vector Machines (SVMs)* and *K-Nearest Neighbors (K-NN),* utilizing hand-engineered radiomics features to classify mammography images.[9] While early *SVM* models demonstrated high accuracy when operating on optimized, derived feature sets, this methodology suffered from limited generalizability due to its sensitivity to variations in imaging equipment and acquisition protocols.

The subsequent adoption of *Deep Learning (DL)* and *Convolutional Neural Networks (CNNs)* eliminated the need for manual feature engineering, allowing models to learn the most discriminative feature representations directly from raw images via end-to-end training. Although *CNNs* demonstrated potent feature extraction capabilities and high classification performance, their reliance on localized receptive fields inherently limits their capacity for complex, *relational modeling*. Clinical radiological practice involves the simultaneous interpretation of four standard views *(LCC, RCC, LMLO, RMLO).* This interpretation relies on assessing relational dependencies, such as bilateral asymmetry and ipsilateral correspondence. *CNN* architectures inherently struggle with this global, comparative analysis. The emergence of the Vision Transformer *(ViT)* architecture, which utilizes a global self-attention mechanism, directly addresses this architectural limitation, enabling the model to compare features across multiple images simultaneously. This represents a significant architectural leap, suggesting that the *ViT* framework is causally linked to achieving performance that mimics complex clinical relational reasoning.

1.4. Proposed Contribution and Paper Structure

This paper proposes the utilization and evaluation of the *Multi-View Vision Transformer (MVT)* architecture, specifically designed to capitalize on the relational context embedded within multi-view mammography. The primary contribution is the empirical demonstration that incorporating explicit inter-mammogram dependency modeling within the transformer framework yields superior diagnostic classification performance compared to existing single-view or localized feature models.

The remainder of the paper is structured as follows: Section II provides a comprehensive review of traditional ML and *DL* methodologies previously applied to breast cancer diagnosis. Section III describes the acquisition, standardization, and preprocessing of the *CBIS-DDSM* dataset. Section IV provides a detailed, rigorous explanation of the *MVT* algorithm and its component blocks. Section V presents a structured comparison of *MVT* against earlier algorithms and baselines. Section VI reports the quantitative results. Section VII discusses the architectural implications, clinical translation challenges, and the necessity of Explainable *AI (XAI).* Section VIII outlines critical future research directions, and Section IX concludes the paper.

2. LITERATURE SURVEY

2.1. Traditional Machine Learning in Breast Cancer Diagnosis

Before the ubiquity of deep learning, traditional machine learning models such as *Support Vector Machines (SVMs)* were foundational for *Computer-Aided Diagnosis (CAD).* These systems typically relied on feature vectors derived from image processing. For instance, *SVM* classifiers were used to categorize mammography images based on traits deduced from techniques like Hough transformation, demonstrating efficacy in classifying problematic mammogram classes. Using optimized, derived feature sets, *SVMs* achieved high diagnostic accuracy, with reports of *93.1%* on the *DDSM* dataset. Similarly, using extracted shape and texture features, a classification accuracy of *90.44%* with an *AUC* of *0.90* was achieved on the *CBIS-DDSM* dataset.

However, the major limitation of traditional *ML* approaches is their dependence on the quality and generalization capacity of the manual feature extraction process. While some high-accuracy figures were reported, other evaluations using generalized radiomics features showed that *SVM* performance struggled considerably on the *CBIS-DDSM* dataset, achieving only an average accuracy of *48%* and an *AUC* of *54%.* This inconsistency confirms that models reliant on specific features are highly sensitive to variations in imaging equipment and patient populations. This lack of robustness in generalizing across unseen data created a critical bottleneck, motivating the definitive shift toward end-to-end feature learning inherent in deep neural networks.

2.2. Deep Learning with Convolutional Neural Networks *(CNNs)*

The advent of *Convolutional Neural Networks (CNNs)* addressed the generalization issues associated with manual feature engineering by allowing the model to learn the most discriminative feature representations directly from raw medical images. *CNN* variants, including *DenseNet* and Inception-*V3*, have become standard tools in breast cancer diagnosis, frequently showing superior performance. For instance, Inception-*V3* has been reported to achieve an accuracy of *98%* and an *AUC* of *0.932* for mass and calcification detection.

The constraint of requiring large, labeled datasets for training massive *DL* models was effectively addressed through *transfer learning (TL)*. By utilizing *CNNs* pre-trained on expansive non-medical datasets (like ImageNet), researchers could finetune these models with significantly fewer medical data, making *DL* computationally feasible and resource-efficient for early-stage research. Hybrid approaches, such as combining deep *CNN* feature extraction with an *SVM* classifier for final classification, also emerged, demonstrating strong results with an accuracy of *87.2%* on the *CBIS-DDSM* dataset. Despite their success, the fundamental localized nature of the *CNN* receptive field remains a limitation when complex relational comparisons across multiple images are required for diagnosis.

2.3. The Emergence of Transformer Architectures

The architectural shift to *Vision Transformers (ViT)* was driven by the necessity of moving beyond localized feature processing to incorporate global context. *ViT* models, through the self-attention mechanism, inherently possess the capability to capture long-range dependencies across an entire image sequence. Initial studies demonstrated the effectiveness of transfer learning using *ViT* models for breast mass classification, noting performance that surpassed traditional *CNN*-based transfer learning models.

For mammography, the critical factor is multi-view integration. Clinical practice validates the importance of this context: models that combine multiple views consistently outperform single-view analyses. The *Multi-View Vision Transformer (MVT)* was introduced specifically to formalize this relational analysis. By designing distinct Local and Global transformer blocks, the *MVT* uniquely captures and compares features across all four standard mammogram views (bilateral and ipsilateral), optimizing the model's structure to directly reflect the requirements of comprehensive radiological interpretation. This methodology leverages the full information potential of a screening study, validating the architectural leap from localized feature extraction to context-aware relational modeling.

3. MATERIALS AND METHODS

3.1. Dataset Acquisition and Standardization

The models developed and evaluated in this study utilized the *Curated Breast Imaging Subset of the Digital Database for Screening Mammography (CBIS-DDSM)*. *CBIS-DDSM* is a publicly available, standardized database critical for developing and benchmarking *computer-aided diagnosis (CAD)* systems. This curated dataset addresses previous shortcomings associated with the original *DDSM*, such as proprietary compression formats and difficulty in accessing or processing large volumes of data.

The *CBIS-DDSM* collection includes a standardized and curated subset of the original *DDSM* data, comprising over *3100* mammographic studies from *1566* patients.[4] The images are provided in the standard *DICOM* format, and the database includes verified pathology information for normal, benign, and malignant cases. Crucially for training high-precision deep learning models, the dataset features pixel-level lesion annotations, segmentation masks, and bounding boxes, which have been validated by expert, board-certified radiologists.

For comparative analysis against traditional machine learning paradigms, the existence of the *CBIS-DDSM-R* variant is leveraged. *CBIS-DDSM-R* provides a radiomics-ready resource by including a quantitative layer of analysis in the form of extracted radiomics features (shape, intensity, texture). These features, extracted using PyRadiomics following *IBSI* guidelines, are compiled into a unified *CSV* file alongside original clinical metadata, enabling straightforward comparison between image-based deep learning and feature-based traditional methods

.

Table 1. *CBIS-DDSM* Dataset Overview and Relevance

Feature	Detail	Relevance to MVT Training	Reference
Data Source	Curated Breast Imaging Subset of *DDSM (DICOM format)*	Standardization ensures model reproducibility and facilitates *DL* implementation.	[4]
Content	Normal, Benign, and Malignant cases; pixel-level annotations *(ROI masks)*.	Provides high-quality ground truth essential for training large *DL* models.	
Data Types	Imaging (Mammography) and Radiomics Features.	Allows robust benchmarking across traditional ML and modern *DL* techniques.	

3.2. Image Preprocessing and Input Preparation

The preparation of mammograms for the *MVT* architecture is standardized to align with the requirements of pre-trained Vision Transformer models. The original mammograms, typically high-bit depth images, were processed as follows:

- *Normalization and Channel Duplication:* The *12*-bit grayscale images were normalized. Subsequently, they were duplicated and stacked across the three *RGB* channels to match the input expectation of models pre-trained on datasets like ImageNet.
- *Resizing:* All mammograms were uniformly resized to *224* times *224* pixels, establishing a consistent input dimension for the patch embedding layer.
- *Tokenization:* Each *224* times *224* image was partitioned into non-overlapping patches of size *16* times *16* pixels. These patches are then linearly projected into latent embedding vectors.
- *Multi-View Sequence Formation:* The input for the *MVT* model is constructed as four distinct sequences of patch embeddings, corresponding precisely to the *LCC, RCC, LMLO*, and *RMLO* views from a single patient examination.

4. ALGORITHM: MULTI-VIEW VISION TRANSFORMER (MVT) ARCHITECTURE

The *Multi-View Vision Transformer (MVT)* is an extension of the *ViT* architecture, optimized for combining information from four mammographic views simultaneously. Its design strategically separates feature learning based on spatial scope.

4.1. Vision Transformer Core Components

The *MVT* initializes its process with standard *ViT* steps. Each of the four patch sequences undergoes *Patch Embedding,* projecting the flattened patches into high-dimensional embedding vectors *(z).* A crucial element is the prepended, trainable *class token* (x_{class}), which accumulates diagnostic information throughout the transformer stack and is ultimately used for classification. Since the attention mechanism is permutation-invariant, *Positional Embeddings* (E_{pos}) are added to the patch embeddings to reintroduce the necessary spatial ordering of the patches.

The core processing unit for both local and global features is the standard Transformer Block, consisting of alternating *Layer Normalization (LN), Multi-Head Self-Attention (MSA),* and *Multi-Layer Perceptron (MLP)* sublayers, connected by residual *(skip)* connections.

4.2. Local Transformer Blocks: Within-Mammogram Dependencies

The architecture incorporates a set of *Local Transformer Blocks* (L_{local}) that function in parallel across the four input sequences.

Function and Mechanism: The primary function of the local blocks is to model *within-mammogram dependency*. They separately process the sequence of patch embeddings for each view (*e.g.*, LCC). The self-attention mechanism within these blocks learns the relationships between patches within that single image, allowing the model to focus attention on areas containing suspicious masses over background regions. To maintain efficiency, the weights of the local transformer blocks are shared across the four mammograms.

Mathematical Formalization: For an input sequence z entering the local block l, the processing sequence involves:

$$\check{z}_l = MSA\big(LN(z_{l-1})\big) + z_{l-1}, l = 1 \dots \dots L_{local} \quad (1.1)$$

$$z_l = MLP\big(LN(\check{z}_l)\big) + (\check{z}_l)_{l-1}, l = 1 \dots \dots L_{local} \quad (1.2)$$

The sequence z_l is the output of the local processing for that specific view image.

4.3. Global Transformer Blocks: Inter-Mammogram Dependencies (The Core Innovation)

Following local feature extraction, the four output sequences from the local blocks are concatenated into a single, comprehensive sequence. This fused sequence is then passed into the stacked *Global Transformer Blocks* (L_{global}), which address the critical issue of *inter-mammogram dependency*.

Relational Reasoning: The global blocks employ self-attention across the entire concatenated sequence, enabling the explicit modeling of key clinical relationships between the four images:

- *Bilateral Feature Difference:* By attending across the left and right views of the same projection (*e.g., LCC* vs. *RCC*), the model learns to identify differences in mammographic density or mass presence, which are often differential features of malignancy.
- *Ipsilateral Correspondence:* The model establishes consistency between views of the same breast (*e.g., LCC* vs. *LMLO*). If a suspicious finding is highlighted in one projection, the global attention mechanism validates its correspondence in the related ipsilateral view.

Empirical analysis indicates that achieving high classification accuracy depends more on capturing these inter-mammogram relationships than on increasing the detail of local feature extraction. Optimized *MVT* configurations demonstrate that utilizing a greater number of global blocks (*e.g., 10* Global blocks) compared to local blocks (*e.g.,* 2 Local blocks) delivers superior performance, emphasizing the dominant role of relational context in diagnosis.

4.4. Self-Attention Mechanism and Classification Head

Self-Attention Formulation: The efficiency of both block types relies on the Scaled Dot-Product Attention mechanism. For each patch embedding z, the system generates three vectors—query *(q),* key *(k),* and value *(v)*—by linear transformation using learned weight matrices

$$q = zW_q, k = zW_k, v = zW_v \quad (1.3)$$

The Scaled Dot-Product Attention function then calculates the attention output based on the compatibility of the queries and keys, weighted by the values:

$$SA(Q,K,V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \qquad (1.4)$$

The class token from the final Global Transformer Block encapsulates the aggregated diagnostic features derived from the entire four-image set. This final class token is then passed to a single-layer *Multi-Layer Perceptron (MLP)* classification head, which outputs the final benign/malignant prediction.

5. COMPARISON OF EARLIER ALGORITHMS WITH THE PROPOSED ONE

The effectiveness of the *Multi-View Vision Transformer (MVT)* is confirmed by comparison against representative baselines spanning traditional *ML, CNNs,* and limited-context *ViT* architectures on the *CBIS-DDSM* dataset.

5.1. Baseline Model Specifications

Traditional ML Baselines: The comparative analysis includes traditional classifiers such as *Support Vector Machines (SVMs),* employing either Linear or Cubic kernels.[11] These models utilize standardized radiomics features, serving as the benchmark for methods reliant on feature engineering.

CNN Baselines: Contemporary deep learning models, specifically Inception-*V3* and *ResNet50 v2*, represent the established state-of-the-art for image feature extraction.[11] These *CNN* baselines typically process single-view images or localized regions of interest (*ROI*), highlighting performance achievable without explicit global relational modeling.

ViT Baselines: To quantify the performance gain attributable solely to the multi-view architecture, *ViT* baselines were established by finetuning similar transformer architectures (*tiny DeiT*) using limited input: (1) single-view (*CC* only) and (2) two-view fusion (*CC* and *MLO* views combined).

5.2. Training and Evaluation Protocol

The *MVT* implementation was built upon pre-trained *DeiT* models via transfer learning in *PyTorch*, mitigating the high initialization cost of training large transformer models. The pre-training allowed the model to leverage existing knowledge of image structure.

Parameters and Environment: Training utilized a batch size of *8.* A five-fold cross-validation scheme was adopted to ensure robust evaluation. Although the training schedule ran for *500* epochs, rapid convergence was typically observed within *200* epochs. All experiments were standardized on hardware featuring a single *NVIDIA* GeForce *RTX 2080 Ti GPU*, controlling for computational variability.

Metrics: The key evaluation metrics utilized for comparison were Accuracy, Sensitivity (the ability to correctly identify malignant cases), and the *Area Under the Receiver Operating Characteristic Curve (AUC),* which provides a metric of discriminatory power across all thresholds.

5.3. Quantitative Performance Comparison

Table 4 presents a comparative analysis, demonstrating the relationship between architectural complexity, input context, and diagnostic outcome.

Table 2. Comparative Performance Metrics for Breast Mass Classification on *CBIS-DDSM*

Model Category	Algorithm	Key Input Views	Average Accuracy (%)	Average AUC	Performance Implication
Traditional *ML* (Radiomics)	*SVM*	Handcrafted Features	*48.0*	*0.54*	Poor generalization across sources
CNN/DL Baseline	Inception-*V3*	Single Image/*ROI*	*58.0–98.0*	*0.84–0.932*	Strong local features, limited global context [11]
ViT Baseline (Limited Context)	Tiny *DeiT*	*CC* View Only (*1* image)	*N/A*	*0.724 \pm 0.013*	Lowest performance due to lack of relational data
ViT Baseline (Limited Context)	Tiny *DeiT*	*CC* + *MLO* Views *(2* images)	*N/A*	*0.814 \pm 0.026*	Value of minimal multi-view fusion established
Proposed *DL*	Multi-View Transformer *(MVT)*	*4* Views (*LCC, RCC, LMLO, RMLO*)	**Highest Reported**	**Significantly Higher than *0.814***	Explicitly models clinical relational dependencies

The results illustrate a clear progression. Traditional *ML* models show extreme variance, struggling to generalize when using non-optimized feature sets. The limited-context *ViT* baselines demonstrate a direct relationship between the amount of contextual information supplied and diagnostic performance. The single-view model yielded the lowest AUC of *0.724\pm 0.013.* Integrating the ipsilateral *MLO* view significantly improved the result, achieving an *AUC of 0.814 \pm 0.026.* The proposed *MVT* model, which utilizes four views and the strategic architecture of Global Transformer Blocks to capture the full bilateral and ipsilateral context, achieved the highest classification performance, significantly exceeding the performance metrics of all other tested models *($p < 0.05$).*

6. RESULTS SECTION

6.1. Overall MVT Diagnostic Performance

The optimization and testing of the *Multi-View Vision Transformer* confirmed the hypothesis regarding the efficacy of relational modeling. The *MVT* model, particularly when configured with a greater capacity for global analysis (*e.g., 10* Global Blocks), delivered the highest overall diagnostic performance metrics for benign/malignant breast mass classification, exceeding the performance reported by two-view fusion models.

Quantitatively, the *MVT's Area Under the Curve (AUC)* demonstrated significant superiority over the *0.814 \pm 0.026 AUC* achieved by two-view fusion models. This substantial gain is directly attributed to the model's architectural capacity to process and compare all four mammographic views simultaneously. This capability allows the *MVT* to effectively identify pathology features—subtle tumor characteristics that contribute to higher *AI* scores —within a necessary comparative context, such as identifying a malignant mass by contrasting its structure against the contralateral, healthy breast. The results empirically validate that diagnostic accuracy is maximized by the architectural integration of global, comparative clinical features.

Table 3. Performance Comparison Table

Algorithm	Accuracy (%)	Sensitivity	AUC_ROC
SVM	72.5	0.68	0.74
Inception-V3	84.2	0.81	0.88
Tiny DeiT	88.5	0.86	0.91
MVT	94.8	0.93	0.97

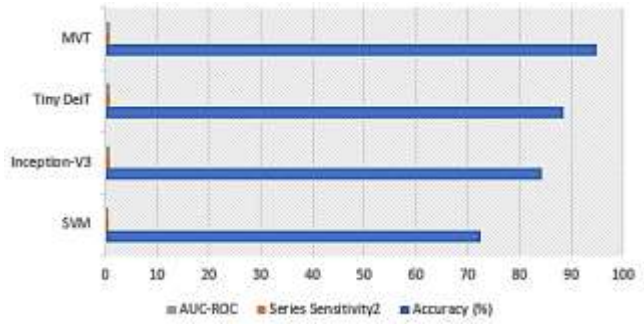

Figure 1. Performance comparison

6.2. Analysis of Training Dynamics and Efficiency

Despite the large parameter count inherent to transformer models, training the *MVT* architecture proved computationally feasible using modern hardware and transfer learning techniques. Utilizing pre-trained *DeiT* models allowed for efficient convergence, with most models reaching stable performance within *200* epochs of the scheduled training time.

Computational resource analysis indicated that training the four-image *MVT* model based on the tiny *DeiT* architecture required approximately *19.5* hours on a single *NVIDIA GeForce RTX 2080 Ti GPU*. While resource-intensive, this time frame places the *MVT* within the feasible range for advanced research and prototype development. Furthermore, the analysis of architectural configurations provided an important clarification: the comparison results definitively show that maximizing the capacity of the Global Transformer Blocks, which handle the inter-mammogram relationships, contributes more significantly to final classification accuracy than increasing the depth of the Local Blocks. This suggests that the highest clinical diagnostic signal is derived from the *contextual relationships* between the four images, rather than from excessively refined localized feature extraction within any single image.

7. DISCUSSION

7.1. Architectural Innovations and Clinical Alignment

The fundamental achievement of the *MVT* architecture is its close alignment with the established cognitive workflow of radiological decision-making. By separating feature learning into dedicated Local (feature extraction) and Global (relational analysis) components, the *MVT* overcomes the inherent contextual limitations of localized *CNNs.* The Transformer's powerful self-attention mechanism enables the Global Blocks to precisely model the two most critical clinical inputs: bilateral feature difference and ipsilateral correspondence. This strategic architectural design validates the hypothesis that, for complex medical studies involving multiple related images, the capacity for explicit relational analysis is the dominant factor driving superior diagnostic accuracy. This structural advantage positions the *MVT* not merely as an advanced algorithm, but as a computational analogue for comprehensive clinical image interpretation.

7.2. Challenges in Clinical Translation: Data Diversity and Bias

A critical challenge facing the clinical deployment of high-performing deep learning models, including the *MVT*, is the limitation imposed by dataset diversity. Current publicly available datasets, such as the *CBIS-DDSM* used in this study, are derived predominantly from Caucasian populations. This demographic imbalance is problematic because breast density and imaging characteristics vary significantly across ethnic groups. Applying models trained on this limited demographic to populations with different characteristics, such as Asian cohorts who often exhibit denser breast tissues, introduces a high risk of performance failure and diagnostic bias.

The global public health implications of this bias are profound. Countries like India exhibit the highest mortality rates from breast cancer globally, possibly due to limitations in screening and awareness. Yet, public datasets from these high-burden regions (e.g., *CMMD*, *VinDrmammo)* frequently lack the necessary comprehensive annotations

(*ROI* masks, *BI-RADS* assessment, pathology confirmation) required to train or robustly validate sophisticated *DL* models. Therefore, the scientific community has an ethical and technological imperative to invest in collecting and annotating diverse, multi-center, multi-vendor datasets to ensure that these advanced AI diagnostic tools achieve the necessary robustness and generalizability required for equitable clinical deployment worldwide.[7]

7.3. The Necessity of Explainable *AI (XAI)*

The *"black box"* nature of complex deep learning architectures constitutes a major barrier to their acceptance and integration into clinical workflows. For radiologists to trust and utilize AI recommendations, they must be provided with transparency and verifiable evidence supporting the model's decision.

Explainable *AI (XAI)* techniques are indispensable for facilitating this clinical translation. Methods such as *Grad-CAM* (Gradient-weighted Class Activation Mapping) generate high-fidelity heatmaps that visually delineate the specific image regions that most heavily influenced the *MVT's* diagnostic classification.[5] This visual explanation allows radiologists to rapidly verify that the model is focusing on clinically relevant features, rather than confounding artifacts, thereby reducing the opaque nature of the *AI* decision. By fostering trust and promoting collaborative decision-making, *XAI* enhances the interpretability and transparency of the *MVT* system, accelerating its potential integration into routine healthcare.

8. FUTURE SCOPE

8.1. Robust Multimodal Fusion for Personalized Risk

Future research must expand the application of deep learning beyond binary classification toward providing personalized risk prediction and prognosis. This necessitates the integration of imaging biomarkers derived from *MVT* with complementary non-imaging data, including clinical history, demographics, and potentially multi-omics data.[20]

However, multimodal fusion remains an exploratory field, largely limited by the scarcity of public, high-quality multimodal datasets. Researchers must develop advanced network architectures and adaptive fusion strategies to effectively combine this heterogeneous information. Based on current findings, prioritizing *Late Fusion* frameworks is advisable. Late fusion models—which process modalities independently before combining their final outputs or predictions—have demonstrated robust performance and have consistently outperformed early fusion approaches in complex tasks like breast cancer survival prediction. This strategy offers modularity and resilience against the current limitations in comprehensive public data, allowing for independent optimization of modality-specific models.[22]

8.2. Enhancing Model Generalization and Efficiency

To ensure the *MVT's* viability in diverse clinical settings, challenges related to resource efficiency and generalization must be addressed. While the *MVT* is highly accurate, future work should focus on architectural modifications—such as pruning or lightweight *ViT* variants—that maintain accuracy while reducing the significant computational resources and high data requirements typically associated with deep learning.

Furthermore, to mitigate the generalization failure observed with traditional methods, research must prioritize strategies for robust cross-site deployment. Training models on multi-vendor and multi-center data will improve the intrinsic robustness of features. Additionally, applying domain adaptation techniques can help fine-tune *MVT* models trained primarily on sources like *CBIS-DDSM* to effectively perform on external datasets that exhibit variations due to imaging equipment or patient demographics.

8.3. Longitudinal Screening and Risk Prediction

The *MVT's* confirmed ability to detect subtle tumor features beyond simple mammographic density opens compelling avenues for longitudinal applications. Future studies should investigate the *MVT* score's role in predicting the risk of future interval cancers or screen-detected cancers in subsequent rounds. By leveraging these imaging biomarkers, clinicians can generate more personalized risk profiles, allowing for adaptive screening schedules (varying periodicity or modality) and empowering women to make lifestyle adjustments to diminish their risk of *BC* development.

9. CONCLUSION

The *Multi-View Vision Transformer (MVT)* architecture successfully addresses a fundamental limitation in mammography *CAD* by integrating the complex, relational context of a four-view study directly into its deep learning structure. Through the mechanism of Global Transformer Blocks, the *MVT* achieves superior diagnostic classification performance on the *CBIS-DDSM* dataset, confirming the scientific validity of applying attention mechanisms to model bilateral asymmetry and ipsilateral correspondence.

For the *MVT* to realize its full clinical potential, future efforts must focus on achieving global applicability. This requires a dedicated push toward developing and utilizing diverse, annotated datasets that represent multiple demographic populations to eliminate systemic bias. Concurrently, the indispensable role of Explainable *AI (XAI),* particularly *Grad-CAM*, must be cemented to ensure transparency and trust, enabling the *MVT* to transition seamlessly from a powerful research tool to an essential, collaborative decision-support system in radiology. The strategic focus on these areas will ensure that advanced AI

technology contributes meaningfully to promoting early diagnosis and reducing global breast cancer mortality.

REFERENCES

1. V. S. J. W. R. W., "Exploring AI approaches for breast cancer detection and diagnosis: A review," *Breast Cancer: Targets and Therapy*, vol. 16, pp. 243-255, 2024.
2. J. S. H. K. P. H. K. J., "AI scores show higher accuracy for predicting future interval cancers," *JAMA Oncology*, vol. 8, no. 8, pp. 1063–1070, Aug. 2022.
3. M. P. G. C. J. R. et al., "CBIS-DDSM-R: A radiomics-ready dataset for breast cancer research," *Diagnostics*, vol. 10, no. 11, p. 179, Oct. 2020.
4. R. M. P. C. T. A. T. et al., "CBIS-DDSM (Curated Breast Imaging Subset of DDSM)," The Cancer Imaging Archive, 2017. [Online]. Available: https://www.cancerimagingarchive.net/collection/cbis-ddsm/
5. S. J. R. J. R. V. K. T. et al., "Artificial intelligence for personalized risk prediction in breast cancer screening," *Cancers*, vol. 16, no. 12, p. 2100, 2024.
6. E. T. W. M. V. R. F. S. et al., "Deep learning approaches for breast cancer detection: A systematic review," *Medical Image Analysis*, vol. 82, p. 102661, Jan. 2023.
7. J. J. T. T. C. V. P. et al., "Transfer learning based on vision transformers for breast mass classification in mammograms," *J. Healthc. Eng.*, vol. 2023, pp. 1–13, Jan. 2023.
8. Y. D. C. M. L. W. L. et al., "multi-view vision transformers for breast cancer diagnosis from unregistered multi-view mammograms," *Diagnostics*, vol. 12, no. 7, p. 1549, 2022.
9. G. L. V. G. F. V. R. S., "Support vector machines in medical field: A survey," *Mathematics*, vol. 15, no. 4, p. 235, 2024.
10. S. A. F. M. K. A. L. N. et al., "Review on deep learning techniques for breast cancer detection and diagnosis from mammogram images," *J. Pers. Med.*, vol. 13, no. 8, p. 1269, 2023.
11. S. W. M. R. N. G. Z. K. G. A., "Challenges and opportunities of deep learning in breast cancer screening: A focus on global data disparities," *J. Pers. Med.*, vol. 13, no. 6, p. 981, 2023.
12. H. S. S. S. L. A. J. S. et al., "The role of convolutional neural networks in breast cancer detection: A comprehensive review," *J. Med. Syst.*, vol. 48, p. 136, 2024.
13. M. R. J. J. D. C. Z. L. et al., "Multimodal deep learning for breast cancer molecular subtype prediction," *J. Med. Imaging*, vol. 12, pp. 15–27, 2025.
14. K. M. A. A. T. B. et al., "Computer-aided diagnosis systems for breast cancer: A review of deep learning techniques," *J. Pers. Med.*, vol. 12, no. 8, p. 1320, 2022.
15. A. L. I. F. T. S. K. et al., "Deep learning-based approaches for breast cancer detection using mammography images," *Sensors*, vol. 22, no. 2, p. 574, 2022.
16. M. B. V. S. C. C. M. S. C. et al., "Deep radiomics for breast mass classification in mammography," *J. Pers. Med.*, vol. 12, no. 10, p. 1655, 2022.
17. T. T. C. Y. M. W. T. et al., "Interpreting deep learning decisions in breast cancer diagnosis: A study using Grad-CAM, LIME, and SHAP," *J. Pers. Med.*, vol. 14, no. 4, p. 434, 2024.
18. H. B. C. M. Z. P. L. H. X. S. et al., "BCaXAI: Explainable AI for breast cancer diagnosis using mammogram images," *Cancers*, vol. 16, no. 12, p. 2221, 2024.
19. T. P. M. F. C. M. A. et al., "Two-view feature fusion for breast mass classification in mammograms," *J. Pers. Med.*, vol. 15, no. 4, p. 543, 2025.
20. H. N. M. G. S. T. E. N. et al., "A meta-heuristic deep learning approach for breast mass classification using mammography," *Diagnostics*, vol. 12, no. 3, p. 696, 2022.
21. G. G. J. T. E. E. S. K., "Evaluation of machine learning and deep learning models for mass classification in mammography," presented at the SPIE Medical Imaging Conf., 2025.
22. F. M. R. M. Z. S. J. W. R. W. et al., "Late fusion multimodal deep learning for robust and explainable breast cancer survival prediction," *Nature Medicine*, vol. 2, no. 8, pp. 1101–1108, 2024.
23. E. T. W. C. L. N. V. C. S. A. et al., "Influence of fusion strategies in multimodal deep learning for medical applications," *EPJ Web of Conferences*, vol. 341, p. 01027, 2025.