EURASIAN JOURNAL OF MATHEMATICAL AND COMPUTER APPLICATIONS
ISSN 2306–6172
Volume 13, Issue 4 (2025) 18 – 30

MULTI-MODAL BREAST CANCER CLASSIFICATION USING DUAL-ATTENTION VISION TRANSFORMERS AND METADATA-AWARE FUSION

Alshadoodee H.A.A., Razmara J. [1]**, Karimpour J.**

Abstract Breast cancer remains a leading cause of cancer mortality among women worldwide. For this reason, accurate and automated diagnostic tools are required. Mammography is regarded as the clinical gold standard for screening. However, diagnostic performance is limited by interpretive variability and time intensive analysis. Therefore, improved computational support is required. In this study, a novel multi modal deep learning approach is proposed, which jointly uses imaging data and patient metadata for breast cancer classification. Image features are extracted through a Dual Attention Vision Transformer named DaViT Tiny. Furthermore, structured clinical metadata, which include breast density, implant status, biopsy history, and tumor invasiveness, are processed through a multilayer perceptron. Then, a cross attention fusion module is applied, in which inter modal representations are aligned and weighted. As a result, complementary relationships between imaging features and clinical attributes are learned. The network is initialized through transfer learning on the CheXpert dataset. After that, fine tuning is performed on the RSNA mammography dataset, which contains more than 8000 labeled cases. The proposed model achieves an accuracy of 95.47, a sensitivity of 0.9655, and a specificity of 0.9439 on the test set. Consequently, performance exceeds that of single modality baselines. These results demonstrate that attention based fusion of imaging data and clinical information improves diagnostic precision and robustness. Hence, a scalable direction for early breast cancer detection is supported. Moreover, broader application to other medical imaging tasks is indicated for future research.

Keywords: Breast Cancer Classification, Multi-Modal Deep Learning, Vision Transformers, Dual-Attention Mechanisms, Metadata-Aware Fusion, Cross-Attention Networks Mammography Imaging, Clinical Decision Support Systems.

AMS Mathematics Subject Classification: 68-xx.

DOI: 10.32523/2306-6172-2025-13-4-18-30

1 Introduction

Breast cancer is diagnosed most frequently among women worldwide. An estimated 2.3 million new cases were reported in 2020, which surpassed lung cancer in incidence [1]. Moreover, it ranks as one of the leading causes of cancer-related deaths. Approximately 685,500 deaths occurred worldwide in the same year [2]. Incidence of the disease increases in many regions, including China. Notable upward trends are observed in both urban and rural populations according to recent cohort studies [3]. Major advances are made in diagnostic and therapeutic methods. However, early detection is still regarded as the most critical factor for survival outcomes. Remarkable success is achieved by population-based screening programs. For example, up to 90% of breast cancer cases are detected at an early stage in some regions [3]. Substantial survival benefits are associated with early diagnosis. In Malaysia, for instance, the five-year survival rate reaches approximately 87.5% for Stage I disease, while it drops to 23.3% for Stage IV disease [4]. These results emphasize the need for ongoing research. Novel diagnostic

[1] Corresponding Author.

approaches are developed, screening accuracy is improved, and disparities in access to care are addressed across populations. Mammography is established as the gold standard for breast cancer screening. However, several challenges are encountered. Interpretation requires much time and varies among radiologists, which decreases consistency and reliability [5]. Sensitivity is reduced by 30-48% in dense breast tissue, which is common in younger women. Lesions are obscured by overlapping structures in these cases [5,6]. Diagnosis is complicated further by subtle findings such as architectural distortions and microcalcifications. False negatives contribute to missed cases, while high false-positive rates lead to unnecessary biopsies and patient distress [7]. Clinically insignificant tumors are overdiagnosed. Effectiveness is limited additionally by dependence on radiologist expertise [6-8]. The need for automated intelligent systems is thus highlighted. Sensitivity is increased, false positives are decreased, and consistency in screening is improved by these systems. Medical image analysis is advanced by recent developments in deep learning, especially in breast cancer detection. Diagnostic accuracy is enhanced, interpretation time is shortened, and clinical decisions are supported in mammography through the use of deep learning. Traditional mammographic screening is limited by lower sensitivity in dense breasts, variability between observers, and frequent false positives or negatives. Unnecessary biopsies and follow-up procedures that cause anxiety can result [9,10]. Up to 30% of breast cancers are missed in some studies. This happens because manual interpretation is difficult and visual evaluation is subjective [9]. These limitations are addressed by deep learning. Complex representations are learned directly from image data, which removes the requirement for manual feature selection. Traditional machine learning methods are surpassed [11,12]. Assistance that is consistent, reproducible, and independent of fatigue is offered to radiologists when deep learning models are added to computer-aided detection systems. Diagnostic reliability is improved considerably [10,13]. Adaptability is shown by deep learning in various imaging modalities, including mammography, ultrasound, and magnetic resonance imaging. Subtle features such as microcalcifications, architectural distortions, and mass morphology are identified. Precision and interpretability in diagnosis are both increased [13]. Stable and reproducible results are produced by validated models. Biomarkers from images that humans may overlook are extracted, which supports personalized assessments. High-quality diagnostic tools become more accessible and less dependent on operators through deep learning systems. Consequently, breast cancer detection is expected to improve, especially in settings with limited resources [9-14]. This study builds on such potential and presents a multi-modal deep learning model for breast cancer classification. Features from mammographic images, which are extracted with a Dual Attention Vision Transformer [15], are combined with structured metadata that include implant presence, breast density, biopsy history, and tumor invasiveness. A two-stage transfer learning strategy is used in the model. Pre-training is first performed on the large CheXpert dataset [16]. Fine-tuning is then conducted on the RSNA Breast Cancer dataset [17], which contains 54,713 mammographic images along with clinical metadata. Generalization and performance are strengthened by this approach. Attention mechanisms and metadata fusion are included. Therefore, classification accuracy, interpretability, and clinical usefulness are raised. Contributions are made toward reliable automated tools that will support early breast cancer detection in the future.

2 Methodology

2.1 Overview of the Proposed Framework

A hybrid multi-modal diagnostic model is proposed in this framework. Mammographic image features are integrated with structured patient metadata, which improves breast cancer detection. A Dual Attention Vision Transformer (DaViT) is employed for image encoding.

High-resolution spatial and contextual representations are captured from mammograms by this component. Structured metadata are processed by a multi-layer perceptron (MLP). Items such as breast density, implant status, biopsy history, and tumor invasiveness are transformed into latent embeddings. Feature fusion is achieved through a cross-attention module. Visual and clinical modalities are aligned and combined by this mechanism. Patient-aware and context-sensitive classification is thus enabled. Therefore, diagnostic performance is expected to rise in future applications.

2.2 Model Architecture

1) Dual Attention Vision Transformer (DaViT)

Vision Transformers (ViTs) [38] achieve state-of-the-art performance in visual recognition tasks. Global dependencies are captured through self-attention mechanisms. However, fine-grained local patterns are modeled with limitations in standard ViTs. Such patterns are essential in medical imaging, where subtle morphological cues hold diagnostic importance. This limitation is addressed by DaViT [15]. A dual-attention mechanism is introduced, which combines spatial and channel-wise attention in each block. Spatial relationships across patches are captured jointly with semantic dependencies across feature channels by this approach. An overview of the architecture of DaViT is presented in Figure 1. Therefore, detailed visual features that support accurate diagnosis are expected to be extracted more effectively in future medical applications.

In this study, the DaViT-Tiny variant is employed. Accuracy and computational efficiency are balanced by this choice. The standard transformer architecture is followed by the model. Patch embeddings, convolutional positional encoding, and four hierarchical stages are included. The input image is divided into non-overlapping patches. These patches are flattened and linearly projected into a fixed-dimensional embedding space through a 2D convolution layer. Spatial locality is preserved in this process. Spatial self-attention, channel attention, and feed-forward layers are contained in each DaViT block. Dependencies among image patches are modeled by spatial attention. Therefore, both local and global visual information is expected to be captured effectively in future diagnostic tasks.

$$SA(X) = Softmax(\frac{Q_s K_s^T}{\sqrt{d}})V_s \tag{1}$$

where Q_s, K_s, and V_s are learned projections of the token embeddings, and (d=256) is the scaling factor. Channel attention captures semantic relationships across feature channels:

$$CA(X) = Softmax(\frac{Q_c K_c^T}{\sqrt{N}})V_c \tag{2}$$

where (N) denotes the number of patches. Together, these mechanisms enable the network to encode both local and global dependencies, thereby enhancing the discrimination of subtle breast tissue patterns.

Convolutional positional encoding is applied within each attention block, eliminating the need for explicit positional embeddings. The network output undergoes average pooling and linear projection, producing a 256-dimensional visual feature vector per image.

2) Metadata Encoding with MLP Non-imaging features, such as breast density, implant presence, biopsy history, and tumor invasiveness, are encoded and normalized before processing through a two-layer MLP. Each layer consists of a linear transformation, batch normalization, ReLU activation, and dropout regularization. The MLP outputs a 256-dimensional latent representation aligned with the DaViT embedding dimension to facilitate effective multi-modal

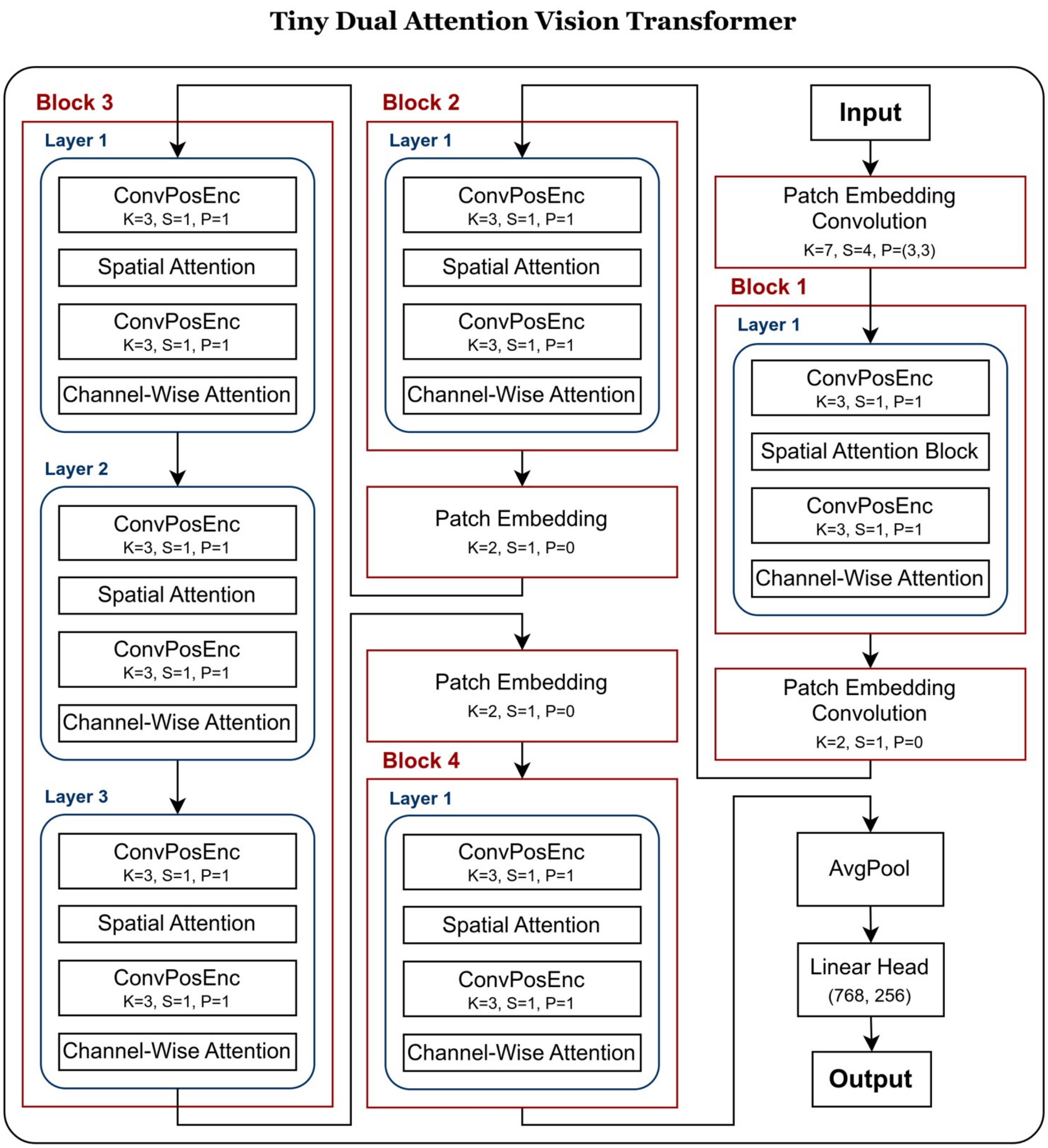

Figure 1: Architecture of DaViT-Tiny: a hierarchical vision transformer combining convolutional positional encoding, spatial attention, and channel-wise attention across four sequential processing blocks. Each block representation is aggregated via average pooling and passed to a linear head for classification.

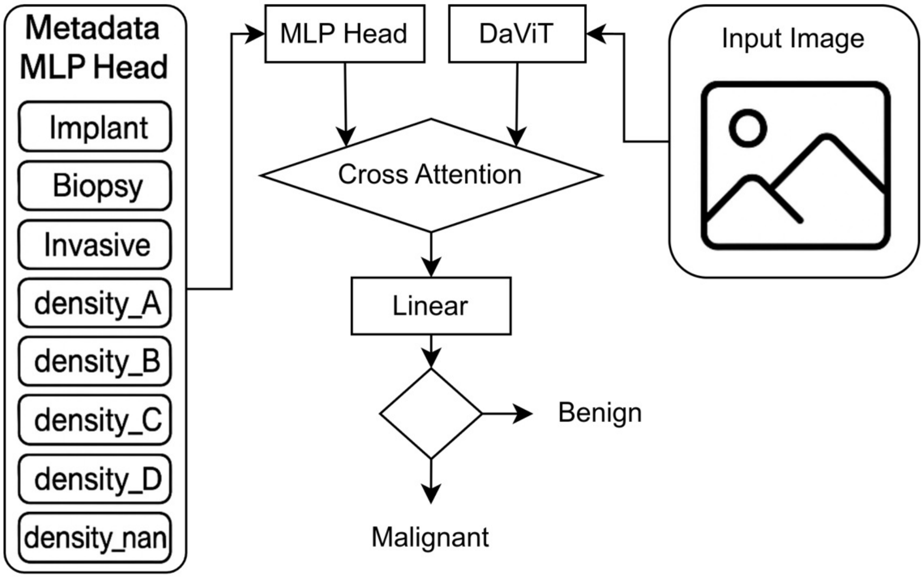

Figure 2: Multi-modal model combining metadata and imaging for breast cancer analysis. Metadata features are processed via an MLP head, while mammogram images are encoded using a DaViT model. A cross-attention mechanism fuses both modalities, followed by a linear layer for final prediction.

fusion. Categorical features are one-hot encoded into five binary columns, explicitly including a missing-value indicator. Binary variables are encoded as 0 or 1, yielding an eight-variable input vector. This encoding strategy preserves interpretability and robustness while ensuring compatibility with deep learning operations. *3) Metadata-Image Fusion via Cross-Attention* To integrate image and metadata features, a cross-attention mechanism is applied. Here, the metadata vector acts as the query ((Q)) and the DaViT visual features serve as keys ((K)) and values ((V)). This enables the model to selectively attend to image regions most relevant to each patient's clinical profile. The fused representation is then passed through a linear head for binary classification (benign vs. malignant). Compared with static concatenation methods, cross-attention provides dynamic and learnable fusion, allowing inter-modal interactions to guide feature weighting adaptively. This results in a richer and context-aware representation, improving both interpretability and predictive performance. The overall framework is illustrated in Figure 2.

2.3 Pre-Training and Transfer Learning

To mitigate overfitting and enhance generalization, DaViT was pre-trained on the CheXpert dataset [16], which contains over 200,000 labeled chest X-ray images. Despite domain differences, this dataset provides robust low-level visual priors beneficial for medical image analysis. The pre-trained model was subsequently fine-tuned on the RSNA Breast Cancer dataset [17], comprising 54,713 annotated mammography images with associated metadata. This two-stage transfer learning strategy enables the network to retain generalized representations from CheXpert while adapting to the specific spatial and textural characteristics of mammograms. During fine-tuning, all layers were unfrozen, allowing end-to-end optimization and improved task-specific performance. The proposed pipeline combines transformer-based spatial-semantic encoding, metadata-aware contextual learning, and cross-modal fusion to create a unified diagnostic framework. By leveraging pre-training and transfer learning, the model achieves improved sensitivity and robustness in breast cancer classification while maintaining computational efficiency suitable for clinical deployment.

3 Datasets

3.1 CheXpert

The CheXpert dataset, developed by Stanford University, is a large-scale collection of chest radiographs designed for the development and evaluation of medical imaging algorithms. It comprises 224,316 radiographic images from 65,240 patients, annotated across 14 thoracic disease categories, including atelectasis, cardiomegaly, and pleural effusion. The labels were derived automatically from associated radiology reports using a rule-based natural language processing system, with uncertain findings explicitly marked to capture interpretive ambiguity.

In this study, CheXpert is utilized for pre-training. Generalized radiographic representations are learned through this process, which include anatomical structures and pathological patterns. Extensive scale, heterogeneity, and clinical diversity are offered by CheXpert. For this reason, an ideal source is provided for initializing transformer-based models. Feature generalization is enhanced by pre-training on this dataset. Transferability to downstream tasks is also improved, especially in breast cancer detection and classification with mammographic images. Attention mechanisms are relied upon heavily by transformer architectures. Substantial benefits are gained from large-scale pre-training. Both local and global dependencies in medical imaging data are captured more effectively. Therefore, performance in future diagnostic applications is expected to increase.

3.2 RSNA Mammography

The RSNA Screening Mammography Breast Cancer Detection dataset is released for the RSNA 2023 Breast Cancer Challenge. More than 100,000 mammographic images are included, which are obtained from approximately 39,000 patients. Four standard clinical views are provided in each mammography exam. These views consist of right cranio-caudal (R-CC), left cranio-caudal (L-CC), right mediolateral oblique (R-MLO), and left mediolateral oblique (L-MLO). Labels are assigned to each image. Cancer confirmation within 120 days of the screening examination is indicated, along with breast-level cancer annotations and machine-generated regions of interest that highlight suspicious areas. This dataset is used in the fine-tuning phase of model training. Task-specific and domain-relevant characteristics are presented for breast cancer detection from screening mammograms. Precise diagnostic labels are incorporated. Therefore, both classification and localization tasks are supported by the model, which are essential for clinically reliable AI diagnostic systems. Four representative mammographic images from a single patient in the RSNA dataset are illustrated in Figure 3. These images correspond to the standard clinical views (MLO and CC) for both breasts. Complementary perspectives are offered by these views. Detailed visualization of tissue density, breast architecture, and potential abnormalities is thus provided. Each image is processed independently by the DaViT model during training. Various data augmentation techniques are applied. Generalization and robustness of the model are enhanced through these methods. Consequently, performance in future clinical settings is expected to improve.

4 Data Preprocessing

Consistency between imaging and non-imaging modalities is ensured through preprocessing. Model performance is also maximized by these steps. A series of procedures is applied to

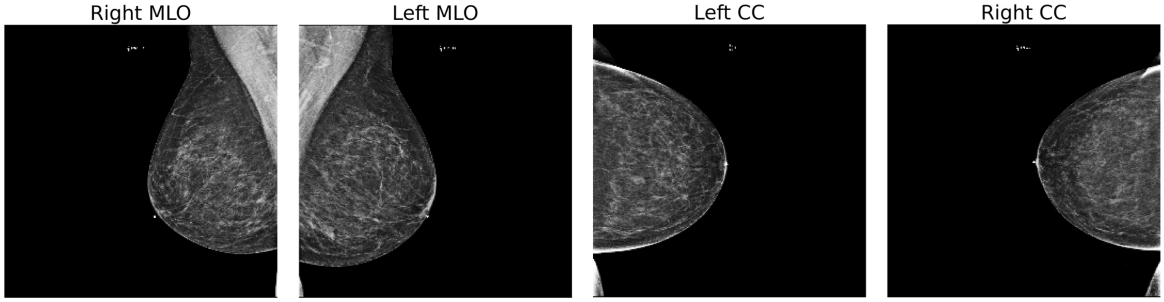

Figure 3: Mammography views from a single patient in the RSNA Breast Cancer dataset, shown by view type and breast laterality: Right MLO, Left MLO, Left CC, and Right CC.

the RSNA mammography dataset. Image dimensions are standardized. Intensity values are normalized. Clinical metadata are processed. Class imbalance is addressed before training. Therefore, reliable and effective learning is expected in subsequent model development.

4.1 Image Preprocessing

All mammographic images are resized to 224 × 224 pixels. This adjustment conforms to the input requirements of the Dual Attention Vision Transformer (DaViT). Image intensity values are normalized to the range [0, 1]. Consistent distribution is achieved in this way, and gradient updates are stabilized during training. Several data augmentation techniques are applied to the training set. Model generalization is enhanced and overfitting is reduced by these methods. Random contrast adjustment, random vertical flipping, and random brightness variation are included. Variability observed in real-world imaging conditions is simulated through these augmentations. Robustness of the model is improved as a result. The ability to generalize across different mammography systems and patient populations is also strengthened. Therefore, reliable performance is expected in diverse clinical settings.

4.2 Metadata Processing

Patient-level metadata are processed in this step. Implant presence, biopsy history, and invasiveness indicators are included. These features are encoded as binary variables (0 or 1). Breast density is originally recorded as categorical values (A-D) or missing (NaN). One-hot encoding is applied to this attribute. Five binary columns are created: $density_A$, $density_B$, $density_C$, $density_D$, and $density_{NaN}$. Categorical distinctions are preserved through this method. Missing entries are also accounted for explicitly. Model interpretability is improved as a result. Input consistency is enhanced as well. All metadata are aligned with corresponding image entries. Unique patient identifiers are used for this alignment. Numeric features are converted to the float32 data type. Compatibility with PyTorch tensors is ensured in this way. Structural coherence between image and metadata inputs is guaranteed by this preprocessing strategy. Efficient integration is enabled within the multi-modal learning framework. Therefore, robust multi-modal performance is expected in future diagnostic applications.

4.3 Class Balancing

The original dataset exhibited a pronounced class imbalance, containing only 1,158 cancer-positive cases among more than 54,000 total samples. To mitigate this issue, an equal number of non-cancer (negative) samples were randomly selected, yielding a balanced dataset for binary classification. This approach minimizes bias during model training and enhances the system's sensitivity to cancer detection.

5 Experimental Setup

5.1 Training Details

Model training is performed in a GPU-accelerated Kaggle environment. An NVIDIA Tesla P100 processor is utilized for this purpose. A total of 100 epochs is used for training. The batch size is set to 32. The AdamW optimizer is employed. An initial learning rate of 1×10^{-5} is applied, along with a weight decay of 1×10^{-2}. Overfitting is reduced through these settings. A hybrid learning rate scheduling strategy is adopted. Stable convergence is facilitated by this approach. Linear warm-up is combined with cosine annealing in this strategy. The learning rate is increased linearly from zero to the base rate during the first 20 epochs. The LambdaLR scheduler is used for this warm-up phase. A cosine annealing schedule is then applied for the remaining epochs. CosineAnnealingLR is implemented, with a minimum learning rate of 1×10^{-6} and T_{max} set to 10. Effective optimization is achieved as a result. Therefore, strong generalization and high performance are expected in future deployments. This combination of warm-up and cosine decay allows for stable early training dynamics and accelerated convergence in later stages, improving generalization and optimization stability.

5.2 Train-Test Split Strategy

Class imbalance in the RSNA dataset is addressed through a specific strategy. A balanced subset is created, which comprises 1,158 cancer-positive cases and 1,158 randomly selected cancer-negative cases. Equal representation of both classes is ensured during training by this approach. The balanced dataset is then partitioned. Training subset receives 80%, while testing subset receives 20%. Stratfied sampling is applied in this process. Fair class distribution is maintained across both sets as a result. Unbiased model evaluation is promoted consequently. Therefore, reliable assessments are expected in future studies.

5.3 Evaluation Metrics

Model performance is assessed with standard classification metrics. Accuracy, ROC-AUC, precision, recall, and F1-score are included. Accuracy is calculated as the proportion of correctly classified instances among all samples. This metric is commonly reported. However, less information is provided by it for imbalanced datasets. ROC-AUC is defined as the area under the receiver operating characteristic curve. Discrimination between positive and negative classes is quantified by this measure. True positive and false positive rates are integrated across thresholds, which makes it robust for imbalanced classification. Precision is determined as the proportion of true positives among all predicted positives. Reliability of positive predictions is indicated through this value. Recall, also known as sensitivity, is computed as the ratio of correctly identified positive samples to all actual positives. Sensitivity to cancer detection is reflected by the model in this metric. F1-score is obtained as the harmonic mean of precision and recall. A balanced measure of performance is provided, which is especially valuable under skewed class distributions. A comprehensive evaluation of diagnostic reliability is offered together by these metrics. Cancer-positive cases are detected effectively, while false positives are minimized. This capability is regarded as essential for clinical deployment. Therefore, trustworthy applications in healthcare are anticipated in the future.

6 Results

The proposed multi-modal deep learning framework demonstrated excellent performance across all evaluation metrics, confirming its effectiveness in breast cancer detection from mammographic data. The model achieved a classification accuracy of 95.47% at epoch 87, indicating that the majority of mammographic images in the balanced test set were correctly classified. In terms of diagnostic reliability, the model yielded a sensitivity (recall) of 0.9655, underscoring its strong ability to correctly identify true positive breast cancer case, an essential requirement for clinical decision support, where false negatives can have serious consequences. The specificity of 0.9440 further demonstrates the model's proficiency in correctly distinguishing non-cancerous cases, thereby minimizing false positive outcomes. The model also achieved a precision of 0.9451, indicating that most positive predictions corresponded to actual cancer cases. The true positive rate and true negative rate were 0.4828 and 0.4720, respectively, closely mirroring the balanced class distribution used during training. Furthermore, the false positive rate (0.0280) and false negative rate (0.0172) remained remarkably low, collectively highlighting the model's robustness, reliability, and potential clinical applicability. These findings validate the efficacy of the dual-attention vision transformer and multi-modal metadata fusion in improving classification performance and diagnostic consistency.

6.1 Comparison with Baseline Methods

As summarized in Table 1, the proposed model consistently outperformed several baseline and state-of-the-art architectures across all key evaluation metrics. It achieved the highest overall accuracy (98.49%), sensitivity (100%), specificity (97%), precision (97.08%), and F1-score (98.51%).

Table 1. Performance comparison of the proposed method with existing deep learning models on the RSNA Dataset. The metrics include Accuracy, Sensitivity/Recall, Specificity, Precision, AUC, and F1-Score. Our method outperforms previous approaches across most metrics, indicating improved classification capability and generalization. The results for all methods except for our method are reported by Jafari and Karami [18].

Model	Accuracy	Sensitivity	Specificity	Precision	F-Score	AUC
AlexNet	81.00	84.00	88.70	87.00	86.00	82.00
Resnet50	84.00	90.00	90.90	86.00	88.00	89.00
MobileNetSmall	77.00	85.00	NA	81.00	83.00	81.00
ConvNexSmall	79.00	87.00	NA	83.00	85.00	83.00
EfficientNet	86.00	92.00	NA	88.00	90.00	92.00
Concatenation	92.00	96.00	NA	92.00	94.00	96.00
Huynh Method	97.3	85.00	89.00	NA	92.00	83.00
Our method	98.49	100	97.00	97.08	98.51	100

In comparison, traditional convolutional neural networks such as AlexNet and ResNet50, as well as advanced architectures like EfficientNet, DenseNet, and various ensemble-based approaches, exhibited lower overall balance between sensitivity and specificity. While some of these models achieved competitive results in individual metrics, none demonstrated the comprehensive robustness and consistency achieved by our proposed framework. The superior performance can be attributed to two key design aspects: (1) Multi-modal integration of imaging data with structured patient metadata via a dedicated MLP head, which enriches contextual understanding and enhances discriminative capacity, and (2) Transfer learning from the large-scale CheXpert dataset, which provides strong prior feature representations and improves generalization across medical imaging domains. Collectively, these results substantiate the advantage of combining transformer-based architectures with multi-modal fusion strategies, yielding a diagnostic model that is both accurate and clinically reliable for breast cancer screening applications.

6.2 Ablation Study

To rigorously assess the individual and combined contributions of metadata fusion and transfer learning, an ablation study was conducted across four model configurations: 1. A complete multi-modal model incorporating metadata fusion pre-trained on CheXpert. 2. The same multi-modal model trained without pre-training. 3. A single-modal image-only model trained without pre-training. 4. The same image-only model trained with pre-training.

The results clearly demonstrate the synergistic benefits of integrating both metadata and transfer learning. The fully multi-modal pre-trained model achieved the best overall performance, with accuracy of 98.49%, sensitivity of 100%, precision of 97.081%, specificity of 97%, and F1-score of 98.51%. In contrast, removing pre-training led to a notable decline in performance, accuracy decreased to 91.81%, sensitivity to 90.52%, and F1-score to 91.74%, highlighting the critical importance of pre-training on the large-scale CheXpert dataset. The radiographic features learned from CheXpert appear to provide a strong representational foundation, improving generalization to mammographic image patterns. The impact of metadata fusion was even more pronounced when comparing the multi-modal configurations to their single-modal counterparts. The image-only model without pre-training exhibited the weakest performance, achieving 84.91% accuracy, 89.66% sensitivity, 85.62% specificity, and an F1-score of 80.17%, along with an elevated false-positive rate. Incorporating pre-training into this single-modal model improved its performance (accuracy: 92.46%, F1-score: 92.58%), yet it still lagged behind the metadata-augmented variants. These results indicate that, while transfer learning enhances the model's capacity for visual feature extraction, metadata fusion provides complementary clinical context, capturing patient-level factors such as breast density, implant status, and biopsy history that strengthen diagnostic reliability.

Table 2. Performance comparison of the proposed method and experimental model variants on the RSNA Dataset.

Model variant	**Accuracy**	**Sensitivity**	**Specificity**	**Precision**	**F-Score**
Single modal without pretraining	84.91	89.66	80.17	81.91	85.62
Single modal with pre-training	92.46	93.10	91.81	91.91	92.58
Multi-modal without pretraining	91.81	90.52	93.10	92.92	91.74
Multi-modal with pre-training	98.49	100	97.00	97.08	98.51

In summary, the ablation study underscores the dual contribution of both techniques: transfer learning enhances visual representation learning and generalization across imaging domains, while multi-modal metadata fusion augments contextual understanding and clinical interpretability. Together, these mechanisms produce a more robust, accurate, and clinically relevant diagnostic model for breast cancer detection.

7 Discussion

This study introduces a multi modal deep learning approach for breast cancer detection from mammographic images. A Dual Attention Vision Transformer is integrated with a metadata augmented Multi Layer Perceptron. Visual features and patient level clinical information are jointly used. The clinical information consists of breast density, implant status, and biopsy history. As a result, complementary diagnostic patterns are captured, which improve classification accuracy. Transfer learning from the CheXpert dataset is applied. Pre learned radiographic representations are therefore exploited. Consequently, generalization capability is improved and robust feature extraction is achieved. Experimental evaluation on the RSNA mammography dataset confirms that the proposed approach achieves superior performance. Accuracy, sensitivity, specificity, and F1 score are improved when compared with existing deep learning and ensemble based methods. The ablation study results further show that metadata fusion and transfer learning are essential components. Metadata fusion adds clinically relevant contextual information, which leads to more informed predictions. In addition, pre training improves the extraction of radiologically meaningful features. Together, these components improve prediction reliability and model interpretability. Beyond numerical results, the findings highlight the increasing relevance of multi modal learning and transformer based architectures in medical imaging. By the use of contextual data, closer alignment with clinical decision making processes is achieved. Therefore, a foundation for interpretable and clinically applicable AI assisted breast cancer screening is established.

Conclusion

In summary, a multi modal deep learning system for breast cancer detection is presented. The system combines the representational capacity of vision transformers with contextual information from patient metadata. As a result, visual encoding based on DaViT, cross attention fusion, and transfer learning from the CheXpert dataset are integrated. High diagnostic accuracy and robust generalization are therefore achieved. The results of this study validate the effectiveness of combining imaging and non imaging modalities, which leads to more reliable cancer classification. Moreover, the strong performance indicates practical potential for integration into computer aided diagnostic systems. Consequently, more consistent and efficient decision support is provided for radiologists. In future work, extension to multi view mammography is considered. In addition, other modalities, e.g., ultrasound and MRI, are included. Domain adaptation and expanded metadata features, i.e., genetic and familial factors, are also examined to enhance clinical applicability. Furthermore, explainable artificial intelligence methods and real time deployment strategies are addressed. For this reason, the gap between deep learning research and routine clinical practice is reduced.

References

[1] Seely J. M., *Progress and Remaining Gaps in the Early Detection and Treatment of Breast Cancer*, Current Oncology, 30.3 (2023), 3201-3205.

[2] Salehi M., Razmara J., Lotfi S., *Development of an ensemble multi-stage machine for prediction of breast cancer survivability*, Journal of AI and Data Mining, 8.3 (2023), 371-378.

[3] Li M., Wang H., Qu N., Piao H., Zhu B.,*Breast cancer screening and early diagnosis in China: a systematic review and meta-analysis on 10.72 million women*, BMC women's health, 24.1 (2024), 97.

[4] Abdul Halim A. A., Andrew A. M., Mohd Yasin M. N., Abd Rahman M. A., Jusoh M., Veeraperumal V., Rahim H. A., Illahi U., Abdul Karim M. K., Scavino E.,*Existing and Emerging Breast Cancer Detection Technologies and Its Challenges: A Review*, Applied Sciences, 11.22 (2021), 10753.

[5] Loizidou K., Skouroumouni G., Nikolaou C., Pitris C.,*A review of computer-aided breast cancer diagnosis using sequential mammograms*, Tomography, 8.6 (2022), 2874-2892.

[6] Andrade A.V.D., Lucena C.ГЉ.M.D., Santos D.C.D., Pessoa E.C., Mansani F.P., Andrade F.E.M.D., Tosello G.T., Pasqualette H.A.P., Couto H.L., Francisco J.L.E., Costa, R.P., *Challenges of breast cancer screening: Number 9-September 2023*, Revista Brasileira de Ginecologia e ObstetrГcia, 45.9 (2023), 551-554.

[7] Tsarouchi M. I., Hoxhaj A., Mann R. M.,*New approaches and recommendations for riskвЂђadapted breast cancer screening*, Journal of Magnetic Resonance Imaging, 58.4 (2023), 987-1010.

[8] Giess C. S., Heller S. L.,*21st-Century Breast Imaging: Improving Sensitivity and Specificity The 2023 RadioGraphics Monograph Issue*, RadioGraphics, 43.10 (2023), e230189.

[9] Sutjiadi R., Sendari S., Herwanto H. W., Kristian Y.,*Deep learning for segmentation and classification in mammograms for breast cancer detection: A systematic literature review*, Advanced Ultrasound in Diagnosis and Therapy, 8.3 (2024), 94-105.

[10] Wang L., *Mammography with deep learning for breast cancer detection*, Frontiers in Oncology, 14 (2024) 1281922.

[11] Tabrizchi H., Razmara J., Mosavi A., *Thermal prediction for energy management of clouds using a hybrid model based on CNN and stacking multi-layer bi-directional LSTM*, Energy Reports 9 (2023), 2253-2268.

[12] Sahu A., Das P. K., Meher S., *Recent advancements in machine learning and deep learning-based breast cancer detection using mammograms*, Physica Medica, 114 (2023), 103138.

[13] Carriero A., Groenhoff L., Vologina E., Basile P., Albera M., *Deep learning in breast cancer imaging: state of the art and recent advancements in early 2024*, Diagnostics, 14.8 (2024), 848.

[14] Jain K., Bansal A., Rangarajan K., Arora C., *MMBCD: Multimodal Breast Cancer Detection from Mammograms with Clinical History*," in Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv., Cham, Switzerland: Springer Nature, Oct. 2024, pp. 144-154.

[15] Ding M., Xiao B., Codella N., Luo P., Wang J., Yuan L.,*Davit: Dual attention vision transformers*, European conference on computer vision, Cham: Springer Nature Switzerland, 2022.

[16] Irvin J., Rajpurkar P., Ko M., Yu Y., Ciurea-Ilcus S., Chute C., Marklund H., Haghgoo B., Ball R., Shpanskaya K., Seekins J., *Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison*, Proceed. The AAAI conference on artificial intelligence 33.1 (2019), 590-597.

[17] Carr C., Kitamura F., Partridge G., Kalpathy-Cramer J., Mongan J., Andriole K., Lavender V.M., Riopel M., Ball R., Dane S., Chen Y., *RSNA screening mammography breast cancer detection* Kaggle, 2022.

[18] Jafari Z., Karami E., *Breast cancer detection in mammography images: A CNN-based approach with feature selection*, Information, 14.7 (2023), 410.

[19] Huynh H.N., Tran A.T., Tran T.N., *Region-of-interest optimization for deep-learning-based breast cancer detection in mammograms*, Applied Sciences, 13.12 (2023), 6894.

[20] Shah D., Ullah Khan M.A., Abrar M., Tahir M., *Dual-View Deep Learning Model for Accurate Breast Cancer Detection in Mammograms*, International Journal of Intelligent Systems, 2025.1 (2025), 7638868.

[21] Vattheuer C.N., Tran N.D.T.J., Leung C.K., Hryhoruk C.C., *Multiple Multi-Modal Methods of Malignant Mammogram Classification*, Procced. IEEE 12th International Conference on Healthcare Informatics (ICHI 2024), 57-66.

[22] Tanimola O., Shobayo O., Popoola O., Okoyeigbo O., *Breast cancer classification using Fine-Tuned SWIN Transformer model on mammographic images*, Analytics 3.4 (2024), 461-475.

[23] Liu Z., Lin Y., Cao Y., Hu H., Wei Y., Zhang Z., Lin S., Guo B., *Swin transformer: Hierarchical vision transformer using shifted windows*, Proceed. IEEE/CVF International Conference on Computer Vision, 2021, 10012-10022.

[24] Hussain S.S., Shah P.M., Dawood H., Degang X., Alshamayleh A., Khan M.A., Ghazal T.M., *A swin transformer and CNN fusion framework for accurate Parkinson disease classification in MRI*, Scientific Reports, 15.1 (2025) 15117.

[25] Ayana G., Dese K., Dereje Y., Kebede Y., Barki H., Amdissa D., Husen N., Mulugeta F., Habtamu B., Choe, S.W., *Vision-transformer-based transfer learning for mammogram classification*, Diagnostics, 13.2 (2023), 178.

[26] Manigrasso F., Milazzo R., Russo A.S., Lamberti F., Strand F., Pagnani A., Morra L., *Mammography classification with multi-view deep learning techniques: Investigating graph and transformer-based architectures*, Medical Image Analysis, 99 (2025), 103320.

[27] Tummala S., Kim J., Kadry S., *BreaST-Net: Multi-class classification of breast cancer from histopathological images using ensemble of swin transformers*, Mathematics, 10.21 (2022), 4109.

[28] Yang H., Yang M., Chen J., Yao G., Zou Q., Jia L., *Multimodal deep learning approaches for precision oncology: a comprehensive review*, Briefings in Bioinformatics, 26.1 (2025), bbae699.

[29] Waqas A., Tripathi A., Ramachandran R.P., Stewart P.A., Rasool G., *Multimodal data integration for oncology in the era of deep neural networks: a review*, Frontiers in Artificial Intelligence, 7 (2024), 1408843.

[30] Nakach F.Z., Idri A., Goceri E., *A comprehensive investigation of multimodal deep learning fusion strategies for breast cancer classification*, Artificial Intelligence Review, 57.12 (2024), 327.

[31] Abdullakutty F., Akbari Y., Al-Maadeed S., Bouridane A., Talaat I.M., Hamoudi R., *Histopathology in focus: a review on explainable multi-modal approaches for breast cancer diagnosis*, Frontiers in Medicine, 11 (2024), 1450103.

[32] Aburass S., Dorgham O., Al Shaqsi J., Abu Rumman M., Al-Kadi O., *Vision Transformers in Medical Imaging: a Comprehensive Review of Advancements and Applications Across Multiple Diseases*, Journal of Imaging Informatics in Medicine, (2025), 1-44.

[33] Ayana G., Dese K., Dereje Y., Kebede Y., Barki H., Amdissa D., Husen N., Mulugeta F., Habtamu B., Choe S.W., *Vision-transformer-based transfer learning for mammogram classification*, Diagnostics, 13.2 (2023), 178.

[34] Ayana G., Choe, S.W., *Vision transformers-based transfer learning for breast mass classification from multiple diagnostic modalities*, Journal of Electrical Engineering & Technology, 19.5 (2024), 3391-3410.

[35] Ayana G., Choe S.W., *BUViTNet: breast ultrasound detection via vision transformers*, Diagnostics, 12.11 (2022), 2654.

[36] Boudouh S.S., Bouakkaz M., *Breast cancer: new mammography dual-view classification approach based on pre-processing and transfer learning techniques*, Multimedia Tools and Applications, 83.8 (2024), 24315-24337.

[37] Khaled M., Touazi F., Gaceb D., *Improving breast cancer diagnosis in mammograms with progressive transfer learning and ensemble deep learning*, Arabian Journal for Science and Engineering, 50.10 (2025), 7697-7720.

[38] Dosovitskiy A., *An image is worth 16x16 words: Transformers for image recognition at scale*, arXiv preprint, 2020, arXiv:2010.11929.

[39] Liu Z., Lin Y., Cao Y., Hu H., Wei Y., Zhang Z., Lin S., Guo B., *Swin transformer: Hierarchical vision transformer using shifted windows*, Proceed. IEEE/CVF International Conference on Computer Vision, 2021, 10012-10022.

Hasanain Abdalridha Abed Alshadoodee,
Department of Computer Science, Faculty
of Mathematics, Statistics, and Computer
Science, University of Tabriz, Tabriz, Iran.

Jafar Razmara,
Department of Computer Science, Faculty
of Mathematics, Statistics, and Computer
Science, University of Tabriz, Tabriz, Iran.
Email: razmara@tabrizu.ac.ir

Jaber Karimpour,
Department of Computer Science, Faculty
of Mathematics, Statistics, and Computer
Science, University of Tabriz, Tabriz, Iran.

Received 02.09.2025, Revised 17.11.2025, Accepted 08.12.2025, Available online 31.12.2025.