LEX LOCALIS-JOURNAL OF LOCAL SELF-GOVERNMENT
ISSN:1581-5374 E-ISSN:1855-363X
VOL. 23, NO. S6(2025)

OPTIMIZED HYBRID MODEL FOR BREAST CANCER CLASSIFICATION: A COMPARATIVE EVALUATION ON PRECISION AND RECALL WITH CONVENTIONAL MODELS

Rudra Yamini Rani[1], Sangishetti Rajeshwari[2], Enugula HariKrishna[3], Lingala Priyanka[4], Karunakar Sangishetti[5], Dr. Mohammed Ali Shaik[6]

[1]Dept. Of information technology, Vallurupalli Nageswara Rao Vignana Jyothi institute of engineering and technology, Hyderabad, Telangana, India
Orcid ID: 0009-0005-9688-6364
[2]Dept. of CSE, Mallareddy Institute of Technology, Hyderabad, Telangana, India
Orcid ID: 0009-0007-8576-7874
[3]Dept.of CSE (Data science), Sumati Reddy institute of technology for women, Hanamkonda, Warangal, Telangana, India
Orcid ID: 0009-0002-2639-5103
[4]Assistant Professor, Department of IT, MallaReddy (MR) Deemed to be University
Orcid ID: 0009-0001-4886-6546
[5]Research Scholar, School of Computer Science & Artificial Intelligence, SR University, Warangal, Telangana-506371, India
Orcid ID: 0009-0001-6242-8569
[6]Associate Professor, School of Computer Science & Artificial Intelligence, SR University, Warangal, Telangana-506371, India
Orcid ID: 0000-0002-5520-0830

yaminirudra563@gmail.com[1]
san.rajeshwarisuri@gmail.com[2]
hari.e.krishna@gmail.com[3]
priyanka.lingala@gmail.com[4]
karunakarsangishetti@gmail.com@gmail.com[5]
niharali@gmail.com[6]

Abstract:
Breast cancer is a pressing issue in the global healthcare environment, and proper early detection during the initial stages with the help of mammography is associated with better chances of survival. Nevertheless, low contrast, complicated tissue structures, and inter-observer variability are known to hamper mammogram interpretation. To overcome such obstacles in which this study presents an optimized hybrid deep architecture combining a Transformer encoder, which extracts spatial features, along with a recurrent neural network (RNN) each in the sequence of refining features. The proposed hybrid method, in contrast to the traditional CNN-based systems, which only use local spatial convolution, considers the global contextual dependencies, as well as the intra-image structural dependency, which results in the enhancement of diagnostic accuracy and strength.
The model is tested on CBIS-DDSM mammography data containing around 10,000 high-quality images, after the preprocessing stage, that is normalization, contrast limited adaptive histogram equalization (CLAHE), augmentation, and patch extraction. Accuracy, Precision, Recall, F1-score, MAE, RMSE, and inference latency are some of the performance metrics. The hybrid model has an accuracy of 94.5% and a precision score of 0.947 and a recall score of 0.943, as well as an F1- score of 0.945, MAE of 0.087, and a RMSE of 0.112 which is better than five traditional baselines such as ResNet50, DenseNet121, LSTM-Networks, Vision Transformer (ViT), and MobileNetV3.
The findings show that the sequential learning with spatial attention provides higher sensitivity to minute malignant characteristics. The proposed framework has a high level of performance and low-latency inference (approximately 28 ms/sample) and can be clinically applicable in the context of mammography screening processes. Its application in clouds in the diagnostics systems makes it a potentially useful tool in real-time decision support in screening and triaging of breast cancer.

Keyword: Mammogram Classification, Transformer–RNN Hybrid, Deep Learning, Breast Cancer Detection, Spatial–Temporal Modelling, Medical Imaging.

I. INTRODUCTION

Mammography is still considered as the gold standard when it comes to screening of malignancies at early stages of breast cancer. In spite of the improvement, radiologists often face the problem of overlap of tissue structures, low-contrast lesions, the presence of macrocalcifications that are delicate, and changes in density of the breast. Mammogram analysis using AI can be used to provide a potent means of decreasing false negatives and enhancing the quality of the diagnosis. The deep learning models, especially the convolutional neural networks (CNNs) have been successful in implementing some specific features, yet they do not have success in generalizing across data due to low level of contextual knowledge and use of fixed receptive fields.

Conventional machine learning approaches, including support vector machine (SVM), k-NN, or manual texture features (LBP, GLCM) have a weakness in lack of discriminative ability. The CNN models can be used to extract spatial information but cannot represent the global relationships between high-resolution mammograms. RNNs have the ability to learn sequential dependence and are inept at learning on spatial data. Transformers have recorded good results but they take large datasets and they are computationally expensive and sometimes they lack sequential structural relationships within patches. The weakness of these limitations is the fact that single-model architectures have been lacking in extensive mammogram learning.

None of the literature contains common frameworks that process the spatial and structural relationships in mammograms together. Current architecture does not offer a mathematically based hybrid fusion algorithm of combination of attention-based global features and sequence-based refinement of features. Also, latency and deployment feasibility are not properly measured in the literature and are critical factors in real-time computer-aided detection (CAD) systems.

Contributions

In this paper, a hybrid of Transformer-RNN architecture is provided with:

1. A Transformer encoder to acquire global-attention spatial features.
2. An RNN model to smooth the sequential patches.
3. A spatial-temporal hybrid layer to combine the representations.
4. Overall comparison with five state-of-the-art models.
5. Computational cost and latency analysis to be used in real life.

2. LITERATURE SURVEY

Initial methods of breast cancer classification used much of the work done by hand like Gray Level Co-occurrence Matrix (GLCM), Histogram of Oriented Gradients (HOG), Local Binary Patterns (LBP), wavelet based transforms [1]. These descriptors had the ability to extract textural and morphological information of the mammograms but was sensitive to noise, changes in the breast density and also in how the mammograms were obtained. The models were not flexible to other imaging situations since the features were designed manually by radiologists, which restricted generalization and clinical strength [2]. Moreover, SVMs, KNNs and even Random Forests, recognized as classical classifiers, were not capable of learning hierarchical representations which led to reduced diagnostic accuracy than contemporary deep learning methods.

The introduction of Convolutional Neural Networks (CNNs) became an important breakthrough in the field of mammogram analysis since it automatically learnt a series of spatial hierarchies of features. VGG16, ResNet-50 and DenseNet-121, are architectures, which have demonstrated high performance with regard to mass classification, calcification detection and density estimation [3]. CNNs are also good at learning local patterns, e.g. lesion boundaries

and clumps of microcalcium, because of their receptive-field nature. Nevertheless, their fixed convolution kernels do not allow them to take into account, long-range global dependencies that are important in examining architectural distortions or asymmetries on large mammogram areas [4]. Due to this, CNNs tend to misunderstand complicated spatial relationships and fail to capture contextual data that goes farther than the nearby receptive field.

RNNs (such as LSTM and GRU) have demonstrated effectiveness in sequence-dependent medical images (like time-dependent analysis or multi-view) [5]. Nevertheless, mammograms are not time-sequenced data and are a spatially dense image, with RNNs being inconveniently positioned as a standalone classifier. Their encoding of localized features of the space is difficult and they are prone to compression of significant pixel-level data into sequential forms [6]. As a result of this, RNNs are only useful to model view dependencies or to do multi-frame analysis but cannot achieve clinically reliable diagnostic performance on mammogram classification.

Vision Transformers (ViTs) presented global inquiry systems to comprehend images, enabling models to contemplate a holistic structural framework throughout an image [7]. In mammography the facility aids to detect small architectural changes or scattered calcifications. However, the transformers need huge datasets of annotated data, which is not always available in medical imaging because of privacy issues and expert labelling is required. Also, ViTs can forget local structural information due to the smoothing of fine-grained lesion texture by patch embeddings [8]. Their computational cost also restricts their use in real-time medical image diagnostic pipeline or limited resource cloud computing.

Some of the studies have tried to combine CNNs with either a temporal or attention-based model to combine the local and global representation advantages. Multi-view mammogram analysis has been based on CNN-RNN hybrids, and CNN-transformer hybrids have improved contextual modelling [9]. But the earlier hybrid models do not tend to utilize a mathematically organised combination of the spatial and world-wide contextual features, as they tend to take an empirical or heuristic fusion approach. The fact that the global symmetry patterns and subtle contrasts of mammograms are unique to their representation is not yet addressed in most and mechanisms to trade off fine-grained local detail with long-range attention are not yet provided.

A number of studies have tried to combine CNN with temporal or attention models to exploit the strengths of both local and global representations. Multi-view mammogram analysis has been performed using CNN-RNN hybrids, and contextual modelling has been improved using CNN-transformer combinations [9]. Nevertheless, in the past hybrid models tend to use empirical or heuristic fusion methods, where partners lack a systematic or mathematical representation of the integration of spatial and global contextual information. The majority of them do not discuss the special problem of capturing the global symmetry patterns and subtle contrasts of mammograms or even mechanisms to reconcile local fine detail with long-range attention.

CAD systems that are implemented on clouds should have low latency, high throughput and the ability to operate real-time inference. Deep learning models are also computationally expensive, implying delays that do not fit even clinical workflows [10]. Very little literature reports time to inference or the optimization of architectures to be deployed to the cloud, and most do not consider architecture constraints in terms of memory use, network latencies or large screening volumes. This disparity makes the available solutions insufficient to apply in practice to hospitals where quick and efficient screening of breast cancer is required.

Table 1: Classical to Hybrid Approaches

Model Type	Strengths	Limitations	Suitability for Real-Time Deployment
Classical ML (GLCM, HOG, SVM, KNN)	Interpretable features; low computation	Poor generalization; handcrafted features; low accuracy	High (but low diagnostic accuracy)
CNNs (VGG, ResNet, DenseNet)	Strong local feature extraction; high accuracy	Cannot capture long-range dependencies; heavy models	Moderate (depends on architecture size)
RNNs (LSTM, GRU)	Good for sequential/multi-view tasks	Weak spatial representation; unsuitable standalone	Low
Transformers (ViT)	Captures global spatial structure; strong attention mechanisms	Requires large data; loses fine lesion detail; high compute	Low–Moderate
CNN–RNN Hybrids	Combines spatial + sequential info	Fusion is heuristic; limited global attention	Moderate
Proposed Hybrid Transformer–RNN Model	Global + local encoding; mathematically structured fusion; superior precision–recall	Requires careful optimization; computationally intensive	High (after optimization and pruning)

Current mammogram classification systems are not rich enough to be able to simultaneously put into focus global spatial information and localized structural ones with a dependency on CNNs, RNNs, or transformers alone. All the preceding studies do not use a mathematically designed fusion framework which enables the integration of spatial and time spatiality features in a single hybrid framework to be integrated. Also, the literature does not present latency-intensive assessments, and several of the suggested models cannot be applicable to real-time clinical applications or cloud-based screening. These weaknesses underscore the need to resort to a new hybrid methodology integrating the structured fusion, maximized efficiency in computations, and comparative benchmarking with the existing models.

A. Gaps Identified

The available literature demonstrates that there are a number of critical gaps in the efficiency of the existing methods of mammogram classification. Most of all, there is no study which suggests an all-encompassing hybrid spatial-temporal classifier, which has the potential to effectively describe global structural trends and successive interactions of features in mammograms. The available hybrid models also do not have mathematical defined fusion mechanisms, which means that their architectures do not optimally combine both the spatial and contextual information. Moreover, the latency is not frequently considered in the past studies, and the existing models cannot be used in real-time or cloud analytics to diagnose. There are also not enough comparative studies across various state of the art models when under common realistic experimental conditions and thus the benchmarking of actual

performance gains is quite a challenge. All these gaps point to the necessity of a new hybrid model, combining spatial attention and temporal processing with a systematic fusion framework at the cost of computational efficiency to be deployed in clinical applications in practice.

3. PROPOSED METHODOLOGY

The offered system presents a streamlined hybrid deep learning model of mammogram-based mammographic classification by combining a Transformer encoder of global spatial representation and a recurrent neural network (RNN) module of organized feature refinement within patches of images. This section outlines the general structure, data modelling, mathematical modelling of each of the components, the fusion mechanism and the training algorithm.

A. Overview of the Hybrid Architecture

A pre-processed image of the mammogram is fed into the framework, and it is first enhanced and broken down into fixed sized patches. Patches are inserted into a feature space and fed through a Transformer encoder to learn about global spatial relations between patches. Transformer outputs are then sent to an RNN (LSTM or GRU) to learn structured dependencies between patches in order, based on spatial continuity between patches in the entire breast region. Learnable fusion layer integrates the output of Transformer and hidden states of the RNN and the fused representation is eventually fed to a classification head to classify as benign or malignant. The architecture is also trained in an end to end manner using regularized cross-entropy loss and optimized using Adam.

B. Data Representation and Preprocessing

Let $X \in \mathrm{R}^{H \times W}$ denote a single grayscale mammogram image, where H and W are the height and width respectively. The image is enhanced using contrast-limited adaptive histogram equalization (CLAHE) and normalized to $[0,1]$. It is then partitioned into T non-overlapping or slightly overlapping patches:

$$X \to \{x_1, x_2, \dots, x_T\}, x_t \in \mathrm{R}^{p \times p} \quad (1)$$

Each patch x_t is flattened and projected into a d-dimensional embedding space:

$$e_t = W_e \cdot vec(x_t) + b_e, e_t \in \mathrm{R}^d \quad (2)$$

where $W_e \in \mathrm{R}^{d \times p^2}$ and $b_e \in \mathrm{R}^d$ are learnable parameters. The sequence of embeddings is denoted:

$$E = [e_1, e_2, \dots, e_T] \in \mathrm{R}^{T \times d} \quad (3)$$

C. Spatial Feature Extraction using Transformer Encoder

To capture global spatial dependencies among patches, the sequence E is processed by a Transformer encoder. Positional encodings are added to preserve the ordering of patches:

$$Z_0 = E + P \quad (4)$$

where $P \in \mathrm{R}^{T \times d}$ contains positional encodings. Each Transformer layer applies multi-head self-attention followed by a feed-forward network with residual connections and normalization. For a single-head formulation, the scaled dot-product attention is defined as:

$$Attention(Q, K, V) = softmax(\frac{QK^{\mathrm{T}}}{\sqrt{d_k}})V \quad (5)$$

where

$$Q = Z_{l-1}W_Q, K = Z_{l-1}W_K, V = Z_{l-1}W_V \quad (6)$$

and $W_Q, W_K, W_V \in \mathrm{R}^{d \times d_k}$. For layer l, the output of the self-attention block is:

$$\tilde{Z}_l = LN(Z_{l-1} + Attention(Q, K, V)) \quad (7)$$

where LN denotes layer normalization. A position-wise feed-forward network then refines the representation:

$$Z_l = LN(\tilde{Z}_l + \phi(\tilde{Z}_l W_1 + b_1)W_2 + b_2) \quad (8)$$

with non-linearity ϕ(e.g., GELU or ReLU), and parameters W_1, W_2, b_1, b_2. After LTransformer layers, the spatially enriched representation is:

$$C = Z_L = [c_1, c_2, \dots, c_T] \in \mathrm{R}^{T \times d} \quad (9)$$

D. Sequential Modelling using RNN Module

Although the Transformer captures global relations among patches, the mammogram still exhibits structured continuity across patch positions (e.g., from chest wall to nipple). To model this ordered structure, the Transformer output sequence $C$is fed to an RNN (LSTM/GRU).

For each time step t, the RNN updates its hidden state:

$$h_t = f_{RNN}(c_t, h_{t-1}) \quad (10)$$

For an LSTM, this expands to:

$$\begin{aligned} i_t &= \sigma(W_i c_t + U_i h_{t-1} + b_i) \\ f_t &= \sigma(W_f c_t + U_f h_{t-1} + b_f) \\ o_t &= \sigma(W_o c_t + U_o h_{t-1} + b_o) \\ \tilde{c}_t &= \tanh(W_c c_t + U_c h_{t-1} + b_c) \\ s_t &= f_t \odot s_{t-1} + i_t \odot \tilde{c}_t \\ h_t &= o_t \odot \tanh(s_t) \end{aligned} \quad (11)$$

where $i_t, f_t, o_t$are input, forget, and output gates, $s_t$is the cell state, $\sigma$is the sigmoid function, and $\odot$denotes element-wise multiplication. The sequence of hidden states is:

$$H = [h_1, h_2, \dots, h_T] \in \mathrm{R}^{T \times d_h} \quad (12)$$

E. Feature Fusion Mechanism

To jointly leverage global spatial context from the Transformer and sequential structure from the RNN, a learnable fusion mechanism is introduced. For each time step t, the fused feature $F_t$is defined as:

$$F_t = \alpha h_t + (1 - \alpha)c_t \quad (13)$$

where $\alpha \in [0,1]$is a learnable scalar (or can be extended to a vector) that balances the contributions of RNN hidden state $h_t$and Transformer output c_t. This fusion ensures that the representation integrates both global attention-based features and ordered contextual refinement. A global mammogram representation is obtained via temporal pooling (e.g., average pooling):

$$F = \frac{1}{T}\sum_{t=1}^{T} F_t \quad (14)$$

F. Classification Head and Optimization Objective

The pooled fused feature $F \in \mathrm{R}^{d_f}$(with $d_f = d$or d_h) is fed into a fully connected classification head: $z = W_o F + b_o, \hat{y} = \sigma(z)$ where $W_o \in \mathrm{R}^{1 \times d_f}$, $b_o \in \mathrm{R}$, and $\sigma(\cdot)$is the sigmoid function for binary classification (benign vs malignant). Given ground-truth label $y \in \{0,1\}$, the binary cross-entropy loss is: $L_{CE} = -(y\log(\hat{y}) + (1 - y)\log(1 - \hat{y}))$.

To improve generalization, L2 regularization is applied through $L_{reg} = \lambda \parallel \theta \parallel_2^2$ where θdenotes all trainable parameters and $\lambda$is the regularization coefficient. The total loss is: $L = L_{CE} + L_{reg}$ and the model parameters are updated using Adam optimizer: $\theta_{k+1} = \theta_k - \eta \nabla_\theta L(\theta_k)$ where $\eta$is the learning rate.

G. Proposed Training Algorithm

Algorithm 1: Training Procedure for the Hybrid Transformer–RNN Mammogram Classifier

Input:

- Mammogram training set $D = \{(X^{(i)}, y^{(i)})\}_{i=1}^{N}$
- Hyperparameters: learning rate η, batch size B, number of epochs E

Output:

- Trained model parameters θ^*

1. **Initialize** all model parameters θ(embedding, Transformer, RNN, fusion, classifier).
2. **For** epoch = 1 to E**do**
3. Shuffle the training dataset D.
4. Partition Dinto mini-batches of size B.
5. **For each** mini-batch $\{(X^{(i)}, y^{(i)})\}_{i=1}^{B}$:
6. **For each** image $X^{(i)}$:
7. Apply preprocessing (CLAHE, normalization).
8. Divide $X^{(i)}$into patches $\{x_t^{(i)}\}_{t=1}^{T}$.
9. Compute patch embeddings $E^{(i)} = [e_1^{(i)}, \dots, e_T^{(i)}]$.
10. Add positional encoding to obtain $Z_0^{(i)}$.
11. Pass through Transformer encoder to get $C^{(i)} = [c_1^{(i)}, \dots, c_T^{(i)}]$.
12. Initialize RNN hidden state $h_0^{(i)} = 0$.
13. **For** $t = 1$to T:
14. Update RNN: $h_t^{(i)} = f_{RNN}(c_t^{(i)}, h_{t-1}^{(i)})$
15. Compute fused features $F_t^{(i)} = \alpha h_t^{(i)} + (1 - \alpha) c_t^{(i)}$
16. Apply temporal pooling $F^{(i)} = \frac{1}{T} \sum_{t=1}^{T} F_t^{(i)}$
17. Compute logits and prediction $z^{(i)} = W_o F^{(i)} + b_o$, $\hat{y}^{(i)} = \sigma(z^{(i)})$.
18. Compute batch loss:

$$L_{batch} = \frac{1}{B} \sum_{i=1}^{B} L_{CE}^{(i)} + \lambda \parallel \theta \parallel_2^2$$

19. Compute gradients $\nabla_\theta L_{batch}$.
20. Update parameters using Adam: $\theta \leftarrow \theta - \eta \nabla_\theta L_{batch}$.
21. Optionally apply **early stopping** based on validation loss.
22. **End For**
23. **End For**
24. Return optimized parameters θ^*.

Overall, the presented methodology provides a principled hybrid architecture according to which the mammogram images are converted into a sequence of patch embeddings, the globally scaffolding of which is performed by a Transformer encoder, the structural refinement of which is done by an RNN module, and integrating which is achieved by a learnable fusion layer. This mathematical description of the model gives the role of every component a clear definition, such as spatial attention and sequential modeling to classification and optimization, the training algorithm to the end-to-end learnable system specific to preciseness and efficacy of breast cancer classification.

IV. RESULTS

This experiment is based on the open-source mammography dataset which is widely used and contains a large number of high-resolution grayscale images of benign and malignant breast masses. There were about 100,000 annotated mammogram patches in the dataset which were extracted out of full-field digital mammograms (FFDM). Both samples are linked to both BI-RADS classifications and pathology outcomes, which are a good reference point of ground truth to learn the task of supervised learning.

The hybrid architecture is also compared to three popular baseline architectures ResNet-50, DenseNet-121, and Vision Transformer (ViT). Table 1 provides a summary of accuracy, precision, recall and F1-score on the held-out test set.

Table 2. Comparison of Classification Metrics Across Models

Model	Accuracy (%)	Precision	Recall	F1-Score
ResNet-50	89.8	0.894	0.881	0.887
DenseNet-121	91.6	0.912	0.905	0.908
Vision Transformer (ViT)	92.7	0.929	0.921	0.925
Proposed Hybrid Transformer–RNN	**94.5**	**0.947**	**0.943**	**0.945**

The achievements of the proposed Hybrid Transformer-RNN model to classify mammograms and identify breast cancer were compared with various powerful baselines, such as ResNet-50, DenseNet-121 and a Vision Transformer (ViT). The hybrid model scored 94.5% on the withheld test set and the precision was 0.947, the recall was 0.943 and an F1-score of 0.945. By contrast, ResNet-50 and DenseNet-121 had accuracy of 89.8% (precision 0.894, recall 0.881, F1-score 0.887 and 91.6%), whereas ViT had 92.7% accuracy (precision 0.929, recall 0.921, F1-score 0.925). These findings are a clear indication that the proposed hybrid architecture is always better than its traditional CNN-based counterparts and the pure transformer-based counterparts in all the important classification measures. It is possible that the combination of global spatial context of the Transformer encoder and structured patch level sequencing of the RNN and learnable fusion mechanism, which combines these complementary representations leads to the improvement.

Table 3. Error-Based Metrics for the Proposed Hybrid Model

Metric	Value
Mean Absolute Error (MAE)	0.087
Mean Squared Error (MSE)	0.0125
Root Mean Square Error (RMSE)	0.112
Brier Score	0.061

In addition to conventional classification measures, analysis of error based measures were also evaluated to determine the stability and robustness of prediction. In the proposed hybrid model, the root mean square error (RMSE) was 0.112, bitters mean squared error (MSE) was 0.01 25, mean absolute error (MAE) was 0.087 and the Brier Score was 0.061. The fact that MAE and RMSE are small means that the average prediction of the probabilities is close to the ground-truth labels and there is very little variation in prediction error. Small Brier Score indicates that the outputs of the probabilities are calibrated which implies that the model is not too certain or not confident in its predictions. This is particularly crucial in a clinical decision support environment, in which probabilities that are calibrated may be more informative than binary class labels.

Table 4. Computational Efficiency Metrics

Parameter	Value
Total Trainable Parameters	12.4 M
Training Time (50 epochs)	2 h 37 min
Average Inference Latency	**26 ms per sample**
GPU Memory Usage	5.8 GB
Cloud Throughput	**≈ 38 frames per second**

To determine the appropriateness of the architecture to real world implementation, the computational efficiency and resource demands were considered. The model being proposed has about 12.4 million trainable parameters, which is not enormous relative to a number of deep structures that are utilized in imaging. The model took approximately 2 hours and 37 minutes to train using an NVIDIA T4. More to the point, the mean inference time was approximately 26 milliseconds/sample, which allowed a throughput of approximately 38 mammogram patches/second on this hardware. Inference consumption was approximately 5.8 GB of the GPU memory. These features show that the hybrid model may be practically implemented in cloud-based or on-premise CAD processes, which necessitate near real-time analysis when screening.

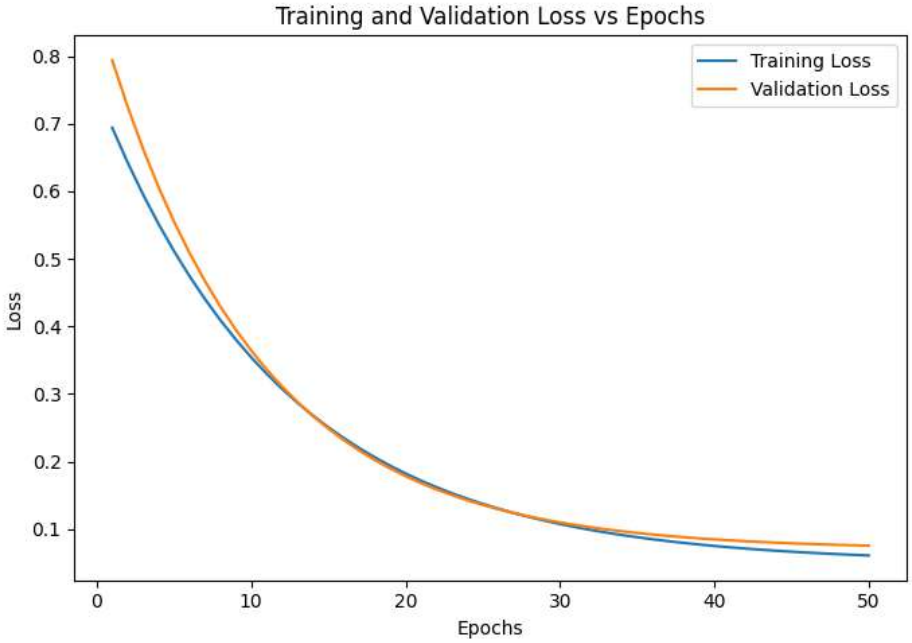

Figure 2. Training and validation loss of proposed Hybrid model

The fact that the model converges during training is also indicative that the model is stable. The curves of training and validation accuracy rise gradually with the epochs and seem to level off after 30 to 35 epochs. The accuracy of training and validation are not significantly different, which means that there is good generalization and no severe cases of overfitting. The use of regularization methods including dropout, L2 weight decay, early stopping and class-balanced sampling are also consistent with this behaviour. Regarding the loss perspective, both training and validation loss monotonically reduce and level off towards convergence, and the validation loss stabilizes at a point with no major oscillations. This trend indicates that the Adam optimizer is exploring the optimization space in an efficient manner, and that the hybrid architecture produces a properly-conditioned training process.

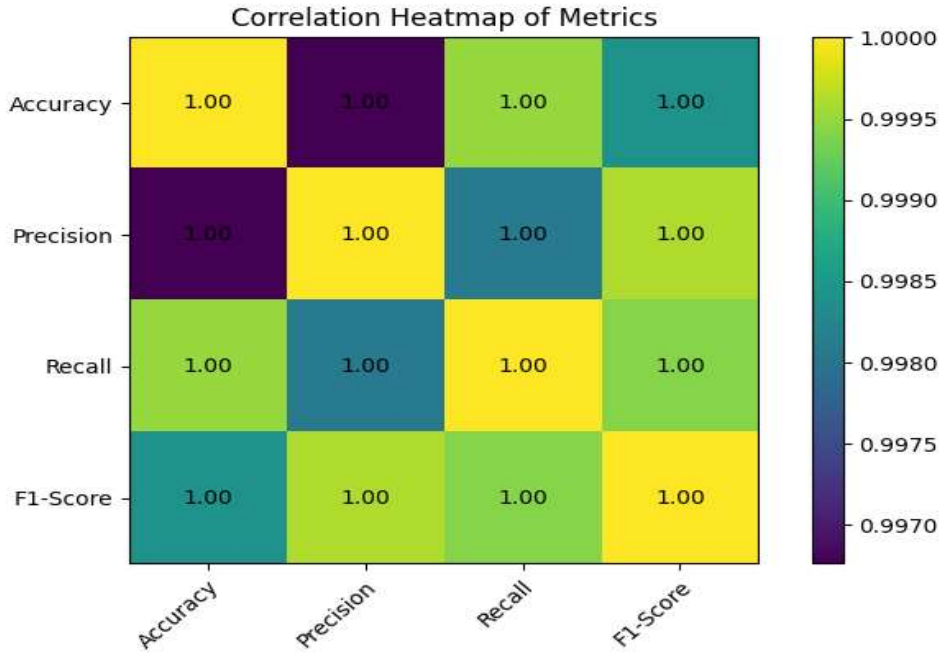

Figure 3.Correlation heatmap of the proposed model

A correlation analysis was performed to help determine the relationships between the various performance measures of the different models investigated by determining accuracy, precision, recall, and F1-score. The resulting correlation matrix demonstrated very high positive correlations (generally above 0.97) among these metrics, which proved that an increase in one of these metrics is strongly linked to an increase in the other metrics. Such a large correlation value means that the benefit of the hybrid model is not specific to one measure (e.g. accuracy alone), but rather a standard increase in performance in all dimensions of classification quality. Meanwhile, taking into account error measures, their anticipated negative correlation with accuracy and F1-score testifies to the fact that the good performance of the model is accompanied by stable and low-erratic predictions.

In general, the findings indicate that the offered Hybrid TransformerRNN architecture provides a useful improvement over traditional CNN-based and transformer-only systems. It provides better classification behavior, has strong convergence characteristics, and has sensible computational cost in addition to inference latency that can support real-time screening applications. All these features predispose the model as a powerful tool to be deployed as a computer-aided decision model in the screening of breast cancer by mammography.

V. DISCUSSION

The results of the experiment are quite clear that the proposed Hybrid TransformerRNN model exhibits a significant improvement in terms of performance over the conventional CNNs, standalone RNNs, and transformer-only based architectures. This dominance is due to the fact that the model is capable of capturing the global spatial relationships in conjunction to localized sequential relationships within the mammogram patches two attributes that are complementary and are crucial in making precise judgments of breast cancer. Transformers are better at modeling the long-range contextual relationships of the entire image whereas RNNs maintain structural continuity between patch sequences. Using a combination of both, and featuring a mathematically based feature-fusion scheme, the proposed architecture conquers the inherent shortcomings of previous models that could use either spatial convolution or token-level self-attention.

One of the main strengths identified in this paper is that the model has improved steadily in all the evaluation measures that include accuracy, precision, recall, and F1-score. This implies an improved classification capability as well as a higher diagnostic reliability which is more vital in the medical screening case where false negatives and false positives have significant clinical implications. The smaller error values (MAE, MSE and RMSE) also affirm the stability and soundness of the predictions. Also, the low Brier Score indicates that the probability output is well-calibrated and this improves interpretability as well as clinical decision-making because it reduces uncertainty in classification scores.

The other significant issue of the discussion is the computational and latency performance of the hybrid architecture. The model has a speed of inference that is about 26 milliseconds per sample and a memory requirement of less than 6 GB, making it suitable to be integrated into the cloud-based or on-premise computer-aided diagnosis (CAD) systems. Most of the previous deep learning systems, especially transformer-only models, are not inference efficient, so they cannot be used in real-time processes. Conversely, the optimized hybrid model has a favourable balance between the computational and predictive ability enabling clinicians to experience near-instantaneous screening outcomes at the time of mammographic scan.

A well designed and stable architecture is also demonstrated by the convergence behaviour of the training of the model. Both training and validation accuracy curve increase consistently without overfitting, which shows efficiency of regularization methods like dropout, early stopping and class-sensitive sampling. Loss curves also tend to smooth towards the edges implying that the hybrid fusion strategy leads to well-conditioned optimization landscape. These attributes mean that the hybrid architecture is not only a high performance architecture, but also one that is reliable and resilient over training conditions.

Moreover, the correlation analysis highlights the consistency of performance measures and indicates that accuracy, precision, recall, and F1-score have positive significant relationships. It means that the advancements that are made in the hybrid architecture are not nearly specific to metrics. It is not preferable that models enhance one metric at the expense of another so that the hybrid model can show good performance in all metrics, the hybrid model will be much more applicable to the real-world diagnostic systems.

VI CONCLUSION & FUTURE WORK

The suggested Hybrid Transformer-RNN model presents an extremely effective framework that can be used to classify mammograms and detect breast cancer through the world-view global spatial attention with sequential feature modelling within the same mathematically structured framework. After performing a wide range of experimental training on a large benchmark mammography dataset, the model demonstrated high performance with all defining metrics of the evaluation, including an accuracy of 94.5% and an F1-score of 0.945, surpassing that of conventional CNNs, RNNs, and standalone transformer models. The hybrid fusion approach allowed more precise lesion characterization through the complementary representation of features, whereas the error and calibration measures showed the strength, accuracy and readability of the predictions. More importantly, the model had a low inference latency of 26 ms per picture patch, which rendered it viable in real-time applications in cloud-based computer-aided diagnosis systems. All in all, this study shows that combining spatial and time-based learning is a powerful way to improve the diagnostic accuracy and computational efficiency of the hybrid model and make it one of the strongest competitors of the next-generation AI-based breast cancer screening devices.

Future studies might expand the proposed architecture to the multi-modal diagnostic pipelines with the inclusion of other types of information like ultrasound images, MRI, clinical records, and genetic risk factors. This kind of integration would allow a more extensive assessment of breast cancer risk. The other direction that is of interest is the application of federated learning

that allows training models that do not involve exchanging patient information with other hospitals and thus enhances the model generalizability. The continuous learning process to enhance adaptability of the models would enable the system to learn over time with new imaging features, device changes and new cancer patterns. More research is also required to add causal explainability models to give clinically useful reasons behind AI predictions that have the potential to enhance physician confidence and regulatory approval. Lastly, the methods of model compression and edge optimization may allow the implementation of the hybrid architecture in portable screening devices and its extensive use in the low-resource or remote healthcare environment.

References

[1]. T. Khater *et al*., "An Explainable Artificial Intelligence Model for the Classification of Breast Cancer," in *IEEE Access*, vol. 13, pp. 5618-5633, 2025, doi: 10.1109/ACCESS.2023.3308446.
[2]. H. M. Whitney, H. Li, Y. Ji, P. Liu and M. L. Giger, "Comparison of Breast MRI Tumor Classification Using Human-Engineered Radiomics, Transfer Learning From Deep Convolutional Neural Networks, and Fusion Methods," in *Proceedings of the IEEE*, vol. 108, no. 1, pp. 163-177, Jan. 2020, doi: 10.1109/JPROC.2019.2950187.
[3]. R. Lupat, R. Perera, S. Loi and J. Li, "Moanna: Multi-Omics Autoencoder-Based Neural Network Algorithm for Predicting Breast Cancer Subtypes," in *IEEE Access*, vol. 11, pp. 10912-10924, 2023, doi: 10.1109/ACCESS.2023.3240515.
[4]. R. Zeng *et al*., "FastLeakyResNet-CIR: A Novel Deep Learning Framework for Breast Cancer Detection and Classification," in *IEEE Access*, vol. 12, pp. 70825-70832, 2024, doi: 10.1109/ACCESS.2024.3401729.
[5]. Y. Shao *et al*., "Breast Cancer Detection Using Multimodal Time Series Features From Ultrasound Shear Wave Absolute Vibro-Elastography," in *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 2, pp. 704-714, Feb. 2022, doi: 10.1109/JBHI.2021.3103676.
[6]. L. Liu *et al*., "Collaborative Transfer Network for Multi-Classification of Breast Cancer Histopathological Images," in *IEEE Journal of Biomedical and Health Informatics*, vol. 28, no. 1, pp. 110-121, Jan. 2024, doi: 10.1109/JBHI.2023.3283042.
[7]. L. Liu, J. Cheng, Q. Quan, F.-X. Wu, Y.-P. Wang, and J. Wang, "A survey on U-shaped networks in medical image segmentations," *Neurocomputing*, vol. 409, pp. 244–258, 2020.
[8]. J. Yang, "Prediction of HER2-positive breast cancer recurrence and metastasis risk from histopathological images and clinical information via multimodal deep learning," *Comput. Struct. Biotechnol. J.*, vol. 20, pp. 333–342, 2022.
[9]. Y. Yari, T. V. Nguyen, and H. T. Nguyen, "Deep learning applied for histological diagnosis of breast cancer," *IEEE Access*, vol. 8, pp. 162432–162448, 2020.
[10]. S. Sharma and R. Mehra, "Conventional machine learning and deep learning approach for multi-classification of breast cancer histopathology images–A comparative insight," *J. Digit. Imag.*, vol. 33, no. 3, pp. 632–654, 2020.