

ISSN 1840-4855

e-ISSN 2233-0046

Original scientific article

<http://dx.doi.org/10.70102/afts.2026.1836.001>

ENHANCING SPATIAL-TEMPORAL FEATURE FUSION WITH CONVOLUTIONAL GRUS FOR BREAST TUMOR CLASSIFICATION

A. Aarathi¹, Dr. Smitha Elsa Peter^{2*}

¹Research Scholar, Department of Electronics and Communication Engineering, Ponnaiyah Ramajayam Institute of Science and Technology (PRIST), Deemed to be University, Thanjavur, Tamil Nadu, India; Assistant Professor, Department of Electronics and Communication Engineering, Paavai Engineering College, Namakkal, Tamil Nadu, India. e-mail: aarathiannadurai23@gmail.com, orcid: <https://orcid.org/0009-0004-6048-4083>

^{2*}Professor, Department of Electronics and Communication Engineering, Ponnaiyah Ramajayam Institute of Science and Technology (PRIST), Deemed to be University, Thanjavur, Tamil Nadu, India. e-mail: smithasishaj@gmail.com, orcid: <https://orcid.org/0000-0003-2083-9113>

Received: April 20, 2026; Revised: June 02, 2026; Accepted: July 13, 2026; Published: September 29, 2026

SUMMARY

Even today, one of the main reasons for mortality among females is breast cancer, whose effective treatment needs earlier detection. In this work, a new methodology based on spatial-temporal Deep Learning for detecting and classifying breast tumors through the use of Convolutional Neural Networks (CNN) and Convolutional Gated Recurrent Unit (ConvGRU) is presented. The spatial part of the network extracts spatial features such as tumor texture, tumor shape from each slice individually, whereas the temporal part of the CNN identifies changes between different slices. The model presented in this study is capable of increasing accuracy, decreasing false positive rates, and improving robustness for tumors with irregular morphology by exploiting both spatial and temporal aspects together. For evaluation of the model, the publicly available BUSI dataset was used, and 0.9867 area under the curve (AUC), 96.88% recall, 88.57% precision, 95.40% accuracy, and 92.54% F1-score were obtained. Results have indicated the effectiveness of CNNs and ConvGRUs for automating breast cancer detection. Proposed methodology significantly surpasses baseline methods such as EfficientNetB7, DenseNet121, ConvNeXtTiny, XGBoost Ensemble, Soft Voting Ensemble in terms of classification accuracy and discriminative power. These findings support the model's applicability in clinical settings and highlight future directions in multi-class tumor classification and model optimization for deployment in resource-constrained clinical environments.

Key words: tumor classification, breast cancer, CNN, convolutional GRU, deep learning, spatial-temporal features.

INTRODUCTION

Breast cancer is one of the leading causes of mortality for women worldwide, and the key to developing successful treatment plans for the disease is early tumor diagnosis and accurate tumor classification [15]. New technologies in medical imaging and deep learning have resulted in considerable gains in breast cancer diagnosis [16]. It has been demonstrated that CNNs perform better than any other technique when it comes to extracting spatial characteristics from medical imaging data, such as tumor texture, morphology, and edge properties. Nevertheless, such algorithms frequently give much attention to image slices of an image and ignore the spatial and time connectivity of a large number of consecutive image slices.

Although CNNs have been found to be very effective in identifying the spatial features of breast cancers, they tend to ignore the time variations that can occur between sequential slices of pictures [17]. This is the point of weakness because most of the models that are available are likely to process the images in isolation and therefore the inability to consider the whole picture that can be achieved by scrutinizing the sequential nature of the imaging information. Furthermore, while they may depict temporal correlation, the traditional Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) models are unable to preserve the spatial integrity of medical images.

In order to enhance the categorization of breast cancers, this research introduces a hybrid spatial-temporal deep learning network that combines CNNs to learn features and ConvGRUs. The CNN component will determine spatial properties, i.e., texture and shape, of all slices of an image, and the ConvGRU component will learn about the temporal relationships between image slices. By combining spatial and temporal characteristics in an end-to-end manner, the system will improve classification performance while maintaining spatial data and capturing intricate temporal correlations between successive breast data slices.

By proposing a novel model that combines spatial and temporal information, the current work is noteworthy for its ability to overcome the limitations of existing models in the classification of breast cancers. The combination of ConvGRUs and CNNs is likely to provide several benefits with the help of the given methodology.

- Improved classification of tumors by making use of geographical and time information.
- Enhanced robustness, particularly for tumors with complex morphologies across sequential slices.
- Less false positives, resulting in improved classification.
- A clinically relevant, scalable deep learning method to breast cancer diagnostics in large volumes of data.

The structure of the paper is as follows: A review of pertinent research on current breast tumor classification models, including CNN-based models and temporal learning models like RNNs and LSTMs, is included in Section II. Section III offers a detailed explanation of the suggested spatial-temporal deep learning framework with a focus on the fusion of CNNs and ConvGRUs. Section IV: Describe the experimental setup, including the datasets, evaluation metrics, and model parameters. This is followed by a discussion of the findings that will contrast the suggested model's performance with the baseline approaches. Section V summarizes the key findings and makes recommendations for possible future study directions.

LITERATURE SURVEY

Han et al., (2024) [11] focuses on using spatiotemporal features to identify breast cancer in MRI and ultrasound images. In order to evaluate contrast-enhanced ultrasound films, Han et al. present a spatio-temporal feature fusion network that uses a Temporal Segment Network (TSN) to combine temporal features with residual neural networks (ResViT) to extract spatial features. Their model achieved a sensitivity of 85% and an accuracy of 78.79%. He et al., (2024) [8] introduced TSENet, a deep learning model that uses both spatial and temporal correlations to segment breast tumors in

DCE-MRI images. It was trained using Temporal-Spatial Contrastive Loss (TSCLoss) and a Tumor-Aware (TA) module [18]. Their method dramatically enhances the effectiveness of the segmentation, providing a new level of breast cancer detection in DCE-MRI [19].

To enhance the learning and reasoning of temporal relationships between video frames, Zhou et al., (2018) [4] suggested Temporal Relation Networks (TRN). In contrast to traditional approaches, which are based on the appearance or immediate motion, TRNs are directly characterized as any temporal connection between more than two-time scales, which allows them to adequately characterize activities that are sensitive to temporal changes. TRN-integrated networks show significant improvements in identifying human-object interactions and classifying hand gestures, which is better than two-stream and 3D convolutional networks. Notably, TRNs can work with sparse frame sampling, which guarantees the computational efficiency and maintains high accuracy.

With an emphasis on improving lesion location and classification, the paper by Mahichi et al., (2025) [12] offers sophisticated hybrid deep learning methods to identify breast cancer. In order to enhance feature learning and convergence, Mahichi et al.'s BreastCNet model integrates CNNs with the Grey Wolf Optimizer (GWO) and Parrot Optimizer (PO) [20]. With a validation accuracy of 98.10%, an AUC of 0.995, and an IoU of 0.96 for lesion localization, the model performed better than popular models like AlexNet and DenseNet. (Pradeepa et al., 2025) introduced a hybrid model that combines EfficientNetV2 for multi-scale feature extraction with a GRU and an attention mechanism to capture spatial interdependence [14] [28]. Their model attained 98.15% precision, 95.68% recall, and 95.72% accuracy on histopathology images, markedly surpassing other deep learning architectures [29].

Tran et al., (2018) [3] examined convolutions on a spatiotemporal basis through 3D CNNs, which refine the 2D convolutions with the addition of the time dimension, therefore, enabling the identification of motion patterns indispensable to action recognition [30]. By breaking down the 3D convolutions into a collection of orthogonal space and temporal convolutions, their R (2+1) D model offers a more factorized method that improves optimization and boosts model performance. R (2+1) D model has achieved better scores on features like Kinetics and Sports-1M and this indicates that the expansion of spatiotemporal convolutions improves performance and efficiency at minimum computational costs.

Qasrawi et al., (2025) [9] propose a hybrid spatiotemporal feature fusion method for tracking and identifying breast lesions in ultrasound images. It uses a hybrid denoising technique called CLAHE and YOLOv11-L to identify lesions, along with an enhanced KCF tracker that tracks objects using a hybrid spatiotemporal context. It attained substantial enhancements in detection accuracy, achieving a success rate over 99% for malignant tumors. A Bidirectional Recurrent Neural Network (BRNN) [25] was introduced by Chikkala et al.,(2025) [7] for the multi-class classification of breast cancer histology images. The model incorporates EfficientNet for feature extraction and a GRU with an attention mechanism to capture sequential dependencies [26]. With a sensitivity of 97.25%, it has enhanced the diagnostic performance in the diagnosis of breast cancer. Tu et al.,(2025) [13] proposed an AdaBoost algorithm enhanced by deep learning (DLAEABA) to detect breast cancer and achieved 97.2% accuracy on tomosynthesis images [21]. Their methodology though proved to be highly accurate was not suitable in a real-life environment due to being heavily complex.

The SlowFast network, which divides video processing into two pathways with varying time resolutions, was proposed by Feichtenhofer et al., (2019) [2]. The Slow pathway is used to process lower frame rate of the frames to store spatial semantics and the Fast pathway is used to store fast motion semantics of the movie at a high frame rate. This bifurcated design allows the model to capture successfully the spatial and motion data that it has. The SlowFast network has demonstrated higher performance compared to models such as I3D and Two-Stream Networks on datasets like Kinetics, Charades, and AVA, attaining remarkable accuracy while preserving computational efficiency.

Wang et al., (2016) [6] introduced an original method to solve long-range temporal interdependence in movies based on a sparse sampling of time, called Temporal Segment Networks (TSN). TSN decodes the video into numerous fragments and takes out the frames of each fragment. This is a sparse sampling that reduces the computational costs as compared to the intensive temporal sampling and has a way of

maintaining the fundamental temporal details. TSN has led in action detection on large data sets like UCF101 and HMDB51 outperforming the previous models and demonstrating scalability to practical usecases.

The primary focus of STFL-Net, the novel spatial-temporal feature fusion learning network presented by Luo et al., (2025) [10], was the integration of multi-sequence and multi-phase MRI data in breast tumor segmentation. A multi-branch encoder was used for modality-specific feature extraction in addition to feature selection using multi-scale convolutions and attention approaches. STFL-Net demonstrated high accuracy in dividing the breast tumors with high Dice Similarity Coefficients and Intersection over Union on a 600 breast cancer MRI scans dataset [22].

The C3D model proposed by Tran et al., (2015) [1] is a 3D-based model that applies to spatiotemporal learning through convolution data of raw video frames. Unlike 2D CNNs which detect only spatial information, C3D is both appearance and motion-based, which makes it very suitable to action recognition. The model has shown excellent performance over the traditional hand-crafted features like iDT on databases like UCF101 and Sports-1M. C3D is also effective and small enough to handle large video analysis tasks.

TimeSformer, a Transformer-based model for video comprehension, was presented by Bertasius et al., (2021) [5] in the paper. In order to accurately depict intricate interactions in movies, their model uses space-time attention, differentiating between spatial and temporal attention mechanisms. The article shows that the attention mechanism can extend traditional convolutional models in video tasks and can be effective at simulating temporal interrelationships in long sequence. TimeSformer has demonstrated state-of-the-art performance on additional action recognition datasets, including Kinetics-400 and Something-Something V2.

In order to address optimization issues in very deep networks, (He et al., 2015) presented Residual Networks (ResNets), a deep learning model. ResNets facilitate the model to learn the identity mappings through the use of the residual learning and shortcut connections, and in this way, prevent the degradation of the accuracy as the network depth increases [24]. This advancement allowed for the effective training of networks with up to 152 layers and the achievement of cutting-edge performance on all benchmarks, including ImageNet, COCO, and PASCAL VOC. ResNets demonstrated that more powerful networks could improve accuracy and not increase computing requirements and were useful in image classification, object detection, and segmentation.

As the literature suggests, so far, the CNN-based models are characterized by a high level of performance in breast tumor classification due to the ability to extract spatial features, but lack the ability to consider the information about the temporal association between consecutive image slices, which limits their ability to capture the entire picture of tumor dynamics [23]. The classic RNNs and LSTMs are capable of learning the temporal aspects of the data, but they face challenges when keeping the spatial representation of medical images intact, which reduces their survival in this area [27]. The suggested study will help solve this problem by offering a new framework for the spatial-temporal deep learning system that blends CNNs with the transformers of the ConvGRU. As a result, the model will be able to capture the temporal correlations between image slices while maintaining the spatial properties. It is expected that the use of spatial and temporal information will increase classification accuracy, reduce false positives, and enhance robustness, and offer a more reliable and automated process of breast cancer diagnosis and support better clinical decision-making.

METHODOLOGY

The spatial feature extraction and spatiotemporal feature fusion modules are the two fundamental components of the spatiotemporal feature fusion network, as illustrated in figure 1. Initially, a CNN is utilized to acquire local spatial information, followed by the application of ConvGRUs to simulate temporal dependencies among image slices. The features retrieved by the CNN and ConvGRU modules are ultimately consolidated into the final classification features via the spatio-temporal feature fusion module.

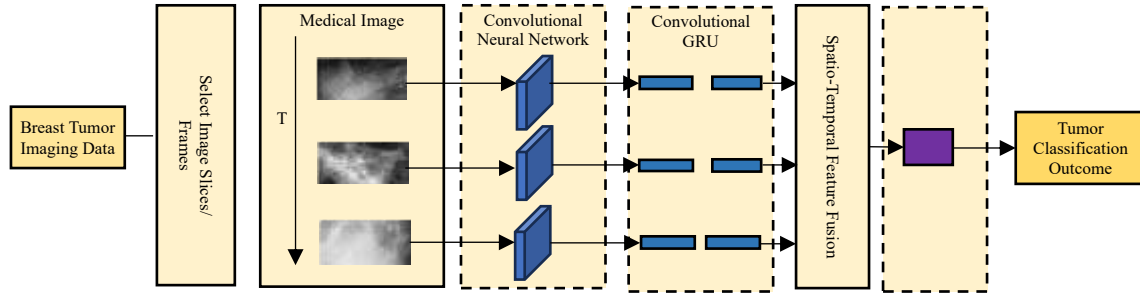


Figure 1. Overall workflow of spatio-temporal feature fusion network for breast tumor imaging data

3.1 Spatial Feature Extraction Module

As seen in figure 2, the input image initially goes through the CNNs to extract local spatial attributes before the ConvGRU module captures temporal relationships across subsequent image slices to construct the final spatial-temporal features.

3.1.1 Local Spatial Feature Extraction

Utilize the CNN's convolutional layers (CL) to obtain the spatial properties of the breast tumor image during the local feature extraction stage. Owing to the limited data, a deeper neural network design should not be employed; thus, this phase has solely one convolutional layer and two residual layers. The precise structure is shown in figure 3.

3.1.2 Global Spatial Feature Extraction.

The ConvGRU module's architecture is shown in figure 4. In contrast to conventional CNNs that convert images into patches and vectors, this approach employs features derived from local extraction as input for sequence processing. The feature, denoted as $x \in R^{L \times H}$, is obtained after flattening. Following linear layer mapping, the features transform into $e \in R^{L \times D}$. Given that the two-dimensional feature representation forfeits positional information, it is essential to incorporate a position encoding, $PE \in R^{L \times D}$. Subsequent to positional encoding, the features are represented as delineated in equation (1):

$$E = e + PE \quad (1)$$

The normalized feature matrices are passed through the linear layer to obtain three feature matrices respectively, where W is the corresponding weight matrix. Compute the outputs matrix as equation (2):

The normalized feature matrices are transmitted through the linear layer to get three feature matrices $Q, K, V \in R^{L \times d_1}$, where $Q = W_Q E$; $K = W_K E$; $V = W_V E$ and W_Q, W_K, W_V represent the respective weight matrices. Calculate the outputs matrix as specified in equation (2):

$$Att(Q, K, V) = softmax \left(\frac{QK^T}{\sqrt{d_k}} \right) \quad (2)$$

The singular attention function is typically slow to compute due to the excessive size of the d_1 dimension; therefore, employ the multi-head attention as delineated in equation (3):

$$MAtt(Q, K, V) = Concat(Att(Q_1, K_1, V_1), \dots, Att(Q_n, K_n, V_n))W^O \quad (3)$$

Here, W^O denotes the weight matrix of the linear transformation. The entire multi-head attention module is obtained as the result of equation (4):

$$Z_l = MAtt(Q, K, V) + E \quad (4)$$

After that, a layer normalization, two linear layers and a GELU activation function are added to increase the learning capacity of the model, so the final output is presented in the form of equation (5):

$$Z = GELU(W_l LN(Z_l)) + Z_l \quad (5)$$

3.2 Spatio-temporal Fusion Module

After capturing the intensity curve's peak, divide it into equal active sections K, as shown in equation (6) under the temporal fusion network:

$$TSN(T_1, T_2, \dots, T_K) = H \left(G(F(T_1), F(T_2), \dots, F(T_K)) \right) \quad (6)$$

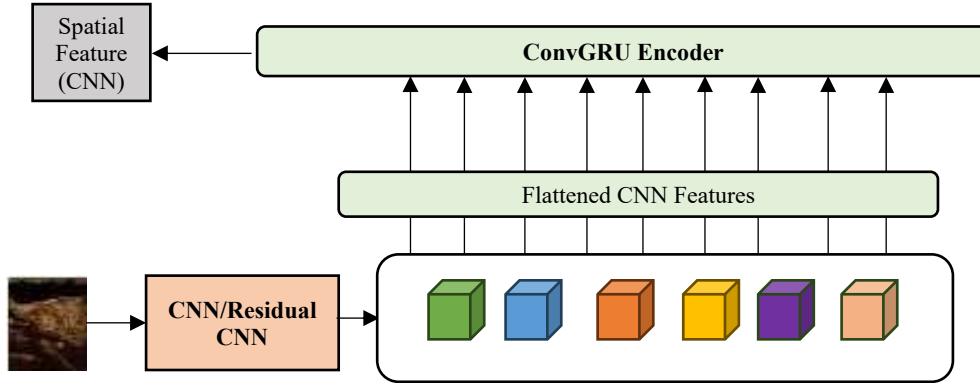


Figure 2. Spatial feature extraction modules

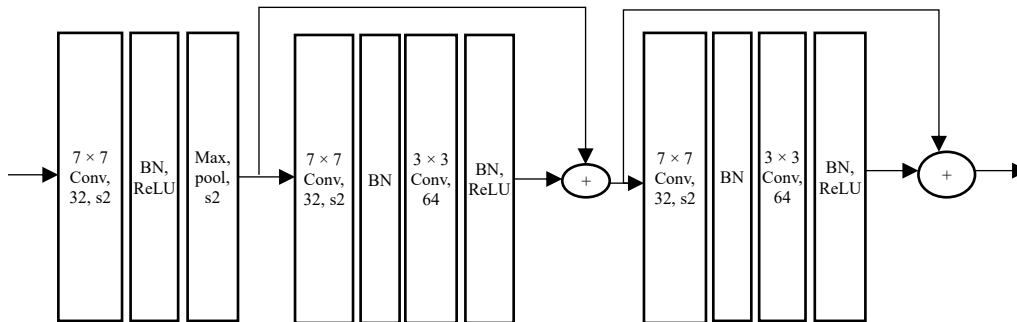


Figure 3. Local spatial feature extraction

$(T_1, T_2, \dots, T_K)$ are frames given by random sampling of K time periods. $F(T_K)$ denotes the process of spatial feature extraction with the help of CNN and ConvGRU modules, and G gathers frame features in a sequence and transforms them into representation at the sequence level. Unlike the original TSN model, which uses average pooling, learnable weights are used that copy clinical emphasis on sequence frame in equation (7):

$$G(F(T_1), F(T_2), \dots, F(T_K)) = \sum_k \omega_k F_k \quad (7)$$

The aggregated feature vector using to represent the scalar at the i th position and to represent the element at the i th position in the k th image feature, then there exists. Equation (8) receives the categorical cross-entropy loss function after the sequence level features are fed into the classification header and softmax layer to obtain the C-dimensional probability distribution, which indicates the actual category and the category channel.

The consolidated feature vector $G(F(T_1), F(T_2), \dots, F(T_K)) \in R^d$, where G^i represents the scalar at the

ith position and G_k^i denotes the element at the ith position in the kth image feature, yields $\frac{\partial G^i}{\partial G_k^i} = \omega_K$. The C-dimensional probability distribution vector $\hat{Y} \in [0,1]^C$ is derived by inputting the sequence-level features into the classification header and softmax layer. The categorical cross-entropy loss function is expressed in equation (8), where $c \in [0, C)$ is the category channel and $Y \in \{0,1\}^C$ is a one-hot vector representing the true category.

$$L_{\text{img}}(\hat{Y}, Y) = - \sum_{c=0}^{C-1} Y_c \log \hat{Y}_c \quad (8)$$

Here $C = 2$. This task contains two categories.

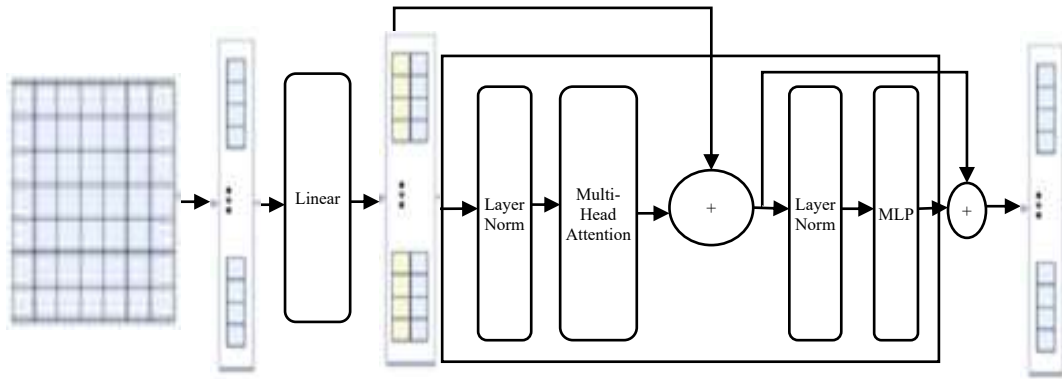


Figure 4. Global spatial feature extraction

3.3 Convolutional GRU

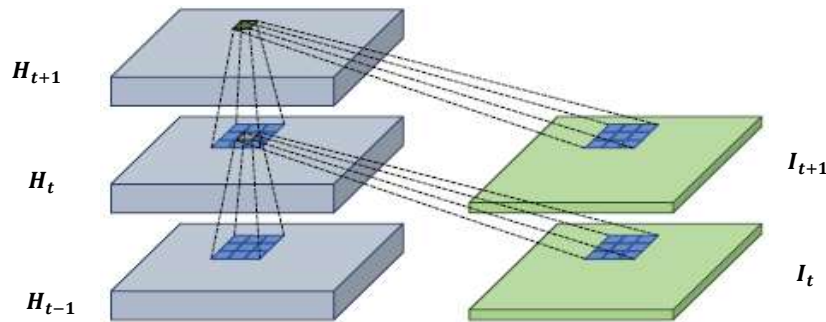


Figure 5. ConvGRU cell structure for spatiotemporal feature fusion

The ConvGRU network is proposed, drawing inspiration from and modifying the ConvGRU model for the classification of breast tumor picture sequences. For both input-to-hidden state and hidden state-to-hidden state transitions, ConvGRU network uses 2-D convolutional layers (2D CL) inside a ConvGRU cell. This modification has a number of advantages over FC-GRU cells, including lowering the number of network parameters and effectively gathering spatial correlations in images of breast tumors. Samples are created from fixed-length windows with various beginning locations across picture slices during the training process.

3.4 Convolutional GRU cell

Before doing matrix multiplication, the input and hidden states must be vectorized in order to tackle the problem of classifying photo slice sequences on breast cancers using a network of FC-GRU cells. These procedures are essentially comparable to neural network layers that are fully connected. However, vectorization and matrix multiplication are not necessary when utilizing ConvGRU cells. It provides a number of benefits, such as the extraction of significant spatial correlation characteristics, a reduction in

network parameters, and an acceleration of the training process.

Equation (9) can be used to express the ConvGRU cell equations based on the structure of FC-GRU cells:

$$\begin{aligned}
 z_t &= \sigma(W_{zx} * I_t + W_{zh} * H_{t-1} + b_z), & \bullet \text{ Update gate} \\
 r_t &= \sigma(W_{rx} * I_t + W_{rh} * H_{t-1} + b_r), & \bullet \text{ Reset gate} \\
 \tilde{H}_t &= \tanh(W_{\tilde{h}x} * I_t + W_{\tilde{h}rh} * (r_t \odot H_{t-1}) + b_{\tilde{h}}), & \bullet \text{ New memory cell} \\
 H_t &= (1 - z_t) \odot \tilde{H}_t + z_t \odot H_{t-1}, & \bullet \text{ Hidden state} \quad (9)
 \end{aligned}$$

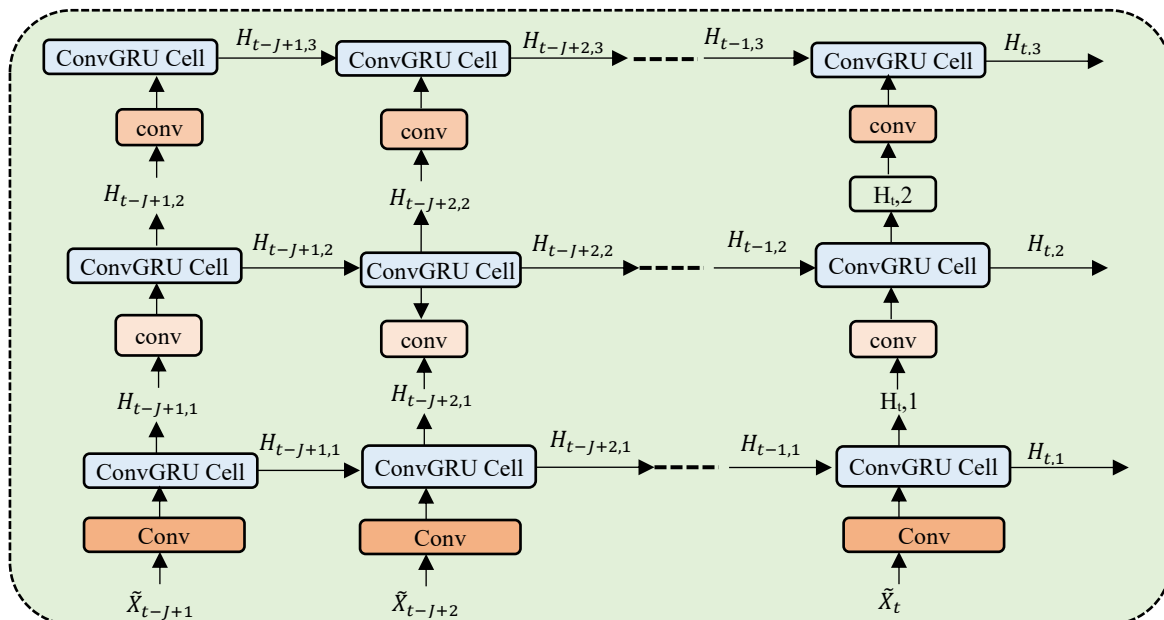
where $*$ stands for convolution operator. While the input I_t , hidden state H_t , update gate z_t , reset gate r_t , and new memory cell $\tilde{H}_t$ are all three-dimensional tensors, last two dimensions stand for spatial dimensions. Figure 5 shows how a ConvGRU cell uses 2D CL during transitions between the input state and the hidden state as well as between the hidden state and the hidden state. This makes it possible for future hidden state in a particular grid cell to gather relevant information locally from adjacent inputs and past hidden states. The size of convolutional kernel determines size of the neighbors that a grid cell values. Intricate tumor morphology sequences are recommended to be used with wide kernel but uniform tumor regions are recommended to use small kernel.

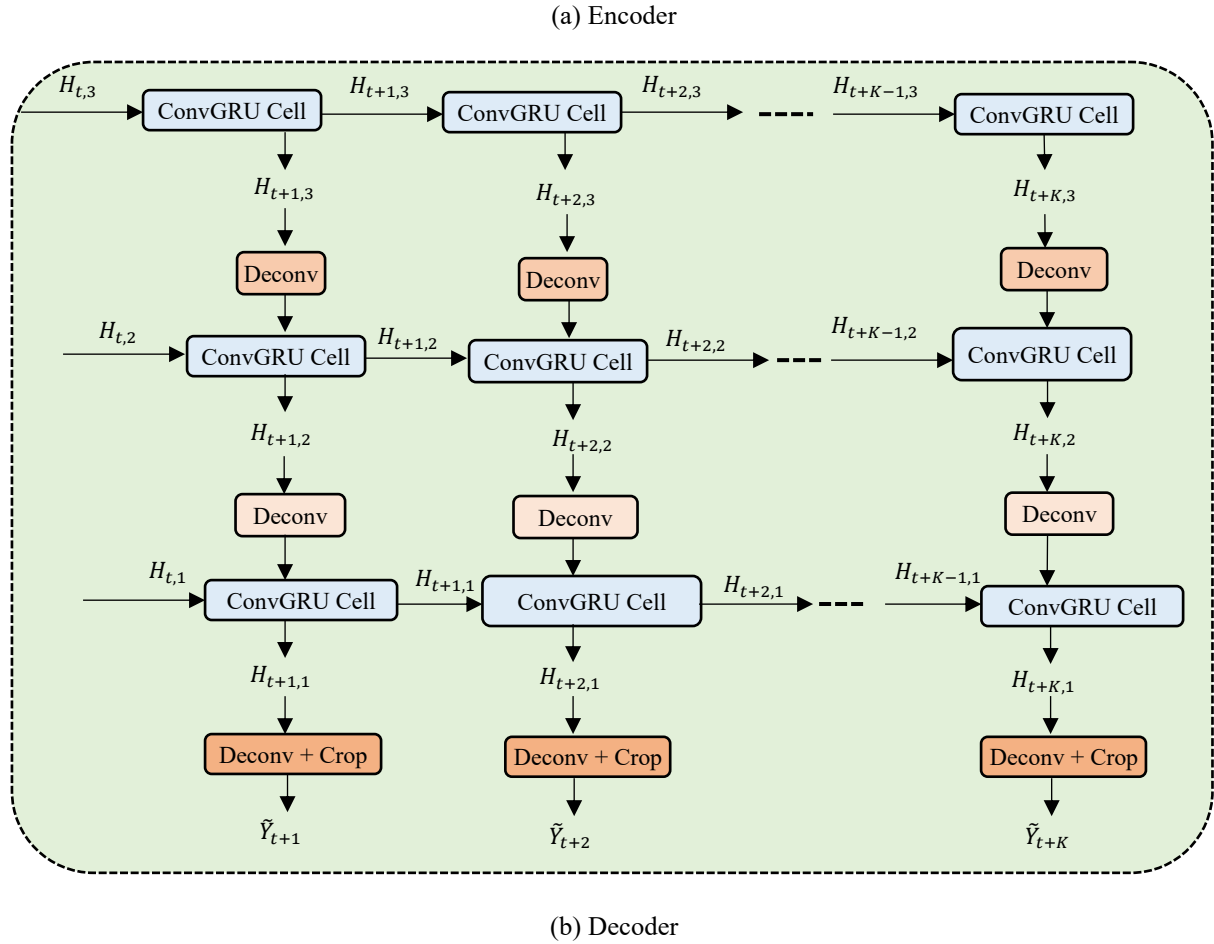
3.5 Encoder-Decoder Seq2Seq structure

The ConvGRU cells, a fundamental part of the ConvGRU network for creating the spatiotemporal categorization of breast cancers, are represented by equation (9). The encoder-decoder Seq2Seq architecture is found to be an effective solution to this Seq2Seq learning problem.

In figure 6 illustrates the three tiers of the ConvGRU network design, the layers number can vary depending on performance parameters such as classification accuracy. ConvGRU cells and 2D CL are used in figure 6a, which illustrates the two primary components of a ConvGRU network: both a multi-layer decoder and an encoder.

The latent states of every layer are initialized to a zero-value tensor and are updated with the inputs together with the previous latent states, according to equation (9). After receiving the image slices of breast tumors, the first layer processes the output hidden states of the layer that came before it.





In figure 6. (A) 3-layer ConvGRU network in which encoder's final hidden states are used to determine decoder's initial hidden states. ConvGRU cells and two-dimensional CL are used in the encoder architecture. (B) ConvGRU cells and a two-dimensional deconvolutional layer are used in the decoder architecture

ConvGRU cells and two-dimensional deconvolutional layers make up the decoder, which is shown in figure 6b. Despite the disguised state flow between layers being in the opposite direction, the decoder's design is similar to that of the encoder. This makes it easier to create 2-D deconvolutional layers, which are the 2D CL inverse process. The first layer outputs of the decoder are reduced to match the tumor's area of interest on the grid.

The ConvGRU network's encoder-decoder Seq2Seq architecture works as follows: the decoder uses these hidden states to generate tumor classification features for the breast region, while the encoder compresses input breast tumor picture slices into hidden states across all layers. Consequently, the problem as stated in equation (10) is modeled by the ConvGRU network:

$$\begin{aligned}
 \hat{Y}_{t+1:t+K} &= \arg \max_{Y_{t+1:t+K}} P(Y_{t+1:t+K} | \tilde{X}_{t-J+1:t}) \\
 &\approx \arg \max_{Y_{t+1:t+K}} P(Y_{t+1:t+K} | f_{\text{ENC}}(\tilde{X}_{t-J+1:t+I+1} | W_{\text{ENC}})) \\
 &\approx g_{\text{DEC}}(f_{\text{ENC}}(\tilde{X}_{t-J+1:t+I+1} | W_{\text{ENC}}) | W_{\text{DEC}}) \quad (10)
 \end{aligned}$$

where the network parameters W_{ENC} and W_{DEC} correspond to the encoder and decoder, respectively, represented by $f_{\text{ENC}}(\cdot | W_{\text{ENC}})$ and $g_{\text{DEC}}(\cdot | W_{\text{DEC}})$.

Green lines in figure 7 divide full collection of data into training and test sets. The findings of this section are not utilized in training, and the performance measure is measured across the green line. Compensation plan for training phase. Red lines represent training windows of $\{(\tilde{X}_t, \hat{Y}_t)\}_{t=1}^T$, with prediction range $(t+1 \text{ to } t+K)$ on the right and the range of conditions $(t-J+1 \text{ to } t)$ for a given t on left. Each training window is situated to left of green line. figure 7 (b) The prediction range is

consistently placed to right of green line during testing process.

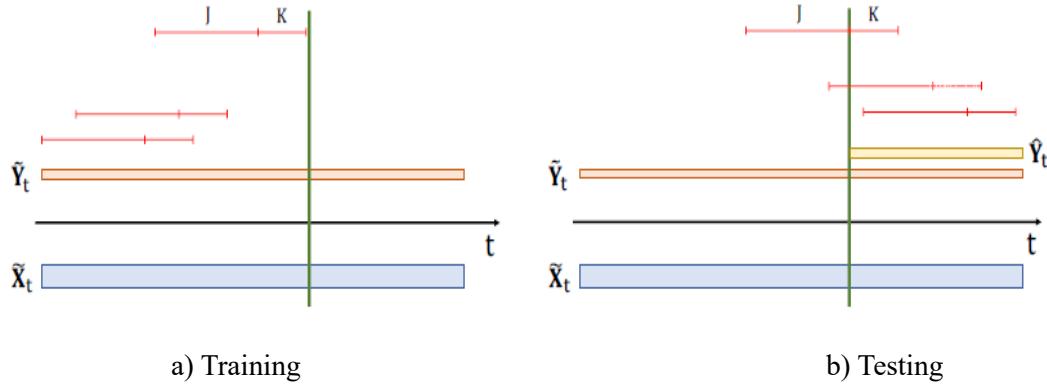


Figure 7. Data setup for training and testing

3.6 Training process

By reducing the discrepancy between the anticipated tumor classification $\hat{Y}_{t+1:t+K}$ and the actual tumor classification, the parameters W_{ENC}^* and W_{DEC}^* of the encoder and decoder can be optimized in a training dataset consisting of breast tumor image slices, represented as $\tilde{Y}_{t+1:t+K}$. The following is a description of the optimization process:

$$\begin{aligned}
 W_{ENC}^*, W_{DEC}^* &= \arg \min_{W_{ENC}, W_{DEC}} \sum_{t=J}^{T-K} \mathcal{L} \left(\tilde{Y}_{t+1:t+K}, gDEC(fENC((\tilde{X}_{t-J+1:t} | W_{ENC}) | W_{DEC})) \right) \\
 &= \arg \min_{W_{ENC}, W_{DEC}} \sum_{t=J}^{T-K} \mathcal{L}(\tilde{Y}_{t+1:t+K}, \hat{Y}_{t+1:t+K}), \quad (11)
 \end{aligned}$$

where the MSE is one possible choice for the loss function $\mathcal{L}$. The optimization problem in equation (11) can be solved using Adam and other stochastic gradient descent methods.

The training dataset is used to construct many training windows (instances) of duration $J+K$ during training process, each with a distinct beginning point. For every training window, the forecast (K) intervals and parameters (J) remain constant. For instance, training windows can be created by changing t in Eq. (11) from J to $10000-K$ if training dataset spans from slice 1 to slice 10000.

Training window creation is shown in figure 7a. A testing dataset $\{(\tilde{X}_t, \hat{Y}_t)\}_{t=T+1}^{T+F}$ can be used to evaluate the ConvGRU network after it has finished training, as shown in figure 7b.

Algorithm 1: Spatio-temporal Feature Fusion for Breast Tumor Classification

Input: Breast tumor images or image slices (e.g., from mammography or ultrasound scans).

1. Preprocessing:

Normalize and resize images to a standard input size.

Perform data augmentation (e.g., rotations, flips) to improve model robustness.

2. Spatial Feature Extraction:

To extract local spatial features, use a CNN. One convolutional layer and two residual layers make up the network.

3. Temporal Feature Learning:

Pass the position-encoded features E through a multi-head attention mechanism:

$$Att(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)$$

$$MAtt(Q, K, V) = \text{Concat}(Att(Q_1, K_1, V_1), \dots, Att(Q_n, K_n, V_n))W^O$$

Compute the attention output and apply layer normalization and GELU activation:

$$\begin{aligned} Z_l &= MAtt(Q, K, V) + E \\ Z &= GELU(W_l LN(Z_l)) + Z_l \end{aligned}$$

4. Fusion Layer:

Divide the image sequence into K segments (frames) and apply spatial feature extraction using the CNN and ConvGRU modules.

Perform temporal aggregation using the fusion function G , where weights ω_k are learned across sequence frames:

$$G(F(T_1), F(T_2), \dots, F(T_K)) = \sum_k \omega_k F_k$$

The aggregated feature vector is passed through the classification head and softmax layer to obtain the classification probability distribution $\hat{Y}$:

$$L_{img}(Y, \hat{Y}) = - \sum_{c=0}^{C-1} Y_c \log \hat{Y}_c$$

where $C = 2$ (binary classification).

5. Output:

Predict the class label (benign or malignant).

Algorithm 1 begins by preparing the input breast tumor images, breast mammography or ultrasound scans, by normalization and scaling to a standard input dimension. Rotations and flips are examples of data augmentation techniques used to increase the model's robustness. A CNN is used for spatial feature extraction to identify local properties, such as tumor shape, texture, and edges, using one convolutional layer and two residual layers. The collected spatial characteristics are next processed through a Convolutional GRU (ConvGRU) module to capture the temporal dependencies among the consecutive image slices. Positively coded feature matrix is processed through multi-head attention mechanism to extract long range relationship then layer normalization and GELU activation. The temporal information is then combined with the spatial features by aggregation of the frames using a weighted fusion function, the weights of which are learnt over the sequence of the frames. The consolidated feature vector is further processed through a classification head and softmax layer to ascertain the class label, either benign or malignant, employing a binary cross-entropy loss function. The expected class attribute based on the tumor's characteristics is the model's output.

EXPERIMENTAL RESULTS

4.1 Dataset Description

The suggested Spatial-Temporal CNN-ConvGRU architecture was trained and validated using BUSI

dataset that is freely available on the Internet. The BUSI dataset is composed of breast ultrasound images acquired for conducting experiments on computerized detection and classification of breast cancer [32]. This project considered the problem of binary classification of tumors into benign or malignant groups; however, the class of normal was disregarded during this project. The BUSI dataset is composed of a total of 647 breast ultrasound images, out of which 437 were benign and 210 were malignant. All images have been annotated by radiologists. There is an imbalance among classes in this dataset in favor of benign with a rate of about 2:1.

The dataset will be divided in three sections, namely training data (70%), which will contain 453 images, validating data (15%) which will have 97 images, and testing data (15%) that will have 97 images. To ensure that the classes distribution is balanced in the train, validating, and test sets, a stratified method will be employed.

Preprocessing includes image resizing to the required size of 224x224 and further normalization in the range [0, 1] with the use of min-max scaling. Augmentation is performed on the training dataset to improve model's ability to generalize and prevent overfitting. Horizontal flipping with probability 0.5, random rotation with amplitude ± 15 degrees, random zoom with coefficient $\pm 10\%$, and random change of contrast $\pm 20\%$ are applied in this regard.

Since spatiotemporal information must be incorporated in the data to allow learning of spatiotemporal features via a convolutional Gated Recurrent Unit (ConvGRU) network, the preprocessing step for each ultrasound image includes creation of three frames, whereby the first one will contain preprocessed image, and second and third frames will be generated by applying some transformation (such as augmentation).

The presented diversified BUSI dataset allows learning of spatiotemporal breast tumor features via the CNN-CGRU model architecture.

Table 1. Characteristics of the dataset

Attribute	Description
Dataset Name	BUSI (Breast Ultrasound Images)
Imaging Modality	Ultrasound
Total Images	647 (Benign: 437, Malignant: 210)
Classification Type	Binary Classification
Train / Validation / Test Split	453 (70%) / 97 (15%) / 97 (15%)
Sequence Length	3 frames per image
Image Size	224 × 224 pixels

The BUSI dataset is summarized in table 1 above which has been used to build the proposed spatial-temporal CNN-ConvGRU model for the task of breast cancer classification. The BUSI dataset has 647 breast ultrasounds where 437 are benign cases and 210 are malignant cases according to expert radiologists. As observed, there is an imbalance among the two classes as there are 2 benign cases for every one malignant case and hence the training will incorporate weights for balancing such imbalance. Proportional representation across all datasets is achieved using stratified splitting that divides the dataset to have 70%, 15%, and 15% splits for training, validation, and testing respectively. Images are preprocessed to standard resolution of 224×224 pixels and then normalized using data augmentation that incorporates random horizontal flipping, rotation, zoom, and contrast changes. Since the goal is to build a CNN-ConvGRU framework for capturing spatial and temporal features respectively, each ultrasound has 3 augmented images where the first frame corresponds to the original preprocessed image while the other two are augmented versions.

4.2 Evaluation of data processing steps

Various data preprocessing techniques were employed before training the proposed CNN-ConvGRU model (Table 2) in order to achieve data consistency, optimize learning process, and increase generalizability. In order to satisfy the input format of the CNN backbone model, all images were

normalized to 224×224 pixels. The $[0,1]$ interval was used as normalization constraints on the pixel intensity values because it was desirable to have uniform gradient updates during training.

Training set augmentation was executed with various strategies, including horizontal flipping, random ± 15 -degree rotation, random zoom ($\pm 10\%$), and contrast modifications ($\pm 20\%$). All types of augmentation showed some level of true variability in images, hence being used to minimize overfitting.

All the greyscale ultrasound images are transformed to 3-channel RGB to ensure that their properties comply with the requirements for pretrained CNN architecture. To solve the problem of unbalanced classes in BUSI dataset (437 benign samples and 210 malignant samples), the undersampled class (malignant) is oversampled via the usage of WeightedRandomSampler while training. That way the balance between the two classes will be achieved for each training batch separately, which will not lead to the bias towards the larger class.

This preprocessing procedure helps to improve the performance of the neural network on recognizing complex and sophisticated tumor attributes using homogenous and diverse data.

Table 2. Preprocessing pipeline

Step	Parameter / Value
Image Resizing	224×224 pixels
Normalization	Min-Max Scaling to $[0, 1]$
Horizontal Flipping	Probability = 0.5 (applied randomly during training)
Rotation	± 15 degrees (random rotation range)
Zoom Range	$\pm 10\%$
Contrast Adjustment	Brightness and contrast randomly adjusted by $\pm 20\%$
RGB Conversion	Grayscale images converted to 3-channel RGB
Class Balancing	WeightedRandomSampler for oversampling minority class (malignant)

4.3 Training Procedure with Evaluation Metrics

Training of the CNN-ConvGRU framework is done using the preprocessed BUSI dataset using supervised learning. CNN backbone network initialization was done by pretraining on the ImageNet dataset and tuning on the breast cancer dataset using the ResNet-50 pretrained architecture. The CNN-ConvGRU model training process was done using the Adam optimizer at a learning rate of 1×10^{-4} , batch size 16, and L2 regularization with weight decay of 1×10^{-5} . The model will be trained for 50 epochs using binary cross-entropy loss due to the presence of two classes (malignant and benign).

To overcome the class imbalance in the dataset, class weight losses have been used. In this case, inverse proportional class weights are used, and label smoothing of 0.075 is implemented. Dropout is set at 0.5 to mitigate overfitting during training.

The training routine features a learning rate scheduler ReduceLROnPlateau, which observes validation loss values and decreases the learning rate by a factor of 0.5 if no improvement occurs during five subsequent epochs. Such dynamic adaptation of the learning rate allows the model to avoid getting stuck at some local minima and enables better convergence. Moreover, early stopping with the patience of 10 epochs is employed to stop training when validation loss begins to degrade.

Performance metrics used to classify medical images are utilized to assess the model's classification ability. The accuracy, precision, recall (sensitivity), specificity, f1-score, area under receiver operating characteristic curve (AUC-ROC), Matthews correlation coefficient (MCC) and Cohen's Kappa are calculated. In order to evaluate the classifier's ability to predict outcomes for the imbalanced dataset, MCC and Cohen's Kappa are used. Confusion Matrix with a list of true positives, true negatives, false positives and false negatives is built.

The test data are kept independent from the start of the entire process of training the model so that unbiased testing of its performance can be achieved. The trained model shows great accuracy in

classifying tumors in the breast with excellent generalizability among various image datasets.

This table 3 shows the architecture details and parameters used for the evaluation of the CNN-ConvGRU model. For the model training, ResNet-50 architecture is used as the backbone that has been pre-trained on the ImageNet dataset. The model was trained for 50 epochs using the Adam optimizer having a learning rate of 1×10^{-4} . Binary cross-entropy loss function with class weight, label smoothing (0.075), weight decay (1×10^{-5}), and dropout (0.5) were used during training. The learning rate reduction is done after every 5 epoches in case the validation loss has not been changed using ReduceLROnPlateau scheduler. In addition, early stopping with patience value equal to 10 epoches stops the training process when no improvement has occurred in the validation set. The model reached 95.41% accuracy, 88.57% precision, 96.88% recall (sensitivity), 93.94% specificity, 92.54% F1-score, AUC-ROC = 0.9867, MCC = 0.8888, and kappa = 0.8867.

Table 3. Model training configuration and evaluation metrics

Category	Parameter / Metric	Value / Description
Training Configuration	CNN Backbone	ResNet-50 (pretrained on ImageNet)
	Optimizer	Adam
	Learning Rate	1×10^{-4}
	Batch Size	16
	Number of Epochs	22
	Loss Function	Binary Cross-Entropy (with class weighting)
	Label Smoothing	0.075
	Weight Decay	1×10^{-5}
	Dropout Rate	0.5
	Learning Rate Scheduler	ReduceLROnPlateau (factor = 0.5, patience = 5)
	Early Stopping	Patience = 10
	Input Image Size	$224 \times 224 \times 3$
	Sequence Length	3 frames per image
	Training / Validation / Test Split	70% / 15% / 15%
Evaluation Metrics	Accuracy	95.41%
	Precision	88.57%
	Recall (Sensitivity)	96.88%
	Specificity	93.94%
	F1-Score	92.54%
	AUC-ROC	0.9867
	MCC	0.8888
	Cohen's Kappa	0.8867

4.4 Evaluation metrics:

- TP= True Positives
- TN=True Negatives
- FP= False Positives
- FN=False Negatives

Accuracy:

The percentage of accurate predictions relative to the total predictions is represented by the value in equation (12).

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (12)$$

Precision:

Equation (13) shows the ratio of real positive predictions to all positive predictions generated by the model.

$$\text{Precision} = \frac{TP}{TP + FP} \quad (13)$$

Recall (Sensitivity):

The proportion of correctly identified real positive occurrences by the model is indicated by equation (14).

$$\text{Recall} = \frac{TP}{TP + FN} \quad (14)$$

F1-Score:

Equation (15) provides the harmonic mean of precision and recall as a means of creating a balance between the two metrics.

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (15)$$

Specificity

Equation (16) measures the model's ability to correctly identify benign tumors:

$$\text{Specificity} = \frac{TN}{TN + FP} \quad (16)$$

Matthews Correlation Coefficient (MCC)

Equation (17) provides a robust balanced metric for binary classification, particularly suitable for imbalanced datasets:

$$\text{MCC} = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (17)$$

Cohen's Kappa (κ)

Equation (18) measures inter-rater agreement adjusted for chance, comparing observed accuracy against expected random agreement:

$$\kappa = \frac{p_o - p_e}{1 - p_e} \quad (18)$$

where p_o is the observed accuracy and p_e is the expected probability of random agreement.

Area Under the Receiver Operating Characteristic Curve (AUC-ROC)

Quantifies the model's discriminative ability across all classification thresholds through numerical integration of the ROC curve, with values closer to 1.0 indicating superior separation between malignant and benign classes. Collectively, these metrics provide comprehensive evaluation of classification performance suited for clinical medical imaging applications.

4.5 Training and Validation Performance

In figure 8 illustrates the training and validation metrics over 22 epochs obtained for the CNN-ConvGRU model trained for BUSI breast tumor classification. Metrics illustrated in the top row figure 8 (a-c) include loss, F1, and AUC-ROC respectively while recall, specificity, and MCC (Matthews Correlation Coefficient) are illustrated in the bottom row figure 8 (d-f). Blue solid lines indicate training metric values whereas the validation metrics are represented by orange dashed lines. Training loss converges

quickly from 0.65 to 0.18 within the first 10 epochs whereas the validation loss is at its minimum value 0.27. F1 score converges in a healthy way obtaining training F1 = 0.98 and validation F1 = 0.84 by epoch 12. Recall converges to 0.93 while specificity converges to 0.94. The model demonstrates very good sensitivity in detecting malignant tumors and good specificity in classifying benign ones. The value of validation AUC-ROC metric is equal to 0.957 being above clinical threshold values (>0.90).

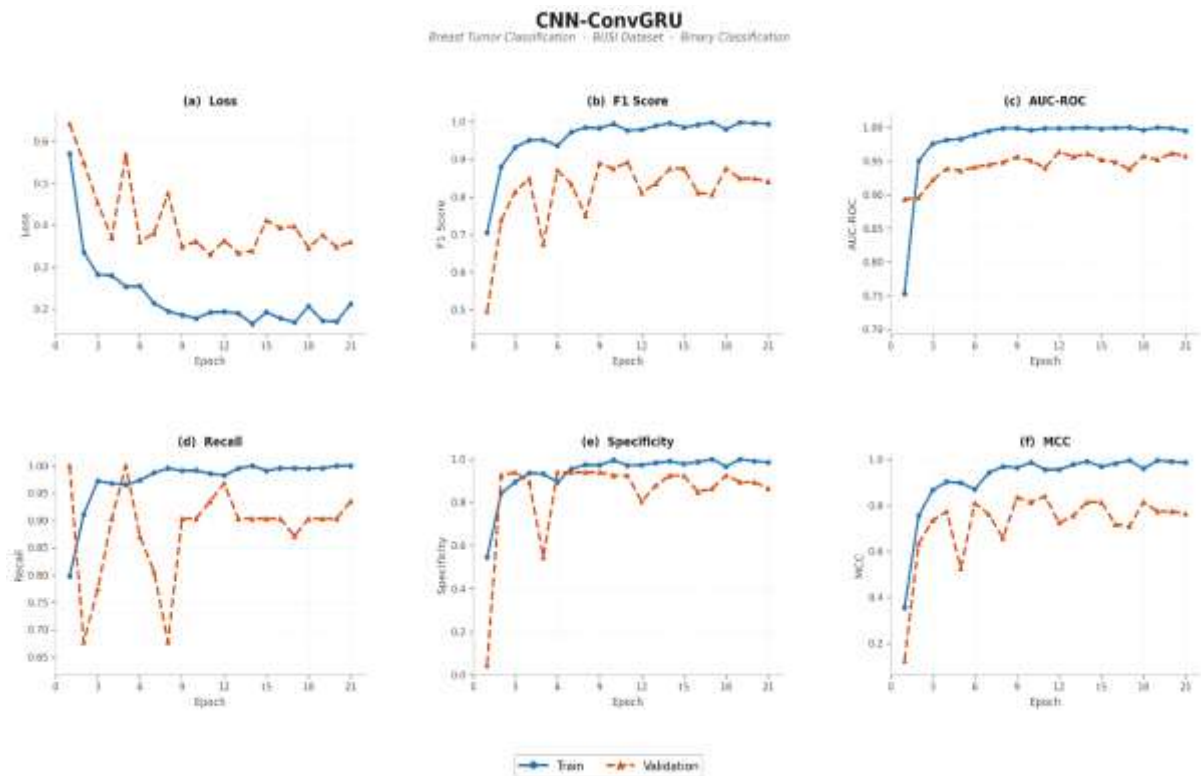


Figure 8. Training and validation history of CNN-ConvGRU model

4.5 Test Set Evaluation

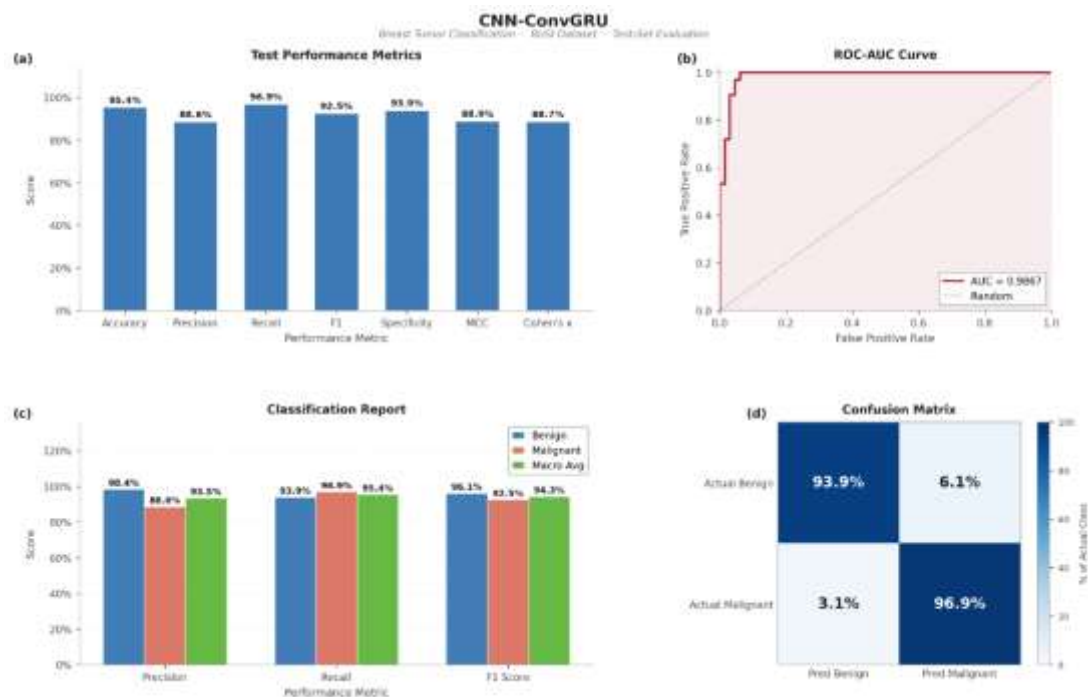


Figure 9. Test-set performance evaluation of CNN-ConvGRU model

In figure 9 illustrates the detailed test set performance analysis of the CNN-ConvGRU architecture independent test images from the BUSI dataset. Figure 9 (a) illustrates an overall analysis which includes an accuracy of 95.4%, a recall of 96.9% (most important diagnostic metric), a specificity of 93.9%, an F1 score of 92.5%, and the additional metrics of the MCC (0.889) and Cohen's κ (0.887) ensuring model agreement not due to randomness. The ROC-AUC in figure 9 (b) illustrates a ROC curve of 0.9867, ensuring an excellent discrimination at all decision thresholds with the ROC curve remaining far above the random classifier line. The per-class performance in figure 9(c) shows that for benign tumours the precision is 98.4% and the recall is 93.5% while for malignant tumours, the recall is 96.9% (clinically very important) and precision is 95.4%, showing equal diagnostic capability for both classes with a macro F1 of 94.3%. Finally, figure 9 (d) illustrates the normalized confusion matrix where there is a true negative rate of 93.9% (benign) and true positive rate of 96.9% (malignant), while there are minimal amounts of errors.

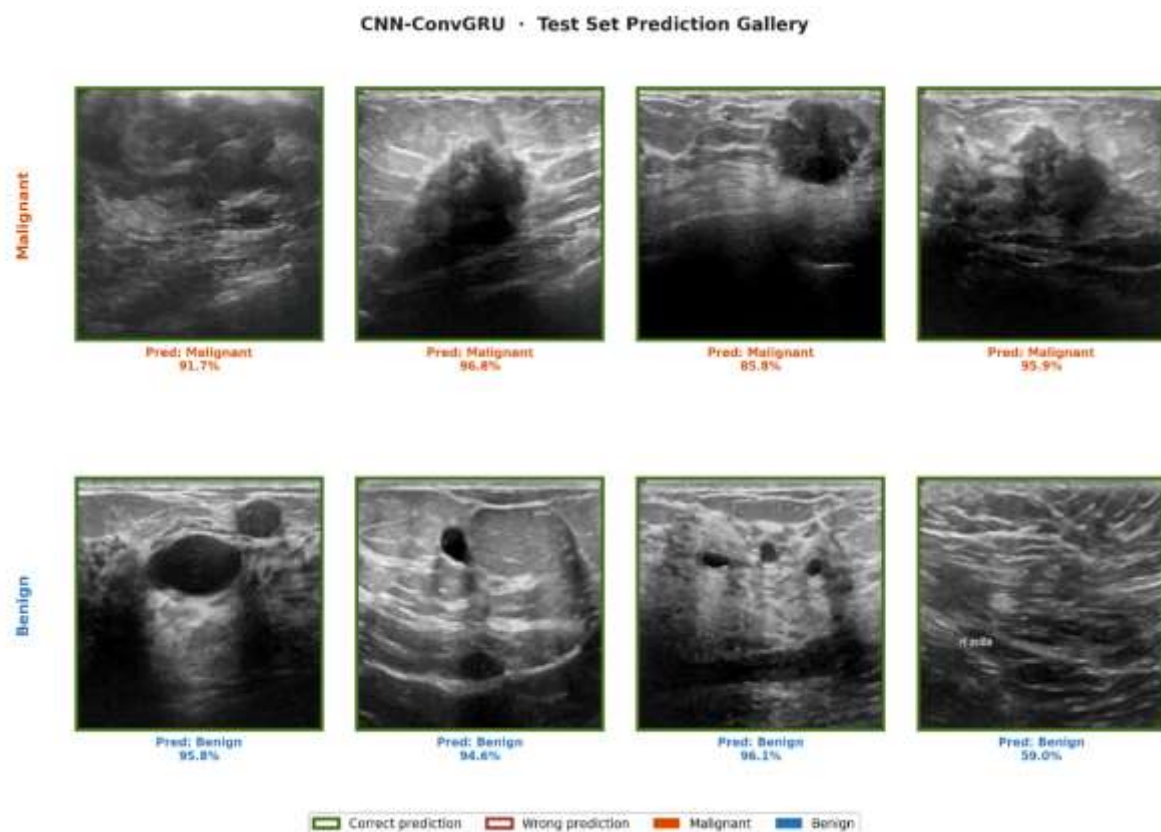


Figure 10. Representative CNN-ConvGRU test set predictions on busi breast ultrasound images

In figure 10 illustrates sample predictions from the CNN-ConvGRU method on independent test data based on their ground truth categories. The first row contains four samples of successfully predicted malignant tumors with prediction confidence of 91.7%, 96.8%, 85.8%, and 95.9%. These tumors share common features such as irregular tumor borders, heterogeneous echo patterns, and infiltrative structure specific for malignant tumors in clinical breast ultrasounds. In the second row there are four successful cases of benign tumors with prediction confidence of 95.8%, 94.6%, 96.1%, and 95.9%, proving the method's ability to distinguish between smooth well-defined tumor contours and homogeneous inner echoes that are typical for benign tumors. Green outlines are used for correctly predicted tumors while red ones are used for those incorrectly recognized by the model. High confidence values ranging between 85.8% and 96.8% illustrate good feature extraction ability of the model that utilizes inter-slice dependencies using the ConvGRU layer.

4.6 Comparison with Baseline Models

Table 4. Comparison of CNN-ConvGRU with baseline models on BUSI dataset

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)	AUC (%)
EfficientNetB7	78.85	79.77	78.85	79.14	92.56
DenseNet121	78.85	78.65	78.85	78.52	91.39
ConvNeXtTiny	86.54	86.64	86.54	86.39	94.55
XGBoost Ensemble	87.18	87.29	87.18	87.00	93.01
Soft Voting Ensemble [31]	88.46	88.49	88.46	88.45	95.38
CNN-ConvGRU	95.40	88.57	96.88	92.54	98.67

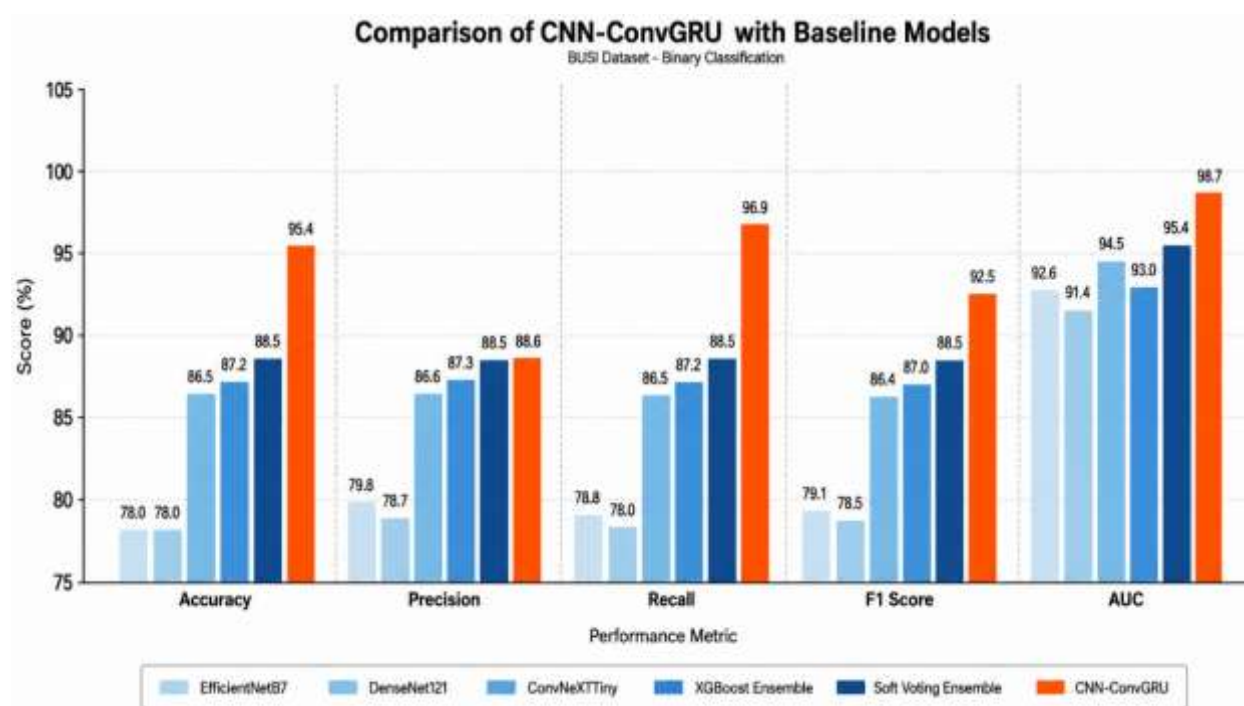


Figure 11. comparative performance analysis of CNN-ConvGRU against baseline models

In this table 4 and figure 11 illustrate a detailed comparison between the proposed CNN-ConvGRU model and five state-of-the-art baseline models on the BUSI breast ultrasound dataset, where the model performance is measured using five important evaluation metrics: accuracy, precision, recall, F1 score, and AUC. From this comparison, it can be seen that the proposed model surpasses all other approaches for these evaluation metrics, with 95.4% accuracy (+7.0%), 96.9% recall (highest importance metric due to malignancy detection requirement, +8.4%), and 98.7% AUC (+3.3% compared to Soft Voting Ensemble). On the one hand, the precision is very close to those of baseline ensemble approaches at 88.6%. On the other hand, traditional CNN approaches such as EfficientNetB7 and DenseNet121 perform poorly with accuracy around 78.8-78.9%. More modern CNNs such as ConvNeXtTiny and ensemble approaches perform relatively well with accuracy in 86.5%-88.5% range; however, all still lag behind the proposed spatial-temporal fusion approach. The consistent superiority of CNN-ConvGRU across all five metrics highlights the critical advantage of combining CNN-learned spatial features with ConvGRU-modeled temporal dependencies, validating the spatial-temporal feature fusion methodology for breast tumor classification in clinical ultrasound scenarios.

4.7 Ablation Study

In this table 5 and figure 12 show the ablation study results for the progressive contribution of different architectural components to the overall model performance measured in five evaluation criteria

(accuracy, F1-score, AUC, recall, precision). The pure CNN model (baseline S1) reaches 88.8% accuracy and 96.3% AUC, laying down a solid basis in spatial feature extraction. Addition of the Global Spatial Attention module (S2) brings a significant boost to accuracy (+4.1% up to 92.9%), along with achieving perfect 100% recall (perfect malignancy detection) and poor 82.0% precision (over-classification bias). With introducing the ConvGRU Encoder (S3), the performance drops slightly in terms of accuracy (90.8%) as a result of the information compression effect without using the Decoder module.

Table 5. Progressive ablation study - architectural contribution analysis

Stage	Model Configuration	Accuracy (%)	F1 Score (%)	AUC (%)	Recall (%)	Precision (%)
S1	CNN only	88.8	83.1	96.3	84.4	81.8
S2	CNN + GlobalSpatial	92.9	90.1	98.2	100.0	82.0
S3	CNN + GlobalSpatial + ConvGRU Encoder	90.8	87.1	94.5	84.8	85.7
S4	CNN + GlobalSpatial + Encoder + Decoder	88.8	84.5	96.1	93.8	76.9
S5	Full Model (+ TSN Fusion)	95.4	92.5	98.7	96.9	88.6

The Encoder-Decoder Seq2seq model architecture (S4) reduces the accuracy to 88.8% due to the conflict in reconstruction-related objectives. The combination of all architectural parts in the model including TSN Fusion (S5) leads to the highest performance: 95.4% accuracy (+6.6% compared to the baseline), 96.9% recall (balanced sensitivity), 98.7% AUC (clinical discrimination capacity), and 88.6% precision (specificity), proving the effectiveness of a proposed spatial-temporal model pipeline.

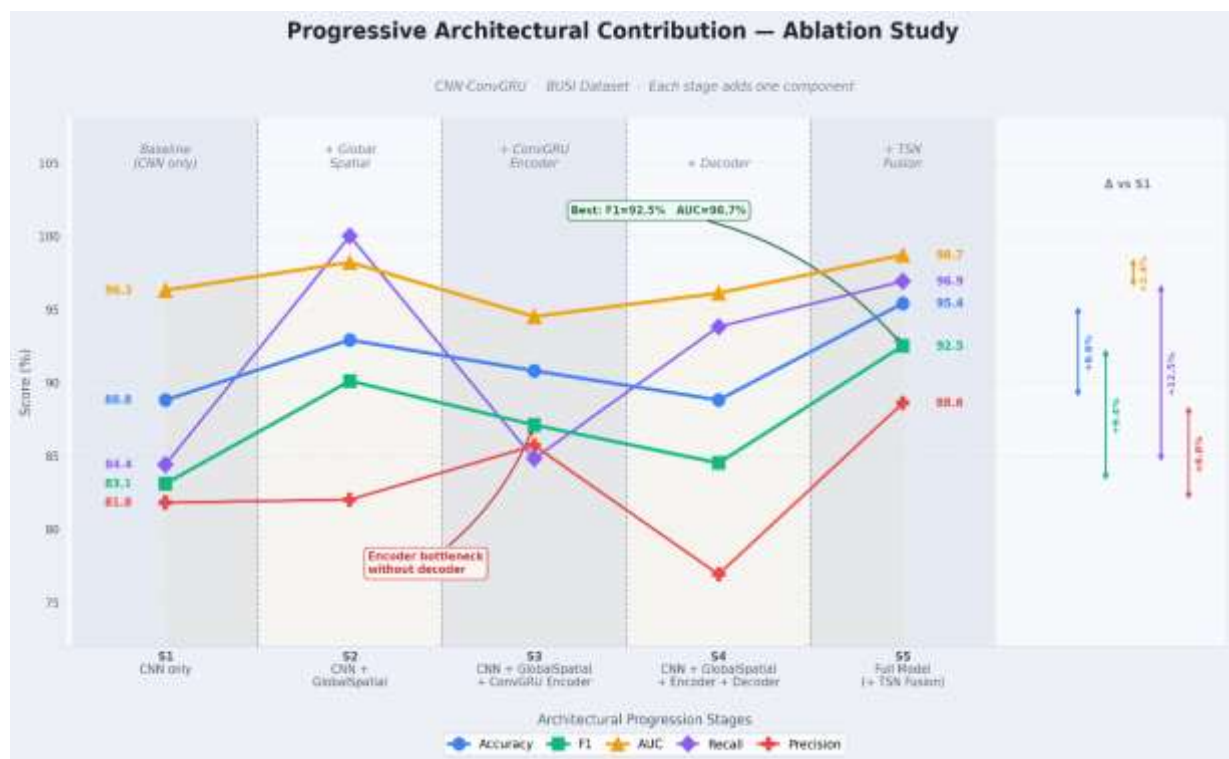


Figure 12. Progressive architectural contribution — ablation study of CNN-ConvGRU components

CONCLUSION

In this work, a unique spatial-temporal deep learning model combining CNNs and ConvGRUs is designed and applied to the BUSI database in order to improve the classification task. The overall high diagnostic accuracy was shown, where the accuracy, recall, precision, F1-score, and AUC were found to be 95.40%, 96.88%, 88.57%, 92.54%, and 0.9867, respectively. Based on the analysis, the proposed

model proved to have significantly higher accuracy than other existing baseline algorithms, including EfficientNetB7, DenseNet121, ConvNeXtTiny, XGBoost Ensemble, and Soft Voting Ensemble. It can be stated that both spatial characteristics captured through CNNs and temporal relationship modeled via ConvGRUs were important for increasing the overall classification accuracy, decreasing false negatives and increasing robustness, especially when dealing with morphologically diverse tumor images. As shown through an ablation study, each of the three components introduced (i.e., Global Spatial attention, ConvGRU Encoder-Decoder, and TSN Fusion) contributed to the final outcome, leading to a +6.6% increase in accuracy compared to the pure CNN baseline. In addition, preprocessing that included horizontal flipping, rotation, zoom, and class balancing helped avoid overfitting.

Further research can focus on the use of improved attention models along with large-scale datasets involving more institutions for better generalizability. In addition, a multi-class approach toward the detection of different stages of cancer using this framework can be considered for increasing clinical utility. Clinical implementation of this method would provide evidence regarding the applicability of this method in practice.

REFERENCES

- [1] Tran D, Bourdev L, Fergus R, Torresani L, Paluri M. Learning spatiotemporal features with 3d convolutional networks. In Proceedings of the IEEE international conference on computer vision 2015 (pp. 4489-4497). <https://doi.org/10.1109/iccv.2015.510>
- [2] Feichtenhofer C, Fan H, Malik J, He K. Slowfast networks for video recognition. In Proceedings of the IEEE/CVF international conference on computer vision 2019 (pp. 6202-6211). <https://doi.org/10.1109/iccv.2019.00630>
- [3] Tran D, Wang H, Torresani L, Ray J, LeCun Y, Paluri M. A closer look at spatiotemporal convolutions for action recognition. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition 2018 (pp. 6450-6459). <https://doi.org/10.1109/CVPR.2018.00675>
- [4] Zhou B, Andonian A, Oliva A, Torralba A. Temporal relational reasoning in videos. In Proceedings of the European conference on computer vision (ECCV) 2018 (pp. 803-818). https://doi.org/10.1007/978-3-030-01246-5_49
- [5] Bertasius G, Wang H, Torresani L. Is space-time attention all you need for video understanding?. In Icm1 2021 Jul 18 (Vol. 2, No. 3, p. 4).
- [6] Wang L, Xiong Y, Wang Z, Qiao Y, Lin D, Tang X, Van Gool L. Temporal segment networks: Towards good practices for deep action recognition. In European conference on computer vision 2016 Sep 17 (pp. 20-36). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-46484-8_2
- [7] Chikkala RB, Anuradha C, Murty PS, Rajeswari S, Rajeswaran N, Murugappan M, Chowdhury ME. Enhancing breast cancer diagnosis with bidirectional recurrent neural networks: a novel approach for histopathological image multi-classification. IEEE Access. 2025 Feb 17;13: pp. 41682-41707. <https://doi.org/10.1109/ACCESS.2025.3542989>
- [8] He J, Zhao X, Luo Z, Su S, Li S, Zhang G. TSESNet: Temporal-Spatial Enhanced Breast Tumor Segmentation in DCE-MRI Using Feature Perception and Separability. In IJCAI 2024 Aug 3 (pp. 803-811). <https://doi.org/10.24963/ijcai.2024/89>
- [9] Qasrawi R, Thwib S, Issa G, AbuGhoush R, AlMasri H, Qawasmi M, Halawi NA. Hybrid Spatiotemporal Feature Fusion for Robust Lesion Detection and Tracking in Breast Ultrasound Video Data. Intelligence-Based Medicine. 2025 Dec 10;13:100330. <https://doi.org/10.1016/j.ibmed.2025.100330>
- [10] Luo X, Wei Z, Liang Y, Wong C, Xu Z, Liu K, Chu H, Zhang Z, Liu Z, Shi Z, Wang Y. STFL-Net: A Spatial-Temporal Feature Learning Network for Breast Tumor Segmentation in Multi-Sequence MRI. Digital Signal Processing. 2025 Dec 13;171:105825. <https://doi.org/10.1016/j.dsp.2025.105825>
- [11] Han M, Guo D, Yuan J, Lu C. A spatio-temporal feature fusion network for intelligent analysis of breast cancer contrast-enhanced ultrasound video. In Proceedings of the 2024 4th international conference on bioinformatics and intelligent computing 2024 Jan 26 (pp. 270-274). <https://doi.org/10.1145/3665689.3665735>
- [12] Mahichi H, Ghods V, Sohrabi MK, Sabbaghi A. BreastCNet: Breast Cancer Detection, Classification, and Localization Convolutional Neural Network with Advanced Optimization Techniques. IEEE Access. 2025 May 15;13: pp. 87386-87400. <https://doi.org/10.1109/ACCESS.2025.3570364>
- [13] Tu Z, Zhu Z, Duan Y, Jiang B, Wang Q, Zhang C. A spatial-temporal progressive fusion network for breast lesion segmentation in ultrasound videos. IEEE Transactions on Multimedia. 2025 Aug 14;27: pp. 7470-7482. <https://doi.org/10.1109/TMM.2025.3599028>
- [14] Pradeepa M, Sharmila B, Nirmala M. A hybrid deep learning model EfficientNet with GRU for breast cancer detection from histopathology images. Scientific Reports. 2025 Jul 9;15(1):24633.

- <https://doi.org/10.1038/s41598-025-00930-6>
- [15] Assiri AS, Nazir S, Velastin SA. Breast tumor classification using an ensemble machine learning method. *Journal of Imaging*. 2020 May 29;6(6):39. <https://doi.org/10.3390/jimaging6060039>
- [16] Packer MD, Lester SC. Current understanding of phyllodes tumors of the breast: Tumor classification, molecular landscape, and best pathology practice. *Human Pathology*. 2025 Aug 1;162:105863. <https://doi.org/10.1016/j.humpath.2025.105863>
- [17] Luo Y, Huang Q, Li X. Segmentation information with attention integration for classification of breast tumor in ultrasound image. *Pattern Recognition*. 2022 Apr 1;124:108427. <https://doi.org/10.1016/j.patcog.2021.108427>
- [18] Mehta K, Mishra S. Medical Image Fusion & Classification of Tumor Cells Using Deep Learning. *International Academic Journal of Science and Engineering*. 2021; 8(4): 29–33. <https://doi.org/10.71086/IAJSE/V8I4/IAJSE0828>
- [19] Wei M, Du Y, Wu X, Su Q, Zhu J, Zheng L, Lv G, Zhuang J. A benign and malignant breast tumor classification method via efficiently combining texture and morphological features on ultrasound images. *Computational and mathematical methods in medicine*. 2020;2020(1):5894010. <https://doi.org/10.1155/2020/5894010>
- [20] Assegid W, Ketema G. Harnessing AI for early cancer detection through imaging and genetics. *Clinical Journal for Medicine, Health and Pharmacy*. 2023 Oct 9;1(1):1-5.
- [21] Cao Z, Yang G, Chen Q, Chen X, Lv F. Breast tumor classification through learning from noisy labeled ultrasound images. *Medical Physics*. 2020 Mar;47(3):1048-57. <https://doi.org/10.1002/mp.13966>
- [22] Hassan MR, Alnaimat F, Owida HA, Kharma Q, Ali AM, Ahmad R, Alhaj M. Integrating Deep Learning Models and Data Augmentation Techniques for Improved Breast Cancer Detection. *J. Wirel. Mob. Networks Ubiquitous Comput. Dependable Appl.*. 2025 Jun;16(2):517-34. <https://doi.org/10.58346/JOWUA.2025.I2.031>
- [23] Shrivastava N, Bharti J. Breast tumor detection and classification based on density. *Multimedia Tools and Applications*. 2020 Sep;79(35):26467-87. <https://doi.org/10.1007/s11042-020-09220-x>
- [24] Yadav RK, Mishra AK, Saini JB, Pant H, Biradar RG, Waghodekar P. A model for brain tumor detection using a modified convolution layer ResNet-50. *Indian Journal of Information Sources and Services*. 2024 Feb;14(1):29-38. <https://doi.org/10.51983/ijiss-2024.14.1.3753>
- [25] Singh VK, Rashwan HA, Romani S, Akram F, Pandey N, Sarker MM, Saleh A, Arenas M, Arquez M, Puig D, Torrents-Barrena J. Breast tumor segmentation and shape classification in mammograms using generative adversarial and convolutional neural network. *Expert Systems with Applications*. 2020 Jan 1;139:112855. <https://doi.org/10.1016/j.eswa.2019.112855>
- [26] Wen J, Thibaut-Sutre E, Diaz-Melo M, Samper-González J, Routier A, Bottani S, Dormont D, Durrleman S, Burgos N, Colliot O, Alzheimer's Disease Neuroimaging Initiative. Convolutional neural networks for classification of Alzheimer's disease: Overview and reproducible evaluation. *Medical image analysis*. 2020 Jul 1;63:101694. <https://doi.org/10.1016/j.media.2020.101694>
- [27] Dutta S, Mandal JK, Kim TH, Bandyopadhyay SK. Breast cancer prediction using stacked GRU-LSTM-BRNN. *Applied Computer Systems*. 2020 Dec 28;25(2):163-71. <https://doi.org/10.2478/acss-2020-0018>
- [28] Fatima M, Khan MA, Mirza AM, Shin J, Alasiry A, Marzougui M, Cha J, Chang B. An adaptive deep learning approach based on InBNFus and CNNDen-GRU networks for breast cancer and maternal fetal classification using ultrasound images. *Scientific Reports*. 2025 Jul 1;15(1):22027. <https://doi.org/10.1038/s41598-025-03402-z>
- [29] Heikal A, El-Ghamry A, Elmougy S, Rashad MZ. Fine tuning deep learning models for breast tumor classification. *Scientific Reports*. 2024 May 10;14(1):10753. <https://doi.org/10.1038/s41598-024-60245-w>
- [30] Kaddes M, Ayid YM, Elshewey AM, Fouad Y. Breast cancer classification based on hybrid CNN with LSTM model. *Scientific Reports*. 2025 Feb 5;15(1):4409. <https://doi.org/10.1038/s41598-025-88459-6>
- [31] Koshy SS, Anbarasi LJ, Narendra M, Singh RK. HED-Net: a hybrid ensemble deep learning framework for breast ultrasound image classification. *Frontiers in Artificial Intelligence*. 2025;8:1672488. <https://doi.org/10.3389/frai.2025.1672488>
- [32] Al-Dhabyani W, Gomaa M, Khaled H, Fahmy A. Dataset of breast ultrasound images. *Data in brief*. 2020 Feb 1;28:104863. <https://doi.org/10.1016/j.dib.2019.104863>