© JUN 2026 | IRE Journals | Volume 9 Issue 12 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I12-1718691

Temporal and Spatial Fusion for Breast Lesion Classification in Ultrasound Videos using Hybrid UNet-LSTM Architecture

SAKTHI YAZHINI JOTHI LATHA
MSc Advanced Computer Science (Artificial Intelligence)

Abstract- ***This research involves the design, implementation, and evaluation of a Hybrid UNet-LSTM architecture aimed at enhancing the classification of breast lesions in ultrasound videos, which is critical for the early and accurate diagnosis of breast cancer. The model integrates Convolutional Neural Networks (CNNs) with Long Short-Term Memory (LSTM) networks inspired from the U-Net architecture to capture both spatial and temporal features from video sequences. This approach addresses the inherent challenges of ultrasound imaging, such as speckle noise, operator dependency, and variability in lesion appearances, which often complicate diagnosis. The Hybrid UNet-LSTM model processes ultrasound videos, employing intra-video and inter-video fusion techniques to improve classification accuracy. The model is built using the latest advancements in deep learning, ensuring a robust and reliable system for medical diagnostics. The implementation of this Hybrid UNet-LSTM architecture serves as a foundational model that can be further improved upon to develop more advanced models in the future, enhancing breast cancer diagnostics by providing a more accurate and reliable method for lesion classification. This groundwork supports ongoing innovation in the field, with the potential to improve patient outcomes and assist radiologists in clinical decision-making.***

I. INTRODUCTION

1.1 Context

Breast cancer is a significant health problem all over the world especially among women, which is why the early classification of the disease and a precise diagnosis are the keys to successful treatment and increasing the survival rates. Ultrasound imaging is widely used in breast cancer diagnostics due to its non-invasive nature, absence of ionizing radiation, and high performance in dense breast tissues where mammography might be less effective (Berg et al., 2008; Giger et al., 2013). However, the complex appearance of breast tissue in ultrasound images often leads to interpretation challenges, contributing to misdiagnosis (Jemal et al., 2011).

In recent years, artificial intelligence and computer vision breakthroughs have provided new ways to automate the medical image analysis. This could help radiologists to identify breast lesions more accurately (Litjens et al., 2017). Traditional machine learning techniques have have been replaced by sophisticated deep learning methods, like hybrid architectures. These architectures combine convolutional neural networks (CNNs) with recurrent neural networks (RNNs), such as Long Short-Term Memory (LSTM) networks (Shen et al., 2017). They are designed to capture both spatial and temporal features from ultrasound videos, improving the accuracy of lesion classification (Ronneberger et al., 2015).

Building on the foundational work presented in the MICCAI 2022 conference paper by Lin et al., which introduced a novel dataset and the CVA-Net architecture for breast lesion classification in ultrasound videos, this research aims to enhance existing methodologies. By developing a Hybrid UNet-LSTM model that integrates intra-video and inter-video fusion, the research seeks to improve classification accuracy and robustness, contributing to more reliable breast cancer diagnostics.

In addition to the technical goals, this research addresses the practical application of AI-driven diagnostic tools in clinical settings, aiming to bridge the gap between automated systems and human expertise in radiology (Topol, 2019).

1.2 Research Aim

The aim of this research is to develop and evaluate a Hybrid UNet-LSTM architecture for the

classification of breast lesions in ultrasound videos. The primary goal is to use spatial and temporal information from ultrasound sequences, using intra-video and inter-video fusion, to improve lesion classification accuracy, thereby assisting in the early and reliable classification of breast cancer.

1.3 Objectives

1. Design and implement a Hybrid UNet-LSTM model that integrates spatial and temporal information from ultrasound videos using intra-video and inter-video fusion.
2. Implement and compare three different models CNN, Transformer, and the Hybrid UNet-LSTM, using the dataset provided in the CVA-Net paper.
3. Evaluate the performance of the proposed models in terms of accuracy, precision, recall, and other relevant metrics.
4. Document the development process, experimental setup, and results in a comprehensive research report.

1.4 Deliverables

1. A GitHub repository containing the source code for the implemented models, including the CNN, Transformer, and Hybrid UNet-LSTM models.
2. A detailed developer documentation providing an overview of the architecture, technologies used, and instructions for setting up and deploying the models.
3. The MSc dissertation report that includes the background research, methodology, experimental results, and conclusions.

1.5 Ethical, legal, and social issues

The data used in this study is sourced from the MICCAI 2022 paper written by Lin et al., which has undergone a medical ethical review and the approval of its use in research (Lin et al., 2022). Securing the patient's privacy is of utmost importance and in this data set, the use of patient data is completely anonymised to protect the individual's privacy, in compliance with ethical norms.

The dataset and research also complies with laws regarding data protection, most notably the GDPR. Since medical data is classified as sensitive personal data under GDPR, the anonymisation of patient information in the dataset is a critical legal step. No personal identifiers are used, and any secondary use of the data adheres to the legal framework governing research data (European Parliament and Council, 2016). Furthermore, the research ensures that any data sharing or publication complies with the applicable data sharing agreements stipulated by the dataset's providers.

The development and deployment of AI-driven diagnostic tools for breast cancer classification have the potential to significantly impact healthcare. It particularly improves access to diagnostic services in underserved or resource-limited regions. However, the social implications of deploying such technologies must be carefully managed to ensure they complement human expertise and do not replace them (Topol, 2019). For instance, even though AI systems can help radiologists diagnose breast cancer, they are not standalone solutions. The developed model supports radiologists, rather than replacing their critical assessment.

II. BACKGROUND RESEARCH

2.1 Literature Survey

Breast lesion classification is a major area in medical imaging and computer-aided diagnosis, mainly aimed at enhancing the accuracy and efficiency of breast cancer classification. This disease, which is the second most common cause of death for women all over the world, is on the rise and thus there is a growing demand for the development of new tools which can help radiologists in classifying suspicious lesions (Sree et al., 2011). This chapter delivers a comprehensive review of the available techniques for the classification of breast lesions with a primary focus on ultrasound methods and deep learning algorithms.

2.1.1 Ultrasound Imaging for Breast Lesion Classification

Ultrasound imaging is a highly crucial method when it comes to the early detection of breast cancer because of it’s non-invasive nature, absence of ionising radiation, real-time capabilities, and ability to differentiate between solid masses and cysts (Berg et al., 2008). It is especially helpful for patients with dense breast tissue, where mammography may be less effective (Giger et al., 2013).

© JUN 2026 | IRE Journals | Volume 9 Issue 12 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I12-1718691

However, interpreting ultrasound images pose challenges such as speckle noise, operator dependence, and variability in lesion appearance.
Speckle noise: This inherent noise in ultrasound images degrades image quality and complicates the identification of lesion boundaries.

a b

Figure 2.1 a) Ultrasound image of kidney b) Noisy image. Reproduced from Figure 4 of Mateo and Fernández-Caballero (2009)

Image (a) shows an original ultrasound image of a kidney, while image (b) displays the same image with added speckle noise.

Advanced filtering techniques such as anisotropic diffusion and wavelet-based denoising have been proposed to mitigate this issue (Acharya et al., 2012). But they require careful parameter tuning and can remove important diagnostic information (Guan et al., 2014).

Operator dependence: The quality of ultrasound imaging is highly dependent on the operator's skill and experience. This causes variance in image acquisition and interpretation. The variability can result in inconsistent diagnostic outcomes, hence the need for standardized imaging protocols and automated analysis tools (Wang et al., 2020; Liu et al., 2020; Farina & Sparano, 2013).

Variability in lesion appearance: Breast lesions in ultrasound images can appear quite diverse due to the different attributes like size, shape, and tissue composition. This variance poses significant challenges for both manual and automated classification systems, necessitating robust algorithms capable of handling diverse lesion characteristics (Cao et al., 2019).

Figure 2.2 Variability in lesion appearance in ultrasound images from different patients from the MICCAI dataset.

This figure shows ultrasound frames from three different patients, illustrating the significant variability in breast lesion appearance. The lesions differ in size, shape, and tissue composition.

To address these challenges, computer-aided diagnosis (CAD) systems have been developed. They assist radiologists in analysing ultrasound images and categorizing breast lesions (Zhou et al., 2019). These systems use machine learning and deep learning techniques to enhance accuracy and reduce operator dependence. Recent advancements have also focused on video-based techniques that utilize temporal information to improve lesion classification and classification.

The Clip-level and Video-level Aggregation Network (CVA-Net) represents a significant advancement in ultrasound imaging by processing video sequences rather than static images (Lin et al., 2022). This approach improves classification accuracy by capturing the dynamic nature of breast lesions. CVA-Net's ability to aggregate information across multiple frames addresses the limitations of single-frame analysis and offers a promising direction for future research.

Though CAD systems and video-based techniques have shown promise, there are areas that require improvement. Current systems rely on large annotated datasets for training which may not be readily available in all clinical settings. Also, the interpretability of deep learning models remains a concern since clinicians need transparent decision-making processes to trust automated systems. Future research should focus on developing more interpretable models to reduce data dependency.

2.1.2 Traditional Approaches to Breast Lesion Classification

Early automation of breast lesion classification relied on traditional machine learning techniques and manually designed features. These approaches primarily had four key steps:

1. Image preprocessing:

Enhancing ultrasound image quality is important due to degradation from speckle noise and low contrast. Techniques like anisotropic diffusion filtering and wavelet-based denoising reduce noise while preserving edges (Acharya et al., 2012). Contrast enhancement methods, such as histogram equalization improve visibility of lesions while intensity normalization ensures consistency across different ultrasound machines. However, these methods require careful parameter tuning and may alter diagnostically relevant features.

2. Segmentation:

This step finds lesion boundaries using methods like region growing, active contours, and watershed algorithms (Horsch et al., 2002). Region growing expands areas based on criteria, active contours minimize energy to fit contours and watershed algorithms identify ridges for classification. Despite their effectiveness, these methods struggle with complex shapes and require manual intervention.

3. Feature extraction:

Features based on shape, texture, and intensity are manually extracted from segmented lesions (Sree et al., 2011).

4. Classification:

Algorithms like Support Vector Machines (SVM), Random Forests, and k-Nearest Neighbors (k-NN) classify lesions using the extracted features (Horsch et al., 2002). SVMs use hyperplanes, Random Forests employ ensemble learning, and k-NN relies on nearest neighbors. These are limited by the quality of manual features and may not generalize well to diverse datasets.

Though traditional approaches have laid the groundwork for automated breast lesion classification, they are not without limitations. They require manual feature engineering and pose the challenge of capturing complex patterns in ultrasound images (Acharya et al., 2012). Future research should focus on using advanced image processing techniques to automate feature extraction. Also, developing robust classification algorithms that can handle diverse lesion shapes can enhance the reliability and scalability of automated systems.

2.1.3 Deep Learning for Breast Lesion Classification

Deep learning has revolutionized the field of medical image analysis including breast lesion classification. Convolutional Neural Networks (CNNs) have become the prevailing architecture for this task offering several advantages over traditional methods:

Automatic feature learning:

Unlike traditional methods that rely on manually engineered features, CNNs autonomously learn hierarchical representations from raw image data. This allows CNNs to capture complex patterns and subtle differences in breast lesions that might be overlooked by manual feature extraction (Litjens et al., 2017).

Improved performance:

CNNs consistently outperform traditional approaches in breast lesion classification benchmarks. They also show high accuracy and robustness. This improvement is because of their ability to model complex, non-linear relationships (Shen et al., 2017).

Several notable CNN architectures have been applied to breast lesion classification:

VGGNet: VGGNet has been adapted for breast ultrasound classification tasks because of its simplicity and effectiveness. It's deep architecture facilitates the learning of detailed features (Simonyan and Zisserman, 2014).

ResNet: ResNet allows for the training of deeper networks by introducing residual connections. These deeper networks can capture more complex patterns in ultrasound images. ResNet also addresses the degradation problem that occurs due to increasing network depth (He et al., 2016).

© JUN 2026 | IRE Journals | Volume 9 Issue 12 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I12-1718691

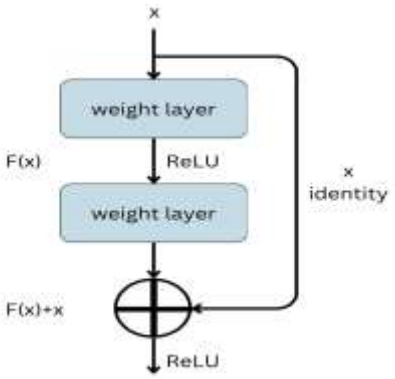

Figure 2.3 Residual learning building block from ResNet architecture. Reproduced from He et al., 2016.

This figure illustrates the basic building block of the ResNet architecture. The block includes two weight layers, each followed by a ReLU activation function, with a shortcut connection that adds the input x directly to the output F(x), creating F(x)+x.

DenseNet: DenseNet promotes feature reuse through dense connections that enhance information flow between layers. (Huang et al., 2017).

While deep learning has significantly advanced breast lesion classification, there are still several challenges that need to be addressed. The requirement of large annotated datasets for training is a major limitation since such datasets are not always available. Additionally, the "black-box" nature of deep learning models raises concerns about interpretability, which is crucial for clinical adoption.

2.1.4 Transfer Learning and Domain Adaptation

Transfer learning has played a key role in the enhancement of the deep learning applications for breast lesion classification, particularly in overcoming the limitations caused by small medical imaging datasets. By using pre-trained models on big natural image datasets like ImageNet, researchers have been able to transfer learned representations to medical imaging tasks (Shin et al., 2016). This approach not only accelerates the training process but also enhances model performance by incorporating generalized features. Common transfer learning strategies are:

Fine-tuning:

This involves adapting pre-trained models to specific tasks by updating all or a subset of the model's weights (Shin et al., 2016). Fine-tuning allows the customization of models to capture domain-specific features pertinent to breast lesions. However, excessive fine-tuning can lead to overfitting, particularly when dealing with small datasets.

Feature extraction:

In this approach, pre-trained models act as fixed feature extractors with only the final classification layers being modified and trained (Tajbakhsh et al., 2016). This method is computationally efficient and reduces the risk of overfitting.

Transfer learning also has a few limitations that need to be overcome. The reliance on pre-trained models from non-medical domains raises concerns about the relevance of transferred features. Also, domain adaptation techniques which address the shift between natural and medical images are still evolving and it requires further exploration to improve generalization in breast lesion classification tasks.

2.1.5 Attention Mechanisms and Interpretability

Attention mechanisms have emerged as a powerful tool to enhance the interpretability of models by highlighting relevant regions in ultrasound images, mimicking the visual attention of radiologists.

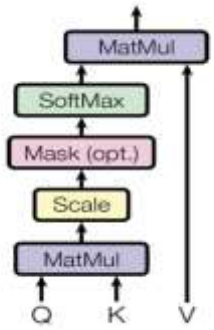

Figure 2.4 Self-attention mechanism used in Transformer models. Reproduced from Vaswani et al., 2017.

This figure illustrates the self-attention mechanism in Transformer models. It does matrix multiplication of three inputs: Query (Q), Key (K), and Value (V),

followed by scaling. An optional mask can be applied to refine the focus. The result is passed through a SoftMax function to generate attention weights, which are then used to weight the Value (V) vectors in the final matrix multiplication.

Vaswani et al. (2017) introduced the transformer architecture which relies entirely on self-attention mechanisms for sequence transduction tasks. This was originally designed for natural language processing, but the principles of self-attention have been adapted for various computer vision tasks, including medical image analysis. The key innovation of Vaswani et al.'s work is the multi-head self-attention mechanism that allows the model to attend to information from different representations at the same time.

Jetley et al. (2018) proposed an end-to-end trainable attention module for convolutional neural networks (CNNs) in image classification tasks. Their approach generates attention maps that highlight regions of interest while suppressing background clutter. The attention module takes 2D feature vector maps as input and outputs a 2D matrix of scores, effectively learning to focus on the most relevant parts of the image. This method demonstrated high generalization over multiple datasets, making it highly applicable to medical imaging tasks.

In ultrasound video analysis for breast lesion classification, the concepts from both Jetley et al. and Vaswani et al. can be adapted and combined. The spatial attention mechanism proposed by Jetley et al. can be used to focus on relevant regions within individual ultrasound frames, while the temporal self-attention from Vaswani et al.'s transformer architecture can be applied to capture dependencies across video frames. This combination of spatial and temporal attention could potentially enhance the model's ability to classify breast lesions.

However, the effectiveness of these mechanisms is heavily dependent on the quality and quantity of training data. Models trained on limited datasets may not generalize well to diverse clinical scenarios. Moreover, the interpretability of attention maps is sometimes questioned, as they may not always align perfectly with human intuition or clinical guidelines.

2.1.6 Multi-modal and 3D Approaches

While 2D ultrasound remains the most common modality for breast lesion classification, researchers have explored multi-modal and 3D approaches as well.

Multi-modal fusion:

Multi-modal fusion involves integrating data from different imaging modalities such as ultrasound, mammography, and MRI, to enhance diagnostic accuracy. This approach utilises the strengths of each modality:

1. Ultrasound provides real-time imaging and differentiation between solid and cystic lesions.
2. Mammography offers high-resolution images ideal for detecting calcifications.
3. MRI provides excellent soft-tissue contrast.

Studies have shown that combining these modalities can improve classification performance by providing complementary information and enhances lesion characterization (Zhou et al., 2019). However, there are challenges in aligning and integrating multi-modal data due to differences in image resolution, orientation, and acquisition parameters. Future research should focus on developing robust algorithms for seamless data integration and exploring the potential of emerging modalities such as photoacoustic imaging (Zhao et al., 2019).

3D ultrasound:

3D ultrasound offers volumetric data. This volumetric data provides additional spatial context, potentially leading to more accurate lesion classification (Zhou et al., 2019). Unlike 2D ultrasound which can miss lesions due to limited imaging planes, 3D ultrasound captures the entire volume of the breast, allowing for more comprehensive analysis. This modality also enables the assessment of lesion shape, size and spatial relationships with surrounding tissues. Despite its advantages, 3D ultrasound is underutilized in clinical practice due to higher costs, longer acquisition times, and the need for specialized equipment and training. Research should focus on optimizing 3D ultrasound techniques for faster acquisition and developing automated tools for volumetric data analysis.

Spatio-temporal analysis:
Spatio-temporal analysis involves incorporating temporal information from ultrasound video sequences to capture dynamic features of breast lesions. This approach recognizes that lesions may exhibit changes in appearance over time, which can provide valuable diagnostic insights. Techniques like the Clip-level and Video-level Aggregation Network (CVA-Net) offer a more holistic view of lesion characteristics (Lin et al., 2022). Though promising, these techniques require large annotated video datasets for training, which are often scarce and the computational complexity of processing video data is difficult for real-time applications. Future work should explore efficient video processing algorithms and investigate the integration of temporal data with other modalities.

While multi-modal and 3D approaches offer significant potential for improving breast lesion classification, several challenges must be addressed. The integration of diverse data sources requires complex algorithms capable of handling variability in image quality and acquisition parameters. The interpretability of complex models remains a concern, as clinicians need to understand the rationale behind automated decisions.

Future research should prioritize the development of interpretable models, automate data integration and investigate the clinical usability of these imaging techniques in diverse patient groups. By addressing these challenges, multi-modal and 3D approaches can provide more comprehensive and accurate diagnostic information.

2.2 Methods and Techniques
This section explores the various methods and techniques that have been employed in breast lesion classification to date, ranging from basic Convolutional Neural Networks (CNNs) to sophisticated attention mechanisms.

2.2.1 Convolutional Neural Networks (CNNs)

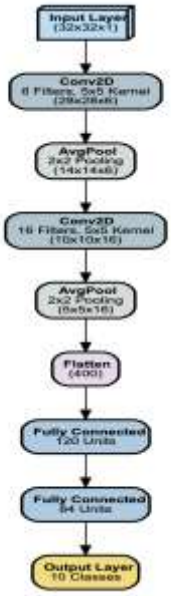

Figure 2.5 Example of a Convolutional Neural Network (CNN) architecture for image classification.

This figure illustrates a simple CNN architecture designed for image classification tasks. The network begins with an input layer for a 32x32 grayscale image, followed by two convolutional layers with 6 and 16 filters respectively, each using a 5x5 kernel. After each convolutional layer, an average pooling layer reduces the spatial dimensions, enhancing computational efficiency. The network then flattens the data into a 1D vector, which is passed through two fully connected layers with 120 and 84 units. The final output layer classifies the image into one of 10 possible classes.

Convolutional Neural Networks (CNNs) can be considered as the building blocks to many modern image classification tasks. They are especially used in medical imaging due to their ability to automatically learn and extract hierarchical features from images (LeCun et al., 2015). A CNN typically consists of several key layers including convolutional layers, pooling layers and fully connected layers. Each layer plays a crucial role in the network's ability to process and classify images.

Convolution:
The core operation in a CNN is convolution. Convolution applies filters (kernels) to the input

© JUN 2026 | IRE Journals | Volume 9 Issue 12 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I12-1718691

image to extract features. This operation is mathematically defined as:

$$Y_{i,j,k} = \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} \sum_{c=0}^{C-1} W_{m,n,c,k} \cdot X_{i+m,j+n,c} + b_k$$

where X is the input feature map,W is the convolutional kernel, b_k is the bias, and Y_(i,j,k) is the output feature map.

The convolution operation involves sliding the kernel over the input feature map and computing a dot product between the kernel and the corresponding input region. This process is repeated across the entire image, producing a set of feature maps. These feature capture the edges, textures and patterns of the input image.

Pooling:
Pooling layers reduce the spatial dimensions of the feature maps, lowering computational load. This enhances the feature invariance. Average pooling is expressed as:

$$Y_{i,j,k} = \frac{1}{P \times Q} \sum_{m=0}^{P-1} \sum_{n=0}^{Q-1} X_{Pi+m,Qj+n,k}$$

where $P \times Q$ is the size of the pooling window and $Y_{i,j,k}$ is the pooled output at position (i,j) for the k-th feature map.

Pooling layers combine information from small neighbourhoods in the feature maps. This leads to a reduction in dimensionality while retaining the most important features.

Fully Connected Layers:
The final layers are fully connected, transforming 2D feature maps into a 1D vector for classification. The output is given by

$$\hat{y} = \sigma(W^{(l)} \cdot h^{(l-1)} + b^{(l)})$$

where $W^{(l)}$ is the weight matrix, $h^{(l-1)}$ is the previous layer's activation, $b^{(l)}$ is the bias and σ is the activation function (e.g., softmax). This final output $\hat{y}$ represents the predicted probabilities of each class.

CNNs might be powerful, but their performance depends on large, annotated datasets, which is the main challenge when dealing with medical data. Also, the "black box" nature of CNNs makes it difficult to adopt for clinical settings. But this lead to the development of interpretability tools like Grad-CAM (Selvaraju et al., 2017) and lightweight architectures like MobileNet (Howard et al., 2017) for deployment in resource-constrained environments.

2.2.2 Residual Neural Networks (ResNet)
Residual Neural Networks (ResNet) were introduced to address the degradation problem in deep networks where increasing depths lead to higher training error (He et al., 2016). ResNet solves this issue by adding residual connections that skip layers.

Residual Block:
The core of ResNet is the residual block, where the input to a layer is added to its output:

$$\boldsymbol{y}_l = \mathcal{F}(\boldsymbol{x}_l, \{W_l\}) + \boldsymbol{x}_l$$

Here, x_l is the input to the l-th layer, $\mathcal{F}(\boldsymbol{x}_l, \{W_l\})$ is the residual mapping (a series of convolutional layers followed by non-linearities) and $\boldsymbol{y}_l$ is the output of the l -th layer after adding the input via a shortcut connection.

Residual Mapping:
The residual mapping, typically a series of convolutions, can be expressed as:

$$\mathcal{F}(\boldsymbol{x}_l, \{W_l\}) = W_2 \cdot \sigma(W_1 \cdot \boldsymbol{x}_l + b_1) + b_2$$

where W_1 and W_2 are the weight matrices for the two convolutional layers, σ is the activation function, and b_1 and b_2 are the bias terms.

© JUN 2026 | IRE Journals | Volume 9 Issue 12 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I12-1718691

The residual connection $\boldsymbol{x}_l + \mathcal{F}(\boldsymbol{x}_l, \{W_l\})$ makes sure that the gradient can flow more easily through the network. This reduces the chances of vanishing gradients, which are common in very deep networks.

Bottleneck Block:
For deeper networks, ResNet uses a bottleneck block to reduce computational complexity. The bottleneck block is structured as follows:

1. A 1x1 convolution reduces the dimensionality.
2. A 3x3 convolution processes the spatial features.
3. Another 1x1 convolution restores the dimensionality.

The parameter count for a bottleneck block is given by:

$$params = (1 \times 1 \times C_{in} \times C_{mid}) + (3 \times 3 \times C_{mid} \times C_{mid}) + (1 \times 1 \times C_{mid} \times C_{out})$$

where C_{in} is the input channel size, C_{mid} is the reduced dimension size and C_{out} is the output channel size.

This structure reduces the number of parameters while maintaining the network's ability to learn complex features. This aspect makes it more effective in medical imaging where high-resolution details are critical.

ResNet's ability to train deep networks effectively makes it vital in medical imaging. But, like CNNs, ResNet requires large annotated datasets and also does not inherently address class imbalance or interpretability issues.

2.2.3 Attention Mechanisms
Attention mechanism is designed to allow models to focus on specific parts of the input by calculating a weighted sum of values. The weights are determined by the relevance of a particular part of the data (Vaswani et al., 2017). This is useful in natural language processing or computer vision where the model needs to selectively emphasize certain features of the input.

Attention mechanism involves three key matrices - query Q, key K, and value V. These are linear transformations of the input data:

$$Q = XW_Q, K = XW_K, V = XW_V$$

Here, X is the input data matrix, W_Q, W_K and W_V are the weight matrices for the query, key, and value projections respectively, Q,K and V are the resulting matrices.

Scaled Dot-Product Attention:
The attention scores are computed by taking the dot product of Q and the transpose of K,

$$Scores = QK^T$$

To prevent large values that could result in small gradients, the scores are scaled by the square root of the key dimensionality d_k

$$Scaled\ Scores = \frac{QK^T}{\sqrt{d_k}}$$

Softmax and Weighted Sum:
The softmax function is then applied to the scaled scores to convert them into a probability distribution:

$$Attention\ Weights = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)$$

Finally, the output of the attention mechanism is computed as a weighted sum of the values:

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

This output allows the model to focus on the most relevant parts of the input, enhancing its capacity to capture important relationships in the data.

2.2.4 Spatio-temporal Analysis
Spatio-temporal analysis is a crucial approach in models like CVA-Net that are designed to capture

© JUN 2026 | IRE Journals | Volume 9 Issue 12 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I12-1718691

both the spatial and temporal dynamics (Lin et al., 2022). This analysis is essential for understanding how lesions evolve over time, providing richer context than static images alone (Simonyan & Zisserman, 2014).

The spatio-temporal feature can be mathematically represented as:

$$f_{st} = \phi(f_s, f_t)$$

where f_s is the spatial feature vector extracted from individual frames, f_t is the temporal feature vector capturing changes across frames and $\phi(\cdot)$ is the fusion function that combines these features into spatio-temporal feature f_{st}.

Spatial Features (f_s) are extracted using convolutional layers by applying a series of filters to the input image to capture spatial hierarchies. For instance, for an image $I \in \mathbb{R}^{H\times W\times C}$, the spatial feature extraction via convolution can be described as:

$$f_s = Conv(I; W) = \sigma(W * I + b)$$

where W is the filter matrix, σ is an activation function, * is the convolution operation and b is the bias term.

Temporal Features (f_t) are extracted to capture how these spatial features evolve over time. This can be done using methods like LSTM layers or Temporal Convolutional Networks (TCNs) (Lea et al., 2016). For LSTM-based temporal feature extraction, given a sequence of spatial features $\{f_{s_1}, f_{s_2}, \dots, f_{s_T}\}$, the temporal feature f_t can be formulated as:

$$f_t = LSTM(f_{s_t})$$

where $LSTM(\cdot)$ represents the LSTM operation that captures dependencies across the temporal sequence.

The Fusion Function $\phi(\cdot)$ can take various forms, such as:

1. Concatenation: $f_{st} = Concat(f_s, f_t)$
2. Addition: $f_{st} = f_s + f_t$
3. Gated Fusion: $f_{st} = g \odot f_s + (1 - g) \odot f_t$,
4. where g is a gate parameter learned during training.

This fusion generates a spatio-temporal feature f_{st} integrating both the static spatial structure and dynamic temporal changes in the lesion.

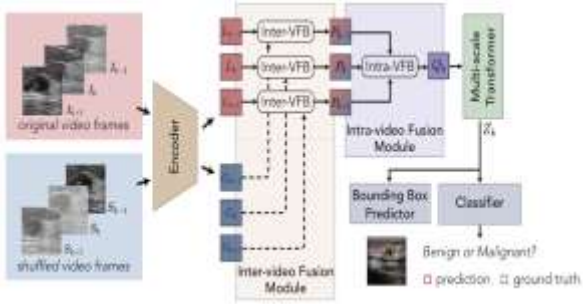

Figure 2.6 Schematic illustration of the CVA-Net architecture for breast lesion detection in ultrasound videos. Reproduced from Figure 2 of Lin et al., 2022.

2.2.5 Transfer Learning and Domain Adaptation

Transfer learning is a technique in which a model developed for a particular task (source domain) is reused as the starting point for a model on a different task (target domain). Transfer learning is useful in medical imaging where annotated datasets are limited (Pan & Yang, 2010).

The primary goal in transfer learning is to adapt a model trained on a large source domain D_s to perform well on a smaller target domain D_t. The total loss function that guides this adaptation can be expressed as:

$$\mathcal{L}_{total} = \mathcal{L}_{source} + \lambda \cdot \mathcal{L}_{domain}$$

© JUN 2026 | IRE Journals | Volume 9 Issue 12 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I12-1718691

where,

- $\mathcal{L}_{source}$ is the loss on the source domain. $\mathcal{L}_{source} = \mathbb{E}_{(x_s,y_s)\sim\mathcal{D}_s}[\ell(f(x_s;\theta_s),y_s)]$
- Here, v is the loss function, f(·;θ_s) is the model with parameters θ_s, and (x_s,y_s) are the source data and labels.
- $\mathcal{L}_{domain}$ is the domain adaptation loss, which ensures the model adapts to the target domain. It is defined as: $\mathcal{L}_{domain} = \mathbb{E}_{x_t\sim\mathcal{D}_t}[\ell(f(x_t;\theta_s),y_t)]$ where (x_t, y_t) are the target data and labels.
- λ is a hyperparameter that controls the trade-off between maintaining performance on the source domain and adapting to the target domain.

In practice, minimizing this combined loss enables the model to retain knowledge from the source domain while also adapting to the new characteristics of the target domain. This is important in medical imaging, where models often need to generalize from datasets with limited diversity.

2.2.6 Long Short-Term Memory (LSTM) Networks

LSTM networks are specialized forms of recurrent neural networks (RNNs) designed to capture long-term dependencies in sequential data. LSTMs address the vanishing gradient problem common in traditional RNNs (Hochreiter & Schmidhuber, 1997). An LSTM network's architecture revolves around its memory cell c_t, which is regulated by three gates: the input gate i_t, forget gate f_t, and output gate o_t. These gates control the flow of information into, out of, and within the cell.

The update equations for an LSTM network are as follows:

- Input Gate: $i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + b_i)$
  The input gate controls how much of the new input x_t should be added to the memory cell c_t

- Forget Gate: $f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + b_f)$
  The forget gate determines how much of the previous cell state c_{t-1} should be retained..

- Cell State Update: $c_t = f_t \odot c_{t-1} + i_t \odot \tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c)$
  The new cell state c_t is a combination of the previous cell state and the new candidate values controlled by the forget and input gates.

- Output Gate: $o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + b_o)$
  The output gate controls the extent to which the cell state affects the hidden state h_t

- Hidden State: $h_t = o_t \odot tanh(c_t)$
  The hidden state h_t represents the output of the LSTM at time t that can be passed to the next LSTM cell in the sequence or used as an output for predictions.

Weight matrices $W_{xi}, W_{hi}, W_{xf}, W_{hf}, W_{xo}, W_{ho}, W_{xc}, W_{hc}$ are learned during training and each gate's operation involves element-wise multiplication ⊙ and the non-linear activation function σ (often the sigmoid function).

LSTM networks are effective in medical imaging applications involving video sequences, such as breast lesion classification. LSTMs can detect subtle temporal changes in lesion appearance by processing ultrasound frames sequentially. For example, the change in size, shape or texture of a lesion across frames can provide important diagnostic clues that static images alone might miss.

2.3 Choice of Methods

2.3.1 Convolutional Neural Networks (CNNs)

The decision to use 3D Convolutional Neural Networks (3D CNNs) was driven by the need to capture both spatial and temporal features from ultrasound video sequences. Unlike traditional 2D CNNs, which only process spatial information from individual frames, 3D CNNs apply convolutions across both the spatial and temporal dimensions. This allows the model to simultaneously capture the structure of the lesion and its temporal evolution across multiple frames.

3D CNNs are specifically designed to process volumetric data. This makes them ideal for handling the spatio-temporal nature of ultrasound videos. By using 3D convolutions, the model can learn features that represent how lesions change over time. The ability to analyze video data as a cohesive unit rather than as isolated frames allows the model to understand context. This is important in medical imaging, where the progression of a lesion over time can provide critical diagnostic information.

2.3.2 Custom Attention Mechanisms

In this research, custom attention mechanisms are implemented within the Inter-Video Fusion Block to integrate the spatial and temporal features extracted from the original and shuffled video sequences. The attention mechanism plays an important role in aligning and fusing these features by computing the importance of different elements within the feature maps. This ensures that the most relevant information from both sequences is emphasized.

The attention mechanism is integral to the Inter-Video Fusion Block, as it computes the similarity between the local features from the original video frames and the global features from the shuffled video frames. It generates a context vector through attention weights and this ensures that the fused representation captures the most critical aspects of both local and global features.

The attention mechanism is specifically designed to facilitate the integration of features from multiple video streams (original and shuffled). This approach enables the model to combine and use complementary information from different sequences, improving the robustness and accuracy of the classification.

Rather than applying the same level of importance across all features, the attention mechanism selectively emphasizes certain features over others. This ensures that the model remains less influenced by noise or irrelevant data. The selective focus is suitable for medical imaging since diagnostic features are often subtle and surrounded by irrelevant information.

2.3.3 Long Short-Term Memory (LSTM) Networks

The Long Short-Term Memory (LSTM) networks are employed within the Intra-Video Fusion Block to capture temporal dependencies within individual ultrasound video sequences. LSTM processes the fused features from both the original and shuffled video frames, enabling the model to learn the temporal dynamics that occur within a single video.

The LSTM is integrated into the Intra-Video Fusion Block to handle the sequential nature of the feature vectors generated from the video frames. This allows the model to understand how features evolve within the context of a single video. LSTM helps in identifying patterns that might indicate the nature of the lesion by modelling these temporal dependencies. LSTM's role in modelling temporal relationships within a single video complements the Inter-Video Fusion Block, where features from original and shuffled videos are combined.

2.3.4 Conclusion

The integration of 3D CNNs, custom attention mechanisms, and LSTM networks within the Hybrid UNet-LSTM framework was designed to tackle the complexities of breast lesion classification in ultrasound videos. The 3D CNNs handle spatio-temporal feature extraction, capturing both the structural and temporal evolution of lesions. Custom attention mechanisms, implemented in the Inter-Video Fusion Block, emphasize the most relevant features from original and shuffled video sequences through the integration of multiple video streams. Meanwhile, LSTM networks in the Intra-Video Fusion Block model temporal dependencies within individual videos that identify patterns crucial for understanding lesion progression. This cohesive approach aligns with the research's objectives and serves as a foundational tool for breast cancer classification.

III. DATASETS AND EXPERIMENTAL DESIGN

3.1 Dataset

This research uses the dataset introduced by Lin et al. (2022) at the MICCAI conference. It stands out as a valuable resource given the limited availability of public medical data. The dataset is notable for its size

© JUN 2026 | IRE Journals | Volume 9 Issue 12 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I12-1718691

and the diversity of imaging conditions, comprising 188 ultrasound videos—113 labeled as malignant and 75 as benign—totaling 25,272 frames. Each video captures the entire lesion, from its initial appearance to its largest section, and finally to its disappearance, providing a rich source of temporal and spatial information.

There are several datasets available for breast lesion classification like the Digital Database for Screening Mammography (DDSM) (Heath et al., 2000) and the Breast Ultrasound Images Dataset (BUSI) (Sahu et al., 2023). However, these datasets mainly consist of static images and lack the temporal dimension for video-based analysis. The DDSM dataset is widely used for mammography but does not provide the dynamic context that ultrasound videos offer (Heath et al., 2000). Similarly, the BUSI dataset includes static ultrasound images without temporal annotations (Al-Dhabyani et al., 2020). In contrast, the dataset by Lin et al. (2022) stands out by providing annotated ultrasound videos, allowing for the examination of spatial-temporal dynamics in lesion progression which is not possible with static datasets. This makes it valuable for training and evaluating models that use temporal information.

Data Collection and Annotation

The dataset was collected using two ultrasound systems—LOGIQ-E9 and PHILIPS TIS L9-3—to ensure a broad range of imaging conditions. Each frame within the videos was annotated by two pathologists, who provided both the rectangular boundaries of the lesions and classification labels. But, the integrity of the dataset was compromised due to the corruption of annotation files as shown in figure 3.1.

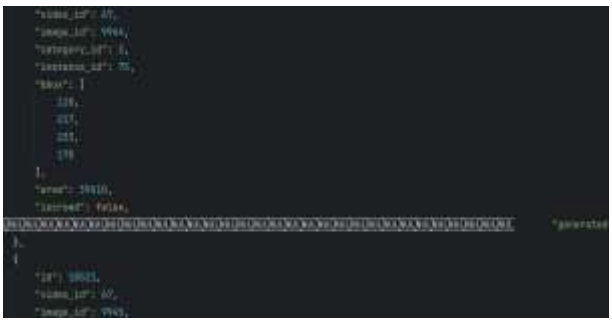

Figure 3.1 Example of a corrupted annotation file in the dataset.

Data Preprocessing and Corrections

The dataset had certain challenges that needed to be addressed to ensure data integrity and reliability. These challenges included duplicate and incorrect data entries, which were rectified following the guidelines provided in the dataset's GitHub repository.

1. Duplicate Video Removal: During the initial review of the dataset, a duplicate video labeled as benign, "rawframes/benign/x66ef02e7f1b9a0ef," was identified as identical to a malignant video. This duplicate was removed to prevent inconsistencies in the dataset.
2. Label Corrections: It was discovered that four videos initially labeled as benign were actually from the same patient and contained overlapping frames. These videos—"benign\x63c9ba1377f35bf6," "benign\x5a1c46ec6377e946," "malignant\2390fbea047347b," and "malignant\7a39ab5d4970bf89"—were corrected to malignant.

Following these corrections, the dataset underwent a series of preprocessing steps to prepare it for model training:

1. Frame extraction: Individual frames were extracted from ultrasound videos. Each video provided a series of frames capturing the progression of the lesion, ensuring that the model could learn from temporal changes within the video sequence.
2. Normalisation: Pixel intensities were normalized to a range of [0, 1]. This sped up the convergence of neural networks during training. Normalization was mathematically performed as,

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

where X is the original pixel value, and X_{min} and X_{max} are the minimum and maximum pixel values in the dataset, respectively.

While this dataset provides a valuable resource for breast lesion classification research, potential limitations include its size and the risk of overfitting due to the relatively small number of videos. Also, biases introduced during annotation could affect model performance. Acknowledging these limitations helps set realistic expectations for the model's capabilities and guides future research directions.

3.2 Experimental Design

This section provides a detailed overview of the experimental design and empirical process used to develop, evaluate, and compare the models implemented in this research. The primary focus of this research is the Hybrid UNet-LSTM model designed for breast lesion classification in ultrasound videos, with secondary investigations into the performance of a 2D CNN and a Transformer-based model. An attempt to incorporate lesion detection through segmentation was also made, which, although not fully successful, offers avenues for future work.

3.2.1. Network Architectures

3.2.1.1 Implementation Challenge and Model Adaptation

Initially, this research aimed to implement the CVA-Net model proposed by Lin et al. (2022) for breast lesion detection in ultrasound videos and to propose a novel model based on it. The model required specific CUDA versions (CUDA>=9.2, GCC>=5.4) for GPU acceleration.

Initial Approach: The first step involved attempting to run the CVA-Net model on my local machine that required precise compatibility between CUDA and PyTorch versions. Despite multiple efforts, persistent CUDA compatibility issues arose due to mismatches between the installed CUDA 11.7 and the required versions. To address these challenges, I transitioned to the university's Advanced Research Computing (ARC) system, which offered a broader range of CUDA versions. However, even on the ARC, the specific combination of CUDA, GPU drivers and dependencies needed for CVA-Net led to similar compatibility issues. This led to unsuccessful model execution.

Impact on Research Direction: Given these challenges, it became evident that an alternative approach was necessary. Drawing inspiration from CVA-Net's strengths in handling spatial and temporal features, I opted to develop a hybrid UNet-LSTM model that was more compatible with the available resources.

Hybrid UNet-LSTM Development: The hybrid UNet-LSTM model combines a UNet inspired architecture with an LSTM networks. This combination enabled the model to capture both spatial features within individual frames and temporal relationships across frames. To ensure compatibility, I selected a PyTorch version that worked seamlessly with CUDA 11.7, the most stable version available on the ARC. This optimization allowed the hybrid model to be trained effectively, avoiding the CUDA-related issues encountered earlier. The initial plan to implement and extend CVA-Net faced significant technical challenges, which led to the creation of a novel hybrid model tailored to the research needs. This model effectively addresses both spatial and temporal aspects of breast lesion detection in ultrasound data.

3.2.1.2 Hybrid UNet-LSTM Model

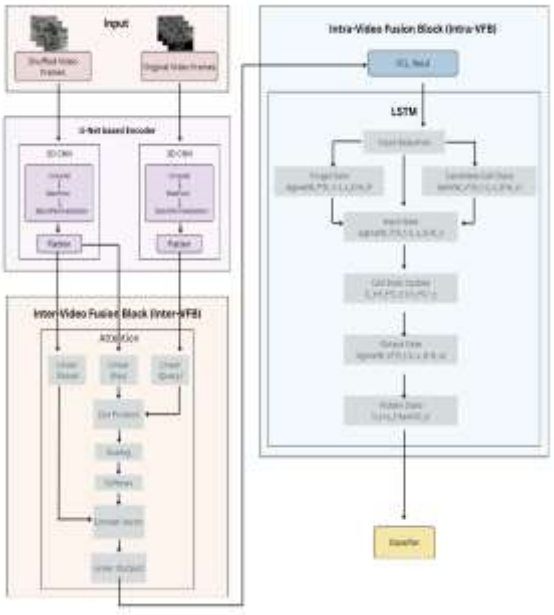

Figure 3.2 Schematic Diagram of the Hybrid UNet-LSTM Model Architecture.

This figure shows the Hybrid UNet-LSTM model for breast lesion classification in ultrasound videos. The

model uses a 3D CNN-based UNet for spatio-temporal feature extraction from original and shuffled frames, which are fused via a custom attention mechanism in the Inter-Video Fusion Block. An LSTM network in the Intra-Video Fusion Block then models temporal dependencies, and the final features are classified to predict lesion malignancy, effectively capturing both spatial and temporal dynamics.

The Hybrid UNet-LSTM model described in this research represents a sophisticated approach to classifying breast lesions in ultrasound videos by combining the strengths of spatial and temporal feature extraction. This model uses the power of a 3D CNN-based UNet for spatio-temporal feature extraction and integrates these features using a custom attention mechanism followed by an LSTM layer, which ultimately feeds into a classifier for lesion classification.

3D CNN-based UNet Backbone for Spatio-Temporal Feature Extraction:
The first stage of the Hybrid UNet-LSTM model involves a 3D CNN-based UNet inspired architecture. It is specifically tailored to handle the spatio-temporal nature of ultrasound video frames. Unlike traditional UNet models that use 2D convolutions, this model employs 3D convolutions to simultaneously capture spatial features and temporal features.

Given an input volume X∈R^(10×64×64×3), where 10 is the depth (temporal dimension), 64×64 represents the height and width, and 3 is the number of channels, the 3D convolution operation can be represented as:

$$Y = \sigma(W * X + b)$$

Here, $W \in \mathbb{R}^{3\times3\times3\times32}$ denotes the 3D convolutional kernel with a filter size of $3 \times 3 \times 3$, 32 is the number of filters, * represents the 3D convolution operation, b is the bias term, and σ is the ReLU activation function. The encoder part of the UNet progressively downsamples the input through convolutional and max-pooling layers, starting from the input shape of $10 \times 64 \times 64 \times 3$ and reducing to $1 \times 14 \times 14 \times 64$.

This model extracts two sets of features: one from the original video frames and one from the shuffled video frames. These two streams are then processed separately through the 3D CNN layers to produce local and global feature representations, respectively.

Inter-Video Fusion via Custom Attention Mechanism:
To effectively fuse the spatial and temporal features extracted from both the original and shuffled video sequences, the model uses an Inter-Video Fusion Block that leverages a custom attention mechanism. This attention mechanism allows the model to focus on the most relevant parts of the feature maps. This is achieved by computing a context vector that aligns features from different video sequences..

Let $F_o \in \mathbb{R}^{1\times14\times14\times64}$ represent the local features from the original video frames and $F_s \in \mathbb{R}^{1\times14\times14\times64}$ the global features from the shuffled video frames. The attention mechanism computes the similarity matrix S as:

$$S = QK^T$$

where $Q = W_Q F_s$, $Q = W_Q F_s$, and $V = W_V F_o$ are the query, key, and value vectors, respectively, with W_Q, W_K, and W_V being learnable weight matrices. The attention weights are obtained by applying a softmax function to the similarity matrix.

$$A = Softmax\left(\frac{S}{\sqrt{d_k}}\right)$$

The context vector C is then computed as:

$$C = AV$$

This context vector fuses the local and global information from the original and shuffled video sequences.

© JUN 2026 | IRE Journals | Volume 9 Issue 12 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I12-1718691

Intra-Video Fusion with LSTM for Temporal Dependency Modeling:
After the Inter-Video Fusion, the model employs an Intra-Video Fusion Block to capture the temporal relationships within a single video sequence. This block first passes the fused features through a fully connected layer (FCL) to reduce the dimensionality from $\mathbb{R}^{12544}$ to $\mathbb{R}^{64}$ and introduce non-linearity.

$$F_{fcl} = ReLU(W_{fcl}F_{fused} + b_{fcl})$$

The output from the FCL is then fed into an LSTM layer, modeling the temporal dependencies across video frames. The LSTM layer operates on the sequence of feature vectors $\{F_{fcl,t}\}_{t=1}^{10}$, where each vector corresponds to a different time step t. The LSTM's dynamics are governed by the following equations:

$$\begin{aligned} i_t &= \sigma(W_i[h_{t-1}, F_{fcl,t}] + b_i) \\ f_t &= \sigma(W_f[h_{t-1}, F_{fcl,t}] + b_f) \\ o_t &= \sigma(W_o[h_{t-1}, F_{fcl,t}] + b_o) \\ c_t &= f_t \odot c_{t-1} + i_t \odot tanh(W_c[h_{t-1}, F_{fcl,t}] + b_c) \\ h_t &= o_t \odot tanh(c_t) \end{aligned}$$

Where i_t, f_t, and o_t are the input, forget, and output gates, respectively; c_t and h_t represent the cell state and hidden state at time step t; and $\odot$ denotes element-wise multiplication. The LSTM's ability to retain and update the cell state over time allows it to capture long-range dependencies critical for understanding lesion progression in ultrasound videos.

Classification Layer:
The output of the LSTM encodes the temporal dynamics of the fused features and then passed to a final fully connected layer for classification. The hidden states across all time steps are aggregated into a single feature vector $F_{agg} \in \mathbb{R}^{64}$. This is then used to predict the likelihood of a breast lesion being malignant:

$$\hat{y} = \sigma(W_{fc}F_{agg} + b_{fc})$$

Here, W_{fc} and b_{fc} are the weights and biases of the fully connected layer, and σ is the sigmoid activation function.

Segmentation and Bounding Box Detection:
An additional feature of lesion detection using segmentation was integrated into the model, where bounding boxes are predicted for lesions in the ultrasound frames. However, this feature encountered challenges in generating accurate bounding boxes for lesions. It resulted multiple boxes that failed to encapsulate the lesion properly. The following section provides a detailed explanation of the model.

- UNet Model for Segmentation:
The UNet model employed in this research is based on a typical encoder-decoder architecture that is widely used for pixel-wise segmentation tasks. The model first encodes the input image into a lower-dimensional feature space and then decodes this representation back to the original image resolution to produce a segmentation mask. Mathematically, each convolutional layer in the UNet model is defined by the following operation:

$$Y = \sigma(W * X + b)$$

where X represents the input feature map with dimensions $H \times W \times C$ (height, width, and channels, respectively), W is the convolutional kernel with dimensions $k \times k \times C \times F$ (kernel size, input channels, and output channels), b is the bias term, and σ is the ReLU activation function. The convolution operation * reduces the spatial dimensions while increasing the depth of the feature maps, effectively capturing spatial features from the input image.

The downsampling in the UNet architecture is achieved through max-pooling layers:

$$Y_{i,j} = \max_{p,q} X_{2i+p,2j+q}$$

where $p, q \in \{0,1\}$ This operation halves the spatial dimensions, allowing the network to capture features at multiple scales.

© JUN 2026 | IRE Journals | Volume 9 Issue 12 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I12-1718691

During the upsampling phase, transposed convolutions (also known as deconvolutions) are employed:

$$Y = W^T * X$$

where W^T represents the transposed weight matrix. This operation restores the spatial resolution, enabling the network to generate a mask that matches the input image's dimensions.

- Bounding Box Prediction:

The bounding box prediction process begins with the segmentation mask generated by the UNet model. The mask, denoted by M, is a probability map where each pixel value indicates the likelihood of that pixel being part of a lesion. To convert this probability map into a binary mask suitable for contour detection, a thresholding operation is applied:

$$M_{binarized}(i,j) = \begin{cases} 1 & if M(i,j) > \tau \\ 0 & otherwise \end{cases}$$

Here, τ is the threshold value, and $M_{binarized}$ represents the resulting binary mask. This binary mask is then used to detect contours that represent the boundaries of potential lesions within the ultrasound image.

The contours are identified using the `find Contours` function in OpenCV, which traces the boundaries of connected regions in the binary mask. For each detected contour C, a bounding box is computed as the smallest rectangle that can fully enclose the contour. This bounding box is mathematically derived as follows:

$$B_k = (x_{min}, y_{min}, w, h)$$

where x_{min} and y_{min} represent the coordinates of the top-left corner of the bounding box, and w and h are the width and height of the bounding box, respectively. These coordinates are determined by finding the extreme points of the contour C:

$$x_{min} = \min_{(x,y)\in C} x, y_{min} = \min_{(x,y)\in C} y$$

$$w = \max_{(x,y)\in C} x - x_{min}, h = \max_{(x,y)\in C} y - y_{min}$$

This process generates a set of bounding boxes, each corresponding to a detected contour in the binary mask.

- Contour-Based Bounding Boxes:

For a given binary mask $M_{binarized}$, let the set of detected contours be denoted as $\{C_k\}_{k=1}^{K}$, where K is the total number of contours. Each contour C_k represents a potential lesion boundary, and the corresponding bounding box B_k is calculated using the extreme coordinates of the contour points:

$$:B_k = \left(\min_{(x,y)\in C_k} x, \min_{(x,y)\in C_k} y, \max_{(x,y)\in C_k} x - \min_{(x,y)\in C_k} x, \max_{(x,y)\in C_k} y - \min_{(x,y)\in C_k} y\right)$$

This formula ensures that the bounding box is the smallest rectangle that can fully contain the contour, theoretically providing an accurate localization of the lesion.

However, in practice, the application of this method to noisy ultrasound data often resulted in inaccurate bounding boxes. The detection of multiple small contours, possibly arising from noise, led to the generation of numerous bounding boxes that did not correspond to actual lesions. The irregular shapes of lesions, combined with the limitations of binary thresholding, also caused the bounding boxes to either overlap or be too large, encompassing areas beyond the lesion.

3.2.1.3 Comparative Models

2D CNN Model

The 2D Convolutional Neural Network (CNN) model serves as the baseline in this study, offering a straightforward approach to breast lesion classification in ultrasound images. The architecture consists of multiple convolutional layers, followed by pooling and fully connected layers, culminating in a binary classification output.

© JUN 2026 | IRE Journals | Volume 9 Issue 12 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I12-1718691

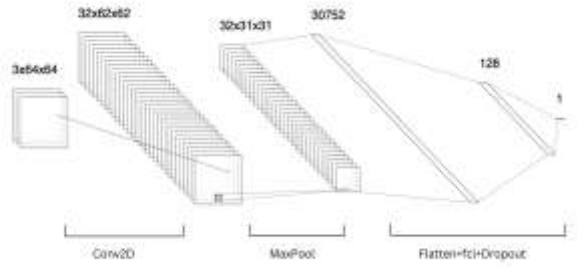

Figure 3.3 Architecture of the 2D CNN Model for Breast Lesion Classification.

The model processes input images of size 3x64x64 through a series of convolutional layers (Conv2D), followed by max-pooling layers to reduce dimensionality while preserving key features. The flattened output is then passed through fully connected layers (FCL) with dropout for regularization, ultimately leading to a single output neuron for classification.

Each convolutional layer in the CNN applies a set of filters to the input image,

$$f(X) = \sigma(W * X + b)$$

where W is the convolutional kernel, * denotes the convolution operation, and σ represents the activation function, typically ReLU.

This operation produces feature maps that capture essential spatial patterns within the image. Max-pooling, is then applied to reduce the spatial dimensions while preserving the most important features.

The output of the convolutional and pooling layers is then passed through fully connected layers, where the final classification is performed. Mathematically, the dense layer operation is represented as:

$$y = W_{fc} \cdot z + b_{fc}$$

where z is the input to the dense layer, W_{fc} are the weights, and b_{fc} is the bias term. The output y is then passed through a sigmoid activation function $\hat{y} = \sigma(y)$, which is defined as: $\sigma(y) = \frac{1}{1+e^{-y}}$.

This function outputs a probability value between 0 and 1, which is used to classify the input image as benign or malignant.

The model is trained using the binary cross-entropy loss function measuring the discrepancy between the predicted probability and the true label:

$$Loss = -\big(y\log(y) + (1-y)\log(1-y)\big).$$

Transformer-Based Model

The Transformer-based model utilises self-attention mechanisms to process ultrasound video sequences, capturing long-range dependencies within the data. The model begins by converting input images into a sequence format using token and position embeddings, where each image X_t is transformed as:

$$Z_t = W_E \cdot flatten(X_t) + PositionEmbedding(t).$$

The core component of the transformer is the multi-head attention mechanism, mathematically defined as:

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

where W_O is the weight matrix for the final linear transformation.

The output of the multi-head attention is passed through a feed-forward network (FFN), consisting of two fully connected layers with a ReLU activation function in between:

$$FFN(Z) = \sigma(ZW_1 + b_1)W_2 + b_2$$

The final output of the transformer is flattened and passed through dense layers to produce the classification output:

$$y = W_{fc} \cdot Flatten(FFN(Z)) + b_{fc}$$

© JUN 2026 | IRE Journals | Volume 9 Issue 12 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I12-1718691

As with the CNN, a sigmoid activation function is applied to the output to obtain a probability value for binary classification: $\hat{y} = \sigma(y)$. The model is trained using the binary cross-entropy loss function, identical to that of the 2D CNN model.

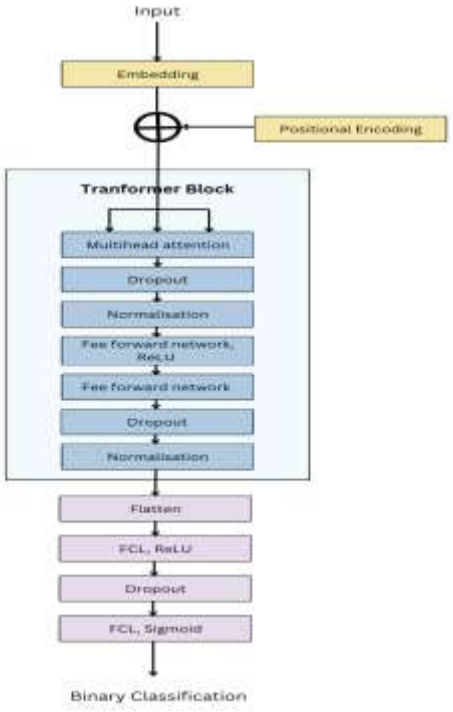

Figure 3.4 Architecture of the Transformer Model for Breast Lesion Classification.

The model begins with an embedding layer, followed by the addition of positional encoding to incorporate sequence information. The core of the model is the Transformer block, which includes multihead attention mechanisms, followed by dropout and normalization layers, and a feedforward network activated by ReLU. The output from the Transformer block is then flattened and passed through fully connected layers (FCL) with dropout, culminating in a binary classification layer.

3.2.3. Experimental Settings

The experiments were conducted on the University of Leeds ARC4 High-Performance Computing (HPC) cluster and locally on a development machine using PyCharm as the IDE. The settings for both environments are summarized in table 3.1.

Table 3.1 Experimental Settings for Model Implementation.

	Experimental Settings
CPU	Intel Xeon E5-2680v4 @ 2.40GHz (ARC4 HPC)
GPU	NVIDIA Tesla P100 (ARC4 HPC)
OS	CentOS 7 (ARC4 HPC), macOS Ventura (Local)
Framework	Tensorflow 2.6.0, Keras
Python Version	Python 3.8.12
IDE	PyCharm (Community Edition)

3.2.4. Training

The training process was designed to optimize model performance using a consistent approach across all models. The dataset was split into training, validation, and testing sets with a ratio of 60:20:20. Each model was trained for 10 epochs using the Adam optimizer, with a learning rate of 1e-4. The binary cross-entropy loss function was employed. Batch sizes were chosen according to the computational demands of each model, as detailed in the table 3.2:

Table 3.2 Model-Specific Training Parameters.

Model	Learning Rate	Batch Size	Epochs
2D CNN	0.001	32	10
Transformer	0.0001	64	10
Hybrid UNet-LSTM	0.0001	32	10

IV. RESULTS OF THE EMPIRICAL INVESTIGATION

4.1 Overview

This chapter presents the detailed results of the empirical investigation into the performance of three models—2D CNN, Transformer, and the proposed

© JUN 2026 | IRE Journals | Volume 9 Issue 12 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I12-1718691

Hybrid UNet-LSTM—on the task of breast lesion classification using ultrasound videos. The primary objective of this study was to evaluate these models across a range of performance metrics, including accuracy, precision, recall, F1-score, Mean Squared Error (MSE), Peak Signal-to-Noise Ratio (PSNR), and G-Mean, to determine their effectiveness in distinguishing between benign and malignant lesions. The experiments were conducted to optimise the classification performance of the proposed Hybrid UNet-LSTM model while comparing it with baseline models, the 2D CNN and Transformer. Although the Hybrid UNet-LSTM is the main model proposed in this research, it faces challenges such as lower accuracy and specificity compared to the simpler CNN and Transformer models. Despite this, the UNet-LSTM's ability to handle temporal information presents it as a promising approach for video-based classification tasks.

In this investigation, the 2D CNN, with its straightforward architecture, achieved the highest accuracy and precision, making it a robust baseline. The Transformer mode offered strong performance but with slightly reduced recall compared to the CNN. The Hybrid UNet-LSTM model, despite its complexity and integration of both spatial and temporal features, showed a balanced F1-score and sensitivity, highlighting its potential for more complex applications that require understanding of temporal dynamics in ultrasound videos.

The results are meticulously detailed in this chapter, with a comparison of key performance metrics provided in tables. The findings are critically analyzed to draw conclusions about the strengths and limitations of each model. Special attention is given to the shortcomings of the Hybrid UNet-LSTM model, particularly in its bounding box segmentation feature, which did not perform as expected.

It is also important to note that the dataset used in this study presents a significant level of variability, including variations in ultrasound image quality, lesion size, and patient demographics, which may have influenced the performance outcomes. This variability underscores the importance of model robustness in real-world clinical settings. The performance of the Hybrid UNet-LSTM model, while not high in all metrics, suggests that further refinement and additional training on more diverse datasets could enhance its ability to generalize across different patient populations and imaging conditions, potentially leading to improved clinical applicability.

4.2 Detailed Results

4.2.1 Model Performance Metrics

Table 4.1 Model Performance Metrics across Implemented Architectures.

Mode	Test Loss	Test Accuracy	Accuracy	Precision	Recall	F1-Score
2D CNN	0.0876	0.9377	0.9377	0.9697	0.9266	0.9477
Transform er	0.1161	0.9373	0.9373	0.9852	0.9107	0.9465
Hybrid UNet-LSTM	0.2304	0.8933	0.8933	0.9091	0.9150	0.9121

Model	MSE	PSNR	Sensitivity	Specificity	G-Mean
2D CNN	0.0317	14.9876	0.5820	0.4180	0.4932
Transformer	0.0374	14.2700	0.5630	0.4370	0.4960

Hybrid UNet-LSTM	0.0722	11.4131	0.6087	0.3913	0.4880

Accuracy:
The 2D CNN model achieves the highest accuracy at 93.77%, closely followed by the Transformer at 93.73% as describe in table 4.1. The Hybrid UNet-LSTM model, despite its more complex architecture designed to capture both spatial and temporal features, achieves a lower accuracy of 89.33%.

This lower accuracy in the Hybrid UNet-LSTM model suggests a potential overfitting to the temporal aspects of the data. The added complexity of LSTM layers designed to capture temporal sequences might not fully contribute to improved accuracy. This might be because of the static nature of the ultrasound images, where temporal changes are less pronounced than in other video-based tasks. This also indicates that while temporal data integration is theoretically advantageous, it requires careful tuning and larger datasets to achieve its full potential.

The accuracy metrics alone suggest that the simpler 2D CNN and Transformer models are more robust for this specific task, where the spatial features of the images dominate the classification performance. However, the trade-offs between these models will be better understood through further analysis of precision, recall, and other performance metrics.

Precision and Recall:
The Transformer model achieves the highest precision at 98.52%, indicating that it is particularly effective at minimizing false positives. This makes it highly valuable in clinical scenarios where the cost of false positives, such as unnecessary biopsies or treatments, must be minimized. However, its recall is slightly lower at 91.07%, suggesting it might miss some true positive cases.

The 2D CNN model demonstrates a balanced performance with precision at 96.97% and recall at 92.66%, resulting in a high F1-score of 94.77%. This balance indicates that the CNN model is robust across both reducing false positives and capturing true positives. This makes it a strong candidate for general use where both types of errors are critical.

The Hybrid UNet-LSTM model, while achieving a slightly lower precision of 90.91%, demonstrates a competitive recall at 91.50%, leading to an F1-score of 91.21%. This suggests that the model is slightly more sensitive to identifying true positive cases compared to the Transformer but does so at the cost of increased false positives. The architecture's ability to integrate temporal features might be contributing to better recall, as it could be more adept at identifying subtle temporal changes that indicate the presence of a lesion.

While the Hybrid UNet-LSTM model shows promise in identifying positive cases, the increased false positive rate (as indicated by lower specificity) could lead to unnecessary procedures. This highlights the need for further tuning, particularly in the decision thresholds, to balance precision and recall more effectively.

Mean Squared Error and Peak Signal-to-Noise Ratio:
The 2D CNN model exhibits the lowest MSE at 0.0317, reflecting that its predictions are closest to the true labels on average. The corresponding high PSNR of 14.99 indicates that the model’s predictions are "clean" and exhibit low noise, making it reliable in its classification outputs.

The Transformer model, with a slightly higher MSE of 0.0374 and a PSNR of 14.27, also demonstrates strong prediction quality, though slightly less precise than the CNN. This slight increase in MSE could be because of the complexity of handling the input data's sequential nature, which may introduce variability in the model’s output.

The Hybrid UNet-LSTM model, showing the highest MSE at 0.0722 and the lowest PSNR at 11.41, suggests that the integration of temporal features may introduce noise or lead to overfitting, particularly given the limited size of the dataset. The higher MSE indicates greater variability in predictions, which could reduce the model’s reliability in clinical practice where accurate predictions are necessary.

These results emphasize the challenges of integrating temporal information in medical image analysis, where the signal-to-noise ratio is critical. The increased noise in the Hybrid UNet-LSTM model's outputs suggests that while temporal data can provide additional context, it also requires more sophisticated handling to prevent degradation in prediction quality.

Sensitivity, Specificity, and G-Mean:
The Hybrid UNet-LSTM model achieves the highest sensitivity at 60.87%, indicating its strong ability to correctly identify true positive cases. This is particularly critical in medical applications where missing a positive case can have severe consequences. However, its specificity is lower at 39.13%, reflecting a higher rate of false positives. This imbalance could result in unnecessary follow-up procedures, which are costly and stressful for patients.

The 2D CNN and Transformer models demonstrate a more balanced performance between sensitivity and specificity, with the Transformer slightly outperforming the CNN in G-Mean (0.4960 vs. 0.4932). This balance is crucial for ensuring that the models not only detect true positives but also correctly identify true negatives, reducing the number of false alarms.

The Hybrid UNet-LSTM model, showing the highest MSE at 0.0722 and the lowest PSNR at 11.41, suggests that the integration of temporal features may introduce noise or lead to overfitting, particularly given the limited size of the dataset. The higher MSE indicates greater variability in predictions, which could reduce the model's reliability in clinical practice where consistent and confident predictions are necessary.

The lower G-Mean in the Hybrid UNet-LSTM model highlights the difficulty in achieving a balanced performance when integrating temporal features. While the model excels in sensitivity, its lower specificity suggests that further optimization, such as adjusting decision thresholds or implementing more sophisticated regularization techniques, is necessary to improve overall performance.

4.2.2 Segmentation and Bounding Box Detection
The additional feature of lesion detection using segmentation was encountered significant challenges in generating accurate bounding boxes for lesions resulting in multiple boxes that failed to encapsulate the lesion properly. This section analyzes the underlying reasons for these issues.

1. Annotation File Corruption:
The presence of corrupted annotation files likely led to poor model performance because the training process could not rely on accurate ground truth data, leading to misaligned model predictions. While manual re-annotation of the files could have been a solution to recover the dataset's usability, this approach was not feasible within the time constraints of this master's dissertation. Manually re-annotating the entire dataset would have required a significant investment of time and resources, which was beyond the scope of this research. As a result, the segmentation model had to be developed and evaluated on a compromised dataset, which limited its effectiveness and the reliability of the results.

2. Multiple and Incorrect Bounding Boxes:
The segmentation approach frequently generated multiple bounding boxes that were unable to encapsulate the lesion effectively. This issue can be due to the model's inability to distinguish between noise and the actual lesion in the ultrasound images. Ultrasound images are inherently noisy, and the model might be picking up on these noise patterns as potential lesions leading to false positives.

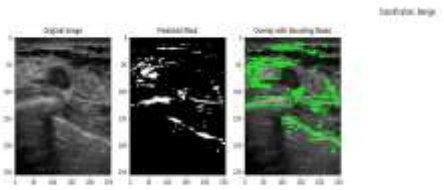

Figure 4.1 Results of the segmentation task showing the original ultrasound image, the predicted mask, and the overlay with bounding boxes.

3. Complexity of Lesion Boundaries:
Breast lesions in ultrasound images often have irregular and diffuse boundaries. This makes it challenging for traditional segmentation models to

© JUN 2026 | IRE Journals | Volume 9 Issue 12 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I12-1718691

accurately delineate the lesion. The UNet model used in this research, while effective in many segmentation tasks, may struggle with the fine granularity required to capture such complex structures accurately.

Despite the difficulties encountered, the segmentation feature represents a promising direction for future work, where more sophisticated techniques could be employed to improve its accuracy and reliability.

4.2.3 Comparative Analysis with Baseline and State-of-the-Art Models

In this section, we compare the proposed Hybrid UNet-LSTM model against a baseline model, CVA-Net (Lin et al., 2022), and a state-of-the-art model, GNet (Daimiel Naranjo et al., 2023) to benchmark the performance of our model and provide a critical evaluation of its strengths and limitations.

Baseline Model: CVA-Net (Lin et al., 2022)

CVA-Net represents a cutting-edge approach to video-based breast lesion detection, designed to exploit both temporal and spatial features effectively. Its architecture is centered around clip-level and video-level feature aggregation through an inter-video fusion mechanism. This inter-video fusion mechanism enables it to capture temporal dependencies and local-global contexts by integrating both original and shuffled frames.

Table 4.2 Performance Metrics of CVA-Net on Breast Lesion Detection in Ultrasound Videos.

This table presents the Average Precision (AP) metrics at various Intersection over Union (IoU) thresholds, including AP, AP50, and AP75, for the CVA-Net model as proposed by Lin et al. (2022)..

Mode	AP	AP50	AP75
CVA-Ne	36.1	65.1	38.5

The metrics provided, AP, AP50 and AP75, reflect CVA-Net's precision in localizing and classifying lesions across different thresholds. This model's ability to maintain high precision across these varying thresholds is indicative of its robustness and reliability.

State-of-the-Art Model: GNet (Daimiel Naranjo et al., 2023)

GNet uses transfer learning by fine-tuning pre-trained networks on large-scale datasets. GNet benefits from generalized feature representations that are crucial for achieving high performance in medical imaging tasks with relatively small datasets.

Table 4.3 Comparative Performance Metrics of GNet and Hybrid UNet-LSTM Models.

This table compares the performance of the GNet model (Daimiel Naranjo et al., 2023) and the Hybrid UNet-LSTM model.

Mode	Accuracy	Sensitivity	Specificity
GNe	88.83%	89.29%	88.57%
Hybrid UNet-LSTM	89.33%	60.87%	39.13%

The comparative analysis of the GNet and Hybrid UNet-LSTM models shows that while the Hybrid UNet-LSTM model marginally outperforms GNet in terms of overall accuracy (89.33% vs. 88.83%), its performance on sensitivity and specificity shows a significant trade-off. Specifically, the Hybrid UNet-LSTM model has a drastically lower sensitivity (60.87% compared to 89.29% in GNet) and specificity (39.13% compared to 88.57% in GNet). This indicates that while the Hybrid UNet-LSTM model is slightly better at making correct overall predictions (accuracy), it struggles significantly with correctly identifying positive cases (sensitivity) and distinguishing between positive and negative cases (specificity). The GNet model, being the state-of-the-art (SOTA) model, demonstrates a more balanced performance across all metrics, suggesting it may be more reliable in practical applications where both sensitivity and specificity are critical. Still the Hybrid UNet-LSTM has an innovative architecture, but requires further tuning.

In conclusion, while the 2D CNN and Transformer models provide robust performance for breast lesion classification, the Hybrid UNet-LSTM model offers a promising direction for future research, particularly if

the challenges of noise and overfitting can be addressed. The integration of temporal features remains a valuable avenue, but it requires careful tuning and advanced techniques to fully harness its potential in clinical diagnostics.

V. VALIDATION OF RESULTS

5.1 Introduction

This chapter delves into the validation of the empirical results obtained from the proposed Hybrid UNet-LSTM model and its comparative evaluation against the 2D CNN, Transformer, and CVANet models.

5.2 Validation Methodology

5.2.1 Data Splitting

The data splitting approach employed in this study adheres to industry-standard practices, with the dataset divided into training, validation, and testing subsets at a ratio of 60%, 20%, and 20%, respectively. Although cross-validation was not utilized in this study, the models demonstrated relatively strong performance, indicating that the chosen data split was effective for both training and evaluation purposes.

Table 5.1 Distribution of the Dataset into Training, Validation, and Testing Subsets.

Subset	Number of Videos	Total Frames
Training	113	15,147
Validation	38	5,054
Testing	37	5,071
Total	188	25,272

5.2.2 Evaluation Metrics

Accuracy: Accuracy measures the overall correctness of the model in classifying both positive and negative cases. While widely recognized, it can be misleading in imbalanced datasets like those in medical diagnostics. In this research, accuracy was tracked during training and testing phases to monitor model performance, but it was supplemented with other metrics for a more nuanced evaluation.

Precision and Recall: Precision and recall are crucial in medical contexts where false positives and false negatives carry significant consequences. For instance, the Transformer model's precision of 98.52% highlights its effectiveness in reducing false positives, ensuring accurate diagnoses.

F1 Score: The F1 score, a harmonic mean of precision and recall, provides a balanced evaluation of the model's performance, especially in managing the trade-offs between false positives and false negatives. The F1 score of 0.947 in the CNN model, for example, indicates a well-balanced performance, critical for minimizing diagnostic errors.

Sensitivity and Specificity: Sensitivity (recall) and specificity are vital in assessing a model's effectiveness in identifying true positives and true negatives. In this research, the Hybrid UNet-LSTM model showed high sensitivity (60.87%), crucial for capturing malignant cases, but lower specificity (39.13%), indicating room for reducing false positives.

G-Mean: G-Mean combines sensitivity and specificity, offering a balanced metric particularly useful in imbalanced datasets. It was used to evaluate overall model performance, with the CNN model achieving a G-Mean of approximately 0.493, indicating a reasonable balance that could be further optimized.

MSE and PSNR: MSE assesses the average squared difference between predicted probabilities and actual labels, providing insight into prediction consistency. PSNR, adapted from image quality assessment, evaluates the "noise" in the model's predictions. In this research, MSE and PSNR were used to ensure the reliability of predictions, with the CNN model achieving an MSE of 0.0317, indicating high confidence in its output.

5.2.3 Comparative Analysis

The Hybrid UNet-LSTM model was rigorously compared against baseline and state-of-the-art models (2D CNN, Transformer, and CVANet). This comparison was essential to determine the effectiveness of integrating temporal dynamics through the LSTM layers and to position the Hybrid

model within the broader research context. These aspects were thoroughly discussed in section 4.2.3.

5.3 Discussion.

Temporal Dynamics and Model Complexity:

The integration of LSTM layers within the UNet framework introduces a level of complexity that enables the model to capture temporal evolution in ultrasound videos, an aspect that simpler models, such as 2D CNNs, inherently miss. This capability is important for tasks where the progression of features over time provides significant diagnostic information. However, the increased complexity comes at a cost. The higher Mean Squared Error (MSE) and lower Peak Signal-to-Noise Ratio (PSNR) observed in the Hybrid UNet-LSTM model suggest that while the model captures more detailed temporal information, it also introduces variability and noise into the predictions. This could be due to the LSTM's sensitivity to input sequence length and quality, where slight variations in temporal sequences can lead to disproportionate changes in output. This finding implies that while the temporal aspect of the data is valuable, its integration into the model requires more sophisticated approaches to avoid introducing noise and reducing prediction confidence.

Trade-offs Between Sensitivity and Specificity:

A significant challenge identified in the validation is the model's difficulty in balancing sensitivity and specificity. The Hybrid UNet-LSTM model demonstrated high sensitivity, indicating its strong ability to detect true positives. However, this high sensitivity is counterbalanced by a notably lower specificity, leading to a higher rate of false positives. This imbalance suggests that the model, in its current form, is biased towards avoiding false negatives at the expense of increasing false positives, which could lead to unnecessary follow-up procedures and patient anxiety. This trade-off highlights the need for further refinement through techniques such as calibration and threshold optimization.

Comparative Model Performance:

When compared to baseline models like the 2D CNN and Transformer, the Hybrid UNet-LSTM model exhibits a mixed performance. The simpler CNN model, with its lower computational demands, outperformed the Hybrid model in terms of accuracy and specificity, which suggests that for tasks predominantly relying on spatial information, complexity might not always translate into better performance. The Transformer model also demonstrated high precision, indicating its effectiveness in minimizing false positives. This comparison underscores a critical point: the necessity of aligning model architecture with the specific characteristics of the task. While the Hybrid UNet-LSTM's ability to process temporal data is theoretically advantageous, in practice, its benefits might only manifest in scenarios where temporal dynamics are truly pivotal to classification. For tasks heavily dependent on spatial resolution or where temporal changes are minimal, simpler models like CNNs or even Transformations with spatial attention mechanisms might be more appropriate and efficient.

VI. CONCLUSIONS AND FUTURE WORK

6.1 Conclusions

Methodology and Findings

This research explored the integration of temporal and spatial data through the Hybrid UNet-LSTM model for breast lesion classification using ultrasound videos. The research methodology involved developing and validating a hybrid model that combines the spatial feature extraction capabilities of UNet with the temporal sequence modeling of LSTM. The model was evaluated against simpler architectures like the 2D CNN and Transformer. The findings revealed that while the Hybrid UNet-LSTM model has the potential to capture temporal dynamics, it did not consistently outperform the simpler models across all metrics. The 2D CNN achieved higher accuracy and precision, particularly in tasks primarily relying on spatial features, while the Transformer model excelled in minimizing false positives. The Hybrid model showed balanced performance in F1-score but faced challenges such as increased Mean Squared Error (MSE) and lower Peak Signal-to-Noise Ratio (PSNR), indicating the introduction of variability and noise through temporal modeling. The segmentation task aimed at detecting lesions with bounding boxes, did not perform as expected, often resulting in inaccurate detections.

Significance

The research contributes to the field of medical image analysis by demonstrating the complexities and potential of hybrid models in ultrasound video classification. The significance of this work lies in its exploration of how temporal information can be leveraged in a clinical setting, potentially leading to more accurate and context-aware diagnostic tools. The findings highlight the importance of aligning model complexity with task requirements and provide insights into the challenges of integrating temporal features in medical imaging. This work serves as a foundation for future research in developing more robust and effective models for breast lesion detection in scenarios where temporal dynamics play a critical role.

Limitations

The Hybrid UNet-LSTM model, while theoretically advantageous in capturing temporal dynamics, introduced significant noise and variability into the predictions. This suggests that the LSTM component, though powerful, requires more sophisticated handling to avoid degrading the overall model performance. Tthe segmentation task using bounding boxes did not yield the desired accuracy, indicating the difficulty of integrating detection and classification tasks within a single model framework in the context of ultrasound imaging. The model's lower specificity also points to the need for further refinement to reduce false positives. Finally, the research was conducted on a relatively limited dataset, which may have constrained the model's ability to generalize. These limitations highlight areas for improvement and set the stage for the future work outlined in the following section.

6.2 Future Work

The findings from this research open several avenues for future research and development:

Improvement of Temporal Feature Integration:

Though the integration of temporal dynamics through LSTM layers was innovative, it introduced variability that may have affected the model's overall performance. Future research could explore alternative temporal modeling techniques, such as Temporal Convolutional Networks (TCNs) (Lea et al., 2016) or Transformer-based models specifically designed for sequential data. These alternatives might offer more stable and interpretable temporal features, leading to improved accuracy. Investigating the interplay between temporal and spatial features within different model architectures could yield insights into more effective ways to capture the temporal aspects of ultrasound videos

Segmentation Task Refinement:

Given the challenges encountered in the segmentation task, future research should investigate more sophisticated segmentation approaches. Techniques like mask-based segmentation models (e.g., Mask R-CNN) or advanced bounding box prediction methods could be explored. Integrating attention mechanisms specifically designed for segmentation could help the model focus on the most relevant image regions, reducing false positives and enhancing lesion localization. Despite the difficulties faced, the segmentation feature represents a promising direction for future work, where more sophisticated techniques could significantly enhance performance.

Improvement of Dataset Quality:

Before implementing more sophisticated models, it is crucial to address the issue of corrupted annotation files. Approaches such as manual re-annotation, semi-automated tools, or collaboration with clinical experts could help reconstruct the dataset accurately. Given the issues with corrupted labels, semi-supervised learning approaches could be explored to utilize vast amounts of unlabeled ultrasound data. Techniques such as pseudo-labeling or consistency regularization could help improve model performance even when working with imperfect labels, potentially increasing the generalizability and reliability of the model.

Multi-Modal Data Integration:

The integration of data from multiple imaging modalities (e.g., combining ultrasound with MRI or mammography) could be a significant enhancement. This approach would allow the model to leverage complementary information from different sources, potentially improving classification accuracy.

Expanding the Dataset and Cross-Domain Validation:

The relatively limited dataset used in this research may have constrained the model's ability to generalize. Future work should aim to expand the dataset, either through data collection or by exploring data augmentation techniques that can create synthetic yet realistic variations of the existing data. Validating the model across different datasets and domains (e.g., different types of ultrasound machines or patient demographics) would be essential to assess its robustness and generalizability.

Conclusion

The challenges faced during the segmentation task underscore the complexities involved in ultrasound image analysis when dealing with noisy data and corrupted annotations. While the initial attempts to integrate segmentation were not fully successful, they provided valuable insights into the limitations of current approaches. Due to time constraints associated with completing a master's dissertation, implementing the above refinements within the given timeframe was not feasible. However, these suggestions represent a strong foundation for future work. By addressing these challenges through the incorporation of advanced models, improvement of dataset quality, and enhancement of post-processing techniques, future work can significantly improve the accuracy and reliability of lesion detection in ultrasound videos.

In conclusion, while this research lays the groundwork for utilizing hybrid models in breast lesion classification from ultrasound videos, significant opportunities for enhancement remain. By addressing the identified limitations and exploring the suggested future directions, the Hybrid UNet-LSTM model can be further developed into a robust tool for clinical diagnostics, potentially contributing to improved breast cancer detection and patient outcomes.

REFERENCES

[1] Lin, Z., Lin, J., Zhu, L., Fu, H., Qin, J., & Wang, L. (2022) 'A New Dataset and a Baseline Model for Breast Lesion Detection in Ultrasound Videos', in *Lecture Notes in Computer Science*. Available at: https://doi.org/10.1007/978-3-031-16437-8_59.

[2] Berg, W.A., Zhang, Z., Lehrer, D., Jong, R.A., Pisano, E.D., Barr, R.G., Böhm-Vélez, M., Mahoney, M.C., Evans, W.P. 3rd, Larsen, L.H., Morton, M.J., Mendelson, E.B., Farria, D.M., Cormack, J.B., Marques, H.S., Adams, A., Yeh, N.M., Gabrielli, G., ACRIN 6666 Investigators (2012) 'Detection of breast cancer with addition of annual screening ultrasound or a single screening MRI to mammography in women with elevated breast cancer risk', *JAMA*, 307(13), pp. 1394-1404. doi: 10.1001/jama.2012.388.

[3] Giger, M.L., Chan, H.P., & Boone, J. (2008) 'Anniversary paper: History and status of CAD and quantitative image analysis: the role of Medical Physics and AAPM', *Medical Physics*, 35(12), pp. 5799-5820. doi: 10.1118/1.3013555.

[4] Jemal, A., Bray, F., Center, M.M., Ferlay, J., Ward, E., & Forman, D. (2011) 'Global cancer statistics', *CA: A Cancer Journal for Clinicians*, 61(2), pp. 69-90. doi: 10.3322/caac.20107.

[5] Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., van der Laak, J.A.W.M., van Ginneken, B., & Sánchez, C.I. (2017) 'A survey on deep learning in medical image analysis', *Medical Image Analysis*, 42, pp. 60-88. doi: 10.1016/j.media.2017.07.005.

[6] Shen, D., Wu, G., & Suk, H.I. (2017) 'Deep Learning in Medical Image Analysis', *Annual Review of Biomedical Engineering*, 19, pp. 221-248. doi: 10.1146/annurev-bioeng-071516-044442.

[7] Ronneberger, O., Fischer, P., & Brox, T. (2015) 'U-Net: Convolutional Networks for Biomedical Image Segmentation', in *Lecture Notes in Computer Science*. Available at: https://doi.org/10.1007/978-3-319-24574-4_28.

[8] Topol, E.J. (2019) 'High-performance medicine: the convergence of human and artificial intelligence', *Nature Medicine*,

25(1), pp. 44-56. doi: 10.1038/s41591-018-0300-7.

[9] European Parliament and Council (2016) *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)*. Official Journal of the European Union, L119, pp. 1-88.

[10] Sree, S.V., Ng, E.Y., Acharya, R.U., & Faust, O. (2011) 'Breast imaging: A survey', *World Journal of Clinical Oncology*, 2(4), pp. 171-178. doi: 10.5306/wjco.v2.i4.171.

[11] Zhou, S.K., Greenspan, H., Davatzikos, C., Duncan, J.S., van Ginneken, B., Madabhushi, A., Prince, J.L., Rueckert, D., & Summers, R.M. (2021) 'A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises', *Proceedings of the IEEE*, 109(5), pp. 820-838. doi: 10.1109/JPROC.2021.3054390.

[12] Acharya, U.R., Ng, E.Y., Tan, J.H., & Sree, S.V. (2012) 'Thermography based breast cancer detection using texture features and Support Vector Machine', *Journal of Medical Systems*, 36(3), pp. 1503-1510. doi: 10.1007/s10916-010-9611-z.

[13] Horsch, K., Giger, M.L., Venta, L.A., & Vyborny, C.J. (2002) 'Computerized diagnosis of breast lesions on ultrasound', *Medical Physics*, 29(2), pp. 157-164. doi: 10.1118/1.1429239.

[14] Simonyan, K., & Zisserman, A. (2014) 'Very Deep Convolutional Networks for Large-Scale Image Recognition', *arXiv preprint*. Available at: https://arxiv.org/abs/1409.1556.

[15] He, K., Zhang, X., Ren, S., & Sun, J. (2016) 'Deep Residual Learning for Image Recognition', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770-778. doi: 10.1109/CVPR.2016.90.

[16] Huang, G., Liu, Z., van der Maaten, L., & Weinberger, K. (2017) 'Densely Connected Convolutional Networks', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. doi: 10.1109/CVPR.2017.243.

[17] Shin, H.C., Roth, H.R., Gao, M., Lu, L., Xu, Z., Nogues, I., Yao, J., Mollura, D., & Summers, R.M. (2016) 'Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning', *IEEE Transactions on Medical Imaging*, 35(5), pp. 1285-1298. doi: 10.1109/TMI.2016.2528162.

[18] Tajbakhsh, N., Shin, J.Y., Gurudu, S.R., Hurst, R.T., Kendall, C.B., Gotway, M.B., & Liang, J. (2016) 'Convolutional Neural Networks for Medical Image Analysis: Full Training or Fine Tuning?', *IEEE Transactions on Medical Imaging*, 35(5), pp. 1299-1312. doi: 10.1109/TMI.2016.2535302.

[19] Jetley, S., Lord, N., Lee, N., & Torr, P. (2018) 'Learn To Pay Attention', *arXiv preprint*. Available at: https://arxiv.org/abs/1804.02391.

[20] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., & Polosukhin, I. (2017) 'Attention Is All You Need', *arXiv preprint*. Available at: https://arxiv.org/abs/1706.03762.

[21] Hochreiter, S., & Schmidhuber, J. (1997) 'Long Short-term Memory', *Neural Computation*, 9(8), pp. 1735-1780. doi: 10.1162/neco.1997.9.8.1735.

[22] Zhao, T., Desjardins, A.E., Ourselin, S., Vercauteren, T., and Xia, W. (2019) 'Minimally invasive photoacoustic imaging: Current status and future perspectives', Photoacoustics, 16, p. 100146. doi: 10.1016/j.pacs.2019.100146.

[23] Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., and Adam, H. (2017) 'MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications'. Available at: arXiv:1704.04861.

[24] Lea, C., Vidal, R., Reiter, A., and Hager, G.D. (2016) 'Temporal Convolutional Networks: A Unified Approach to Action Segmentation'. In: Hua, G., Jégou, H. (eds) Computer Vision – ECCV 2016 Workshops. Lecture Notes in Computer Science, vol 9915. Springer, Cham. doi: 10.1007/978-3-319-49409-8_7.