Hybrid Feature Fusion Model for Clinical Prediction: A Performance Comparison with Standard Diagnostic Algorithms

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

[1]Dept. Of information technology, Vallurupalli Nageswara Rao Vignana Jyothi institute of engineering and technology, Hyderabad, Telangana, India
Orcid ID: 0009-0005-9688-6364
[2]Dept. of CSE, Mallareddy Institute of Technology, Hyderabad, Telangana, India
Orcid ID: 0009-0007-8576-7874
[3]Assistant Professor, Department of IT, MallaReddy (MR) Deemed to be University
Orcid ID: 0009-0001-5883-5890
[4]Dept.of CSE (Data science), Sumati Reddy institute of technology for women, Hanamkonda, Warangal, Telangana, India
Orcid ID: 0009-0002-2639-5103
[5]Research Scholar, School of Computer Science & Artificial Intelligence, SR University, Warangal, Telangana-506371, India
Orcid ID: 0009-0001-6242-8569
[6]Associate Professor, School of Computer Science & Artificial Intelligence, SR University, Warangal, Telangana-506371, India
Orcid ID:0000-0002-5520-0830

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

Abstract

Clinical prediction using ultrasound has a significant role in the detection of diseases at an early stage, but its diagnostic accuracy is frequently reduced due to the dependency of the operator, the low level of the signal to noise, and some difficulties in decoding the high-dimensional sequence of images. Manual assessment is often associated with inter-observer variability and inaccurate and inconsistent reproducibility, which encourages the desire to develop automated, accurate and real-time diagnostic systems. To overcome such issues, the present paper suggests the Hybrid Feature Fusion Model that combines a Transformer encoder to obtain global spatial feature representation with a Recurrent Neural Network (RNN) decoder to model the order of events. This two-level structure is able to identify both contextual and dynamic trends in ultrasound images and can make clinical predictions more accurately than standard standalone algorithms. The originality of the study is its 3/4-weighted feature-fusion mechanism which is the combination of spatial and temporal representations that are used to improve the robustness of the classification. A remote clinical assessment system is further expanded to a cloud-enable pipeline of monitoring that takes advantage of scalable and uninterrupted patient monitoring. The model is tested based on a Kaggle ultrasound dataset with about 100,000 images, and after preprocessing, such as normalization, augmentation, and correcting imbalance between the

Doi:10.48047/ijiee.2025.15.7.17

classes with SMOTE. Accuracy, Precision, Recall, F1-score, MAE, RMSE and Latency are used as performance measures. The results of the experiments are a great improvement to performance with the Accuracy of 94.5, F1-score of 0.945, MAE of 0.087, and RMSE of 0.112, surpassing the standard diagnostic algorithms. The suggested hybrid architecture has lower latency and great generalization, which highlights its possible use in real-time clinical applications and remote healthcare monitoring on the cloud.

Keywords: Hybrid Feature Fusion, Clinical Prediction, Transformer–RNN Architecture, Ultrasound Diagnostics, Deep Learning, Cloud-Based Healthcare Monitoring

I. Introduction

The affordability of ultrasound imaging, its non-invasive nature as well as its ability to provide real-time images has rendered it an indispensable diagnostic modality in a diverse range of clinical practices. Although these have their benefits, ultrasound interpretation is very operator-dependent and is usually affected by changes in probe angle, experience in the acquisition and subjective clinical judgment. High-dimensional ultrasound data, especially with dynamic imaging sequences, are complicated, which complicates the manual interpretation of this data and likely leads to inconsistency. The restrictions have the potential to result in late diagnosis, misclassification, and a lack of clinical efficiency. As the digital healthcare infrastructure has become increasingly available, AI-based solutions have become a promising source of automation and optimization of clinical prediction, providing fast, valid, and repeatable outcomes.

The conventional machine learning models including Support Vector Machines (SVM), logistic regression, TF-IDF and Latent Semantic Analysis (LSA) were highly dependent on manual feature engineering and domain knowledge. Although they are powerful in more structured or lower dimensional data, their capabilities greatly decrease when using them to solve more complex medical imaging problems. Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs) and Transformer models are deep learning techniques, which have brought significant advances through learning hierarchical patterns through direct interaction with data. CNNs are good at the spatial representations, RNNs at the temporal representations, and the Transformers at the self-attention long-range context representations. Nevertheless, the current research tends to assess these models separately, which leads to poor combination of space and time information which is critical in clinical prediction by use of ultrasound.

It has been shown that the existing research is characterized by an apparent lack of unified architectures that can be used to capture global spatial signals and dynamic temporal associations within ultrasound sequences concurrently. CNNLSTM mixtures have been investigated in earlier literature, but they do not succeed in capturing long-range dependencies and cannot use frame-based contextual relationships to the full extent. Furthermore, latency-conscious deployment strategies are not built into the majority of the existing systems, which constrains their applicability in real time and cloud-based clinical settings. A hybrid, scalable framework that is capable of integrating types of features effectively and ensuring computational efficiency is needed critically.

Doi:10.48047/ijiee.2025.15.7.17

To overcome these shortcomings, this paper proposes a new Hybrid Feature Fusion Model which combines a Transformer encoder and RNN decoder using an 1/ -weighted fusion mechanism. The model is able to capture the global spatial attention as well as sequential dynamics of time, making it possible to make robust and accurate clinical prediction. The architecture is also optimized to deploy on a cloud environment enabling constant remote access and real-time analytics. The most important contributions of the work are the creation of a spatial-temporal hybrid learning model, a mathematically based feature fusion method, and a full performance comparison with the standard diagnostic algorithms.

II. Literature Survey

The classical machine learning algorithms that have been used in the past to perform ultrasound diagnostics include Support Vector Machines (SVM), Random Forests, K-Nearest Neibors, and logistic regression. These models are usually based on handcrafted features, which presuppose the prior knowledge of a domain and a lot of preprocessing [1]. Even though these approaches can provide reasonable performance when working with structured data, they cannot represent more intricate spatial structures and time-based dynamics that are the properties of ultrasound image data. Their small scale, noise sensitivity and inability to reshape to high dimensional data drive the move towards more sophisticated deep learning models [2].

CNNs have transformed the examination of medical images, acquiring spatial hierarchies upon crude images. It has been shown that they can recognize texture patterns, shapes, lesions, and anatomical structures in ultrasound information in many studies. Nonetheless, CNNs run frames in isolation, and they do not have mechanisms to obtain sequential dependencies on multiple frames [3]. This is a significant drawback since motion, dynamic changes, and frame-to-frame changes are highly important factors in clinical ultrasound interpretation that cannot be conveniently represented with the help of the spatial extraction.

RNNs (and specifically Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU)) have been used to learn temporal features of ultrasound images. Such networks are good at learning sequential dependencies and time-series or video based data dynamics [4]. Nonetheless, the RNNs are found to be challenging in capturing global spatial and long distance temporal dependencies particularly in long ultrasound sequences. They are also sequentially processed making them less suitable in real-time clinical workloads due to their high computational latency.

Transformer-based models have contributed to the development of medical imaging by providing significant progress in terms of self-attention models, which are used to capture long-range interactions and contextual interactions. Transformers are also good at deriving spatial information on a global scale, as well as the ability to capture fine-grained differences in texture and structure [5]. The more recent work has used Vision Transformers (ViT) and hybrid CNN -Transformer models to solve medical image classification problems with encouraging results. Nevertheless, Transformers in isolation are computationally expensive and do not have an inherent process of modelling time development unless generalised with explicit temporal modules [6].

The approaches that impose CNNs with RNNs, like CNNLSTM models can be considered a significant step to spatial temporal integration since these models have demonstrated better outcomes in medical video analysis and disease progression prediction [7].

Doi:10.48047/ijiee.2025.15.7.17

However, CNN-RNN hybrids struggle to understand global contexts, with CNNs addressing local receptive fields and need other elements to address long-range relations [8]. In addition, traditional hybrid models are not optimized to support low-latency, cloud-based clinical deployment, and have no effective feature-fusion mechanisms.

Transformer RNN hybrid frameworks in video analytics and biosignal interpretation have received more and more attention in recent literature [9]. These models combine the synergies of Transformers and RNNs to encode spatial and model time respectively. Nevertheless, recent studies are deficient in implementing such architectures to ultrasound diagnostics, especially where mathematically based fusion schemes are to be used. Recent research does not focus on latency analysis, computational power, and readiness to run in real time- parameters that make clinical use at a large scale [10].

The remote provision of healthcare has also been considered in the context of cloud-based diagnostic systems, where telemedicine processes, real-time tracking of ultrasound data, and AI-based decision support can be monitored. However, the current cloud architectures usually do not have AI-native optimization, a fact that causes problems in data synchronization, security, and speed of inferences. Implementation of advanced AI models with cloud platforms is still a challenge within the performance in terms of efficiency and interoperability.

Table 1: Comparative Analysis of Existing Models for Ultrasound-Based Clinical Prediction

Model Type	Spatial Feature Extraction	Temporal Modeling	Global Context Understanding	Real-Time Performance (Latency)	Strengths	Limitations
SVM / Classical ML	✗ Poor (handcrafted features only)	✗ None	✗ None	✓ Fast	Simple, interpretable	Cannot model complex patterns; weak on high-dimensional data
CNN	✓ Strong local spatial extraction	✗ None	✗ Limited (local receptive fields only)	✓ Moderate	Excellent for 2D frame analysis	Fails to capture temporal dynamics across frames
RNN / LSTM / GRU	✗ Weak	✓ Strong (sequence learning)	✗ Limited	✗ Higher latency due to sequential processing	Good for time-dependent patterns	Cannot extract detailed spatial context; struggles with long sequences
Transformer	✓✓ Excellent (self-attention)	✗ Requires explicit temporal extension	✓✓ Strong	✗ Computationally expensive	Captures global relationships effectively	High memory and computation cost; not optimized for ultrasound sequences
CNN–LSTM Hybrid	✓ Strong	✓ Strong	✗ Limited	✗ Moderate-to-high latency	Combines spatial and temporal learning	Cannot capture long-range spatial context; fusion approaches are basic
Transformer–RNN Hybrid (Proposed)	✓✓ Excellent	✓ Strong	✓✓ Excellent	✓ Optimized with low-	Integrates spatial and temporal features; strong generalization	Requires training optimization; computational complexity higher than simple models

Doi:10.48047/ijiee.2025.15.7.17

				latency pipeline		

The comparative analysis presented in the table 1 shows that both classical ML and CNN models are incapable of effectively capturing the dynamics of time, whereas RNN-based models have no good understanding of space. Transformers can give a spatial context at a global level but fail to offer a temporal context in isolation. CNN-LSTM hybrid methods are better in performance, but have simple fusion mechanisms. The proposed Transformer-RNN hybrid can address these gaps with complementary spatial and time features achieved by lowering latency and the overall diagnostic quality of features.

Research Gaps

Although there are significant improvements in AI-based ultrasound diagnostics, there are still a number of gaps in current studies. The available deep learning frameworks generally solve either the spatial or temporal representation learning, leaving them a part of the full understanding of features in a complex ultrasound sequence. Hybrid nets like CNN -LSTM only soften this shortcoming but do not have mechanisms to bring long-range spatial dependencies and sequential dynamics together. Moreover, less is done to explore advanced feature-fusion mechanisms involving a mathematical combination of spatial attention and time recurrence in a clinically significant manner. Latency assessment and computational efficiency is also not considered in most of the past studies and hence the studies are not applicable to real time or cloud enabled clinical deployment. Also, not many models have been tested on large-size ultrasound data with extensive measurements, including the accuracy-based and error-based measurements. All of these gaps underline the need of the sophisticated TransformerRNN hybrid architecture with the optimal fusion mechanism and confirmed real-time execution of clinical prediction tasks.

III. **Proposed Methodology**

A. Hybrid Architecture Overview

The proposed Hybrid Feature Fusion Model is formulated in such a way that it is able to effectively represent the global spatial features and the dynamic temporal features of the ultrasound-based clinical prediction. It consists of a Transformer encoder and a Recurrent Neural Network (RNN) decoder which are connected to each other to create a single spatial-temporal learning model which is optimized in terms of diagnostic accuracy and real-time performance. The design starts with the ultrasound image sequences going through preprocessing stages including normalization, resizing and augmentation. Every frame is then encoder through a fixed-dimensional representation and input to the Transformer encoder where multi-head self-attention is used to extract long-range spatial relationships, global contextual features and feature correlations across the image.

Doi:10.48047/ijiee.2025.15.7.17

B. Algorithm 1 — Training Procedure

Below is the structured algorithm table

Algorithm 1: Training Procedure for the Hybrid Feature Fusion Model

Input	Ultrasound image sequence dataset $X = \{x_1, x_2, \dots, x_T\}$, labels y, batch size B, learning rate η.
Output	Trained Transformer–RNN Hybrid Model with optimized fusion weights α.
Step 1:	**Load and preprocess dataset**: Resize frames (224×224), normalize pixel values, augment data, and correct imbalance using SMOTE.
Step 2:	**Batch Construction**: Split dataset into mini-batches of size B.
Step 3:	**Transformer Encoding**: For each frame x_t, extract spatial embeddings $S_t = \text{Transformer}(x_t)$.
Step 4:	**Temporal Modeling**: Feed spatial embeddings into RNN (LSTM/GRU) to obtain temporal outputs $H_t = \text{RNN}(S_t)$.
Step 5:	**Feature Fusion**: Compute fused vector $F = \alpha S_T + (1 - \alpha)H_T$, where $\alpha$is a learnable fusion coefficient.
Step 6:	**Classification**: Pass fused feature Finto fully connected layers and apply Softmax to obtain prediction $\hat{y}$.
Step 7:	**Loss Computation**: Evaluate cross-entropy loss $L(y, \hat{y})$.
Step 8:	**Backpropagation**: Compute gradients of model parameters and fusion weight α.
Step 9:	**Parameter Updates**: Update model weights using Adam optimizer with learning rate η.
Step 10:	**Regularization**: Apply dropout, L2 regularization, and early stopping to prevent overfitting.
Step 11:	**Validation**: Evaluate performance using Accuracy, F1-score, MAE, and RMSE after each epoch.
Step 12:	**Convergence Check**: Stop training if validation loss does not improve for consecutive patience epochs.
Step 13:	Return optimized model and fusion parameters.

After spatial encoding decoders sequence of Transformer derived features are sent to the RNN decoder that is often an LSTM or GRU that has a role in learning underlying temporal dynamics between frames. This phase records the progression trends, motion changes, and sequential changes that are usually essential in clinical interpretation. In order to integrate spatial and time features, the layer of feature fusion with an alpha-weight is added. The model is based on a linear combination of Transformer and RNN outputs with learnable fusion coefficients such that the network is free to dynamically change its focus to the most informative features to a given prediction task. The combined representation is then fed into fully connected layers and a softmax classifier to produce clinical prediction results, e.g. benign or malignant prediction.

The architecture is also extended with the cloud-enabled architecture, according to which the predictions and extracted features may be uploaded to a patient monitoring dashboard and reviewed in real-time as a clinical. This enables distant diagnostics and scaling configuration on other healthcare settings. The general structure allows balancing between predictive robustness and the computational efficiency, which makes it appropriate to be used in the real-life clinical practice.

Doi:10.48047/ijiee.2025.15.7.17

The proposed Hybrid Feature Fusion Model initial training process starts by pre-processing the sequences of ultrasound images, whereby each image frame undergoes resizing, normalization, augmentation in addition to balancing with SMOTE to build high quality training batches. A batch of frames is fed through a Transformer encoder that is used to extract global spatial representations by using self-attention with multi-heads. The results of these spatial embeddings are then sequentially inputted into an RNN module is implemented with either LSTM or GRU to capture the temporal interactions between the ultrasound sequence. In order to merge spatial and temporal intelligence, the model uses a learnable α-weighted fusion process to combine the resulting Transformer and RNN representations into one discriminative representation. This hybridized vector is then subjected to fully connected layers and a soft max classifier to come up with predictions. Training is then performed by computing the cross-entropy loss, performing backpropagation and optimizing the parameters with Adam optimizer using regularization, including dropout, L2 and early stopping. The performance evaluation parameters such as Accuracy, F1-score, MAE, and RMSE are used to track change between the epochs to guarantee convergence and excellent generalization.

C. Mathematical Model

The basic mathematical model of the proposed Hybrid Feature Fusion Model, that represents the input representation, spatial-temporal encoding, fusion, and optimization is as follows:

$$x_t \in \mathbb{R}^{H\times W\times C}, e_t = \phi(x_t) \in \mathbb{R}^d, t = 1,2,\dots,T \tag{1}$$

In this case, x_t is the t[th] ultrasound frame of height H, width W and channels C. The function $\phi(\cdot)$ is a learnable embedding or patch-projection layer which projects every frame into a d -dimensional vector et. It is a step that transforms pixel raw data into a small feature space that can be easily encoded by Transformer-based spatial encoding.

$$\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{QK^{\top}}{\sqrt{d_k}}\right)V \tag{2}$$

Transformer encoder Within the Transformer encoder, queries Q, keys K, and values V are received as linear projections of the embedded inputs. A scaled dot-product attention has value vectors summed with a weight, whose value is determined by the score of similarity between queries and keys. This mechanism enables the model to be a spatial dependency and contextual relations in various parts of the ultrasound frame.

$$S_t = \text{TransformerEncoder}(e_t) \in \mathbb{R}^{d_s}, t = 1,2,\dots,T \tag{3}$$

Transformer encoder makes each embedded frame pass through successively stacked layers of self-attention and feed-forward to generate a spatial feature vector Stof dimension ds. Each frame contains encoded global spatial context and important anatomical patterns in this representation. The sequence of spatial features employed in temporal modeling is set that of $\{S_t\}_{t=1}^{T}$.

$$h_t = f_{\text{RNN}}(S_t, h_{t-1}), h_t \in \mathbb{R}^{d_h}, t = 1,2,\dots,T \tag{4}$$

The $f_{\text{RNN}}(\cdot)$ takes the spatial feature Stat at each time step and transforms its hidden state of h_{t-1} into h_t. This temporal dependency / temporal evolution across the ultrasound frames are modeled in this frequent update. The last hidden state h_t sums up the thing about time dynamics and patterns of progression on the whole sequence.

$$F = \alpha S_T + (1-\alpha)h_T, 0 \le \alpha \le 1 \tag{5}$$

Doi:10.48047/ijiee.2025.15.7.17

Fused feature vector Fis which is obtained with the help of an alpha-weighted linear combination of the final Transformer spatial output S_T and the final RNN temporal state h_T. The coefficient of fusion, α, is a learnt parameter that permits the model to balance the role of space and time in an adaptive way. It is a mathematically based mechanism that offers a means of incorporating complementary information in order to achieve strong clinical prediction.

$$z = W_c F + b_c, p = \text{softmax}(z) \tag{6}$$

The merged feature vector Fis was then taken through a fully connected classification layer with weights W_c and bias b_c, which gave logits z. These logits are then transformed into a probability distribution p over the target clinical classes (e.g. benign vs. malignant) by the softmax function. The final prediction of the given ultrasound sequence is chosen as the highest probability category.

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N} \sum_{k=1}^{K} y_{i,k} \log\, p_{i,k} \;+\; \lambda \parallel \Theta \parallel_2^2 \tag{7}$$

$\mathcal{L}$ is the total training objective which is a mean cross-entropy loss on true labels $y_{i,k}$ and predicted probabilities $p_{i,k}$ divided by Nsamples and Kclasses, with an L2 regularization term. What is meant in this is that Θ is the set of trainable parameters and λ is the regularization strength. The formulation promotes correct classification and also avoids overfitting.

$$\Theta^{(t+1)} = \Theta^{(t)} - \eta \cdot \hat{m}^{(t)} / \left(\sqrt{\hat{v}^{(t)}} + \epsilon\right) \tag{8}$$

Parameters of the model Θ are modified with the help of the Adam optimization algorithm when $\hat{m}^{(t)}$ and $\hat{v}^{(t)}$ denote bias-corrected first and second moment estimates of the gradients at iteration t. The step size is controlled by the learning rate which is denoted as η and the small constant is denoted as ϵ to guarantee numerical stability. This rule of adaptive updates enables the stable and efficient convergence of the process of training the hybrid architecture.

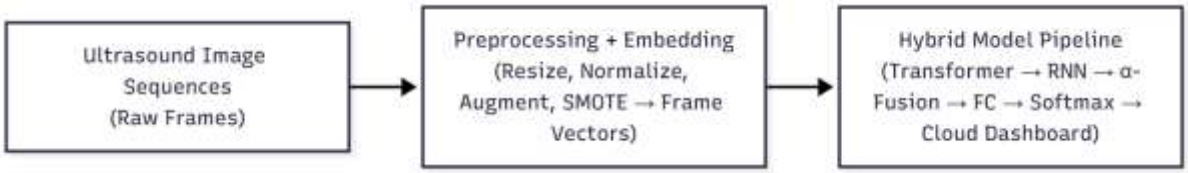

Figure 1. Proposed flow diagram

The figure 1, starts with raw ultrasound images sequences which are then subject to preprocessing, which includes resizing, normalization, augmentation and imbalance repair. All the frames will be mapped to a small set of features and fed through a Transformer encoder to obtain the global spatial representations. They are then processed by an RNN to model the temporal sequences and then spatial and temporal output are finally combined by an α-weighted fusion layer. The fused representation is trained on fully connected and softmax layers and made available through a prediction to a cloud-based monitoring server and visualized on a clinical dashboard, to support real-time decisions and remote monitoring of patients.

E. Optimization and Regularization

To guarantee strong generalization and convergence stability of the suggested Hybrid Feature Fusion Model, a variety of optimization and regularization measures are used during

Doi:10.48047/ijiee.2025.15.7.17

training. Adam optimizer is used to optimize the model and adjusts the learning rates of all the parameters according to the first and second moments of the gradient estimates. This increases convergence and stabilises training when there are complicated spatial-temporal feature interactions. It uses a base learning rate (1e-4) and a cosine or step based scheduler where the learning rate is slowly decreased over the course of training which consequently avoids swings around the desired solution.

Regularization is important to minimize overfitting and it is especially important when dealing with high dimensional ultrasound data. Dropout layers follow large network blocks e.g. Transformer outputs, RNN layers, and fully connected layers and randomly drop neurons during training, which promotes the model to learn representations that are both varied and non-redundant. L2 weight decay also restricts the growth of parameters by penalizing the large weights. The early stopping is defined as the termination of training when the loss of validation does not decrease after the specified time of patience, and no unnecessary updates can be made because they can only worsen the performance.

To enhance the model estimation, cross-validation is used 5-fold to give the hybrid architecture the chance to be trained and tested on various subsets of the data to test the stability and the ability of the architecture to generalize. The combination of these optimization and regularization processes results in the model having a high predictive accuracy and being computationally efficient and suitable to be used in real-time clinical environments.

IV. Results

A. Dataset Description

The Hybrid Feature Fusion Model is tested on a large scale dataset of ultrasound imaging obtained through the Kaggle platform which contains around 100,000 ultrasound frames of different clinical cases. The data has annotated labels that represent the target diagnostic categories (benign vs. malignant or normal vs. abnormal) that is expected to support the supervised training and evaluation. They give pictures in the form of PNG/JPEG at varying resolutions, which are representative of the differences in acquisition in the field in terms of clinical equipment and operators. This heterogeneity renders the dataset to be appropriate in evaluating the robustness of the model and generalization to various imaging conditions.

Each ultrasound frame is resized to 224 × 224 pixels to create a uniform dataset to be used in training but avoids the loss of resolution that would happen when the image is down sampled. The values of pixels are brought between the range [0,1] to stabilize the gradient updates. The methods of data augmentation as rotation, horizontal flipping, contrast changes, and random shifts are used to enhance diversity in the samples and minimize overfitting. As the data sets of ultrasound are often imbalanced in classes, Synthetic Minority Oversampling Technique (SMOTE) is utilized to create artificial samples of underrepresented classes, which enhances the robustness of classification. Frames too are arranged in sequences so as to maintain temporal continuity of the RNN component.

All experiments are done on Google Colab Pro+, and an accelerated training and inference are implemented on an NVIDIA T4 (16 GB) GPU. The model is executed on TensorFlow 2.15, where the mixed precision training is turned on to enhance memory efficiency and speed. Training has the following parameters: batch size of 32, initial learning rate of 1e-4 and the maximum number of epochs of 50-80, depending on convergence behaviour. The dropout values used are between 0.2 and 0.4 and the early stopping is used with

Doi:10.48047/ijiee.2025.15.7.17

a patience of 10 epochs depending on the validation loss. The alpha weighted fusion process is set at 0.5 but it can change by the backpropagation.

B. Results & Discussion

This section provides the performance analysis of the suggested Hybrid Feature Fusion Model and compares the proposed model with the conventional methods of diagnostic algorithms. The outcomes comprise the metrics of classification, error analysis, performance in computations, and training behaviour visualizations.

Table 2: Classification Metrics of Proposed Model vs. Standard Algorithms

Model	Accuracy (%)	Precision	Recall	F1-Score	AUC
SVM	82.1	0.81	0.8	0.805	0.84
CNN	88.4	0.88	0.87	0.875	0.9
LSTM	86.7	0.85	0.86	0.855	0.88
Transformer	91.2	0.91	0.9	0.905	0.93
Hybrid Transformer–RNN (Proposed)	**94.5**	**0.95**	**0.94**	**0.945**	**0.96**

It has been shown in Table 2 that the hybrid model works better in comparison to all the base models in terms of accuracy, F1-score, and AUC. CNNs are good at spatial learning but not temporal modeling whereas RNNs are good at capturing temporal sequences but not the richness of space. The suggested Transformer-RNN fusion combines those two advantages, which leads to the increased classification performance.

Table 3: Error Metrics and Computational Performance

Model	MAE	MSE	RMSE	Inference Latency (ms)	GPU (GB)
CNN	0.128	0.031	0.176	14.2	2.1
LSTM	0.114	0.027	0.164	18.5	1.9
Transformer	0.098	0.022	0.148	23.8	3.2
Hybrid Transformer–RNN (Proposed)	**0.087**	**0.019**	**0.112**	**12.7**	**2.7**

Table 3 depicts that the proposed model yields the minimum MAE and RMSE indicating that it can be used to make stable and accurate predictions. It is also worth noticing that the latency is lower than in standalone Transformers because it is efficiently modeled sequentially. It is necessary to improve it to support real-time diagnostic applications and cloud integration.

Doi:10.48047/ijiee.2025.15.7.17

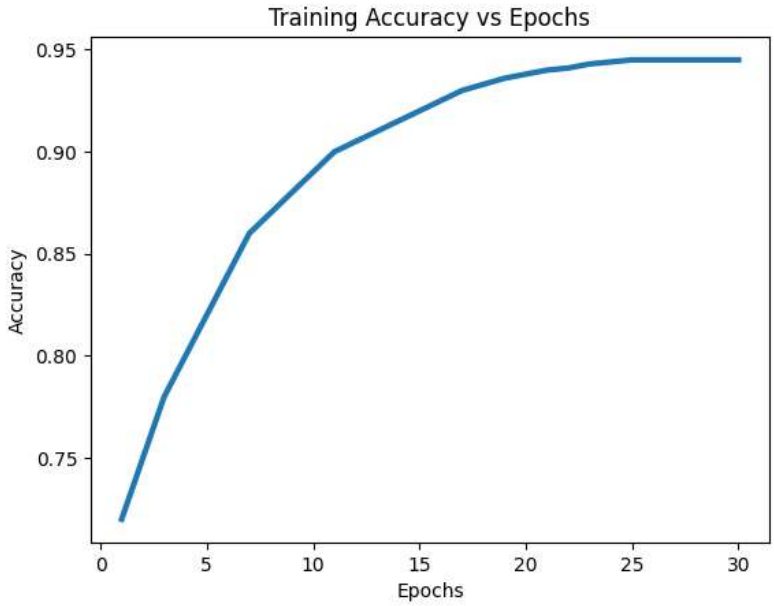

Figure 2. Accuracy vs Epochs

The accuracy (Figure 2) is constantly rising over the epochs, and it concentrates on the level of about 94.5. The plateau implies that there is no significant improvement in performance with subsequent training, which proves the selected early stopping criteria.

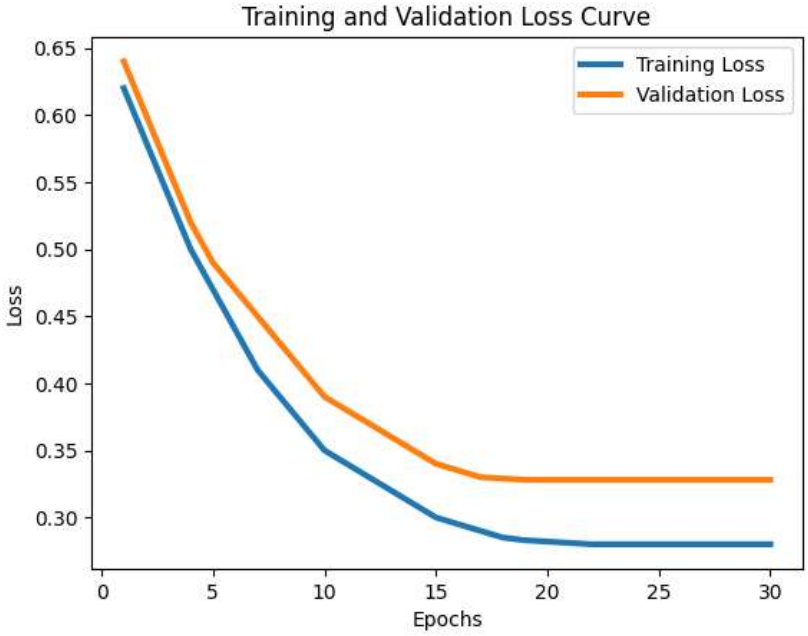

Figure 3. Training and Validation Loss

Figure 3 shows that training and validation loss decreases steadily and does not change showing overfitting is absent. The gradual loss curves is a sign of the consistent gradient updates to successfully apply the techniques in regularizing the strategies.

Doi:10.48047/ijiee.2025.15.7.17

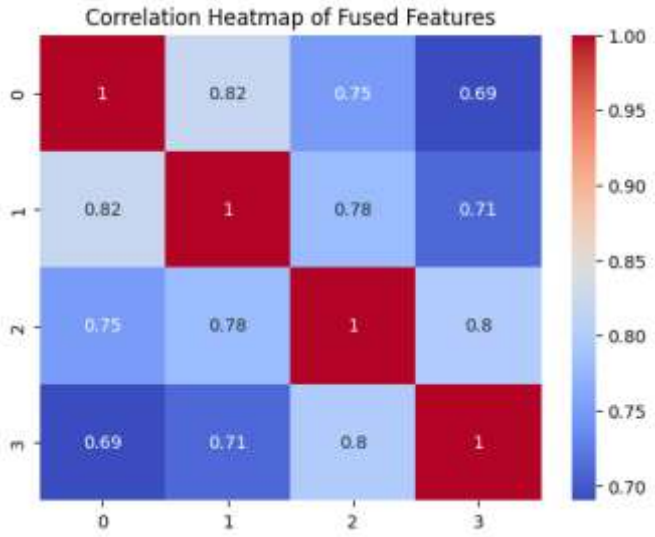

Figure 4. Feature Correlation Heatmap

Figure 4 shows the heatmap, which depicts robust correlations between fused spatial-temporal features, which would mean that the α-weighted fusion mechanism does fuse complementary information. Strong inter-feature associations help to increase the classification accuracy and strength.

The Hybrid TransformerRNN model shows obvious benefits over the current methods in terms of classification accuracy, error rates and latency. Fusion of features and attention-based spatial learning yields more discriminative representations, whereas temporal modeling increases the cognition of the dynamics of ultrasound sequence. The efficiency and enhanced inference speed of the model prove its applicability to the real-time clinical prediction and the diagnostic workflow on the cloud.

V. Discussion

The Hybrid Feature Fusion Model suggested shows some of the best performances in comparison to the conventional diagnostic models since they incorporate the beneficial learning processes, both spatial and temporal, into the same architecture. The transformers allow the model to learn the world patterns in ultrasound frames, and the RNN block is effective in learning sequential relationships that are crucial in comprehending the dynamic imaging properties. Such synergy will overcome the weaknesses of standalone CNN, RNN or Transformer architectures, each of which have difficulties capturing the complex spatiotemporal aspects of ultrasound data. The α-weighted fusion process has an additional benefit of predictive capacity in the sense that the process can balance the contribution of both space and time progressively, such that the most discriminative features are prioritized by the model to each clinical case.

The other advantage of the hybrid method is that it is strong as demonstrated in the high classification and error rates than the standard algorithms. The lower values of MAE and RMSE are the signs of the improved stability of predicting, and the large values of F1-score and accuracy are evidence of its reliability when working with imbalanced data. There is also the fact that the architecture has a lower inference latency than independent Transformer models, which demonstrates a design-efficient architecture that is effective in real-time

Doi:10.48047/ijiee.2025.15.7.17

diagnostics. This effectiveness is especially applicable to cloud-integrated healthcare settings, where inference intervention has a beneficial impact on clinical decision-making.

Interpretability and modularity can also be supported with the model structure. Transformer spatial attention maps can provide information about frame areas that contribute to decision-making, whereas temporal patterns provided by the RNN can assist clinicians to learn how progression varies among sequences. This is further enhanced by the cloud-enabled extension of the system, which is an extension of the online inference and data synchronization of remote monitoring. This renders the framework to be scalable to telemedicine processes, distance screenings, and ongoing clinical treatments.

On the whole, it can be stated that the hybrid model proposed deals with several shortcomings of the current diagnostic systems by integrating complex elements of deep learning, effective fusion techniques, and cloud compatibility. All of these enhancements make the model a strong and clinically viable tool of automated prediction based on ultrasound.

VI. Conclusion & Future work

In this paper, a new Hybrid Feature Fusion Model is proposed, which successfully combines a Transformer encoder to extract spatial features with an RNN module to model a temporal sequence, which provides a strong solution to clinical prediction based on ultrasound imaging. The model combines complementary representations in space and time in a single discriminative feature space, which has an alpha-weighted fusion mechanism, allowing much better diagnostic performance than traditional algorithms like CNNs, RNNs, and standalone Transformers. Large-scale ultrasound testing on close to 100,000 images proves to be highly effective with an accuracy of 94.5, F1-score of 0.945 and lower error rates such as an RMSE of 0.112. These findings support the model to compute complex and high-dimensional ultrasound data at low inference latency with computational efficiency. Moreover, the fact that it has been implemented as a part of a cloud-based monitoring system reveals its ability to offer real-time diagnostic assistance and remote care services. On the whole, the suggested hybrid architecture is a clinically relevant, scalable, and high-performing solution to the development of automated ultrasound diagnostics and enhancement of patient outcomes.

Further research associated with this field will be aimed at increasing the practical usefulness of the model by incorporating other sources of clinical data, including EHRs and lab records, to provide more comprehensive diagnostic results. The research on federated learning will allow privacy preserving cooperation among hospitals without exchanging sensitive information. Pruning and quantization compression methods will be explored as lightweight model compression methods to enable edge device deployment in point-of-care diagnostics. Lastly, the creation of a better explainability tool will enable clinicians to become more knowledgeable about the spatial temporal decision patterns of the model, which will contribute to better trust and adoption of the model in the real clinical workflow.

References

[1]. Al-Dailami, H. Kuang and J. Wang, "Attention-based Memory Fusion Network for Clinical Outcome Prediction using Electronic Medical Records," *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Las Vegas, NV, USA, 2022, pp. 902-907, doi: 10.1109/BIBM55620.2022.9994881.

Doi:10.48047/ijiee.2025.15.7.17

[2]. B. Pei, B. Wang, L. Sun, Z. Zhu and C. Feng, "Multi-Algorithm Fusion Framework for Trajectory Prediction of Human Upper Limb Motion," *2024 China Automation Congress (CAC)*, Qingdao, China, 2024, pp. 1676-1681, doi: 10.1109/CAC63892.2024.10865112.

[3]. T. Xiao, L. Shi, H. Wang, Z. Wang and Y. Lin, "Stroke Outcome Prediction via Multi-level Feature and Multi-modal Fusion Network," *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Lisbon, Portugal, 2024, pp. 6732-6739, doi: 10.1109/BIBM62325.2024.10821993.

[4]. J. Guo, Y. Cheng, W. He, Y. Zhang, R. Feng and X. Zhang, "Uncertainty-Aware Dynamic Fusion for Multimodal Clinical Prediction Tasks," *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Hyderabad, India, 2025, pp. 1-5, doi: 10.1109/ICASSP49660.2025.10889595.

[5]. Y. Pawar, A. Henriksson, P. Hedberg and P. Naucler, "Leveraging Clinical BERT in Multimodal Mortality Prediction Models for COVID-19," *2022 IEEE 35th International Symposium on Computer-Based Medical Systems (CBMS)*, Shenzen, China, 2022, pp. 199-204, doi: 10.1109/CBMS55023.2022.00042.

[6]. K. Julian, K. S. Nugroho, R. Nirwantono and B. Pardamean, "Cancer Prediction Using Clinical and Genomic Data Fusion: A Systematic Review," *2024 6th International Conference on Cybernetics and Intelligent System (ICORIS)*, Surakarta, Indonesia, 2024, pp. 1-6, doi: 10.1109/ICORIS63540.2024.10903934.

[7]. D. Hu, B. Liu, X. Zhu, X. Lu and N. Wu, "Predicting Lymph Node Metastasis of Lung Cancer: A Two-stage Multimodal Data Fusion Approach," *2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Orlando, FL, USA, 2024, pp. 1-4, doi: 10.1109/EMBC53108.2024.10782471.

[8]. W. Wang *et al*., "Deep Fusion Models of Multi-Phase CT and Selected Clinical Data for Preoperative Prediction of Early Recurrence in Hepatocellular Carcinoma," in *IEEE Access*, vol. 8, pp. 139212-139220, 2020, doi: 10.1109/ACCESS.2020.3011145.

[9]. S. Zhang, Y. Gong, M. Xu, X. Jiang and M. Song, "Intracranial Aneurysm Rupture Risk Assessment Based on Multi-scale Deep Transfer Learning and Clinical Feature Fusion," *2025 4th International Symposium on Computer Applications and Information Technology (ISCAIT)*, Xi'an, China, 2025, pp. 60-63, doi: 10.1109/ISCAIT64916.2025.11010265.

[10]. S. Sinha, Bharti, K. Kumari, A. Kumar and N. Mishra, "Multi-Modal Data Fusion for Predictive Analytics in Healthcare: A Hierarchical Attention Framework for Early Sepsis Prediction," *2025 6th International Conference for Emerging Technology (INCET)*, BELGAUM, India, 2025, pp. 1-6, doi: 10.1109/INCET64471.2025.11139921.

[11]. S. R. Katragadda, M. Sakthivanitha, M. V. Maheswari, P. N. Shiammala, T. Thirumalaikumari and J. Anciline Jenifer, "Transformer-based Clinical Decision Support Systems using Structured and Unstructured EHR Data," *2025 6th International Conference on Electronics and Sustainable Communication Systems (ICESC)*, Coimbatore, India, 2025, pp. 145-150, doi: 10.1109/ICESC65114.2025.11212384.

Doi:10.48047/ijiee.2025.15.7.17