Hybrid Deep Learning Framework for Enhanced Cardiovascular Disease Detection Using ECG Signal

Mr. Sanjib Kumar Dhara, Mr. Nilankar Bhanja, Dr. K Venkata Murali Mohan, DR. Mukesh Tiwari,

[1]*Research Scholar, Electronics and Communication Engineering,*
Sri Satya Sai University of Technology & Medical Sciences,
Bhopal-Indore Road, Sehore (M.P), 466001, Madhya Pradesh, India
sanjibkumardhara@gmail.com
[2]*Research Scholar, Electronics and Communication Engineering,*
Sri Satya Sai University of Technology & Medical Sciences,
Bhopal-Indore Road, Sehore (M.P), 466001,Madhya Pradesh, India
nilankarbhanja2005@gmail.com
[3]*Professor of ECE and Principal,*
Teegala Krishna Reddy Engineering College,
kvmmece@gmail.com
[4]*Professor of Electronics and Communication Engineering,*
Sri Satya Sai University of Technology & Medical Sciences,
info@sssutms.co.in

KEYWORDS
Cardiovascular Disease, Electrocardiogram, Deep Learning, Feature Extraction, Transformer, Signal Pre-processing

ABSTRACT
Cardiovascular Disease (CVD) is a serious medical issue in today's society. The electrocardiogram (ECG) is considered the most appropriate non-invasive diagnostic technique for detecting cardiac conditions. However, interpreting an ECG requires specialist experience and is time-consuming. This underscores the need for automated CVD diagnosis using advanced techniques. Many researchers have proposed various techniques to identify CVD. However, current approaches have been inefficient in identifying small differences due to the irregular and complex nature of the ECG rhythms. This research proposes a novel hybrid deep learning (DL) model called CNN (Convolutional Neural Network)-GRU (Gated Recurrent Unit)-Transformer. The spatial features of ECG are retrieved by the CNN, and temporal features are retrieved by the RNN model. Both features are fused and classified for CVD detection using the Transformer network. The fusion of features helps detect the minor changes in ECG and identify CVD with high reliability. The experimental outcome of the proposed model on the PTB-XL database for ECG classification of CVD shows the highest accuracy of 98.8% and the lowest false negative rate (FNR) and false positive rate (FPR) of 1.2% and 0.3%, respectively. The importance of the proposed network architecture is analyzed through an ablation study. Two ablation studies are conducted: first, the CNN is removed, and the GRU features are given to the Transformer for classification; in the second study, the GRU is removed. The ablation study shows accuracies of 95.8% and 97%, which are significantly lower than the proposed model's accuracy. Additionally, the proposed network is compared with existing research. The outcome shows that the proposed network outperforms state-of-the-art techniques in detecting changes in ECG for CVD classification. The analysis of the proposed network suggests that it is a promising tool for detecting CVD at earlier stages with high accuracy rates.

I. INTRODUCTION

CVDs represent an extremely serious threat to human health. They are some of the most severe diseases in the world, responsible for a significant number of deaths each year [1]. The majority of cardiovascular-related deaths occur unexpectedly, leaving patients with insufficient time to seek medical assistance. Hypertension, obesity, high cholesterol, smoking, and poor dietary habits are all risk factors for CVD. Daily integrated modern biosensor monitoring can assist with early detection, prevention, and

the selection of appropriate treatments for heart diseases [2]. It is critical to identify individuals with heart disease early and monitor them regularly to provide optimal healthcare treatment.

The ECG monitors the electrical signals of the heart, collecting essential information to help understand the cardiovascular system's operations [3]. As a non-invasive diagnostic technique, it is frequently employed to monitor and diagnose heart problems. It can detect heart issues in their early stages and assist in providing appropriate therapy. The ECG delivers essential information to cardiologists about the heart's condition, making it an invaluable tool for detecting various cardiac issues. An ECG device uses electrodes attached to the patient's skin to monitor the heart's rhythmic contractions and relaxations. Normal ECG signals include T, P, and QRS waves. The statistical and anatomical properties of ECG waves are key health indicators that can reveal heart problems. For instance, the absence of P waves and an irregular ventricular rhythm in ECG data may indicate atrial fibrillation [4]. Cardiologists regularly perform ECG screenings on patients to detect heart anomalies and provide effective treatment. However, this process requires significant human effort and costly medical procedures. As the population ages, the patients suffering from CVDs is expected to increase dramatically, necessitating rapid, accurate, and low-cost automatic ECG diagnosis.

By implementing automated CVD detection technologies, healthcare practitioners can optimize resource allocation, streamline patient care, and potentially reduce the cost burden on individuals and healthcare systems [5]. Previous approaches to ECG-based CVD diagnosis relied heavily on human interpretation, which might result in subjectivity and unpredictability in diagnoses. The accuracy of traditional techniques depends on the competency of the interpreting healthcare practitioner, and CVD might be misinterpreted or misclassified, potentially resulting in erroneous treatment regimens or missed opportunities for intervention [6]. Furthermore, traditional techniques may struggle to detect certain types of CVD that exhibit complex or atypical patterns. These limitations of traditional CVD detection methods highlight the need for novel procedures that improve accuracy, objectivity, and efficiency in recognizing and categorizing CVD [7]. Thus, there is significant promise in using DL to accurately and automatically identify CVD using ECG signals. Its capacity to automatically learn various patterns and qualities from raw data makes it ideal for assessing ECG readings and detecting CVDs.

In this research, a novel hybrid DL model is proposed, which combines the strengths of three DL networks: CNN, GRU, and Transformer. First, the spatial features of ECG are extracted by CNN, and temporal features are extracted by the GRU. The extracted features are then fused and classified into different types of CVD using the Transformer network. The strength of the proposed model lies in the combination of both spatial and temporal features. This helps to identify the minor variations in ECG, which are crucial for detecting different types of CVD, as each type varies by only minor changes. The dataset for the research is taken from the PTB-XL dataset, and pre-processing is performed to make it ready for the DL model.

The article is organized as follows: Section I discusses CVD and ECG, highlighting the need for a novel method for CVD detection using ECG. Section II reviews the most recent research on CVD. Section III elaborates on the proposed architecture with a flow diagram. Section IV presents the results of the proposed network, along with an ablation study and a comparison with existing research, as well as detailed information about the dataset. Section V concludes the research with future work.

II. Related Work

For automated CVD detection using ECG signals, numerous researchers have already explored various ML and DL models. Some notable works are summarized in Table I, which aids in identifying recent research gaps and addressing these through the proposed novel hybrid DL model. The table outlines the models used, the accuracy achieved, and the limitations of each study. This comprehensive review provides a foundation for the proposed approach in the research.

TABLE I. LITERATURE REVIEW ON ECG-BASED CVD DETECTION

Ref	Model	Accuracy	Limitation
[8]	Learner module using Support Vector Machine and Random Forest	77.4%	Moderate classification accuracy; further improvements in model performance and feature engineering are required.
[9]	Three-Filter Feature Selection approach with Random Forest	85.58%	The novel feature selection technique improved performance but added complexity
[10]	Independent Component Correlation feature selection with Artificial Neural Network	99.6%	Not fully automated feature extraction and impractical for large-scale or real-time applications.
[11]	Customized CNN-2D for ECG image classification	80%	Further work is required to optimize the architecture
[12]	MobileNetV2	95.18	The small database and the absence of a truly independent test group.
[13]	Deep Neural Network (DNN) with XGBoost	78.65%	An unbalanced dataset leads to poor accuracy
[14]	Long Short-Term Memory (LSTM)-based classifier	95%	Training time is not discussed, which impacts the cost-effectiveness of the approach.
[15]	Stacked LSTM and Bi-LSTM	95%	High computational costs due to the complex structure.
[16]	CNN	94%	Synthetic data generation was introduced, but class imbalance remains a challenge.
[17]	DNN with genetic algorithm	94%	Increased computational cost due to robust feature extraction and optimization protocol.
[18]	1D CNN	97.40%	Excessive pre-processing required, long training time, increasing computational cost.

The proposed research aims to address current gaps, such as imbalanced datasets, reducing complexity, and eliminating manual feature extraction, while ensuring the accuracy of CVD detection.

III. PROPOSED METHODOLOGY

The proposed network for CVD detection using ECG signals is detailed in this section. The hybrid CNN-GRU-Transformer network consists of three important modules: CNN, GRU, and Transformer. The CNN and GRU are employed to retrieve the spatial and temporal features from the ECG signal, respectively. Those features are fused and classified by the transformer network. The overall architecture of the proposed hybrid CNN-GRU-Transformer is illustrated in Figure 1. The functioning of each module is detailed in the below subsections.

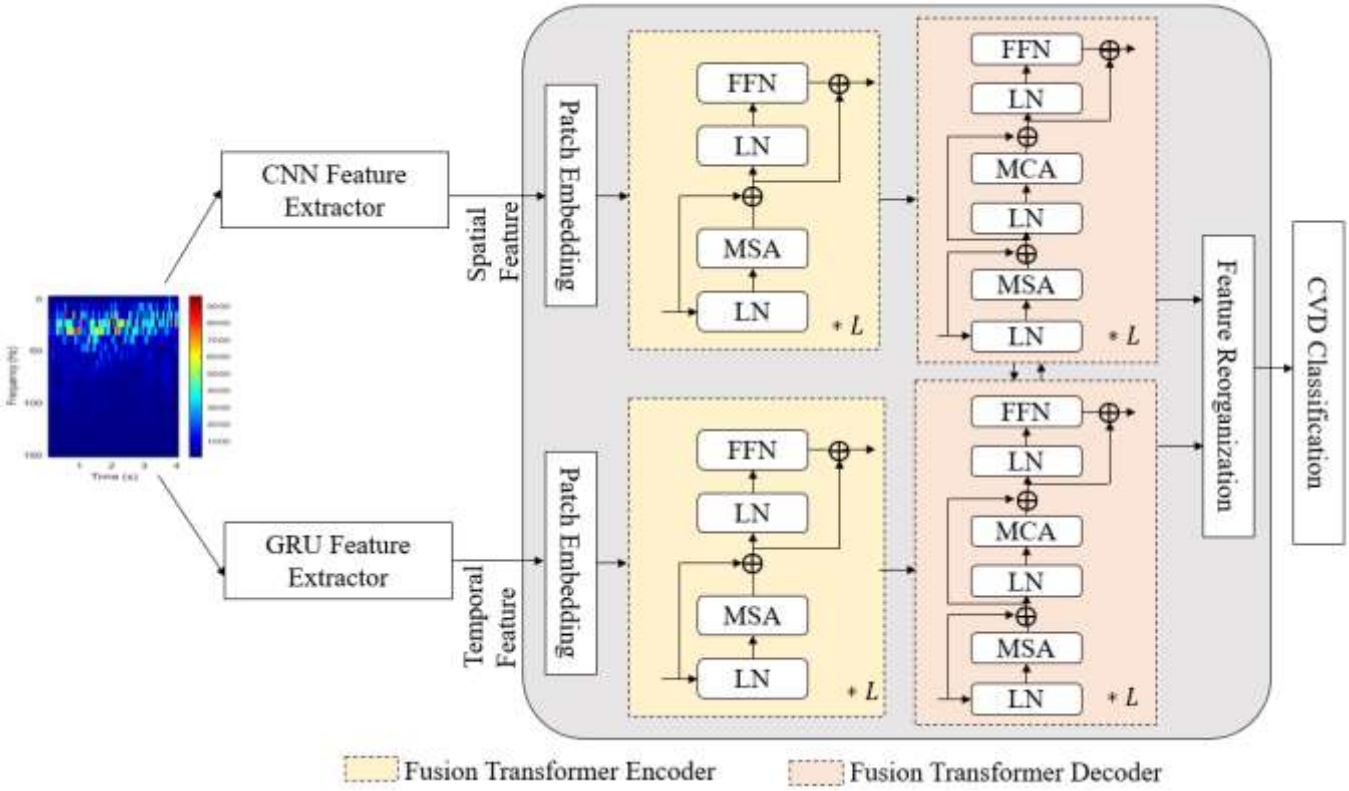


Fig. 1. Proposed CNN-GRU-Transformer Network for CVD Classification

A. CNN

CNNs can be developed with numerous layers, including a convolution layer (CL), a pooling layer (PL), an activation layer (AL), and a classification layer [19]. To retrieve features from the input, the three layers (CL, PL, AL) are required. Because of their high feature extraction capacity, two-dimensional CNNs are commonly used in image processing. The classification layer uses the retrieved features for classification. The CL is the fundamental layer of the CNN framework, and it merges input data with filtering kernels [20]. The network trains the filter to activate when it retrieves the specified features. The mathematical structure can be described in the following equation:

$$y_j^k = K_i^k * x_i^k = \sum_{i \in M_j} x_i^k * w_{ij}^k + b_j^i \quad [1]$$

where y_j^k denotes the outcome of the k-th layer; K_i^k denotes the i-th convolution kernel of the k-th layer; x_i^k denotes the input of the k-th layer; the $*$ symbolizes the convolution process; w_{ij}^k and b_j^i denotes the weight and bias.

The AL often follows the CL, which is an important layer. A neuron's output and input connections are typically defined by its activation function, which is nonlinear. This allows the network to acquire nonlinear characteristics from the input, thereby enhancing its feature extraction performance. The rectified linear unit (ReLU) is selected as an activation function in CNNs, and it is described in the following equation:

$$ReLU(y_j^k) = \begin{cases} 0 \; y_j^k < 0 \\ y_j^k \; y_j^k \geq 0 \end{cases} \quad [2]$$

DNNs use batch normalization (BN) to reduce internal correlation shifts while increasing network training accuracy [21]. Furthermore, BN normalizes the learned parameters (to a range of 0 to 1), which speeds up the model's training. The transformation process of BN is discussed below:

$$\begin{cases} \mu_B = \frac{1}{m}\sum_{i=1}^{m} x_i \\ \delta_B^2 = \frac{1}{m}\sum_{i=1}^{m}(x_i - \mu_B)^2 \\ \hat{x}_i = \frac{x_i - \mu_B}{\sqrt{\delta_B^2 + \varepsilon}} \\ y_i = \gamma \hat{x}_i + \beta \end{cases} \quad [3]$$

Where, m denotes the mini-batch size, μ_B denotes the mean, and δ_B^2 denotes the variance. The network's parameters, γ and β, are learnable. The PL performs a downsampling process to eliminate duplicate features and acquire more detailed ones. The most popular pooling operations are maximum pooling (MP) and average pooling (AP). MP usually outperforms AP in time-series classification tasks, as shown below:

$$p_i^k = max_{(j-1)s+1 \le t \le js}\{a_i^k(t)\} \quad [4]$$

The k-th layer's output features are represented by p_i^k, $a_i^k(t)$ denotes the outcome of the i-th channel's t-neuron of the k-th layer, and s represents the pooling stride. After this, the classification layer is present. The output before the classification layer is considered as the spatial features.

B. GRU

The GRU, a recurrent neural network (RNN), addresses the issue of gradient vanishing in long-term dependencies during time series learning in conventional RNNs [22]. Both GRU and LSTM tackle this issue; however, while they perform equally on a range of DL tasks, GRU requires fewer parameters and computations. This minimize the likelihood of overfitting and conserves computational resources, making it more efficient. The GRU model consists of two fundamental gates: the update gate (UG) and the reset gate (RG) [23]. The RG controls the amount of the past hidden state that affects the candidate state, based on the past hidden state and the current input. The UG decides which historical data from the last hidden state has to be discarded and which data from the present candidate state has to be included in the new hidden state. Equations (7) and (8) provide the update formulas for the candidate and hidden states, respectively, whereas Equations (5) and (6) calculate the RG and UG weights.

$$r_n = \sigma(W_{ir}x_n + W_{hr}h_{n-1}) \quad [5]$$

$$z_n = \sigma(W_{iz}x_n + W_{hz}h_{n-1}) \quad [6]$$

$$c_n = tanh\big(W_{ic}x_n + W_{hc}(r_n \odot h_{n-1})\big) \quad [7]$$

$$h_n = o_n = (1 - z_n)\odot c_n + z_n \odot h_{n-1} \quad [8]$$

Where, x_n represents the input at the (n) moment, and h_{n-1}represents the hidden state at the $(n-1)$ moment. W_{ir} and W_{hr} denotes the RG weight matrices. W_{iz} and W_{hz} denotes the UG weight matrices. W_{ic} and W_{hc} denotes the candidate state weight matrices. At the (n) moment, c_n denotes the candidate state, h_n and o_n denotes the hidden and output state. The $\odot$ denotes element-wise multiplication. The activation functions σ and $tanh$ are calculated using the formulas: $\sigma(x) = \frac{1}{1+e^{-x}}$ and $tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$. These functions enhance the model's nonlinear capabilities. In the GRU network, the output layer is also eliminated, and the temporal features are extracted for further processing.

C. Transformer

The Transformer includes both encoder and decoder structures within its architecture [24]. Fusion Encoder (FTE) leverages both non-local attention and sequential feature learning to capture deep temporal and spatial characteristics from ECG signals. The CNN feature f_h is spatially segregated as a series of two-dimensional patches $\{f_{h_i}\}_{i=1}^{N}$, where $f_{h_i} \in \mathbb{R}^{P^2 * C}$, P represents the patch size, and $N = \frac{H*W}{P^2}$ represents the number of patches. These spatial features, extracted by the CNN, capture the local patterns within the ECG signals. Meanwhile, the GRU extracts temporal features from the ECG signals. This

combination enables the model to learn both spatial and temporal representations effectively. The outputs from both the CNN and GRU are then fused to create a unified representation for further classification. A linear projection layer $E \in \mathbb{R}^{(P^2.C)*D}$ has been trained for mapping flattened patches into the D-dimensional hidden embedding region. For making the architecture adaptive to extracted data, a series of learnable position embeddings, $\{E_{p_i}\}_{i=1}^{N}$, has been included in the patch embeddings, where $E_{p_i} \in \mathbb{R}^{D}$. The stacked transformer encoder receives $\{f_{h_i}E + E_{p_i}\}_{i=1}^{N}$. Each encoder layer has a feed-forward network (FFN) and multi-head self-attention (MSA) mechanism [25]. The encoder layer employs the skip connection, and layer normalization (LN). The aforementioned procedures are organized as follows:

$$x_0 = [f_{h_1}E + E_{p1}; f_{h_2}E + E_{p2}; \ldots; f_{h_N}E + E_{pN}] \quad [9]$$

$$x'_l = MSA(LN(x_{l-1})) + x_{l-1} \quad [10]$$

$$x_l = FFN(LN(x'_l)) + x'_l, \quad l = 1 \ldots L \quad [11]$$

$$[f_{FTE_1}, f_{FTE_2}, \ldots, f_{FTE_N}] = LN(x_L) \quad [12]$$

Equation (10) computes the similarities between the n-th and the other patch embeddings, which are subsequently used as aggregate weights for encoding the n-th patch. MSA acts as a feature extraction technique, as determined by the following formula:

$$q_l^i = LN(x_{l-1})E_q \quad [13]$$

$$k_l^i = LN(x_{l-1})E_k \quad [14]$$

$$v_l^i = LN(x_{l-1})E_v \quad [15]$$

$$Att^i(q_l^i, k_l^i, v_l^i) = softmax\left(\frac{q_l^i {k_l^i}^T}{\sqrt{D_{att}}} v_l^i\right), i = 1,2,\ldots,H \quad [16]$$

$$x_l^i = Concat\left(\{Att^i(q_l^i, k_l^i, v_l^i)\}_{i=1}^{H}\right) E_{out} \quad [17]$$

Where, E_q, E_k and $E_v \in \mathbb{R}^{D*D_{att}}$, the patch embedding dimension is decreased to D_{att} to lessen the computing cost of Equation (16). FTE provides a collection of encoding features $\{f_{FTE_i}\}_{i=1}^{N}$ for all sources.

A fusion transformer decoder (FTD) is utilized to merge the CNN and GRU features globally. The feature point $f_h(i,j)$ at the (i,j) position of the CNN and GRU features uses the MSA module to perform a search across the entire feature map, collecting feature points with corresponding temporal and spatial data. The cosine distance is employed for calculating the similarity matrix (SM) between $f_h(i,j)$, and the features at every position in f_h using Equation (10). The SM is then normalized. Using this normalized matrix, The $f_h(i,j)$ feature points are summed and weighted. Lastly, the features are concatenated with $f_h(i,j)$. The receptive area of features are expanded by FTE, while improving the recognizing strength of individual-source features.

$$y_0^i = [f_{FTE_1}^i + E_{p1}^i; f_{FTE_2}^i + E_{p2}^i; \ldots f_{FTE_N}^i + E_{pN}^i] \quad [18]$$

$$y_l'^i = MSA(LN(y_{l-1})) + y_{l-1} \quad [19]$$

$$y_l''^i = MCA\left(LN(y_l'^i), LN(y_l'^i)\right) + y_l'^i \quad [20]$$

$$y_l^i = FFN\left(LN(y_l''^i)\right) + y_l''^i, \qquad l = 1 \ldots L \quad [21]$$

$$[f_{FTD_1}^i, f_{FTD_2}^i, \ldots, f_{FTD_N}^i] = LN(y_L^i) \quad [22]$$

Equation (20) identifies the similarities between source i ($n-$th patch embedding) and j (all patch embeddings), which are used for aggregation. Therefore, MCA is a feature combination mechanism, which is defined below (single-head cross-attention is given):

$$q_l'^i = LN\left(y_l'^i\right)E_{q'}^i \quad [23]$$

$$k_l'^i = LN\left(y_l'^j\right)E_{k,v'}^i \quad [24]$$

$$v_l'^i = LN\left(y_l'^j\right)E_{k,v'}^i \quad [25]$$

$$Att'^i\left(q_l'^i, k_l'^j, v_l'^j\right) = softmax\left(\frac{q_l'^i k_l'^{jT}}{\sqrt{D_{att}}} v_l'^j\right) \quad [26]$$

$$y_l''^i = Att'^i\left(q_l'^i, k_l'^j, v_l'^j\right)E_{out}^i \quad [27]$$

where $E_q^i, E_{k,v}^j \in \mathbb{R}^{D*D_{att}}$, the patch embedding dimension is deduced to D_{att} for minimizing the computational complexity of Equation (20). Next, FTD produces the decoding feature $\{f_{FTD}\}_{i=1}^N$.

The MCA component does not instantly fuse the patch embeddings $f_{FTE_n}^i$ and $f_{FTE_n}^j$ of sources i and j. The MCA has done a global search on the features in j. Equation (26) uses the cosine distance to determine the SM between $f_{FTE_n}^i$ and each embedding in j and then normalized. The patch embeddings from source j are summed and weighted using the normalized matrix. Next, the features are merged with $f_{FTE_n}^i$. As a result, even if the two sources have semantic biases, FTD combines features from CNN and GRU that include semantically similar information.

The class token is used to interpret both spatial and temporal information. Instead of initializing randomly, class tokens are employed to accelerate convergence. Global AP is performed on CNN and GRU features $f_h \in \mathbb{R}^{H'*W'*D}$ to produce the semantic class token $f_{sct} \in \mathbb{R}^C$, which contributes to the extraction and fusion of features. To compute the FTD and FTE, Equations (9), (12), (18), and (22) are adjusted as shown in Equation (28-31):

$$x_0 = \left[f_{sct}E_s + E_{p_0}; f_{h_1}E + E_{p_1}; \ldots; f_{h_N}E + E_{p_N}\right] \quad [28]$$

$$\left[f_{FTE_0}, f_{FTE_1}, \ldots, f_{FTE_N}\right] = LN(x_L) \quad [29]$$

$$y_0^i = \left[f_{FTE_0}^i + E_{p0}^i; f_{FTE_1}^i + E_{p1}^i; \ldots f_{FTE_N}^i + E_{pN}^i\right] \quad [30]$$

$$\left[f_{FTD_0}^i, f_{FTD_1}^i, \ldots, f_{FTD_N}^i\right] = LN\left(y_L^i\right) \quad [31]$$

Where $E_s \in \mathbb{R}^{C*D}$. The tokens of rearranged features are chosen from the decoded and the fused features $f_F = f_{FTD_0}^{\mathrm{CNN}} + f_{FTD_0}^{\mathrm{GRU}}$. The final fully connected network accepts the fused features as input and generates a prediction map $\tilde{f}_{p_i}$ for each CVD. The formula for predicting CVD is given in Equation (33).

$$\tilde{f}_{p_i} = F_{ti}(Fusion(f_{FTD}^{CNN} + f_{FTD}^{GRU})), \tilde{f}_{p_i} \in [0,1]^{hxwxC} \quad [33]$$

IV. Discussion

This section discusses the data collection and processing steps in CVD classification. The experimental outcomes of the proposed network, ablation study, and metrics comparison with the existing research are also presented.

A. Data Collection and Processing

This study takes ECG signal from the PTB-XL ECG database [26]. The dataset includes 21,837 clinical 12-lead ECGs from 18,885 individuals, each lasting 10 seconds and collected at 500 and 100 Hz

with 16-bit resolution [27]. ECG results are susceptible to contamination by background noise and bioelectrical interference. For accurate evaluation and assessment, unwanted noise must be removed from the ECG. This work uses DWT [28], a widely used denoising method, as a viable choice for denoising ECG data. The research also developed wavelet families for ECG signals, such as Haar, Symlets, Bior, Daubechies, and Coiflet, to identify which wavelet type produced the most effective signal denoising results. Based on the highest signal-to-noise ratios, the Symlet wavelet was chosen as the best DWT parameter for ECG signal denoising. Next, the Fast Fourier Transform (FFT) was employed to map the denoised ECG signal to images [29]. The formula for the FFT of the ECG time series x is defined as follows:

$$Y_{p+1,q+1} = \sum_{j=0}^{m-1} \sum_{k=0}^{n-1} \omega_m^{jp}\, \omega_n^{kp} X_{j+1,k+1} \quad [34]$$

Where, $\omega_m = e^{-2\pi i/m}$ and $\omega_n = e^{-2\pi i/n}$ represents the complex roots, i represents the imaginary unity, q and k represents indices $(0 \text{ to } n-1)$, and p and j represents indices $(0 \text{ to } m-1)$. The sample scalogram images are given in Figure 2.

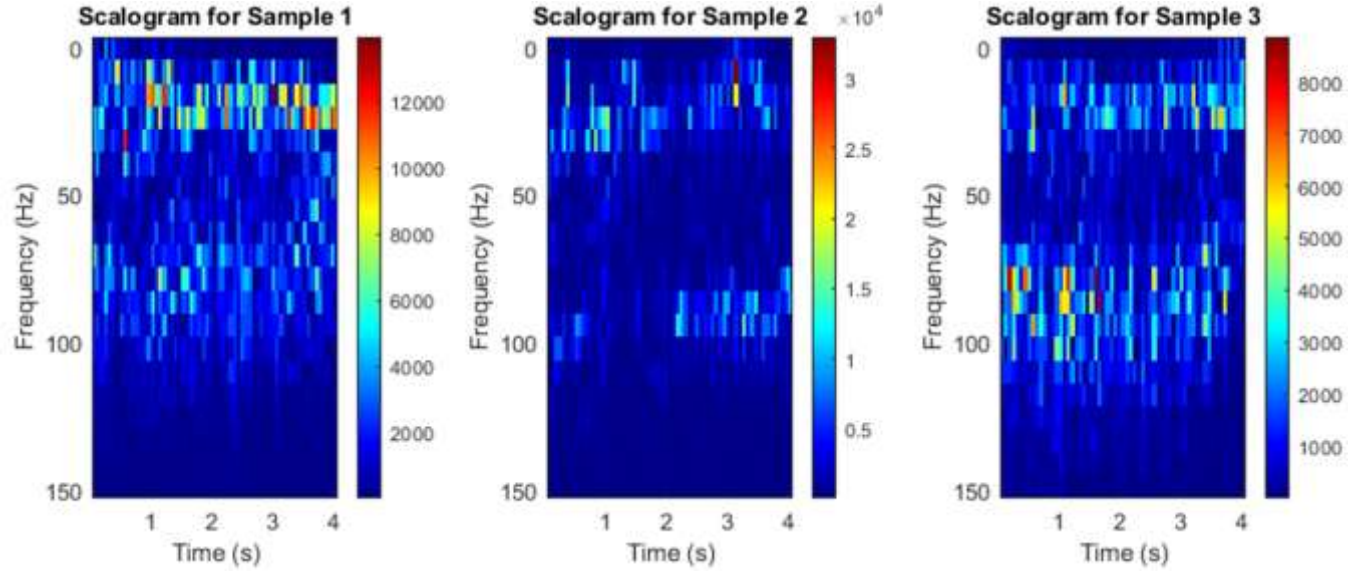


Fig. 2. Scalogram images of ECG signal

The scalogram images are separated into three groups: training, validation, and test. These data were then normalized in the range of 0 to 1 and utilized as inputs to the proposed model for evaluation. Table II provides a detailed description of the PTB-XL dataset before and after balancing, including the data used for training, validation, and testing.

TABLE II. DETAILS OF PTB-XL DATASET

Data	Actual	Balanced	Train	Validation	Test
NORM	7185	1000	700	200	100
CD	3232	1000	700	200	100
STTC	3064	1000	700	200	100
MI	2936	1000	700	200	100
HYP	812	1000	700	200	100

B. *Experimental Outcome*

The research was conducted on Google Colaboratory. The Python programming language was used, and the TPU runtime was selected to implement the network. For CVD identification, ECG signals from the PTB-XL database were collected and processed. The processed signals were fed into the proposed CNN-GRU-Transformer network. A total of 3,500 samples were used for training, and 1,000 samples were used for validating the CNN-GRU-Transformer model. The loss and accuracy plots of the model are shown in Figures 3 and 4.

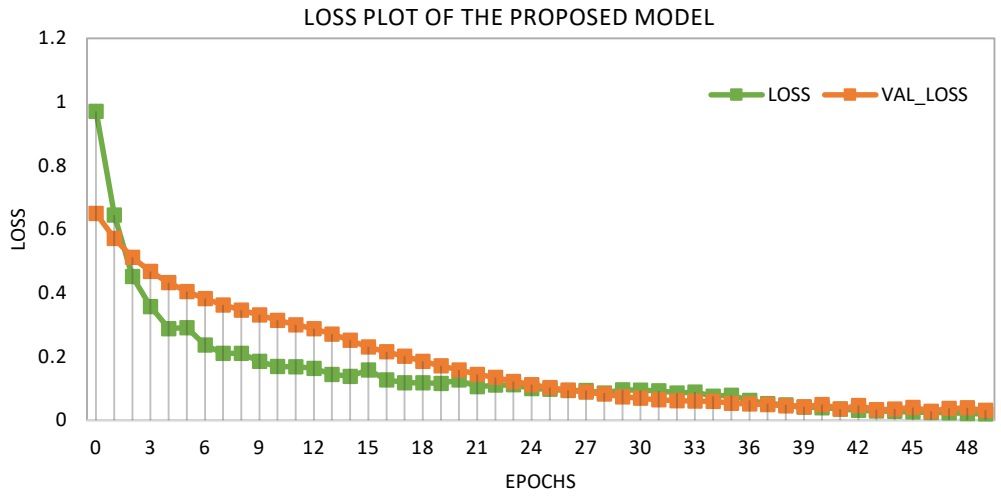


Fig. 3. Loss plot of the proposed network on CVD Classification

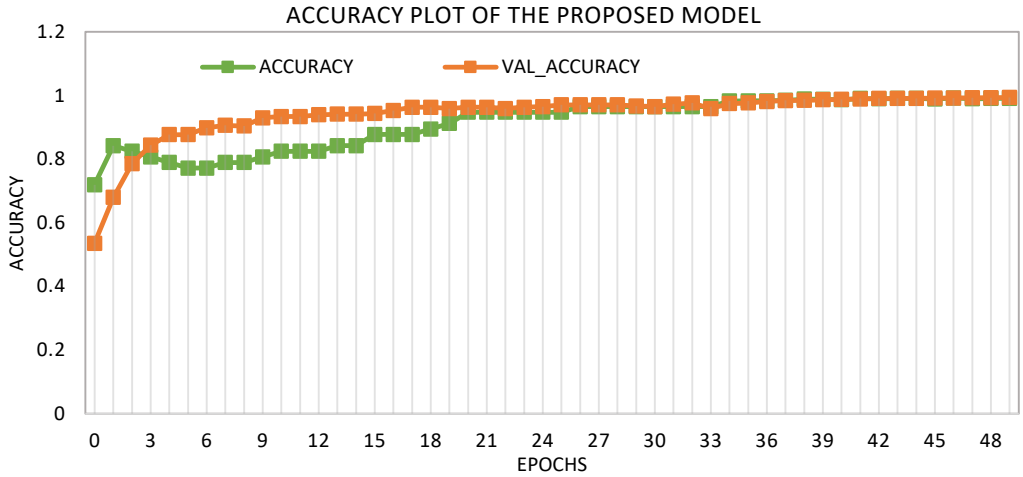


Fig. 4. Accuracy plot of the proposed network on CVD Classification

After training and validation, 100 samples from each CVD category were used for testing the model. The proposed model correctly identified 494 CVD categories and misclassified 6 CVDs out of 500 ECG signals. The confusion matrix of the proposed network is shown in Figure 5. Using the confusion matrix elements, metrics such as accuracy, precision, recall, F1 score, specificity, FNR, and FPR were calculated using Equations (4-9).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad [35]$$

$$Precision = \frac{TP}{TP+FP} \quad [36]$$

$$Recall = \frac{TP}{TP+FN} \quad [37]$$

$$F1 - Score = \frac{2.TP}{2.TP+FN+FP} \quad [38]$$

$$\text{Specificity} = \frac{TN}{TN+FP} \quad [39]$$

$$FNR = \frac{FN}{TP+FN} \quad [40]$$

$$FPR = \frac{FP}{FP+TN} \quad [41]$$

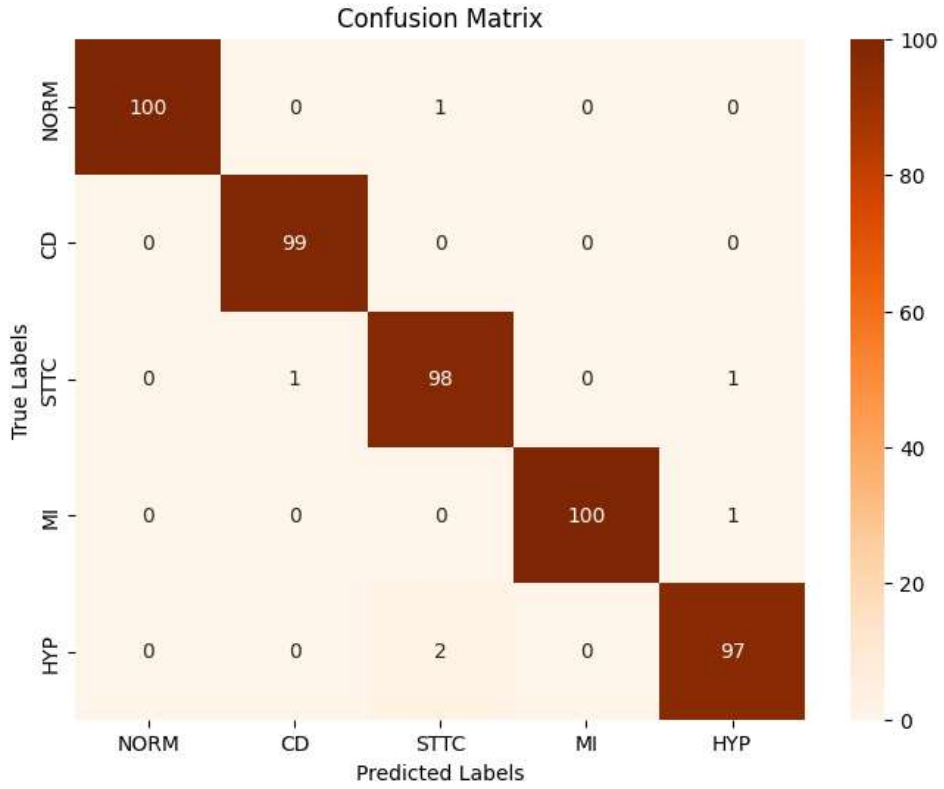


Fig. 5. Confusion Matrix of the Proposed Network

In these equations, TP (True Positive) and TN (True Negative) denotes the correct detection of CVD from the ECG, while FP and FN indicate incorrect detections of CVD. Table III shows the performance evaluation of the proposed model for each CVD category. The proposed network obtained an accuracy of 98.80%. Other metrics such as precision, recall, F1 score, and specificity were 98.80%, 98.80%, 98.80%, and 99.70%, respectively, while FRR and FAR were 1.20% and 0.30%, respectively. The proposed network performed excellently in predicting normal, MI, and CD conditions.

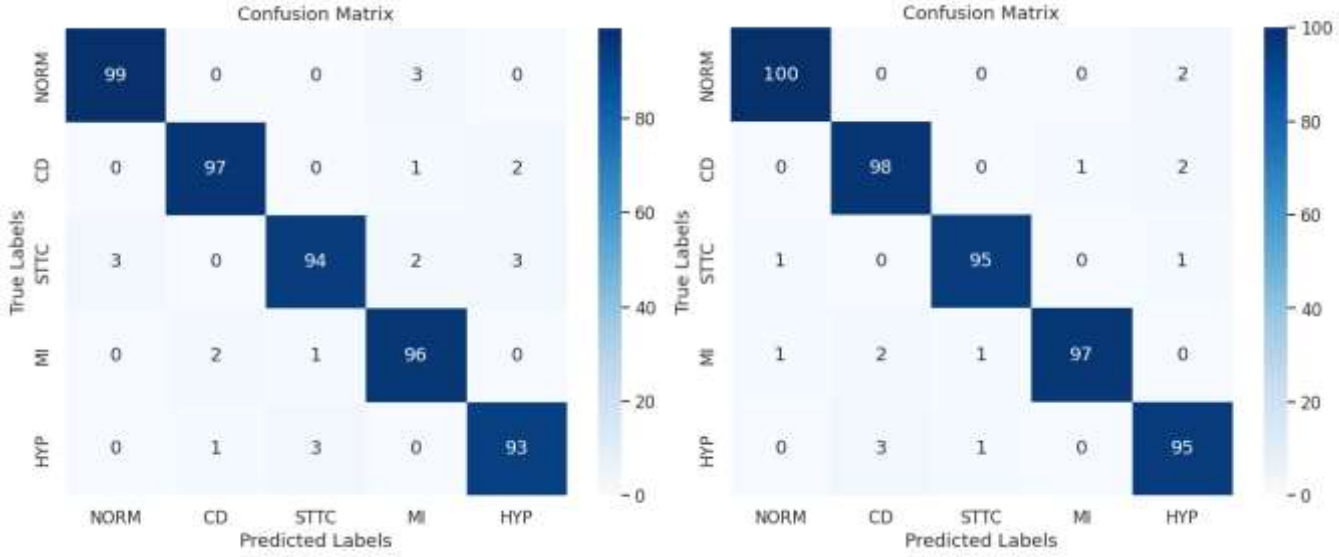
TABLE III. PERFORMANCE COMPARISON OF THE PROPOSED NETWORK

	Accuracy	Precision	Recall	F1	Specificity	FNR	FPR
NORM	98.80	100.00	99.01	99.50	100.00	0.99	0.00
CD		99.00	100.00	99.50	99.75	0.00	0.25
STTC		97.03	98.00	97.51	99.25	2.00	0.75
MI		100.00	99.01	99.50	100.00	0.99	0.00
HYP		97.98	97.98	97.98	99.50	2.02	0.50
Average		98.80	98.80	98.80	99.70	1.20	0.30

An ablation study was conducted on the proposed network. In the proposed network, CNN and GRU features were fused by a transformer to perform classification. In Ablation Study 1, the GRU module was removed from the proposed network, and CNN features were fed to the transformer module for classification. The performance metrics achieved by the model were: Accuracy: 95.80%, Precision: 95.80%, Recall: 95.81%, F1 Score: 95.79%, Specificity: 98.95%, FRR: 4.19%, FAR: 0.30%. Table IV provides the detailed metrics obtained for each category in Ablation Study I.

In the second ablation study, the CNN module was removed from the proposed network, and GRU-retrieved features were fed to the transformer module for the classification of CVD from ECG signals. The metrics achieved by Ablation Study 2 were: Accuracy: 97.00%, Specificity: 99.25%, Precision: 97.02%, Recall: 97.00%, F1: 97.00%, FRR: 3.00%, FAR: 0.75%. Table V provides the detailed metrics obtained for each category in Ablation Study II. By comparing the performance of the ablation studies

with the proposed model, it is clear that each module contributes to improving the accuracy of CVD detection. The confusion matrix of the ablation study without GRU and CNN is given in Figure 6.



a) Without GRU b) Without CNN

Fig. 6. Confusion Matrix of the Ablation Study

TABLE IV. PERFORMANCE COMPARISON OF THE PROPOSED NETWORK WITHOUT GRU FEATURES

	Accuracy	Precision	Recall	F1	Specificity	FNR	FPR
NORM	95.80	97.06	97.06	97.06	99.25	2.94	0.00
CD		97.00	97.00	97.00	99.25	3.00	0.25
STTC		95.92	92.16	94.00	99.00	7.84	0.75
MI		94.12	96.97	95.52	98.50	3.03	0.00
HYP		94.90	95.88	95.38	98.76	4.12	0.50
Average		95.80	95.81	95.79	98.95	4.19	0.30

TABLE V. PERFORMANCE COMPARISON OF THE PROPOSED NETWORK WITHOUT CNN FEATURES

	Accuracy	Precision	Recall	F1	Specificity	FNR	FPR
NORM	97.00	98.04	98.04	98.04	99.50	1.96	0.50
CD		95.15	97.03	96.08	98.75	2.97	1.25
STTC		97.94	97.94	97.94	99.50	2.06	0.50
MI		98.98	96.04	97.49	99.75	3.96	0.25
HYP		95.00	95.96	95.48	98.75	4.04	1.25
Average		97.02	97.00	97.00	99.25	3.00	0.75

The proposed model in this research was compared with existing research work from the literature. Table VI shows the comparison of performance metrics. Among the studies considered for comparison, references [1-4] reported accuracy metrics greater than 90%. The highest accuracy attained was by reference [1], which achieved an accuracy of 98.51%. However, the proposed network achieved the highest accuracy of 98.80%, surpassing all other studies. Other references produced accuracy values below 90%. All the studies used the same PTB-XL database.

TABLE VI. COMPARISON OF PROPOSED NETWORK WITH THE STATE-OF-ART RESEARCH

Ref	Accuracy	Recall	Specificity	Precision	F1
[30]	98.51	97.9	98.12	-	97.95
[31]	93	90	-	92	93
[32]	97.5	97.42	-	97.61	97.52

[33]	91.24	97.82	64.03	90.73	94.13
[34]	88.7	81.7	-	84.9	83.3
[35]	89.84	76.87	93.48	81.83	78.75
[36]	89.87	65.19	-	81.08	71.49
[37]	89.1	69.3	93.4	79.8	-
[38]	89.4	75.6	-	77.8	76.7
Ours	98.80	98.80	99.70	98.80	98.80

The state-of-the-art comparison and the ablation study outcomes further validate the effectiveness of the proposed network for CVD classification using ECG signals.

V. CONCLUSION

The research aims to develop an efficient novel hybrid DL model to detect the type of CVD from ECG signals. The research used the PTB-XL dataset, which contains labeled CVD ECG signals. The collected signals suffered from noise and an imbalanced dataset, which were addressed through pre-processing techniques. The processed signals were converted into spectrogram images, which were then fed into the proposed network for training and testing. The proposed network correctly identified 494 CVD categories and misclassified 6 CVD out of 500 ECG signals, achieving an accuracy of 98.8%. The proposed network architecture was further analyzed through an ablation study. First, the CNN was removed from the network, and it resulted in an accuracy of 97%. Next, the RNN was removed, and the accuracy dropped to 95.8%. The comparison with state-of-the-art research highlights the effectiveness of the proposed network. The power of the proposed network architecture lies in its ability to collect both spatial and temporal information from the ECG signal using the CNN and GRU networks. These features are effectively fused and classified with the help of the Transformer.

The limitation of the research is that only one public dataset was used to validate the model. To assess the generalizability of the network, it should be validated with additional datasets in the future. Furthermore, future work involves deploying the model in real-time. The FPGA will be designed for optimal usage of hardware resources and power, while maintaining a high detection rate of CVD.

REFERENCES

[1] Li, Yan, Gui-ying Cao, Wen-zhan Jing, Jue Liu, and Min Liu. "Global trends and regional differences in incidence and mortality of cardiovascular disease, 1990− 2019: findings from 2019 global burden of disease study." *European journal of preventive cardiology* 30, no. 3 (2023): 276-286.

[2] Ouyang, Mengxing, Dandan Tu, Lin Tong, Mehenur Sarwar, Arvind Bhimaraj, Chenzhong Li, Gerard L. Cote, and Dino Di Carlo. "A review of biosensor technologies for blood biomarkers toward monitoring cardiovascular diseases at the point-of-care." *Biosensors and Bioelectronics* 171 (2021): 112621.

[3] Faruk, Nasir, Abubakar Abdulkarim, Ifada Emmanuel, Yusuf Y. Folawiyo, Kayode S. Adewole, Hammed A. Mojeed, Abdukareem A. Oloyede et al. "A comprehensive survey on low-cost ECG acquisition systems: Advances on design specifications, challenges and future direction." *biocybernetics and biomedical engineering* 41, no. 2 (2021): 474-502.

[4] Johnson, Linda SB, Anders P. Persson, Per Wollmer, Steen Juul-Möller, Tord Juhlin, and Gunnar Engström. "Irregularity and lack of p waves in short tachycardia episodes predict atrial fibrillation and ischemic stroke." *Heart rhythm* 15, no. 6 (2018): 805-811.

[5] Ullah, Muneeb, Shah Hamayun, Abdul Wahab, Shahid Ullah Khan, Mahboob Ur Rehman, Zia Ul Haq, Khalil Ur Rehman et al. "Smart technologies used as smart tools in the management of cardiovascular disease and their future perspective." *Current Problems in Cardiology* 48, no. 11 (2023): 101922.

[6] Schläpfer, Jürg, and Hein J. Wellens. "Computer-interpreted electrocardiograms: benefits and limitations." *Journal of the American College of Cardiology* 70, no. 9 (2017): 1183-1192.

[7] Bhanja, Nilankar, Sanjib Kumar Dhara, and Prabodh Khampariya. "Heuristic-Assisted Adaptive Hybrid Deep Learning Model With Feature Selection For Epilepsy Detection Using EEG Signals." *Biomedical Engineering: Applications, Basis and Communications* 35, no. 06 (2023): 2350036.

[8] Gupta, Aashuli, Arnob Banerjee, Disha Babaria, Kunal Lotlikar, and Hema Raut. "Prediction and classification of cardiac arrhythmia." In *Sentimental Analysis and Deep Learning: Proceedings of ICSADL 2021*, pp. 527-538. Springer Singapore, 2022.

[9] Singh, Namrata, and Pradeep Singh. "Cardiac arrhythmia classification using machine learning techniques." In *Engineering Vibration, Communication and Information Processing: ICoEVCI 2018, India*, pp. 469-480. Springer Singapore, 2019.

[10] Sarfraz, Mohammad, Ateeq Ahmed Khan, and Francis F. Li. "Using independent component analysis to obtain feature space for reliable ECG Arrhythmia classification." In *2014 IEEE international conference on bioinformatics and biomedicine (BIBM)*, pp. 62-67. IEEE, 2014.

[11] Aversano, Lerina, Mario Luca Bernardi, Marta Cimitile, Debora Montano, and Riccardo Pecori. "Characterization of Heart Diseases per Single Lead Using ECG Images and CNN-2D." *Sensors* 24, no. 11 (2024): 3485.

[12] Mhamdi, Lotfi, Oussama Dammak, François Cottin, and Imed Ben Dhaou. "Artificial intelligence for cardiac diseases diagnosis and prediction using ECG images on embedded systems." *Biomedicines* 10, no. 8 (2022): 2013.

[13] Karthik, S., M. Santhosh, Muthu Subash Kavitha, and A. Christopher Paul. "Automated Deep Learning Based Cardiovascular Disease Diagnosis Using ECG Signals." *Computer Systems Science & Engineering* 42, no. 1 (2022).

[14] Rana, Amrita, and Kyung Ki Kim. "ECG heartbeat classification using a single layer lstm model." In *2019 International SoC Design Conference (ISOCC)*, pp. 267-268. IEEE, 2019.

[15] Hiriyannaiah, Srinidhi, Siddesh GM, Kiran MHM, and K. G. Srinivasa. "A comparative study and analysis of LSTM deep neural networks for heartbeats classification." *Health and Technology* 11, no. 3 (2021): 663-671.

[16] Acharya, U. Rajendra, Shu Lih Oh, Yuki Hagiwara, Jen Hong Tan, Muhammad Adam, Arkadiusz Gertych, and Ru San Tan. "A deep convolutional neural network model to classify heartbeats." *Computers in biology and medicine* 89 (2017): 389-396.

[17] Hammad, Mohamed, Abdullah M. Iliyasu, Abdulhamit Subasi, Edmond SL Ho, and Ahmed A. Abd El-Latif. "A multitier deep learning model for arrhythmia detection." *IEEE Transactions on Instrumentation and Measurement* 70 (2020): 1-9.

[18] Wu, Mengze, Yongdi Lu, Wenli Yang, and Shen Yuong Wong. "A study on arrhythmia via ECG signal classification using the convolutional neural network." *Frontiers in computational neuroscience* 14 (2021): 564015.

[19] Zhao, Xia, Limin Wang, Yufei Zhang, Xuming Han, Muhammet Deveci, and Milan Parmar. "A review of convolutional neural networks in computer vision." *Artificial Intelligence Review* 57, no. 4 (2024): 99.

[20] Zoumpourlis, Georgios, Alexandros Doumanoglou, Nicholas Vretos, and Petros Daras. "Non-linear convolution filters for cnn-based learning." In *Proceedings of the IEEE international conference on computer vision*, pp. 4761-4769. 2017.

[21] Ziaee, Amir, and Erion Çano. "Batch Layer Normalization A new normalization layer for CNNs and RNNs." In *Proceedings of the 6th International Conference on Advances in Artificial Intelligence*, pp. 40-49. 2022.

[22] Zargar, S. "Introduction to sequence learning models: RNN, LSTM, GRU." Department of Mechanical and Aerospace Engineering, North Carolina State University (2021).

[23] Salem, Fathi M., and Fathi M. Salem. "Gated RNN: The Gated Recurrent Unit (GRU) RNN." *Recurrent Neural Networks: From Simple to Gated Architectures* (2022): 85-100.

[24] Han, Kai, An Xiao, Enhua Wu, Jianyuan Guo, Chunjing Xu, and Yunhe Wang. "Transformer in transformer." *Advances in neural information processing systems* 34 (2021): 15908-15919.

[25] Heidari, Moein, Reza Azad, Sina Ghorbani Kolahi, René Arimond, Leon Niggemeier, Alaa Sulaiman, Afshin Bozorgpour et al. "Enhancing Efficiency in Vision Transformer Networks: Design Techniques and Insights." *arXiv preprint arXiv:2403.19882* (2024).

[26] Wagner, Patrick, Nils Strodthoff, Ralf-Dieter Bousseljot, Dieter Kreiseler, Fatima I. Lunze, Wojciech Samek, and Tobias Schaeffter. "PTB-XL, a large publicly available electrocardiography dataset." *Scientific data* 7, no. 1 (2020): 1-15.

[27] Bhanjaa, Mr Nilankar, and Prabodh Khampariya. Design and Comparison of Deep Learning Model for ECG Classification using PTB-XL Dataset. 2023.

[28] Singh, Pratik, Gayadhar Pradhan, and S. Shahnawazuddin. "Denoising of ECG signal by non-local estimation of approximation coefficients in DWT." *Biocybernetics and Biomedical Engineering* 37, no. 3 (2017): 599-610.

[29] Safdar, Muhammad Farhan, Robert Marek Nowak, and Piotr Pałka. "A denoising and fourier transformation-based spectrograms in ECG classification using convolutional neural network." *Sensors* 22, no. 24 (2022): 9576.

[30] Selvam, Immaculate Joy, Moorthi Madhavan, and Senthil Kumar Kumarasamy. "Detection and classification of electrocardiography using hybrid deep learning models." *Hellenic Journal of Cardiology* (2024).

[31] Sinha, Nidhi, MA Ganesh Kumar, Amit M. Joshi, and Linga Reddy Cenkeramaddi. "DASMcC: Data Augmented SMOTE Multi-class Classifier for prediction of Cardiovascular Diseases using time series features." *IEEE Access* (2023).

[32] Bhanjaa, Mr Nilankar, and Prabodh Khampariya. Design and Comparison of Deep Learning Model for ECG Classification using PTB-XL Dataset. 2023.

[33] Alqahtani, Hamed, Ghadah Aldehim, Nuha Alruwais, Mohammed Assiri, Amani A. Alneil, and Abdullah Mohamed. "Leveraging electrocardiography signals for deep learning-driven cardiovascular disease classification model." *Heliyon* 10, no. 16 (2024).

[34] Geng, Quancheng, Hui Liu, Tianlei Gao, Rensong Liu, Chao Chen, Qing Zhu, and Minglei Shu. "An ecg classification method based on multi-task learning and cot attention mechanism." In *Healthcare*, vol. 11, no. 7, p. 1000. MDPI, 2023.

[35] Ayano, Yehualashet Megersa, Friedhelm Schwenker, Bisrat Derebssa Dufera, Taye Girma Debelee, and Yitagesu Getachew Ejegu. "Interpretable Hybrid Multichannel Deep Learning Model for Heart Disease Classification Using 12-leads ECG Signal." *IEEE Access* (2024).

[36] Elyamani, Haneen A., Mohammed A. Salem, Farid Melgani, and N. M. Yhiea. "Deep residual 2D convolutional neural network for cardiovascular disease classification." *Scientific Reports* 14, no. 1 (2024): 22040.

[37] Ojha, Jaya, Hårek Haugerud, Anis Yazidi, and Pedro G. Lind. "Exploring Interpretable AI Methods for ECG Data Classification." In *Proceedings of the 5th ACM Workshop on Intelligent Cross-Data Analysis and Retrieval*, pp. 11-18. 2024.

[38] Yang, Zicong, Aitong Jin, Yu Li, Xuyi Yu, Xi Xu, Junxi Wang, Qiaolin Li, Xiaoyan Guo, and Yan Liu. "A coordinated adaptive multiscale enhanced spatio-temporal fusion network for multi-lead electrocardiogram arrhythmia detection." *Scientific Reports* 14, no. 1 (2024): 20828.