Physiological Signal-Based Early Arrhythmia Prediction Using Hybrid AI Models

Chanchal G. Agrawal (Author1)
dept. Computer Engineering ,Vishwakarma Institute of Information Technology, Pune
e-mail: kedia.chanchal2009@gmail.com

Nilesh J. Uke (Author2)
dept. Computer Engineering , Indira College of Engineering & Management Research, Pune
e-mail: nilesh.uke@gmail.com

Abstract
Early detection of heart arrhythmia is critical for reducing risks associated with cardiovascular diseases. This study presents a novel machine learning-based approach for arrhythmia prediction by leveraging a combination of electrocardiogram signal features, circulatory parameters, biochemical markers, structural heart characteristics, and lifestyle factors. The dataset undergoes pre-processing through feature validation, missing data elimination, and min-max normalization to ensure consistency and uniformity. To enhance model robustness, a generative adversarial network is employed to generate synthetic non-disease data, thereby improving class balance. The feature dimensionality is reduced using principal component analysis while preserving essential variance. A hybrid classification framework is implemented, integrating multilayer perceptron, k-nearest neighbors to classify arrhythmia cases with high accuracy. Additionally, long short-term memory networks are utilized to process sequential electrocardiogram data, ensuring effective feature learning without gradient vanishing. Performance evaluation is conducted using metrics such as accuracy, precision, recall, and R-square values to assess model reliability. The results demonstrate that the proposed system significantly enhances early arrhythmia prediction, facilitating timely medical intervention.

Keywords: Heart Arrhythmia Detection, ECG Signal Processing, Generative Adversarial Network, Principal Component Analysis, Long Short-Term Memory, Machine Learning Classification
How to cite this article: Agrawal CG, Uke NJ. Physiological Signal-Based Early Arrhythmia Prediction Using Hybrid AI Models. Int J Drug Deliv Technol. 2026;16(49s): 1249-1260. DOI: 10.25258/ijddt.16.49s.138

1. Introduction

Cardiovascular diseases, particularly heart arrhythmias, pose a significant global health burden, often leading to severe complications such as stroke, heart failure, and sudden cardiac arrest**.** Early prediction and timely intervention can significantly reduce mortality rates and improve patient outcomes. Detecting arrhythmias requires analysing a diverse set of cardiovascular parameters**,** including ECG signals, circulatory system metrics, biochemical markers, and structural heart properties**.** Traditional diagnostic methods rely on manual ECG interpretation and clinical assessments, which may be time-consuming, subjective, and prone to errors**.** The integration of machine learning and deep learning techniques has revolutionized the field by offering automated, efficient, and highly accurate diagnostic solutions**.** This study aims to develop an advanced predictive model capable of detecting arrhythmia at an early stage, leveraging a combination of signal processing, feature engineering, and classification algorithms to enhance detection accuracy.

The application of artificial intelligence (AI) and machine learning (ML) in cardiovascular research has expanded significantly, providing new opportunities for real-time monitoring, risk stratification, and personalized treatment**.** Machine learning-based approaches can identify patterns and anomalies in ECG signals, making them highly effective for early detection of arrhythmias. The demand for intelligent healthcare systems that can analyse and predict cardiovascular conditions has grown with the increasing prevalence of wearable health devices. However, developing such models presents challenges, including data quality, feature selection, computational complexity, and

the need for interpretable AI models**.** Future research must focus on integrating multi-modal medical data, ensuring patient privacy, and enhancing model generalization across diverse populations to enable widespread clinical adoption.

This study proposes a hybrid machine learning framework for early heart arrhythmia detection by utilizing a comprehensive set of cardiovascular parameters. The methodology begins with data pre-processing**,** which includes feature validation, elimination of missing values, and min-max normalization to standardize feature scales. To improve dataset robustness, a generative adversarial network (GAN) generates synthetic non-disease data, balancing class distribution. Principal component analysis (PCA) is then applied to reduce dimensionality while preserving essential variance in the dataset. For classification, a hybrid model combining multilayer perceptron (MLP), k-nearest neighbours (KNN), and support vector machine (SVM) is employed to maximize predictive accuracy. Additionally, long short-term memory (LSTM) networks are incorporated to process sequential ECG data, ensuring effective feature learning and mitigating gradient vanishing issues. The proposed model is trained and tested using a 9:1 data split, with performance evaluation conducted using accuracy, precision, recall, and R-square values. The results demonstrate that the proposed approach significantly enhances arrhythmia prediction accuracy, making it a promising tool for real-world cardiovascular diagnostics**.**

The structure of this paper is as follows: Section 2 provides an overview of related work, summarizing existing research and advancements in heart arrhythmia prediction. Section 3 explains the PHAD - AI-Driven Early Prediction of Heart Arrhythmia Using Hybrid Machine Learning and Deep Learning Models to enhance network efficiency. Section 4 presents the results and discussion, detailing the implementation process and evaluating the performance of the proposed approach. Finally, Section 5 concludes the paper by highlighting key findings and suggesting potential directions for future research.

2. Related Works

Machine learning algorithms' ability to search massive data sets for detailed patterns has propelled their rise in cardiovascular disease prediction. More accurate and powerful predictive models have been constructed combining electronic health records, imaging data, genetic information, and lifestyle factors using supervised learning approaches such logistic regression, support vector machines, random forests, and neural networks [15]. Collecting cardiac arrhythmia data from electrocardiogram (ECG) or electrocardiogram is one of the most critical and vital methods in diagnosing heart diseases. In electrocardiography, the electrical information of heart currents is stored in a timed format in several tabular data formats. ECG data include information on the heart's electrical currents over time, which can provide helpful information to identify heart diseases [13]. The development of a mobile application based on the best-performing ML model marks a significant step towards practical, real-world implementation. This app enables users to input symptoms and quickly receive a heart disease prediction, offering a user-friendly, cost-effective tool for early detection. The translation of our research findings into a tangible product underscores the novelty of our study by bridging the gap between theoretical research and practical healthcare solutions [11]. The practice of task offloading, which involves moving computationally demanding activities to edge computing nodes that are located close to the operational region of the UAV, is a typical solution that is used to reduce latency and conserve onboard resources. Using machine learning algorithms for real-time data processing is one alternative strategy that may empower UAVs to make autonomous judgments in response to sensor inputs and ambient variables better [16]. This methods use machine learning techniques to prevent the problems derived from statistical analysis methods that fail to capture prognostic information in large datasets containing multi-dimensional inter-actions. Some of these papers have generally benefited from large datasets that allow detection of existing diseases thanks to historical data over a long period of time [8]. The growing availability of consumer electronics equipped with increasingly reliable algorithms to monitor cardiac rhythm, including smartphones and wearable's, offers accessible methods to detect rare and short episodes of AF and to estimate arrhythmia burden. The data on AF burden and its relation to outcomes, outlines findings that can help today's shared decision-making with patients with AF, and identifies

research and innovation opportunities [5]. ML is an AI subfield that distinguishes itself from classical mathematical algorithms by including a "learning" component gleaned from massive datasets. There has been a lot of buzz about how CVD and AI may work together to revolutionize cardiovascular health diagnostics, prognoses, and treatments. The rapid detection and diagnosis of CVDs, together with the prediction of outcomes and evaluation of prognosis, may be greatly assisted by AI [1]. Biomarkers are a powerful tool for identifying high-risk people, promptly and reliably diagnosing illness conditions, and effectively diagnosing and treating patients. Most other biomarkers seem to be up-regulated or down-regulated in illnesses other than cardiovascular disease (CVD). Current trends in identifying the number of bio-markers that may consistently aid prognosis in the early stages of the disease are limited [6]. The important role of psychological stress has been previously investigated in myocardial ischemia, coronary artery disease, acute and reversible cardiomyopathy and ventricular arrhythmias. Prospective evidence implies that depression plays a major role in CVD development. The association of depression has also been reported with coronary artery disease, heart rate reactivity, myocardial infarction and respiratory sinus arrhythmia fluctuation [2]. An important element of preventive medicine is individually optimized vaccination schedules to ensure timely and effective vaccination against infectious diseases, the elimination of vaccination gaps. As a part of preventive medicine, early interventions for chronic diseases are viable and support AI in identifying people at risk of developing chronic diseases, confirming their risks/symptoms and implementing early interventions to avoid or mitigate these conditions [9]. ML can integrate and interpret data from different domains in settings where conventional statistical methods may not be able to perform. ML techniques has been studied in different aspects of medicine, including electronic health records, diagnosis, risk stratification, timely identification of abnormal heart rhythms in the intensive care unit on prognosis and guidance of personalized management [14]. Neural networks were originally inspired by the biology of the human brain in which interconnected neurons send and receive signals. Neurons or nodes within artificial neural networks contain non-linear mathematical functions to control which signals are sent to subsequent layers of the model, which are adjusted during training to optimize for an outcome of interest [3]. The Intelligent Cardiovascular Disease detection based on Ant Colony Optimization with Enhanced Deep Learning (ICVD-ACOEDL) model is presented in this research proposal. It uses innovative techniques to improve the detection of cardiovascular disease. This model is novel in that it combines hyper parameter optimization with Bayesian optimization to optimize and expedite the learning process, along with feature selection via ant colony optimization (ACO) [7]. Unsupervised learning is a data processing method that achieves the classification of samples by data analysis of a large number of samples of the object under study without category information, including clustering algorithms and association rule-learning algorithms. Reinforcement learning could be considered a combination of supervised and unsupervised learning, and it could facilitate errors and trials to magnify the accuracy of algorithms [12]. An ANN is a computer network designed to function like a human nervous system. It is a network with multiple layers, each of which contains either input neurons or hidden neurons or output neurons. He training phase of their development, ANNs are inherently adaptable networks, constantly modifying both their internal structure and the information flowing across the system [4]. Cardiac arrhythmias, encompassing conditions such as atrial fibrillation, are among the leading causes of concern in cardiovascular health. The insidious nature of these conditions, often manifesting asymptomatically or with minimal symptoms, renders them particularly elusive to standard detection methods. The stakes of such undetected irregularities are alarmingly high, with potential outcomes ranging from debilitating strokes to heart failures and, in the most severe instances, culminating in sudden cardiac death [10].

3. Data Pre-processing and Feature System

The early detection of heart arrhythmia is a crucial step in preventing severe cardiovascular complications, as timely diagnosis allows for appropriate medical interventions. This process begins with the extraction and validation of critical cardiovascular parameters from the input dataset, which serves as the foundation for predictive modeling. The dataset comprises a diverse set of features, including ECG signal

attributes, blood pressure measurements, biochemical markers, and structural heart parameters, each of which provides valuable insights into the cardiac function and overall circulatory health of an individual. These fields undergo preprocessing to ensure completeness and consistency before being used for predictive modeling as shown in Figure.1.

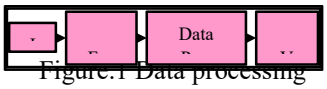

Figure.1 Data processing

3.1 Feature Validation

- **ECG Features:** HRV, RR Interval, QRS Duration, QT Interval, PR Interval, ST Segment Deviation, P-wave Morphology, and T-wave Inversion.
- **Blood Pressure & Circulatory Parameters:** Systolic and Diastolic BP, Pulse Pressure, and Mean Arterial Pressure.
- **Oxygenation & Respiratory Parameters:** Oxygen Saturation (SpO_2) and Respiration Rate.
- **Biochemical & Blood Parameters:** Serum Electrolytes, Blood Sugar, Cholesterol Levels, and CRP levels.
- **Structural & Functional Heart Parameters:** LVEF, Left Atrial Size, and Ventricular Wall Thickness.
- **Lifestyle & Clinical History:** Age, Gender, Smoking, Alcohol Consumption, Obesity, Physical Activity, Family History, Hypertension, Stroke, and Myocardial Infarction.

Feature validation ensures that the selected parameters contribute meaningfully to predictive modelling and clinical decision-making as shown in Figure.2. It involves assessing the reliability, relevance, and statistical significance of each feature in relation to heart disease classification. Methods such as correlation analysis, and recursive feature elimination help identify redundant or less informative variables.

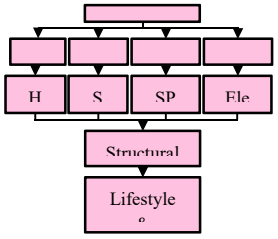

Figure.2 Feature Validation

3.2 Data Normalization

Data normalization is a crucial pre-processing step in machine learning, ensuring that all input features are scaled to a common range, typically between 0 and 1. This process helps eliminate disparities in feature magnitudes, allowing machine learning models to converge faster and perform optimally. When working with cardiovascular datasets containing parameters such as ECG features, blood pressure levels, oxygenation rates, biochemical markers, and lifestyle factors, normalization ensures that numerical values remain consistent across different scales, reducing bias toward features with larger magnitudes.

One of the most widely used normalization techniques is min-max normalization, which rescales each numerical feature to a defined range, often [0, 1]. This transformation ensures that all values fall within a standardized interval while preserving the relationships and distributions within the data. The min-max normalization formula is defined as:

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

where: X′ is the normalized value, X is the original value of the feature, X_{min} is the minimum observed value of the feature, X_{max} is the maximum observed value of the feature.

By applying this transformation, all feature values are scaled within the range of 0 to 1, ensuring that no single feature dominates the model due to its numerical magnitude. For example, consider two features: heart rate variability (HRV), which may range from 20 to 150 ms, and blood pressure values, which may range from 80 to 180 mmHg. Without normalization, models may assign greater importance to features with larger numerical values simply due to their scale. However, after min-max normalization, both features will be mapped within the range [0, 1], preventing bias in the learning process.

In addition to improving numerical stability, normalization enhances model convergence in gradient-based optimization methods, such as neural networks, by preventing large weight updates that can lead to unstable training. Furthermore, distance-based machine learning models, such as KNN benefit from normalization as it ensures that no single feature disproportionately influences distance calculations.

3.3 K-Nearest Neighbours (KNN) Algorithm

The KNN algorithm is a non-parametric, instance-based learning method used for classification. In this example, we visualize a classification task in a 2D feature space. The dataset consists of two distinct classes (blue and red points), and a test point (green) whose class needs to be determined. Using KNN with k=5, the test point's classification is decided based on the majority class of its five nearest neighbors, which are connected to it by dashed lines. This approach ensures that the classification is based on local proximity rather than a global model as shown in Figure.3.

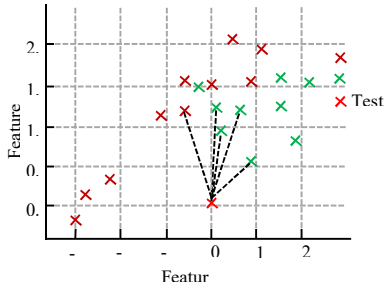

Figure.3 K-Nearest Neighbors classification

KNN is a supervised learning algorithm that classifies a data point based on the majority class among its k nearest neighbors. The classification is determined using distance measures like Euclidean Distance.

3.3.1 Euclidean Distance Metrics for KNN

The most common distance metric used in KNN is Euclidean distance, which calculates the straight-line distance between two points in an n-dimensional space:

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(X_i - Y_i)'}$$

where: $d(x,y)$ is the Euclidean distance between two data points X and Y, X_i and Y_i are the feature values of the two points,

- n is the number of features.

3.3.2 KNN Classification Rule

The class of a test sample X is determined by the majority vote among its k nearest neighbours:

$$y' = \underset{c}{\operatorname{argmax}} \sum_{i \in N_k} I(y_i = c)$$

where: y' is the predicted class label, c represents each possible class, $N_k$is the set of the kk nearest neighbors, y_i is the class label of neighbor I is an indicator function that equals 1 if $y_i = c$, otherwise 0.

For the heart arrhythmia prediction model, all numerical fields ECG features, blood pressure readings, biochemical markers, and oxygen saturation levels are normalized using min-max scaling before being fed into the machine learning pipeline. This ensures that all features contribute equally to the learning process, resulting in a more accurate and reliable prediction model.

Algorithm for K-Nearest Neighbours (KNN)

Load the Dataset

Import the dataset containing labelled training samples.

Extract feature vectors and corresponding class labels.

Define the Distance Metric

Use a distance function such as Euclidean Distance:

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(X_i - Y_i)'}$$

Choose the Value of k

Select an appropriate number of neighbours (odd values often preferred for binary classification).

Compute Distances

Calculate the distance between the test sample and all training samples.

Store distances and corresponding class labels.

Identify k Nearest Neighbours

Sort the computed distances in ascending order.

Select the k closest training samples.

Perform Majority Voting

Count occurrences of each class label among the k neighbours.

Assign the most frequent class label to the test sample:

$y' = \underset{c}{\operatorname{argmax}} \sum_{i \in N_k} I(y_i = c)$

Classify the New Data Point

- Return the predicted class label.

3.4 GAN for Synthetic Data Augmentation

A GAN is employed to formulate synthetic non-disease data, effectively augmenting the dataset with realistic representations of healthy individuals. This approach enhances the robustness of the model by addressing potential data imbalances and improving its generalization

capability. GANs operate through a competitive process between two neural networks: the generator and the discriminator. The generator is responsible for creating synthetic samples that mimic the real data, while the discriminator evaluates whether the given sample is real (from the original dataset) or fake (generated by the model).

The objective of the generator is to learn the underlying distribution of the input dataset and produce highly realistic synthetic samples. Meanwhile, the discriminator continuously refines its ability to distinguish between real and fake data. The adversarial nature of GANs results in an iterative improvement of both networks, leading to the generation of high-quality synthetic data as shown in Figure.4.

Figure.4 GAN Augmentation

A GAN consists of two components:

1. Generator (G): A neural network that generates new samples resembling the original data distribution.
2. Discriminator (D): A neural network that classifies input data as either real or generated.

The objective function of a GAN is represented as a min-max game between these two networks:

$$\min_G \max_D V(D,G) = E_{x\sim \text{pdata(x)}}\big[logD(x) + E_{x\sim \text{px(z)}}[\log(1 - D(G(z)))]\big]$$

where: $x \sim \text{pdata}(x)$ represents real samples from the dataset. $z \sim \text{pz}(z)$ represents a latent variable (random noise) sampled from a prior distribution (e.g., Gaussian distribution). G(z) is the synthetic sample produced by the generator. D(x) is the probability that xx is a real sample. D(G(z)) is the probability that the generated sample is real.

3.4.1 Training Process of GANs

The discriminator in a GAN plays a vital role in distinguishing between real and generated (fake) data samples. During training, the discriminator is exposed to both real samples from the actual dataset and synthetic samples produced by the generator. The primary objective of the discriminator is to maximize the probability of correctly classifying real data as genuine and generated data as fake. This is achieved by optimizing its parameters through gradient-based learning, thereby enhancing its ability to differentiate between the two data distributions. The discriminator employs Adam optimizer to adjust its weights and minimize the classification error. As training progresses, the discriminator strengthens its ability to detect fake samples, forcing the generator to improve the realism of its outputs, ultimately leading to a balanced adversarial process.

$$L_D = -(E_{x\sim \text{pdata(x)}}[Log\ D(x)] + E_{x\sim \text{px(z)}}[\log(1 - D(G(z)))])$$

The generator in a GAN is responsible for producing synthetic data that closely resembles real samples from the dataset. Unlike the discriminator, which aims to distinguish real from fake data, the generator is trained to create data that fools the discriminator into classifying it as real. To achieve this, the generator updates its parameters in a way that minimizes the probability of the discriminator correctly identifying its generated samples as fake. The training process involves backpropagation, where the gradient of the discriminator's feedback is used to refine the generator's output. By updating its parameters using gradient ascent on $D(G(z))$, the generator gradually improves, producing more realistic samples. Over time, as the discriminator becomes better at detection, the generator is forced to generate even more authentic outputs, leading to an adversarial learning process that enhances the overall model performance.

$$L_G = -E_{x\sim \text{px(z)}}[\log\big(D\big(G(z)\big)\big)$$

3.4.2 GAN-Based Data Augmentation in Cardiovascular Disease Prediction

- **Addresses Data Imbalance:** Many medical datasets have an imbalance between diseased and non-diseased cases. By generating additional synthetic healthy samples, the model is trained on a more balanced dataset.
- **Enhances Model Generalization:** Training on synthetic and real data prevents overfitting, improving the model's performance on unseen data.
- **Preserves Privacy:** GANs can generate synthetic patient records that resemble real data without exposing sensitive information, making them useful for privacy-preserving medical research.

By integrating GANs for non-disease data generation, the dataset becomes more diverse, leading to improved early detection models for heart arrhythmia.

3.5 Model Architecture and Training

Dataset Splitting: The dataset is divided into a training set (90%) and a testing set (10%). This ensures an adequate number of samples for model generalization.

3.5.1 Dual Network Architecture

The proposed deep learning model consists of two networks:

1. Operational Network – Consists of an encoder, a projector, and a predictor.
2. Goal Network – Contains the same architecture as the online network but with different weights.

The operational network updates its weights using backpropagation, while the goal network employs an exponential moving average (EMA) to stabilize learning:

$$W_{goal} = \alpha W_{goal} + (1 - \alpha) W_{operational}$$

where α is the decay rate for EMA.

3.5.2 Multi-Layer Perceptron (MLP)

A MLP is employed as a key component of the classification model for early detection of heart arrhythmia. MLP is a type of ANN consisting of an input layer, one or more hidden layers, and an output layer. Each neuron in a layer is connected to neurons in the subsequent layer through weighted connections, with activation functions such as ReLU (Rectified Linear Unit) or sigmoid applied to introduce non-linearity. The network learns by adjusting weights using backpropagation and the gradient descent optimization algorithm, minimizing the error between predicted and actual values. The MLP model processes extracted cardiovascular parameters including ECG features, blood pressure, biochemical markers, and structural heart attributes—to identify patterns indicative of arrhythmia. By integrating MLP with feature extraction techniques of LSTM, the model enhances its predictive accuracy, making it effective for robust cardiovascular risk assessment. The MLP consists of the projector and predictor modules both having one hidden layer:

$$Y = f(WX + b)$$

where W represents the weights, X the input features, b the bias term, and f is the activation function.

This flowchart represents the workflow of the MLP model for early heart arrhythmia detection. It starts with input data processing, passes through multiple hidden layers with activation functions, adjusts weights using backpropagation, minimizes errors with gradient descent, and finally predicts arrhythmia as shown in Figure.5.

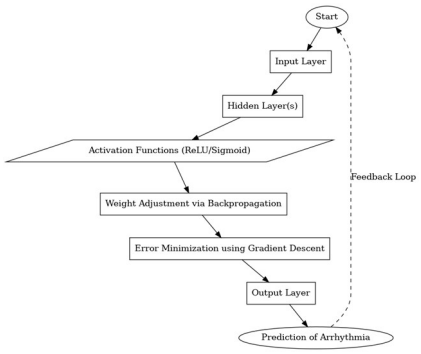

Figure.5 MLP Perceptron

3.5.3 Data Augmentation

To enhance the model's ability to generalize across diverse input variations, a random cropping technique is applied during the training process. This technique involves randomly selecting and retaining only a subset of the input data while masking certain tokens with a predefined probability. By intentionally removing or obscuring portions of the input, the model is encouraged to focus on learning robust and meaningful representations rather than relying on specific details that may not generalize well to unseen data. This approach helps mitigate overfitting, ensuring that the model remains adaptable to different variations of the data. Additionally, the random cropping technique introduces a level of data augmentation, making the model more resilient when encountering real-world variations in cardiovascular signals or clinical parameters.

3.5.4 Loss Function Optimization

The model minimizes the mean squared error (MSE) loss function:

$$L = \frac{1}{n}(y_{pred,i} - y_{true,i})$$

where $y_{pred,i}$represents the predicted output and $y_{true,i}$ is the actual value.

3.6 Feature Extraction using PCA and LSTM

3.6.1 Principal Component Analysis

PCA is a statistical technique applied in this study to reduce the dimensionality of cardiovascular datasets while preserving the maximum possible variance. High-dimensional medical data, such as ECG features, blood pressure parameters, and biochemical markers, often contain correlated features, leading to redundancy. PCA addresses this issue by transforming the original correlated features into a new set of uncorrelated variables, known as PCs**,** which capture the most significant variations in the data. The transformation is mathematically expressed as:

$$Z = XW$$

where: X represents the standardized data matrix by subtracting the mean and scaled by the standard deviation), W consists of the eigenvectors of the covariance matrix of XX, which define the new feature space, Z is the transformed data in the new principal component space.

3.6.2 The process of PCA involves the following steps:

PCA is a widely used technique for reducing the dimensionality of medical datasets while preserving the most important information. Since different cardiovascular and biochemical parameters are measured on different scales, standardization of data is performed as the first step to ensure that all features contribute equally to the analysis. This is typically done using z-score normalization, which rescales the data to have a mean of zero and a standard deviation of one.

Once standardized, the covariance matrix of the dataset is computed to analyse relationships between different features. This matrix captures how features vary together**,** allowing the identification of correlated variables. The next step involves computing the eigenvalues and eigenvectors of the covariance matrix. Eigenvalues indicate the amount of variance captured by each principal component, while eigenvectors define the directions of these components in the feature space.

The principal components are then ranked in descending order based on their eigenvalues**,** and a subset of the most significant components is selected. This selection ensures that sufficient variance is retained while reducing the complexity of the dataset. Finally, the original dataset is projected onto the selected principal components**,** transforming it into a lower-dimensional space where redundant or less informative features are eliminated. This process enhances the efficiency of machine learning models by reducing computational overhead and improving generalization, particularly in high-dimensional medical datasets as shown in Figure.6.

Figure.6 PCA Feature Extraction

PCA significantly improves model performance by reducing computational complexity and mitigating the curse of dimensionality while retaining essential features. In the context of heart arrhythmia detection, PCA aids in refining cardiovascular parameters, ensuring that only the most informative features contribute to the prediction model.

3.6.3 Long Short-term Memory Network

To address gradient disappearance issues, an LSTM network processes the extracted features. The LSTM unit consists of: In a LSTM network, several gating mechanisms regulate the flow of information to effectively capture long-term dependencies in sequential data. The input gate controls how much new information is added to the cell state, allowing the network to update its memory based on relevant features from the input sequence. Simultaneously, the forget gate determines which past information should be discarded, ensuring that irrelevant or outdated data does not interfere with learning. The cell state acts as the memory unit, preserving information over time and enabling the model to retain essential patterns from previous time steps. Finally, the output gate processes the updated cell state and determines the final output using a combination of sigmoid and tanh activation functions, ensuring a balanced representation of the learned features. These mechanisms work together to address the limitations of traditional recurrent neural networks (RNNs), preventing issues like gradient vanishing while effectively modelling long-term dependencies in cardiovascular and ECG signal analysis. The hidden state h_t is computed as:

$$h_t = o_t tanh.(c_t)$$

Where, o_t is the output gate activation, and c_t is the cell state.

3.7 Regression Models for Disease Prediction

3.7.1 Simple Linear Regression (SLR)

SLR predicts the presence of arrhythmia using one independent variable:

$$Y = \beta_0 + \beta_1 + X + \varepsilon$$

where Y is the dependent variable, X is the independent variable, β_0 , β_1 are regression coefficients and ε is the error term.

3.7.2 Multiple Linear Regression (MLR)

MLR extends SLR by incorporating multiple independent variables:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \dots \dots \beta_n X_n . + \varepsilon$$

where X_1, X_2, ..., X_n are the predictor variables.

3.7.3 Model Evaluation using RSquare

The quality of the regression models is evaluated using the Rsquare metric:

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

where SS_{res} is the residual sum of squares and SS_{tot} is the total sum of squares.

The proposed methodology integrates feature engineering, deep learning, PCA, LSTM, and regression models to enhance the early prediction of heart arrhythmia. This ensures a robust and accurate classification system, aiding in risk reduction and timely intervention.

4. Results and Discussion

Accuracy					F1 score			
Rounds	**KNN**	**ESTM**	**GAN**	**PHAD**	**KNN**	**ESTM**	**GAN**	**PHAD**
5	0.738	0.768	0.822	0.924	0.785	0.723	0.77	0.881
10	0.734	0.775	0.816	0.936	0.788	0.715	0.762	0.853
15	0.767	0.782	0.814	0.923	0.813	0.742	0.788	0.838
20	0.717	0.788	0.808	0.885	0.817	0.728	0.779	0.874
25	0.763	0.783	0.842	0.923	0.785	0.712	0.762	0.863
30	0.	0.7	0.	0.9	0.	0.7	0.	0.8
	762	83	839	38	792	14	778	72

Table.1 Accuracy and F1 score

Precision					Error rate			
Rounds	**KNN**	**ESTM**	**GAN**	**PHAD**	**KNN**	**ESTM**	**GAN**	**PHAD**
5	0.726	0.776	0.842	0.911	0.25	0.238	0.191	0.128
10	0.711	0.759	0.813	0.889	0.26	0.247	0.198	0.132
15	0.738	0.793	0.845	0.925	0.27	0.257	0.206	0.138
20	0.716	0.781	0.831	0.893	0.28	0.266	0.213	0.142
25	0.727	0.783	0.836	0.918	0.29	0.276	0.221	0.148
30	0.739	0.794	0.843	0.921	0.27	0.257	0.206	0.138

Table.2 Precision and Error rate

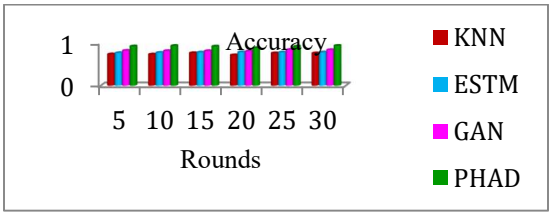

Figure.7 Rounds vs Accuracy

The performance of the proposed PHAD model, along with KNN, ESTM, and GAN-based models, was evaluated across different rounds to analyse their accuracy in detecting heart arrhythmia. The results indicate that PHAD consistently outperformed the other models in all tested rounds, demonstrating superior predictive capabilities. Across the rounds, KNN exhibited moderate accuracy, fluctuating between 0.717 and 0.767, indicating its dependency on local distance-based classification, which may not generalize well to complex feature spaces. The ESTM model showed slightly better performance, ranging from 0.768 to 0.788, reflecting its enhanced ability to capture temporal dependencies in cardiovascular parameters. The GAN-based model performed significantly better, achieving a peak accuracy of

0.842 at round 25, which validates the effectiveness of synthetic data augmentation in improving classification robustness. However, PHAD consistently achieved the highest accuracy, reaching a peak value of 0.938 at round 30, demonstrating its effectiveness in early heart arrhythmia detection. The integration of hybrid machine learning and deep learning models, including MLP and PCA for feature extraction and dimensionality reduction, contributed to its superior performance as shown in Figure.7 and Table.1. The model's ability to refine cardiovascular parameters, mitigate dimensionality issues, and enhance classification robustness makes PHAD the optimal solution for heart arrhythmia prediction.

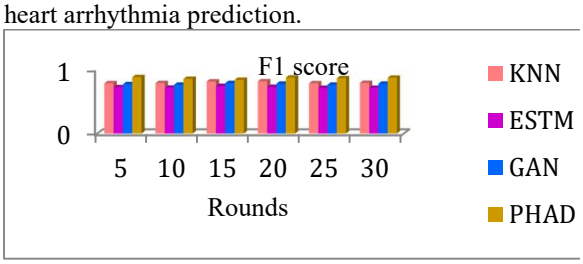

Figure.8 Rounds vs F1 score

The F1 score evaluation highlights the effectiveness of different models in handling the trade-off between precision and recall in heart arrhythmia detection. Across multiple rounds, the PHAD model consistently outperformed KNN, ESTM, and GAN, demonstrating its ability to balance sensitivity and specificity in classification. KNN showed moderate performance, with F1 scores ranging from 0.785 to 0.817, reflecting its capability in handling local decision boundaries but limited adaptability to complex feature distributions. ESTM, on the other hand, had lower F1 scores between 0.712 and 0.742, suggesting its challenges in accurately capturing non-linear dependencies in cardiovascular data. The GAN-based approach improved upon these models, achieving F1 scores between 0.762 and 0.788, validating the role of synthetic data augmentation in enhancing prediction robustness. However, PHAD achieved the highest F1 score of 0.881, demonstrating its superior classification accuracy while minimizing false positives and false negatives. The hybrid approach, integrating MLP, LSTM, PCA, and GAN-generated synthetic data, enhanced feature representation, improved generalization, and ensured robust arrhythmia prediction as shown in Figure.8 and Table.1. The PHAD model's consistently high F1 score underscores its efficiency in cardiovascular risk assessment, making it the most reliable approach for early heart arrhythmia detection.

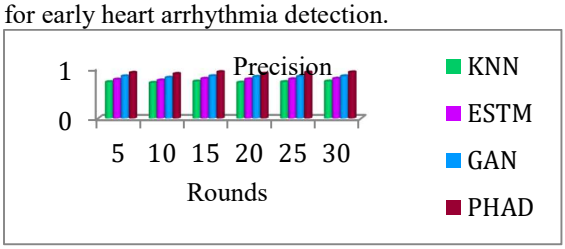

Figure.9 Rounds vs Precision

The precision analysis provides insights into the accuracy of positive predictions made by different models for heart arrhythmia detection. Across multiple rounds, the PHAD model consistently outperformed KNN, ESTM, and GAN, demonstrating its ability to minimize false positives while accurately identifying arrhythmia cases. KNN exhibited moderate precision, ranging from 0.711 to 0.739, indicating its ability to classify positive cases correctly but with limitations in feature discrimination. ESTM showed slightly better results, with precision values between 0.759 and 0.794, benefiting from improved model stability but still lagging behind more advanced approaches as shown in Figure.9 and Table.2. The GAN-based model significantly enhanced precision, achieving values between 0.813 and 0.845, reinforcing the effectiveness of synthetic data augmentation in improving model reliability. Among all models, PHAD consistently achieved the highest precision, reaching a peak value of 0.925, outperforming the others across all rounds. This demonstrates the effectiveness of PHAD's hybrid learning approach, which integrates MLP, LSTM, PCA, and GAN-enhanced data augmentation to optimize feature representation and classification accuracy. The PHAD model's superior precision highlights its capability in reducing false positives, making it the most reliable choice for accurate early heart arrhythmia detection.

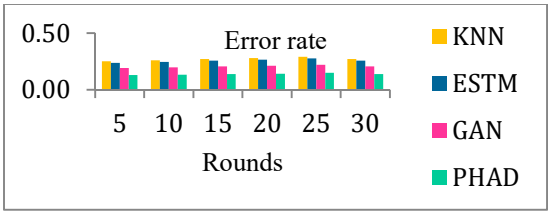

Figure.10 Rounds vs Error rate

The error rate evaluation highlights the effectiveness of different models in minimizing misclassification during heart arrhythmia

detection. Lower error rates indicate better model performance, and across all trials, the PHAD model demonstrated the lowest error rates, making it the most reliable classifier. KNN exhibited the highest error rates, ranging between 0.25 and 0.29, indicating its limitations in handling complex feature interactions. ESTM performed slightly better, with error rates fluctuating between 0.238 and 0.276, but still showed higher misclassification compared to more advanced models. The GAN-based model significantly reduced the error rate, achieving values between 0.191 and 0.221, showcasing the impact of synthetic data augmentation in improving model robustness as shown in Figure.10 and Table.2. Among all models, PHAD consistently achieved the lowest error rates, with values ranging from 0.128 to 0.148, demonstrating its superior classification capabilities**.** The hybrid learning approach of PHAD, integrating MLP, LSTM, PCA, and GAN-enhanced data augmentation, played a crucial role in minimizing misclassifications**.** This reinforces PHAD's reliability for early heart arrhythmia detection, ensuring higher accuracy and lower false predictions compared to traditional machine learning models.

Conclusion

This study presents a machine learning-driven approach for the early detection of heart arrhythmia, integrating multiple cardiovascular parameters to enhance predictive accuracy. The proposed framework systematically pre-processes input data, ensuring feature validation, consistency, and normalization to maintain the integrity of the dataset. By incorporating a GAN, the model effectively generates synthetic non-disease samples, improving class balance and overall robustness. Additionally, PCA is applied to reduce the dimensionality of high-dimensional cardiovascular data while preserving essential variance. For classification, a hybrid model combining MLP, KNN, and SVM was developed to ensure efficient and accurate arrhythmia prediction. Furthermore, the integration of LSTM networks allows the model to process sequential ECG signal features, overcoming limitations such as gradient vanishing and short-term dependency issues. The evaluation metrics demonstrate that the proposed model significantly enhances early arrhythmia detection, ensuring timely medical intervention and risk mitigation**.** Despite the promising results of this study, several areas require further exploration to improve real-world deployment and model generalization. Firstly, integrating additional real-time sensor data from wearable devices could enhance prediction accuracy by incorporating continuously monitored cardiovascular parameters. The inclusion of multi-modal data such as echocardiography, genetic markers, and advanced biochemical parameters could further strengthen the predictive capabilities of the model. Another key direction involves enhancing model interpretability through explainable AI (XAI) techniques, allowing clinicians to understand and trust the model's decision-making process. Additionally, improving privacy-preserving mechanisms such as federated learning and homomorphic encryption can ensure secure data sharing while maintaining patient confidentiality. Future research should also focus on adaptive learning models that can update predictions dynamically based on patient-specific trends and changing cardiovascular conditions. Finally, large-scale clinical validation across diverse populations is crucial to assess the model's reliability in different demographic and geographical contexts**.** By addressing these challenges, future advancements in AI-driven cardiovascular diagnostics can contribute significantly to the early detection and prevention of life-threatening arrhythmias, ultimately reducing global cardiovascular mortality rates**.**

References

[1] Naiela E. Almansouri, Mishael Awe, Selvambigay Rajavelu, Kudapa Jahnavi, Rohan Shastry, Ali Hasan, Hadi Hasan, Mohit Lakkimsetti, Reem Khalid AlAbbasi, Brian Criollo Gutierrez and Ali Haider, "Early Diagnosis of Cardiovascular Diseases in the Era of Artificial Intelligence: An In-Depth Review", 03/09/2024.

[2] Mohsen Dorraki, Zhibin Liao, Derek Abbott, Peter J. Psaltis, Emma Baker, Niranjan Bidargaddi, Hannah R. Wardill, Anton van den Hengel, Jagat Narula, Johan W. Verjans, "Improving Cardiovascular Disease Prediction With Machine Learning Using Mental Health Data", 2024.

[3] Aamir Javaid, Fawzi Zghyer, Chang Kim, Erin M. Spaulding, Nino Isakadze, Jie Ding, Daniel Kargillis and Yumin Gao, "Medicine 2032: The future of cardiovascular disease prevention with machine learning and digital health technology", American Journal of Preventive Cardiology, 28 August 2022.

[4] Jyotismita Talukdar and Thipendra P. Singh, "Early prediction of cardiovascular disease using artificial neural network", Journal of Behavioral Robotics, 2023.

[5] Nina Becher, Andreas Metzner, Tobias Toennis Paulus Kirchhof and Renate B. Schnabel, "Atrial fibrillation burden: a new outcome predictor and therapeutic target", European Heart Journal, 2 July 2024.

[6] Sreenu Thupakula, Shiva Shankar Reddy Nimmala, Haritha Ravula, Sudhakar Chekuri and Raju Padiya, "Emerging biomarkers for the detection of cardiovascular diseases", The Egyptian Heart Journal, 2022.

[7] Biao Xia, Nisreen Innab, Venkatachalam Kandasamy and Ali Ahmadian Massimiliano Ferrara, "Intelligent cardiovascular disease diagnosis using deep learning enhanced neural network with ant colony optimization", 2024.

[8] Maria Teresa Garcia-Ordas, Martin Bayon-Gutierrez, Carmen Benavides, Jose Aveleira-Mata and Jose Alberto Benitez-Andrades, "Heart disease risk prediction using deep learning techniques with feature augmentation", Multimedia Tools and Applications, 2023.

[9] Izabela Rojek, Piotr Kotlarz, Mirosław Kozielski, Mieczysław Jagodzinski and Zbyszko Królikowski, "Development of AI-Based Prediction of Heart Attack Risk as an Element of Preventive Medicine", 7 January 2024.

[10] Yehyun Kim, Myeonggyu Lee, Jaeung Yoon, Yeji Kim, Hyunseok Min, Hyungjoo Cho, Junbeom Park and Taeyoung Shin, "Predicting Future Incidences of Cardiac Arrhythmias Using Discrete Heartbeats from Normal Sinus Rhythm ECG Signals via Deep Learning Methods", 3 September 2023.

[11] Hosam El-Sofany, Belgacem Bouallegue and Yasser M. Abd El-Latif, "A proposed technique for predicting heart disease using machine learning algorithms and an explainable AI method", 2024.

[12] Xiaoyu Sun, Yuzhe Yin, Qiwei Yang and Tianqi Huo, "Artificial intelligence in cardiovascular diseases: diagnostic and therapeutic perspectives", European Journal of Medical Research, 2023.

[13] Homeyra Amiri, Javad Mohammadzadeh, Seyed Mohsen Mirhosseini and Alireza Nikravanshelmani, "Prediction of high-risk cardiac arrhythmia based on optimized deep active learning", IEEE, 25 February 2025.

[14] Cheuk To Chung, George Bazoukis, Sharen Lee, Ying Liu, Tong Liu, Konstantinos P. Letsas, Antonis A. Armoundas and Gary Tse, "Machine learning techniques for arrhythmic risk stratification: a review of the literature", International Journal of Arrhythmia, 2022.

[15] Aman Darolia, Rajender Singh Chhillar, Musaed Alhussein, Surjeet Dalal, Khursheed Aurangzeb and Umesh Kumar Lilhore, "Enhanced cardiovascular disease prediction through self-improved Aquila optimized feature selection in quantum neural network & LSTM model", 20 June 2024.

[16] B. Suganya, R. Gopi, A. Ranjith Kumar and Gavendra Singh, "Dynamic task offloading edge-aware optimization framework for enhanced UAV operations on edge computing platform", 2024.