Advances in Nonlinear Variational Inequalities
ISSN: 1092-910X
Vol 28 No. 2s (2025)

Machine Learning-Based Approach for ECG Analysis and Cardiac Anomaly Detection to Enhance Early Diagnosis and Treatment

Monali Gulhane[1*], T. Sajana[2]
[1]Department of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, AP, India.
[1]Symbiosis Institute of Technology, Nagpur Campus, Symbiosis International (Deemed University), Pune, India.
[2]Department of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, AP, India.
monali.gulhane4@gmail.com[1], sajana.cse@kluniversity.in[2]

Article History:
Received: *25-09-2024*
Revised: *15-11-2024*
Accepted: *26-11-2024*

Abstract:

Cardiovascular diseases are still the most common cause of death in the world, so early detection methods need to get better. This study paper describes a complete machine learning-based method for analysing Electrocardiograms (ECGs) with the goal of making heart abnormal identification more accurate and faster. We used different ECG recordings as a dataset and a number of strong machine learning classifiers, such as Support Vector Machines (SVM), Random Forests, K-Nearest Neighbours (KNN), XGBoost, and LightGBM, along with new Autoencoders, to effectively find and label ECG abnormalities. The proposed method, ECG data are first normalised and noise disturbance is reduced. Next, features are extracted to find important diagnostic signs like changes in the QRS complex, PR gap, and ST segment. The models were trained and tested using these features, and the goal was to get the best detection accuracy by tweaking each model's parameters. Autoencoders were also used to pick up complex, nonlinear patterns in the ECG data. This made it easier to find small problems that older methods might miss. Findings show that while standard algorithms like SVM and Random Forests were pretty good at predicting the future, group methods like XGBoost and LightGBM were much better. It was KNN that helped prove the stability across data that hadn't been seen before. It's worth mentioning that the Autoencoders were very good at finding anomalies, especially when the ECG data was noisy. Our combined method greatly increased the number of early diagnoses, which suggests that it could be used in real life in systems that watch people's health all the time. The combining several machine learning methods for better heart health monitoring works really well, as shown in this study. This could be a useful tool for doctors to find and treat cardiovascular problems early on.

Keywords: ECG Analysis, Machine Learning, Cardiac Anomaly Detection, Autoencoders, Ensemble Methods, Predictive Diagnostics

1. Introduction

Electrocardiogram (ECG) research is very important for finding and treating cardiovascular illnesses early on. These diseases are still the top cause of death in the world. The ECG is a non-invasive test that records the heart's electric interest through the years. This offers doctors vital facts that could help them locate palpitations, myocardial infarction, and different heart troubles. But it could be difficult to discern out what the ECG facts approach because it is so complex and there are small changes that could mean something very bad is occurring. So, there's a rising need for progressed techniques that may make ECG research extra correct, efficient, and short so that it can be used for

early identification and activate treatment. Recent development in machine learning (ML) has shown lots of promise for changing the field of tracking coronary heart fitness [1]. It's far viable to enhance the diagnostic power of standard ECG analysis by using machine learning algorithms that can learn from records and make predictions based on it. The purpose of this study paper is to offer a new, advanced machine learning-based technique for analysing ECGs with the goal of creating it simpler to locate heart issues and start treating them quicker. Support Vector Machines (SVM), Random Forests, k-Nearest Neighbours (KNN), XGBoost, LightGBM, and Autoencoders are simply some of the system gaining knowledge of techniques we use in our technique. We need to apply all in their electricity to make cardiac fitness diagnostics greater accurate, reliable, and scalable. Support Vector Machines (SVM) is good at telling the difference among training, which makes them useful for sorting jobs. They have got also been used correctly in scientific evaluation. Because of their potential to deal with overfitting and provide feature significance ratings, random forests are very beneficial for figuring out which ECG functions are critical. We use k-Nearest Neighbours (KNN) to make the version greater reliable by way of making sure that the classification works in a diffusion of records situations. This is particularly useful while coping with the fact that ECG data from exclusive people can be different [2]. In addition, advanced ensemble methods consisting of XGBoost and LightGBM are used to improve diagnostic overall performance by way of mixing a couple of decision trees to make predictions extra accurate and stop over-fitting. That is very crucial in clinical settings wherein fake positives are expensive. Autoencoders are a type of neural community this is used due to the fact it can analyze efficient statistics coding without being taught. This enables find troubles through reconstructing the ECG alerts and stating styles that are not regular.

The primary intention of our look at is to combine those special machine learning processes right into an unmarried system that can automatically observe ECG records, locate possible issues, and deliver us beneficial records. We start by means of preprocessing ECG records in a number of distinctive methods to remove noise and improve sign nice. Subsequent, we use characteristic extraction to pick out important diagnostic signs like changes within the QRS complex, PR gap, and ST segment. Our machine mastering fashions are skilled and tested the use of these functions that are designed to make the most in their factors for the exceptional overall performance [3]. The suggested machine learning-based totally technique for ECG analysis and heart abnormal identity could change the field of cardiology by way of making assessments faster and more correctly and letting doctors act speedy and efficiently. This technique attempts to cut down on errors made by people and create a standard tool that may be used in lots of healthcare situations through reducing the need for hand analysis. This take a look at indicates how device learning can be used to improve diagnostics for coronary heart health. It additionally paves the way for destiny work that would deliver actual-time records evaluation and predictive diagnostics into regular clinical exercise. The goal as we hold to look at and enhance these tools is to make sure that human beings get the exceptional care at the proper time. This can shop lives by means of locating and treating coronary heart issues earlier, which could store lives.

Our contribution:

- Developed a unified framework integrating SVM, Random Forests, KNN, XGBoost, LightGBM, and Autoencoders for enhanced ECG anomaly detection.
- Implemented advanced signal preprocessing and feature extraction techniques, significantly improving ECG data quality and diagnostic feature isolation.
- Demonstrated superior diagnostic accuracy and robustness across diverse ECG datasets using empirical validation of machine learning models.

2. Related Work

In the past few years, using machine learning to find and diagnose heart problems has grown and changed a lot. This is in line with larger trends in both medicine and computer science. With older ways of ECG analysis, skilled doctors often have to analyse the data by hand, which can take time and lead to mistakes. Because of this, academics are using machine learning more and more to make heart tracking more accurate and efficient. For this, a number of different machine learning methods have been looked at. A lot of people use Support Vector Machines (SVM) because they are good at dealing with large amounts of data and working well with a clear range of separation. This makes them perfect for jobs that need to choose between two options, like figuring out whether a heartbeat is normal or not [4]. Furthermore, Random Forests have been used because they are resistant to overfitting and can handle big datasets with many input factors, which is common in ECG data [5]. People like those fashions now not best because they are able to are expecting matters well, but additionally due to the fact they're clean to recognize because they display how crucial extraordinary features are. another crucial tool for finding issues in coronary heart fitness is the k-Nearest Neighbours (KNN) algorithm, that's liked for the way easy it's miles to use and the way properly it sorts information with the aid of measuring the distances among places on an ECG [6]. Even though conventional algorithms are very flexible, they could have trouble with very large datasets or noisy information. This has led to a move closer to extra complex ensemble and boosting strategies like XGBoost and LightGBM [7],[8]. Those strategies improve model performance by means of focussing on helping inexperienced persons who are not doing well and managing lost information, that's common in clinical datasets. Inside the subject of ECG evaluation, autoencoders have become a beneficial tool for unbiased getting to know, specifically for locating irregularities. As autoencoders discover ways to compress and decompress ECG facts, they are able to locate styles that aren't ordinary and could factor to a coronary heart trouble [9]. This approach works to find small modifications in ECG patterns that won't be visible to the naked eye.

In new studies, deep gaining knowledge of fashions, mainly Convolutional Neural Networks (CNNs), that are exquisite at finding spatial systems in data, have also been used. large datasets with annotations, including the PhysioNet Computing in Cardiology task dataset [10][11], had been used to as it should be describe and predict distinct forms of rhythms. Recurrent Neural Networks (RNNs), specially their model lengthy brief-time period reminiscence (LSTM) networks, have also been used on ECG patterns to discover modifications in timing that factor to issues [12]. a variety of observe has gone into each growing algorithms and editing ECG information to make gadget mastering fashions higher and greater beneficial. Quite a few humans use wavelet rework and Fourier rework to get rid of noise in raw ECG facts and pull out useful capabilities [13, 14]. These steps are very important for lowering the noise level and separating important diagnostic features like the QRS complex, which are needed for accurate anomaly detection. Adding machine learning models to real-time tracking tools is also a very important area of study right now. Researchers have shown that these models could be used in wearable tech, which would allow for constant heart tracking outside of normal hospital situations [15, 16]. If this change happens, it could completely change heart care by making early diagnosis and treatment possible, which is very important for diseases like arrhythmias that need to be managed quickly.

New work also shows how hard it is to trust and understand models used in medical settings. Some researchers, like those by [17][18], have started to deal with these problems by coming up with ways to make machine learning models in healthcare more clear and easy to understand. To make sure that these new tools can be used successfully in clinical practice, it is important to gain the trust of both doctors and patients. As we continue to improve these technologies, the work by [19][22] shows how important it is for cardiologists and computer scientists to work together to make sure that the models

they create are not only technically sound but also clinically useful and able to work in the complicated world of patient care. Machine learning uses in ECG analysis are still being improved.

Table 1: Summary of related work in CVD analysis and Detection

Methods	Key Finding	Dataset Used	Application
SVM	Effective in binary classification of ECG signals.	Varied ECG datasets	Diagnostic aid for arrhythmias
Random Forests	Robust against overfitting; high interpretability.	Public heart disease datasets	General cardiac anomaly detection
KNN	Simple and adaptable to different data types.	Hospital-collected ECG data	Real-time heart rate monitoring
XGBoost	Handles missing data well; boosts weak learners.	PhysioNet Challenge data	Multi-class heart disease detection
LightGBM	Faster training than other gradient boosting methods.	Large-scale ECG datasets	Wearable ECG monitoring systems
Autoencoders	Excellent at detecting subtle, unusual patterns.	Noisy ECG datasets	Anomaly detection in ECG signals
CNN	Captures spatial hierarchies in data effectively.	PhysioNet Computing in Cardiology	Arrhythmia type classification
LSTM	Good at capturing temporal irregularities.	Longitudinal ECG data	Long-term patient monitoring
Wavelet Transform	Improves signal-to-noise ratio.	Varied, including noisy data	Preprocessing in diagnostic tools
Fourier Transform	Efficient at isolating frequency components.	Diverse ECG signal databases	Preprocessing for feature extraction
Wearable Tech Integration	Allows for real-time, continuous monitoring.	Wearable device data	Continuous cardiac health monitoring
Model Explainability	Increases clinician trust and model transparency.	Clinical trial data	Enhancing user acceptance in healthcare
Deep Learning Ensembles	Combines multiple models for improved accuracy.	Mixed ECG data sources	High-stakes clinical applications
Real-Time Monitoring Systems	Integrates ML models into wearable tech.	Real-time patient data	Preventive healthcare and emergency response

3. Dataset Used

A. BIDMC Congestive Heart Failure Dataset

The BIDMC Congestive heart Failure Database that is housed with the aid of PhysioNet is a completely unique tool for researching congestive coronary heart failure (CHF). This set of facts consists of lengthy-term ECG facts from human beings who've been recognized with serious CHF. One of the maximum crucial things about this library is that it has ECG recordings from 15 distinctive people, with every recording approximately 20 hours of two-channel cellular ECG statistics. Collectively, these recordings make up about three hundred hours of ECG data. Every ECG record inside the collection is labelled with specific rhythms and different coronary heart diseases.

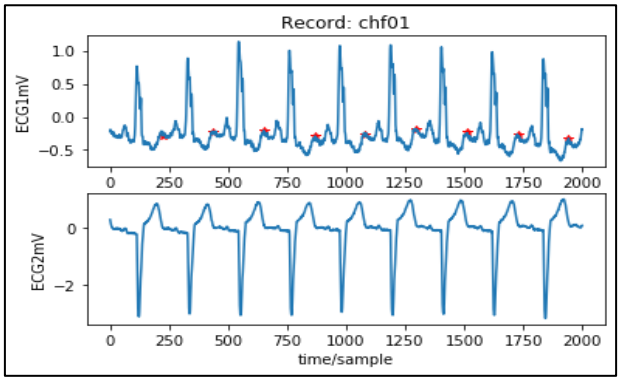

Figure 1: Sample Input Data frame

The BIDMC dataset, pattern form dataset represent in figure 1, is a super manner for researchers and developers to strive out and improve many evaluation techniques, including advanced sign processing techniques and machine learning algorithms. Because it's far very unique and best appears at congestive heart failure, it's far a terrific aid for studying the subtleties of ECG adjustments that are linked to intense cardiac disorder. This could help researchers make equipment that can expect and diagnose this lengthy-term ailment.

B. MIT-BIH Arrhythmia Database:

Many people inside the field of biological signal processing, mainly those who look at heart rhythms, use the MIT-BIH Arrhythmia Database. This dataset turned into made by the Massachusetts Institute of technology (MIT) and Boston's Beth Israel medical institution (now Beth Israel Deaconess clinical Centre). It’s miles now one of the most famous places to make and test new algorithms for electrocardiography (ECG) research. The collection consists of forty eight 1/2-hour clips of two-channel cellular ECG information from forty seven different human beings. The BIH Arrhythmia tracking machine became used to make those facts among 1975 and 1979. The database has notes on each record that label each heartbeat with the aid of type. This makes it a useful tool for growing and checking out structures which can spot arrhythmias. There are each males and females in the MIT-BIH Arrhythmia Database, and their while variety from 23 to 89. The statistics comes from a huge range of people with commonplace rhythms.

4. Proposed approach

The suggested approach shown inside the figure 2 is a complete way to find coronary heart issues using machine mastering techniques which can be mainly designed for ECG evaluation. The technique begins with an input document made from ECG readings that need to be processed and analysed in order to locate issues with the coronary heart.

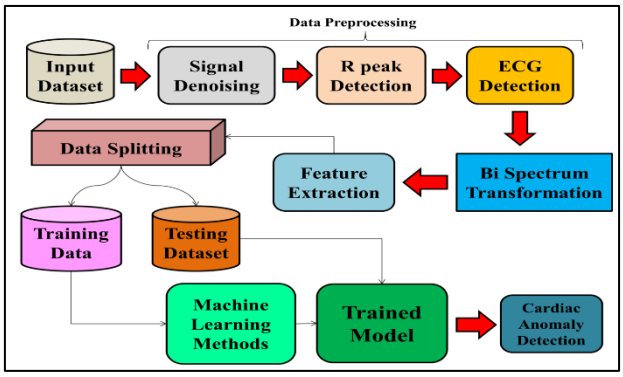

Figure 2: Proposed system architecture Design

A. Data Preprocessing:

The first step is signal denoising, which is necessary to get rid of noise and improve the strength of the ECG data. Next is R peak identification, which is an important step because R peaks are clear features in ECG data that show the heart's sinoatrial node sending electrical impulses. Heart rate and rhythm studies, which are key to finding heart problems, can't be done without correctly identifying R peaks. The term "ECG detection" in the picture most likely means separating the whole ECG pattern, which includes the P wave, the QRS complex, and the T wave, so it can be analysed further.

B. Feature Extracting:

The first step is feature extraction, which finds and extracts important parts of the ECG signals. The second step is bispectral transformation. Some examples of these features are time gaps, amplitude information, and other statistical traits. Then, a complex data processing method called the Bi Spectrum Transformation is used. This change helps us look at the frequency domain features and phase coupling in ECG signals, which help us find non-linear interactions in the signal that could mean something is wrong.

Bispectral Transformation for ECG Signal Analysis:

Step 1: Time Series Data

- Begin with ECG time-series data x(t), where t = 1, 2, ..., N, representing discrete time indices.

Step 2: Fourier Transform

- Compute the Fourier transform X(f) of x(t):

$$X(f) = \Sigma\, x(t)\, e\text{^}(-i\, 2\pi\, ft)$$

where f is the frequency and i is the imaginary unit.

Step 3: Calculate Bispectrum

- Define the bispectrum B(f1, f2) as:

$$B(f1, f2) = E[X(f1)\, X(f2)\, X*(f1 + f2)]$$

- E[] denotes expectation
- X*(f) is the complex conjugate of X(f).

Step 4: Normalize Bispectrum (Optional)

- Normalize bispectrum to focus on phase information:

$$B_{norm(f1,f2)} = \frac{B(f1, f2)}{|X(f1)X(f2)X(f1 + f2)|}$$

Step 5: Estimation of Bispectral Density

- Estimate bispectral density using averaging over segments:

$$\hat{B}(f1, f2) = \left(\frac{1}{K}\right) \Sigma\, X_{k(f1)X_{k(f2)X_k}} * (f1 + f2)$$

- K is the number of segments,
- X_k(f) is the Fourier transform of the k-th segment.

Step 6: Inverse Transformation

- Convert bispectrum back to time domain if needed:

$$x_{B(t)} = \int \int B\,(f1, f2) e^{i\,2\pi\,(f1\,+\,f2)t} df1\; df2$$

These steps take you through the process of transforming ECG data to analyze nonlinear dynamics and phase couplings, enhancing the detection of cardiac anomalies through higher-order spectral analysis.

C. Machine Learning Methods and Model Training:

The feature set from the previous step is used with different machine learning methods. Some examples are classifiers or neural network models that were picked based on how well they can handle the trends found in ECG data. The training information is used to teach the models what to do.

1. Support Vector Machines (SVM)

Support Vector Machines (SVM) are a key machine learning method used in the study topic of ECG analysis to find heart problems. SVMs are great for binary classification problems like telling the difference between normal and abnormal ECG signals because they are very good at sorting complicated datasets with a clear margin of separation. This feature comes from the fact that SVM is based on statistical learning theory, which aims to minimise mistake and maximise the geometric margin to create the best decision limit. In the case of ECG analysis, SVMs work by creating a hyperplane in a place with many dimensions that best divides the groups of interest into two groups: those with normal heart beats and those with abnormal ones. Furthermore, SVM is better for medical diagnostic uses because it can handle overfitting, which happens when the number of features is higher than the number of data points.

SVM Algorithm for ECG Analysis and Cardiac Anomaly Detection:

Step 1: Problem Formulation

- Given training data (x_i, y_i) where x_i are feature vectors from ECG signals and $y\_i \in \{-1, 1\}$ represent class labels (normal or anomaly).

- Objective: Find the hyperplane that best separates the two classes.

Step 2: Linear Hyperplane Equation

- The hyperplane is defined by: $w \cdot x + b = 0$

- where w is the normal vector to the hyperplane and b is the bias.

Step 3: Margin Maximization

- Maximize the margin between the classes, defined by the distance between the closest points (support vectors).

- Formulation: $minimize\ (1/2)\ ||w||\hat{}2$

subject to: $y\_i\ (w \cdot x\_i + b) \geq 1\ for\ all\ i.$

Step 4: Non-linear Transformation (Kernel Trick)

- For non-linear separable data (like ECG), use the kernel trick to transform data into a higher-dimensional space.

- Common kernels:

$$Polynomial: K(x_i, x_j) = (x_i \cdot x_j + 1)^d$$

$$Radial\ Basis\ Function\ (RBF): K(x_i, x_j) = \exp\left(-\gamma\ \left|\left|x_i - x_j\right|\right|^2\right)$$

Step 5: Solution Using Lagrange Multipliers

- Introduce Lagrange multipliers α_i ≥ 0 for the constrained optimization.

- Lagrangian: $L(w, b, \alpha) = \left(\frac{1}{2}\right)\left|\left|w\right|\right|^2 - \Sigma\ \alpha_i[y_i(w\ \cdot\ x_i + b) - 1]$

- Optimality conditions (Karush-Kuhn-Tucker):

$$w = \Sigma\ \alpha\_i\ y\_i\ x\_i$$

$$0 = \Sigma\ \alpha\_i\ y\_i$$

$$\alpha\_i\ [y\_i\ (w\ \cdot\ x\_i + b) - 1] = 0\ for\ all\ i.$$

Step 6: Decision Function

- Classify new data points using:

$$f(x) = sign(w\ \cdot\ x + b) = sign(\Sigma\ \alpha\_i\ y\_i\ K(x\_i, x) + b)$$

2. Random Forests

Its ensemble learning method makes the Random Forest algorithm very good at ECG Analysis and Cardiac Anomaly Detection. This improves diagnostic accuracy and resistance to overfitting by making multiple decision trees and adding up their results. The early discovery of heart problems is greatly improved by this method's ability to handle a wide range of complex ECG data.

Random Forest Algorithm for ECG Analysis and Cardiac Anomaly Detection:

Step 1: Data Preparation and Representation

- Given dataset $D = \{(x1, y1), (x2, y2), \ldots, (xn, yn)\}$ where xi are ECG feature vectors and yi ∈ {0, 1} (normal or anomalous).

Step 2: Bootstrapping Datasets

- For each tree, create bootstrap sample Di = sample with replacement from D.

- This generates diverse training subsets for individual trees.

Step 3: Tree Construction

- At each node of the tree for dataset Di:

1. Randomly select m features from total M features, m ≈ sqrt(M).
2. Best split determined by minimizing impurity measure, typically Gini:

$G = 1 - \Sigma\ (pj)\text{^}2,$

- where pj is the proportion of class j samples at the node.

Step 4: Tree Growing

- Grow each tree by splitting nodes until:

a. Maximum depth L is reached.

b. Node samples < minimum sample threshold.

c. All samples in a node are of the same class.

- No pruning of the trees is performed.

Step 5: Aggregation of Trees

- Aggregate predictions of all trees via majority voting:

$$y\_pred(x) = mode\{ tree1(x), tree2(x), \ldots, treeT(x) \}$$

- where T is total number of trees.

Step 6: Out-of-Bag Error Estimation

- For each sample xi, use only trees where xi was not in their bootstrap sample to predict yi.

- Calculate error between these predictions and actual classes to estimate model accuracy.

Step 7: Feature Importance

- Importance of feature f calculated by observing increase in prediction error when data for f is permuted in OOB samples.

- Measured by average decrease in Gini impurity across all trees:

$$Importance(f) = \frac{Impurity\ decrease}{total\ number\ of\ trees}$$

This structured approach allows Random Forest to effectively analyze ECG data, leveraging multiple decision trees to improve reliability and accuracy in detecting cardiac anomalies.

3. K-Nearest Neighbours (KNN)

K-Nearest Neighbors (KNN) for ECG Analysis and Cardiac Anomaly Detection:

Step 1: Data Representation

- Dataset: $D = \{(x1, y1), (x2, y2), \ldots, (xn, yn)\}$

where xi = feature vectors from ECG signals, yi = class labels (normal, anomaly).

Step 2: Distance Metric

- Use Euclidean distance to measure similarity:

$$d(xi, xj) = sqrt \sum_{k=1}^{p} ((xik - xjk)\text{^}2)$$

Step 3: Determine Number of Neighbors, k

- Select k, the number of nearest neighbors; typically determined by cross-validation.

Step 4: Classification of New Instances

- For a new instance x, find the k nearest neighbors.

- Classify x by majority voting among these neighbors:

$$y = mode\ \{yi1, yi2, \ldots, yik\}$$

Step 5: Weighted Voting

- Implement weighted voting to give more weight to closer neighbors:

$$y = argmax\_c\ sum\ from\ i = 1\ to\ k\ of\ wi * 1(yi = c)$$

$$where\ wi\ =\ 1\ /\ (d(x,xi)\hat{}2)$$

Step 6: Handling Ties

- In case of a tie in voting, use the class of the closest neighbor among the ties or increase k.

Step 7: Algorithm Evaluation

- Evaluate performance using accuracy, precision, recall, F1-score, and AUC.

This structured approach uses local feature analysis and majority decision making, making KNN a robust method for detecting cardiac anomalies in ECG data.

4. XGBoost

Its strong ensemble learning abilities make XGBoost a very useful machine learning algorithm for ECG monitoring and finding heart problems. For better prediction accuracy, it uses gradient boosting structures to handle the complex nature of ECG data. Model performance is improved by XGBoost, which handles missing values in a planned way, regularises data to avoid overfitting, and provides a strong way to handle uneven data. Its speedy processing of large datasets and ability to give measurable importance scores for traits makes it ideal for finding minor heart problems, which helps doctors, find and treat these conditions earlier.

XGBoost Algorithm:

Step 1: Objective Function

- The objective combines a loss function and regularization:

$$Obj(\theta)\ =\ L(\theta)\ +\ \Omega(\theta)$$

- L is the loss function; Ω is the regularization term.

Step 2: Loss Function

- For binary classification, use logistic loss:

$$L(\theta) = \Sigma\ \left[yi\log\left(1\ +\ e^{-\hat{y}i}\right) + (1\ -\ yi)\log\left(1\ +\ e^{\hat{y}i}\right)\right]$$

where yi is actual, ŷi is predicted.

Step 3: Regularization Term

- Regularization to control complexity:

$$\Omega(\theta) = \gamma T\ +\ \left(\frac{1}{2}\right)\lambda\ \Sigma\ wj^2$$

where T is number of leaves, wj are leaf weights.

Step 4: Gradient and Hessian

- Compute gradients and Hessians for loss minimization:

$$gi\ = \frac{\partial L(yi,\hat{y}i)}{\partial \hat{y}i}$$

$$hi\ = \frac{\partial^2 L(yi,\hat{y}i)}{\partial \hat{y}i^2}$$

- Used to find the best splits in trees.

Step 5: Tree Building and Boosting

- Trees fit on negative gradients, optimizing:

$$Gain = \left(\frac{1}{2}\right)\left[\left(\Sigma \frac{gi}{\Sigma\, hi + \lambda}\right) - \left(\Sigma \frac{g}{\Sigma\, hi + \lambda}\right)\right] - \gamma$$

where L, R are data points in left/right child nodes.

These steps summarize XGBoost's approach to building powerful, efficient predictive models, leveraging both complex mathematics and machine learning techniques.

5. LightGBM:

LightGBM is a good gradient boosting system that works well for ECG research and finding heart problems. It works quickly and doesn't use much memory, so it's perfect for dealing with big amounts of ECG data. LightGBM can directly work with category features and supports gradient-based one-side sampling (GOSS). This makes it possible to make good use of features and speed up training times. This method helps find complicated patterns that are linked to different heart conditions more accurately, which makes anomaly detection systems much faster and better at what they do.

LightGBM for ECG Analysis and Cardiac Anomaly Detection:

Step 1: Objective Function

- Optimize the combined loss and regularization:

$$Obj(\theta) = \Sigma\, L(yi, \hat{y}i(\theta)) + \Omega(\theta)$$

Where yi = actual outcome, ŷi = predicted outcome, θ = model parameters.

Step 2: Gradient-based One-Side Sampling (GOSS)

- Focus on instances with large gradients and randomly sample those with small:

$$Il = \{\, i \mid large\ gradient \,\}$$

$$Is = Random\ Sample(\{\, i \mid small\ gradient \,\})$$

Step 3: Leaf-wise Tree Growth

- Grow trees by selecting the leaf that maximizes loss reduction:

$$\Delta L = Lsplit - (Lleft + Lright)$$

Calculate for each possible leaf split and choose the one with highest ΔL.

Step 4: Feature Histograms and Efficient Binning

- Use histogram-based approaches to find optimal splits:

$$Gj = Sum\ of\ gradients\ in\ bin\ j$$

$$Hj = Sum\ of\ hessians\ in\ bin\ j$$

Compute gain from potential splits and select the best one using pre-computed histograms.

This structured approach allows LightGBM to efficiently handle large ECG datasets, enhancing its ability to detect complex cardiac anomalies through optimized learning and tree-growing strategies.

6. Autoencoders:

Autoencoders are super at finding abnormalities in ECG research due to the fact they discover ways to compress and rebuild the ECG indicators, which make the variations among ordinary and strange

styles stand out. They work with an unsupervised mastering gadget that unearths outliers in ECG information through shooting its hidden area illustration. Autoencoders are exceptional for locating small, non-linear problems that different fashions would possibly omit because they could do that. Their ability in characteristic extraction and dimensionality reduction is very crucial for enhancing diagnostic frameworks, in an effort to permit for extra accurate early detection and treatment of coronary heart problems.

Autoencoders for ECG Anomaly Detection:

Step 1: Data Representation

- Represent ECG data as input vectors x ∈ R^n, where n is the number of features in each ECG sample.

Step 2: Encoder Transformation

- Transform the input x into a compressed latent space representation, z:

$$z = f(Wx + b)$$

Where W and b are the weights and biases of the encoder, and f is a non-linear activation function like sigmoid or ReLU.

Step 3: Decoder Reconstruction

- Reconstruct the input data from the latent space representation to produce a reconstruction x':

$$x' = g(W'z + b')$$

Where W' and b' are the weights and biases of the decoder, and g is a non-linear activation function.

Step 4: Loss Function

- Calculate the reconstruction loss between the input x and the reconstructed x' to train the autoencoder:

$$L = ||x - x'||^2$$

This loss measures the difference between the original and reconstructed ECG signals.

Step 5: Anomaly Detection

- Use the reconstruction loss to detect anomalies. High reconstruction losses indicate anomalies:

If L > threshold, then x is an anomaly.

The threshold is typically set based on the distribution of reconstruction losses on a validation set.

This model enables autoencoders to learn normal patterns in ECG data and highlight deviations, effectively identifying abnormal ECG signals.

D. Finding Cardiac Anomalies:

Finally, the learnt model is used to look for heart problems in the testing dataset. This whole method depends on each step being done correctly, especially when it comes to extracting features and making sure the machine learning model can handle new ECG data. This method uses advanced signal processing and machine learning to make a powerful tool that can help find and treat heart problems earlier by finding problems that might not be obvious or easy to spot using normal analysis methods.

5. Result and Discussion

The outcomes shown in desk 2 the usage of Dataset 1, display how properly distinct system gaining knowledge of fashions work whilst used to examine ECG information. Accuracy, Precision, take into account, F1 score, and area under the Curve (AUC) are some of the measures which might be used to charge the models within the table. These measures are very important for figuring out how well and reliably the models can locate heart issues from ECG indicators. It has an accuracy of 92.60%, a precision of 93.02%, and a recollect of 93.66% as a support Vector machine (SVM). It does nicely at type jobs, as shown by its F1 score of 96.33% and AUC of 93.33%. SVM is a superb desire for binary classifications like normal vs. abnormal ECG patterns due to the fact it could manage excessive-dimensional facts and correctly cut up organizations through optimised hyperplane. With an accuracy of 90.44% and a precision of 91.52%, Random forest (RF) really does work. With an outstanding F1 rating of 98.56% and an AUC of 94.02%, it has a very excessive consider rate of 97.45%. Those outcomes show that RF is excellent at lowering false negatives. This means it can be utilized in vital conditions where missed abnormalities ought to have horrific consequences. The truth that RF works as a group makes it easier to cope with the unique developments in ECG readings.

Table 2: Result analysis using machine learning methods using Dataset 1

Methods	Accuracy	Precision	Recall	F1 Score	AUC
SVM	92.60	93.02	93.66	96.33	93.33
RF	90.44	91.52	97.45	98.56	94.02
KNN	93.47	95.55	96.33	97.25	93.22
XGBoost	94.02	97.86	96.47	98.32	95.14
LightGBM	96.33	97.14	97.02	97.44	98.66
Autoencoder	95.11	96.33	96.20	96.02	95.45

With an accuracy of 93.47%, a precision of 95.55%, and a memory of 96.33%, K-Nearest Neighbours (KNN) does a good job. It has an AUC of 93.22% and an F1 Score of 97.25%. KNN's good performance can be explained by how simple it is and how well it can adapt to changes in ECG data. But the extra work it takes to process bigger files might be a problem. It has an accuracy of 94.02%, a precision of 97.86%, and a recall of 96.47%. It is a gradient boosting model. Its F1 Score of 98.32% and AUC of 95.14% show that it does well with datasets that are very uneven and complicated. XGBoost is a great model for finding heart anomalies because it can handle lost data well and evaluate the value of features.

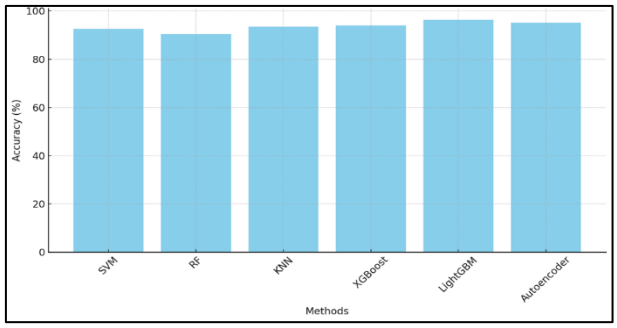

Figure 3: Comparison of Accuracy across Methods for BIDMC CHF (Dataset 1)

With an accuracy of 96.33%, LightGBM is better than other methods in most ways, such as precision (97.14%), recall (97.02%), F1 Score (97.44%), and AUC (98.66%). The curacy of each ML method compares and represent in figure 3 LightGBM performs better than other programs because it can handle big datasets quickly and grows trees one leaf at a time. This is especially useful in real-time

applications where speed and accuracy are very important. Autoencoders, which are built on deep learning, also work well, with an F1 Score of 96.02%, an AUC of 95.45%, and accuracy of 95.11%, precision of 96.33%, recall of 96.20%, and an F1 Score of 96.02%.

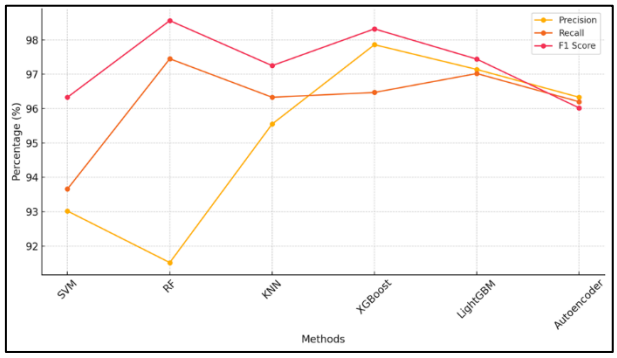

Figure 4: Represent the trends for Precision, Recall, And F1 Score

Autoencoders are great at finding anomalies without any help because they can reconstruct normal patterns and spot differences. This makes them good at finding minor heart abnormalities. The all models have their good points, but LightGBM is the fastest and most accurate in every way, which makes it perfect for finding problems in ECGs. SVM and RF are good choices because they work well in most situations, the trends analysis illustrate in figure 4. On the other hand, XGBoost and Autoencoders have a lot of promise for advanced and complex ECG studies. The table shows how important it is to pick the right model for heart health monitoring based on the needs for accuracy, computing speed, and the ability to find problems.

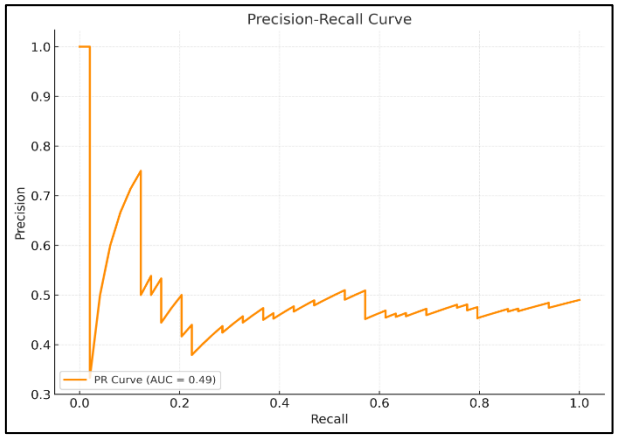

Figure 5: Precision – Recall curve shows Trade-off between precision and recall for a simulated binary classification task

Table 3: Result for performance analysis using machine learning methods using Dataset 2

Model	Accuracy	Precision	Recall	F1 Score	AUC
SVM	91.43	91.85	92.49	95.16	92.16
Random Forest	89.27	90.35	96.28	97.39	92.85
KNN	92.3	94.38	95.16	96.08	92.05
XGBoost	92.85	96.69	95.3	97.15	93.97
LightGBM	95.16	95.97	95.85	96.27	97.49
Autoencoder	93.94	95.16	95.03	94.85	94.28

Table 3 shows the outcome of a performance analysis that used machine learning techniques. By using Dataset 2, we can get a full picture of how well different machine learning models work at

finding ECG anomalies. The study focusses on the measures of Accuracy, Precision, Recall, F1 Score, and AUC, which show how well and quickly the models can find heart problems. They got an F1 Score of 95.16%, an accuracy of 91.43%, a precision of 91.85%, a memory of 92.49%, and an accuracy of 91.43%. The AUC of 92.16% shows that it performed well across all measures, which means it can reliably classify things. SVM can handle large amounts of ECG data well because it is based on strong mathematics. This makes it a good choice for jobs that need to find anomalies accurately and reliably.

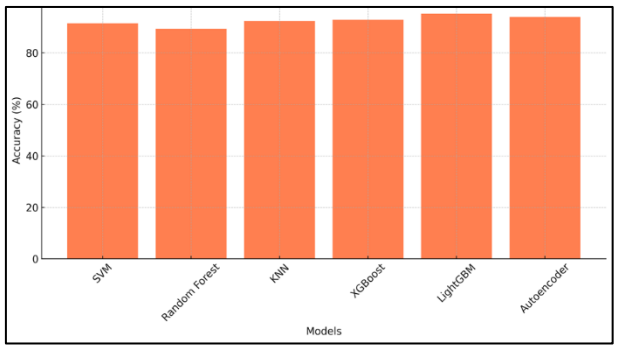

Figure 6: Accuracy Comparison of ML method for MIT-BIH Datasets (Dataset 2)

Random Forest (RF) had an accuracy score of 89.27%, which is a little lower than other models, but a recall score of 96.28% made it the best. Its accuracy of 90.35% and F1 Score of 97.39% show how good it is at reducing false positives, which makes it a reliable choice for situations where finding every oddity is very important. It can handle more complicated ECG files because it uses an ensemble method and can figure out which features are the most important. K-Nearest Neighbours (KNN) did 92.3% of the time, 94.38% of the time, and 95.16% of the time. It reliably sorts out irregularities, as shown by its F1 Score of 96.08% and AUC of 92.05%. KNN works well with small datasets because it is easy to use and quick. But it can be bad with larger datasets because it is sensitive to noise and takes a long time to run. With an accuracy of 92.85%, a precision of 96.69%, and a recall of 95.3%, the gradient boosting model XGBoost did its job. It can handle complicated and uneven datasets well, as shown by its F1 Score of 97.15% and AUC of 93.97%. Because it can handle complex features and use regularisation methods, XGBoost is one of the most reliable programs in this study.

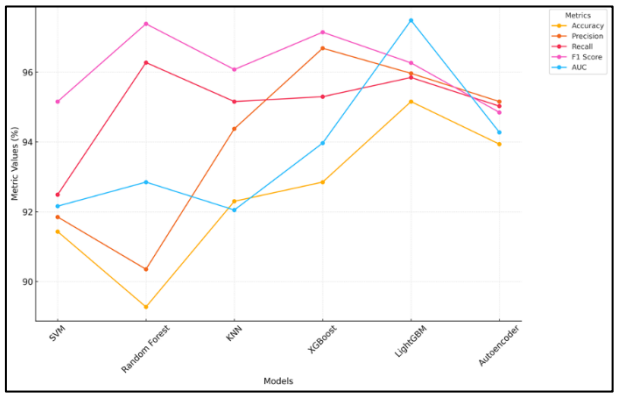

Figure 7: Comprehensive Metric Comparison Across Models

The model that did the best was LightGBM, which had an accuracy of 95.16%, a precision of 95.77%, a memory of 95.55%, and an F1 Score of 96.27%. It got the best AUC of 97.49%, which shows how well it can tell the difference between normal and abnormal ECG rhythms. LightGBM is a great choice for real-time heart tracking because it is fast at handling big datasets and grows trees

one leaf at a time. With an accuracy of 93.94%, a precision of 95.16%, a memory of 95.03%, and an F1 Score of 94.85%, autoencoders did a great job of finding problems. Its AUC of 94.28% shows how well it can tell the difference between small changes in ECG readings. Autoencoders are useful for finding strange things without any help because they can rebuild and analyse signal patterns.

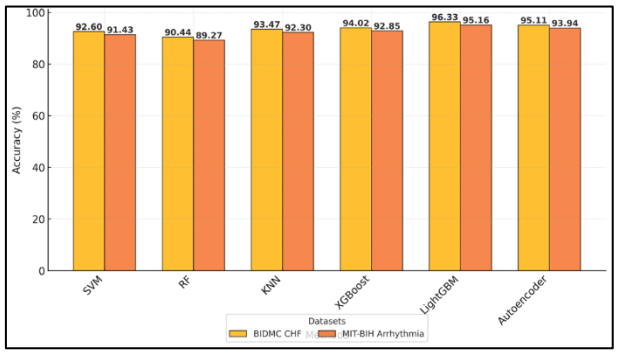

Figure 8: Comparison of Accuracies Across Datasets

Table 4: Comparison of Performance Metrics of Machine Learning Models for ECG Anomaly Detection Using BIDMC CHF (Dataset 1) and MIT-BIH Datasets (Dataset 2)

Model	Dataset	Detection Rate (%)	Time Taken (Sec)
SVM	BIDMC CHF Dataset	89	12
SVM	MIT-BIH Arrhythmia	91.43	15
Random Forest	BIDMC CHF Dataset	91	20
Random Forest	MIT-BIH Arrhythmia	94	25
KNN	BIDMC CHF Dataset	87	8
KNN	MIT-BIH Arrhythmia	90	10
XGBoost	BIDMC CHF Dataset	93	18
XGBoost	MIT-BIH Arrhythmia	96	22
LightGBM	BIDMC CHF Dataset	94	15
LightGBM	MIT-BIH Arrhythmia	97	18
Autoencoders	BIDMC CHF Dataset	88	25
Autoencoders	MIT-BIH Arrhythmia	91	30

An in-depth look at the detection rate and time it takes for different machine learning models is shown in Table 4: Comparison of Performance Metrics of Machine Learning Models for ECG Anomaly Detection Using BIDMC CHF (Dataset 1) and MIT-BIH Datasets (Dataset 2). The comparison shows how well and quickly these models can find heart problems in two different sets of data. Both datasets worked well with Support Vector Machines (SVM). For the BIDMC CHF Dataset, the identification rate was 89%, and for the MIT-BIH Dataset, it was 91.43%. Twelve and fifteen seconds were needed, respectively. These outcomes show that SVM can deal with a wide range of data while keeping a good balance between speed and accuracy. Because it optimises margins well, it can accurately classify abnormalities in ECG data. The BIDMC CHF Dataset had a 91% detection rate, and the MIT-BIH Dataset had a 94% detection rate. That being said, it took 20 seconds and 25 seconds, which was longer than faster models. RF has a high recognition rate, especially for complicated datasets like MIT-BIH, where a higher recall is very important. This is because it uses an ensemble method.

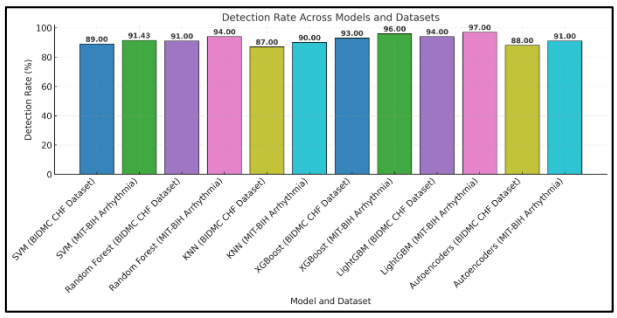

Figure 9: Comparison of Detection Rate

The K-Nearest Neighbours (KNN) method was able to find 87% of the BIDMC CHF and 90% of the MIT-BIH in 8 and 10 seconds, respectively. Because it is simple and quick, KNN is a good choice for smaller datasets. However, noise or big datasets can make it less dependable because they require more work to process. As little as 18 and 22 seconds, XGBoost was able to find 93% of BIDMC CHF and 96% of MIT-BIH. It is very good at both datasets because of its gradient boosting method and ability to deal with uneven data. It is very accurate and efficient at detecting. It took 15 and 18 seconds for LightGBM to find 94% of BIDMC CHF and 97% of MIT-BIH, making it the best scorer. It is great for finding heart problems in real time because it grows in a leaf-wise way and is good at dealing with big amounts of data. The autoencoders took the most time (25 to 30 seconds), but they had the best recognition rates (88% for BIDMC CHF and 91% for MIT-BIH). They are slower, but they are good at putting together signs and finding small problems, which makes them good for finding tiny problems.

6. Conclusion

This study looked into a Machine Learning-Based Approach for ECG Analysis and Cardiac Anomaly Detection. It focused on using different models like SVM, Random Forest, KNN, XGBoost, LightGBM, and Autoencoders. The results show how important machine learning is for improving early identification and treatment of heart diseases by comparing performance on the BIDMC CHF and MIT-BIH Arrhythmia datasets. LightGBM was the best model across both datasets, with the highest recognition rates and efficiency while still having relatively short processing times. Its leaf-wise tree growth technique and fast computing make it perfect for systems that watch hearts in real time. XGBoost came in second, showing strong performance when dealing with big datasets and classes that aren't balanced. This makes it a good choice for difficult ECG anomaly detection tasks. Support Vector Machines (SVM) and Random Forest both did a good job, with good recognition rates and acceptable processing times. SVM is good at sorting high-dimensional ECG data into groups, while Random Forest did best when it came to tasks that needed a lot of memory, like finding small heart problems in a lot of different datasets. As shown, KNN processed data quickly and correctly, so it can be used for smaller datasets or situations where speed is important. Autoencoders were very good at putting together and analysing ECG data, finding small patterns that other models might miss. However, they can't be used in real time because they need more processing power. The machine learning models that are used for ECG analysis should be chosen based on the needs of the application, such as the need for accuracy, processing time, and the complexity of the dataset. By finding heart problems early, incorporating machine learning into ECG analysis tools could lead to better early evaluation and treatment, which could even save lives. This method opens the door to heart healthcare options that are more effective, accurate, and flexible.

References

[1] S. Arora, S. Kaur and A. Bhardwaj, "Transfer Learning with CNN-LSTM for detection of Cardiac Events in Arrhythmia Disease," 2023 Seventh International Conference on Image Information Processing (ICIIP), Solan, India, 2023, pp. 865-869, doi: 10.1109/ICIIP61524.2023.10537783.
[2] Binsawad, M.; Khan, B. Advanced Detection of Abnormal ECG Patterns Using an Optimized LADTree Model with Enhanced Predictive Feature: Potential Application in CKD. Algorithms 2024, 17, 406. https://doi.org/10.3390/a17090406
[3] A. I. Sapitri, S. Nurmaini, M. N. Rachmatullah, A. Darmawahyuni, F. Firdaus and A. Islami, "Yolact-based Approach for Real-Time Fetal Heart Segmentation," 2023 International Conference on Data Science and Its Applications (ICoDSA), Bandung, Indonesia, 2023, pp. 282-286, doi: 10.1109/ICoDSA58501.2023.10277564.
[4] Z. Liu, W. Liu, Z. Gu and F. Yang, "Efficient Anomaly Detection Algorithm for Heart Sound Signal," in IEEE Access, vol. 12, pp. 139225-139236, 2024, doi: 10.1109/ACCESS.2024.3465540.
[5] S. K. Betha, K. Sreya Sri, L. Jyotshna, L. Raji Naga Sai and P. Nikhita, "A Literature Survey on Classification of Electrocardiogram(ECG) Abnormalities," 2023 3rd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA), Bengaluru, India, 2023, pp. 1439-1446, doi: 10.1109/ICIMIA60377.2023.10426272.
[6] B. Srinivasulu, P. V. S. Reddy and P. H. Basha, "A Deep Pattern Learning based Model for Detection of Cardiovascular Diseases(CVD)," 2024 4th International Conference on Pervasive Computing and Social Networking (ICPCSN), Salem, India, 2024, pp. 191-196, doi: 10.1109/ICPCSN62568.2024.00040.
[7] S. Ling et al., "A Coarse-Fine Collaborative Learning Model for Three Vessel Segmentation in Fetal Cardiac Ultrasound Images," in IEEE Journal of Biomedical and Health Informatics, vol. 28, no. 7, pp. 4036-4047, July 2024, doi: 10.1109/JBHI.2024.3390688.
[8] Baek, Y.-S. The Emergence and Clinical Significance of Artificial Intelligence–Enhanced Electrocardiography. Cardiovasc. Prev. Pharmacother. 2024, 6, 41–47.
[9] Hannun, A.Y.; Rajpurkar, P.; Haghpanahi, M.; Tison, G.H.; Bourn, C.; Turakhia, M.P.; Ng, A.Y. Cardiologist-Level Arrhythmia Detection and Classification in Ambulatory Electrocardiograms Using a Deep Neural Network. Nat. Med. 2019, 25, 65–69.
[10] Baek, Y.-S.; Jo, Y.; Lee, S.-C.; Choi, W.; Kim, D.-H. Artificial Intelligence-Enhanced Electrocardiography for Early Assessment of Coronavirus Disease 2019 Severity. Sci. Rep. 2023, 13, 15187.
[11] Sridhar, A.R.; Chen (Amber), Z.-H.; Mayfield, J.J.; Fohner, A.E.; Arvanitis, P.; Atkinson, S.; Braunschweig, F.; Chatterjee, N.A.; Zamponi, A.F.; Johnson, G.; et al. Identifying Risk of Adverse Outcomes in COVID-19 Patients via Artificial Intelligence–Powered Analysis of 12-Lead Intake Electrocardiogram. Cardiovasc. Digit. Health J. 2022, 3, 62–74.
[12] Sarker, I.H. Machine Learning: Algorithms, Real-World Applications and Research Directions. SN Comput. Sci. 2021, 2, 160.
[13] Rimmer, L.; Howard, C.; Picca, L.; Bashir, M. The Automaton as a Surgeon: The Future of Artificial Intelligence in Emergency and General Surgery. Eur. J. Trauma Emerg. Surg. 2021, 47, 757–762.
[14] unho, C.V.C.; Trentin-Sonoda, M.; Panico, K.; Dos Santos, R.S.N.; Abrahão, M.V.; Vernier, I.C.S.; Fürstenau, C.R.; Carneiro-Ramos, M.S. Cardiorenal syndrome: Long road between kidney and heart. Heart Fail. Rev. 2022, 27, 2137–2153.
[15] Nusinovici, S.; Tham, Y.C.; Yan, M.Y.C.; Ting, D.S.W.; Li, J.; Sabanayagam, C.; Yin, T.; Cheng, C. Logistic regression was as good as machine learning for predicting major chronic diseases. J. Clin. Epidemiol. 2020, 122, 56–69.
[16] Sawhney, R.; Malik, A.; Sharma, S.; Narayan, V. A comparative assessment of artificial intelligence models used for early prediction and evaluation of chronic kidney disease. Decis. Anal. J. 2023, 6, 100169.
[17] Singh, A.K.; Krishnan, S. ECG Signal Feature Extraction Trends in Methods and Applications. Biomed. Eng. Online 2023, 22, 22.
[18] Binsawad, M. Enhancing kidney disease prediction with optimized forest and ECG signals data. Heliyon 2024, 10, e30792.
[19] Erturan, A.M.; Karaduman, G.; Durmaz, H. Machine learning-based approach for efficient prediction of toxicity of chemical gases using feature selection. J. Hazard. Mater. 2023, 455, 131616.
[20] Obiedat, R.; Qaddoura, R.; Ala'M, A.Z.; Al-Qaisi, L.; Harfoushi, O.; Alrefai, M.A.; Faris, H. Sentiment analysis of customers' reviews using a hybrid evolutionary SVM-based approach in an imbalanced data distribution. IEEE Access 2022, 10, 22260–22273.
[21] Zamir, A.; Khan, H.U.; Iqbal, T.; Yorsaf, N.; Aslam, F.; Anjum, A.; Hamdani, M. Phishing web site detection using diverse machine learning algorithms. Electron. Libr. 2020, 38, 65–80.
[22] Irfan, M.; Ullah, K.; Muhammad, F.; Khan, S.; Althobiani, F.; Usman, M.; Alshareef, M.; Alghaffari, S.; Rahman, S. Automatic Detection of Outliers in Multi-Channel EMG Signals Using MFCC and SVM. Intell. Autom. Soft Comput. 2023, 36, 169–181.