Comparative Analysis of Machine Learning Algorithms for Early Heart Disease Detection Using ECG Data

Saroj Kumari[1*], and Raghav Mehra[2]

[1]**Research Scholar, Department of Computer Engineering & Applications, Mangalayatan University, Aligarh, UP, India, saroj.cse10@gmail.com*

[2]*Professor, Department of Computer Engineering & Applications, MangalayatanUniversity,Aligarh, UP, India, dr.raghavm@gmail.com*

KEYWORDS

Cardiovascular disease, Machine learning (ML), Electrocardiogram (ECG), heart disease detection, Feature extraction.

ABSTRACT:

Cardiovascular Diseases (CVDs), is one of the leading causes of death worldwide,highlighting the importance of early identification for timely treatment. This studymarks a comparative analysis based on the Machine Learning (ML) approach targeted toward heart disease detection using the Electrocardiogram (ECG)as its input data. Methods such as Discrete Wavelet Transform (DWT) for feature extraction and Recursive Feature Elimination (RFE) for feature selection were applied to optimize model performance. The ML models evaluated included Vision Transformer (ViT), Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM), and Multi-Layer Perceptron (MLP). These models were trained and tested on a dataset of 986 patients to assess their predictive accuracy. The results showed that MLP achieved the highest accuracy of 99.3%, followed by LSTM and CNN. These findings highlight the capability of ML to improve early detection of heart disease. Future research may include enhancing generalizability by including larger and more diverse datasets, hybrid models, and real-time diagnostic tools to further improve prediction accuracy and extend clinical applications.

Introduction

Heart diseases or CVD is a medical term that integrates several conditions affecting the valves, muscles,and blood arteries of the heart. These conditions might potentially bring about a serious cardiovascular problem, like a heart attack. The CVDs are recognized as one of the world's leading causes of deaths [1]. It is responsible for the highest number on the global mortality chart, approximating a yearly death toll of 17.9 million [2]. WHO conducted the World Health Survey, which estimated that cardiovascular disease (CVD contributes to approximately 17.9×10^6 deaths annually, or 31% of all deaths around the world [3].In 2018, cardiovascular disease and stroke were responsible for around 400,000 fatalities in women, making up 28% of all recorded deaths[4]. The mortality rates associated with different forms of CVD are most prevalent in industrialized nations. The main factors contributing to this life-threatening condition, as seen in Figure 1, are hypertension, obesity, hypercholesterolemia, genetic predisposition, tobacco use, smoking [5], and alcohol use. These risk factors are prevalent throughout many age groups in today's society.

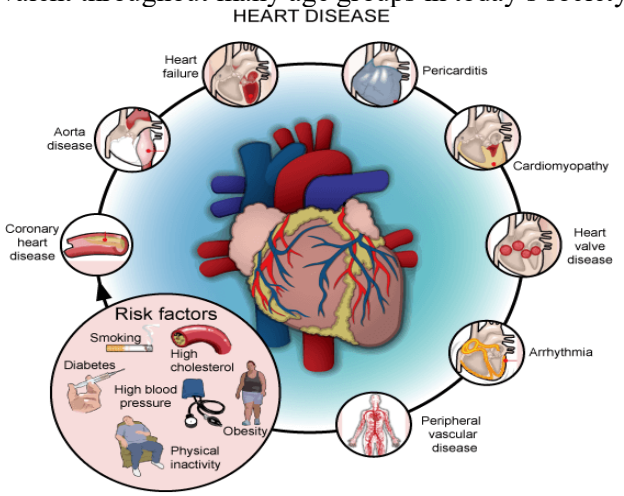


Figure 1. Risk factors associated with heart diseases [6].

Avoiding heart disease is as simple as making little changes in everyday routine, such as eating healthier, avoiding smoking, and increasing physical activities [7-10]. When cardiac disease is detected early, patients may be treated before they get sick, which may represent the difference between life and death [11]. Thus, heartdisease prediction is considered a primary area of study in the field of healthcare analysis [12-14].

The wide range of symptoms associated with CVD makes it difficult for healthcare providers to make an appropriate diagnosis rapidly [15–17]. Since global medical organizations are gathering vast information about heart diseases, healthcare professionals are better approach to know these diseases much better and thus increasing the efficiency of the treatment provided to the patients [18–20]. For proper and fast decision-making based on the right information, the massive amount of raw data from the medical field should be converted into practical information in cardiovascular data analysis. The collected data should also be scanned and processed in detail for efficient extraction of the information [21-23].

Depending on the symptoms and other conditions of the patient, a doctor may ask for further tests to prevent heart disease. In addition to the basic diagnostic procedures such as blood tests [24], chest X-rays [25], and ECGs [26], conventional imaging techniques include cardiac magnetic resonance imaging (MRI) [27] and cardiac Computed Tomography (CT) scans [28] for the diagnosis of cardiac illness [29]. Traditional statistics complicated the execution of such studies on large datasets. Therefore, ML has emerged as the most efficient technique to process data and apply that knowledge for the betterment of health in this current era [30,31]. ML is highly useful for the analysis of medical data and information extraction in the medical domain [32]. It has a combination of great processing skills and enormous capacities for data processing, making it top when it comes to problems that are complex or real-time or offline solutions [33–35]. Heart disease diagnosis has made extensive use of several techniques, such as CNN, Support Vector Machine (SVM), Ensemble Classifiers, and Artificial Neural Networks (ANN), among many others [36]. In addition, there have been several hybrid models that have been released and have been very successful [37-39]. Also, many new models and approaches have been developed that rely on ML and image processing [40,41].

Diagnosing heart disease and providing suitable treatments is becoming more challenging in many underdeveloped nations owing to a shortage of medical experts and ineffective diagnostic equipment. Hospitals and professionals must have reliable methods of diagnosis and forecasting. When developing an intelligent healthcare approach, it is essential to address these critical challenges. Advances in computer technology have made numerous abilities for collecting and storing data in realtime possible. Clinical research greatly benefits from the massive amounts of health data generated. Thus, forecasting the condition of the heart and predicting the disease with the help of ML might play a vital role in preventing errors in diagnosis and reducing the mortality rate due to heart diseases [42,43]. Hence, the exploration and extraction of huge datasets to discover hidden knowledge and patterns have drawn the attention of the researchers.

Therefore, the purpose of this study is to address these problems by developing an algorithm that uses ML algorithms to determine if a patient has heart disease based on their medical history. To achieve this, this study has developed an algorithm to predict the occurrence of cardiac problems using aprimary ECG dataset collected from patients that employ many ML algorithms and other robust methods. The major contributions of this study are:

- Applying DWT for obtaining related features from ECG signals and using RFE to identify the most significant characteristics, hence improving the prediction abilities of the models.
- Exploring a range of ML model architectures and leveraging their unique strengths for comprehensive heart disease prediction.
- Conducting thorough training and evaluation using various performance metrics, along with extensive hyperparameter tuning, to ensure the highest predictive accuracy and robustness of the models.

The restportions of thework arearranged as follows: Section 2 presents the relevant work of various authors on the prediction of cardiac problems using ML. In Section 3, the dataset used for this research, along with the suggested ML framework for the prediction of heart diseases, is described.

1. Review of Related Literature

Many researchers have been attempting to predict heart disease using various ML techniques, leading to significant advancements in this field. **Pachiyannan et al. [44]**proposed the ML-based Congenital Heart Disease Prediction Method (ML-CHDPM) based on advanced ML techniques, capable of accurately categorizing CHD cases in pregnant women, with impressive metrics of up to 94.28% accuracy and 96.25% recall. Similarly, **Alimbayeva et al.[45]** developed an innovative ECG monitoring system that may predict heart disease at an early stage using methods such as isolation forests and CNNs with an accuracy of 92.6%.**Ribeiro et al. [46]**focused on distinguishing between various cardiovascular diseases using ML models trained on non-linear features extracted from ECG signals, showcasing an accuracy range of 73% to 100%. **Utsha et al. [47]** introduced a mobile application that continuously monitors ECG signals, utilizing a pre-trained ANN model to classify heart diseases with an overall accuracy of 94%, while **Baghdadi et al. [48]** developed a Catboost model that achieved an average accuracy of 90.9% and F1-score of 92.30% and demonstrating high classification performance.

Other contributions in this area include **Yilmaz et al. [49]**, which compared various models for classifying coronary heart disease. The Random Forest (RF) model obtained a 92.9% accuracy. **Hossain et al. [50]** applied several ML algorithms to a merged dataset of ECG records and reported SVM as the best model with 85.49% accuracy. **Anuar et al. [51]** used six ML algorithms in a case-control study, and the ANN gives an accuracy of 90% with specificity and sensitivity. **Tyagi et al. [52]** proposed the hybrid CNN architecture that includes the Grasshopper Optimization Algorithm, which provides an average classification accuracy of 99.58%. Finally, **Hammad et al. [53]** proposed the classifier for ECG signals that outperforms others with an average classification accuracy of 99%. All these studies combined reflect significant developments in applying ML for timely prediction of cardiovascular diseases.

Despite the promising results, these studies have certain drawbacks that need attention. Variability and limitations of the datasets used are one common issue, often lacking diversity and not representative of broader populations, which might affect the generalizability of the models. Another point is that feature extraction and selection methods vary widely, and some studies rely on limited or suboptimal features that may not fully capture the complexity of cardiovascular conditions. Hyperparameter tuning is another area where inconsistencies and suboptimal practices are evident that may impact the performance and robustness of the models. Thus, the rectification of these drawbacks may serve as a potential basis for a new study focused on the improvement of cardiovascular disease prediction through ML.

2. Research Methodology

This section discusses the dataset used in this study, the techniques applied, and the proposed approach in detail.

3.1 Dataset Description

The proposed methodology relied on a primary dataset based on patients. This set of data contains information that covers 986 patients and a total of 26 features.However, only 12 of these attributes(as provided in Table1) were utilized for heart disease prediction, as the others were deemed less impactful. Before classification, the dataset underwent a cleaning and filtering process to remove any missing or redundant values. The dataset was then randomly divided into training and testing subsets, with 70% (735 records) used for training and 30% (251 records) for testing. The training data were used fortraining and evaluating the proposed approach.

Table 1: Major attributes of the dataset

Attribute Description	Attribute Name
the slope of the peak exercise ST segment	ST slope
Serum cholesterol (in mg/dl)	Cholesterol
0 = female; 1 = male	Sex
ST depression induced by exercise relative to rest	Oldpeak
Patient's age	Age
Resting BP (in mm Hg)	$BPS_{Resting}$

Kind of Chest Pain	Chest pain
1= Fastingbloodsugar> 120mg/dl; 0= Fastingbloodsugar < 120mg/dl	FBP
Exercise triggered angina (0 = no;1 = yes)	Exercise angina
Maximum heart rate of an individual (in beats/min)	Heart rate $_{Max}$
Resting ECG outcomes	ECG $_{Resting}$
1 = heart attackmight occur; 0 = heart attackmight not occur	Target

3.2 Technique Used

The techniques utilized for various purposes in the proposed study are given as follows:

(i) Discrete Wavelet Transform (DWT)

Reducing the number of attributes used to depict ECG signals is crucial for accurate detection and diagnosis. The ECG data were transformed into time-frequency patterns using the DWT [54]. Recently, the DWT method has been extensively used in processing signals. DWT expresses a signal s(t) as a linear combination of shifted instances of the lowest passing scale operator φ(t) and altered and scaled instances of the basic band-pass wavelet ψ(t).

$$\psi_{j.,k}(t) = 2^{(-y2)}\psi(2^{-j}t - k) \quad (1)$$

$$\varphi_{j,k}(t) = 2^{-j}\varphi(2^{-j}t - k) \quad (2)$$

Where j manages the transformation, k denotes the location of the ψ(t).

One significant benefit of the DWT is its ability to provide high-quality temporal resolution. DWT is capable of accurately identifying the specific temporal and frequency attributes of the input signal due to its exceptional localization capabilities [55].

(ii) Recursive Feature Elimination (RFE)

Feature selection is essential for identifying the most relevant features in various fields [56], including early heart disease detection using ECG data. RFE is a commonly used feature selection method that aims to choose the optimal subset of traits based on model learning and classification accuracy. In large datasets, irrelevant features are common and could impact the efficiency and accuracy of classification algorithms [57]. RFE addresses this by determining the most crucial variables for accurate predictions, reducing dataset dimensionality while preserving valuable characteristics. This iterative method ranks attributes by importance using RF classifiers and retrains the model with the refined feature set to enhance classification accuracy, continually removing less important features until an optimal set is achieved. Figure 2 indicates the workflow diagram of RFE.

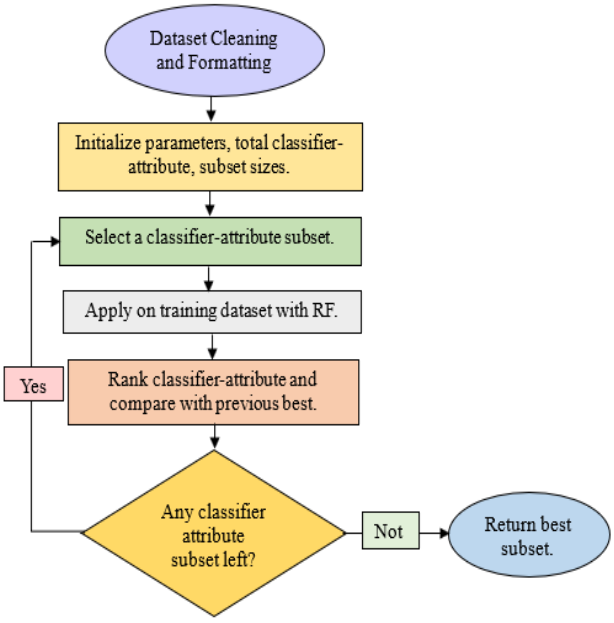


Figure 2. Recursive feature elimination [58].

(iii) Convolutional Neural Network (CNN)

CNN is a neural network that is employed to handle inputs with grid-like topology, such as ECG data. As illustrated in Figure 3, CNNs consist of one or more convolutional layers and are primarily used for tasks like disease detection, classification, prediction and other related data-processing applications.

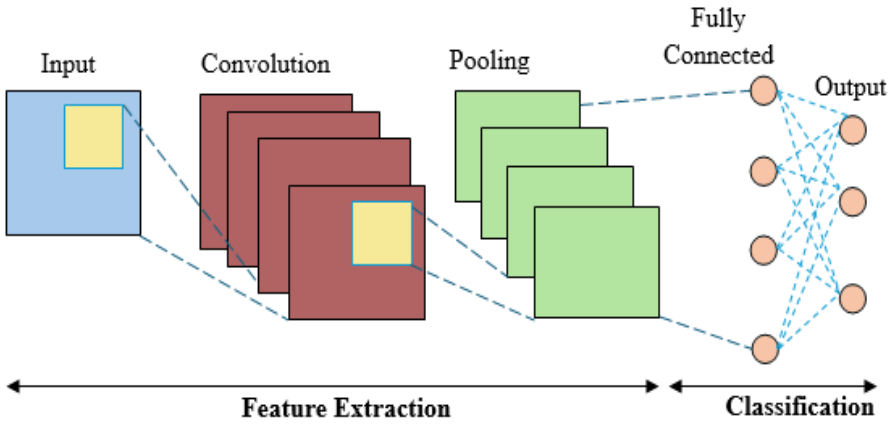


Figure 3. Convolutional Neural Network [59].

Convolution is the method of slipping a filter across an input signal [60] that allows CNNs to analyse the surroundings of a function to presentrecovered or more preciseestimates of its result [61]. This process helps in recognizing features in ECG signals by focusing on smaller sections rather than the entire signal at once. Additionally, these are well-suitedtocommon functional signal processing and segmentation tasks, making them well-suited for early heart disease detection.

(iv) Multi-Layer Perceptron (MLP)

MLPs are among the most commonly used neural network architectures due to their simplicity, structural flexibility, and robust representational capabilities. They are especially effective in the context of heart disease prediction using ECG data. MLPs are feedforward neural networks and universal approximators, typically trained using the backpropagation approach. As supervised networks, they need a preferred response to train, enabling them to understand the transformation of ECG input data into accurate heart disease predictions.

An MLP consists of 3 layers such as input, output and one or more hidden layers, as depicted in Figure 4. With just 1 or 2 hidden layers, MLPs can estimate any input-output relationship, making them highly effective for pattern classification tasks.

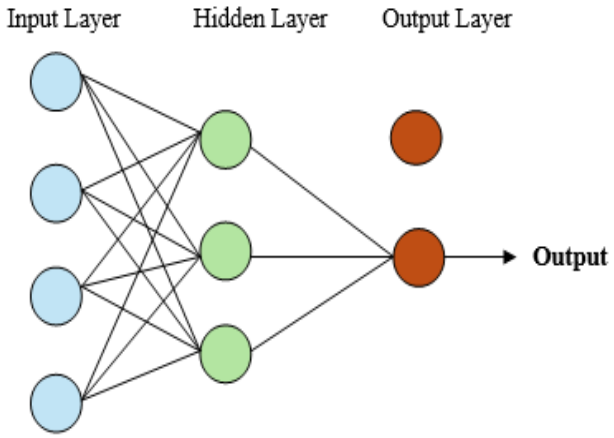


Figure 4. Multi-layer perceptron [62].

The above architecture employs an entirely linked network, in which every neuron in a single layer is coupled to every other neuron in the next layer. This connectivity enhances the MLP's ability to capture complex patterns within ECG data, thereby increasing the precision of heart disease prediction.

(v) Long Short-Term Memory (LSTM)

LSTM was presented to focuson the issue of disappearing or exploding gradients in recurrent neural networks, particularly relevant in heart disease prediction using sequential ECG data. LSTM networks are equipped with internal memory cells accessed by forget and input gate networks, as demonstrated in Figure 5.

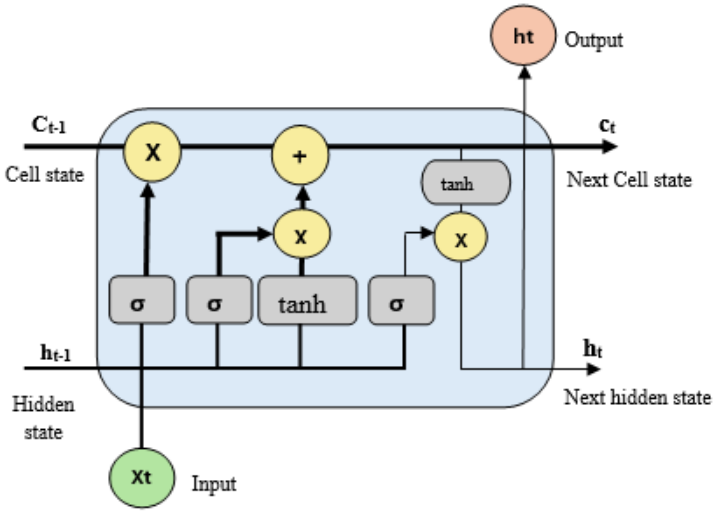


Figure 5. Functional block diagram of LSTM [63].

The forget gate in an LSTM layer regulates the frequency at which memory is passed on to the subsequent time step, whereas the input gate adjusts the amount of incoming input being fed into the memory cells. The LSTM model can capture and depict both short- or long-term relationships in sequential data, relying on the current state of both gates [64], making it particularly effective for analyzing and predicting heart disease from ECG signals.

(vi) Vision Transformer (ViT)

ViT is a neural network architecture, designed for image processing tasks, showing higher accuracy compared to traditional models. The model is inspired from theBidirectional Encoder Representations from Transformers(BERT) [65] and Attention is All You Need models [66]. The model makes use of the attention mechanism, which was originally developed for language recognition purposes, and it introduces the concept of the transformer. ViT can be utilized for the prediction of heart disease from ECG data by extracting 2D patches from the images first. Then, 1D arrays are produced that fit the structure of the ViT. A positional encoder is added to aid in remembering the relative position of the patches for processing before the next layer. The inputs are passed on for normalization and given to the transformer block. In the multi-head attention layer, plays a more critical role [67]. It is the layer through which the multi-head applies weights towards the major regions: this layer will lead the network to the most prominent parts of the ECG signals. The multiple-head attention layer produces a weighted sum of each head, hence enhancing the model's ability to accurately forecast cardiac illness by focusing on key characteristics of the ECG data.

3.3 Proposed Methodology

This section presents the proposed approach to predict heart disease with the help of ML algorithms from ECG data. The heart disease prediction process involves a collection of data from a number of patients, and this data includes the type of patient and whether that patient suffers from heart disease or not. The pre-processing of the collected data is done by missing value imputation or deletion followed by normalization of the feature values. Feature extraction will be done using DWT for extracting relevant features from ECG data. Then, the selected most relevant features are used for the application of RFE. It partitions the dataset into training and testing subsets. Various model architectures, including CNN, MLP, LSTM, and ViT, are defined and trained using the training data. The models are evaluated on the testing set using various metrics. Hyperparameter tuning is conducted to optimize model performance. Finally, the best-performing model is used to predict heart disease in new instances, providing an output indicating the presence or absence of the disease. The flowchart given in Figure 6 depicts the proposed methodology.

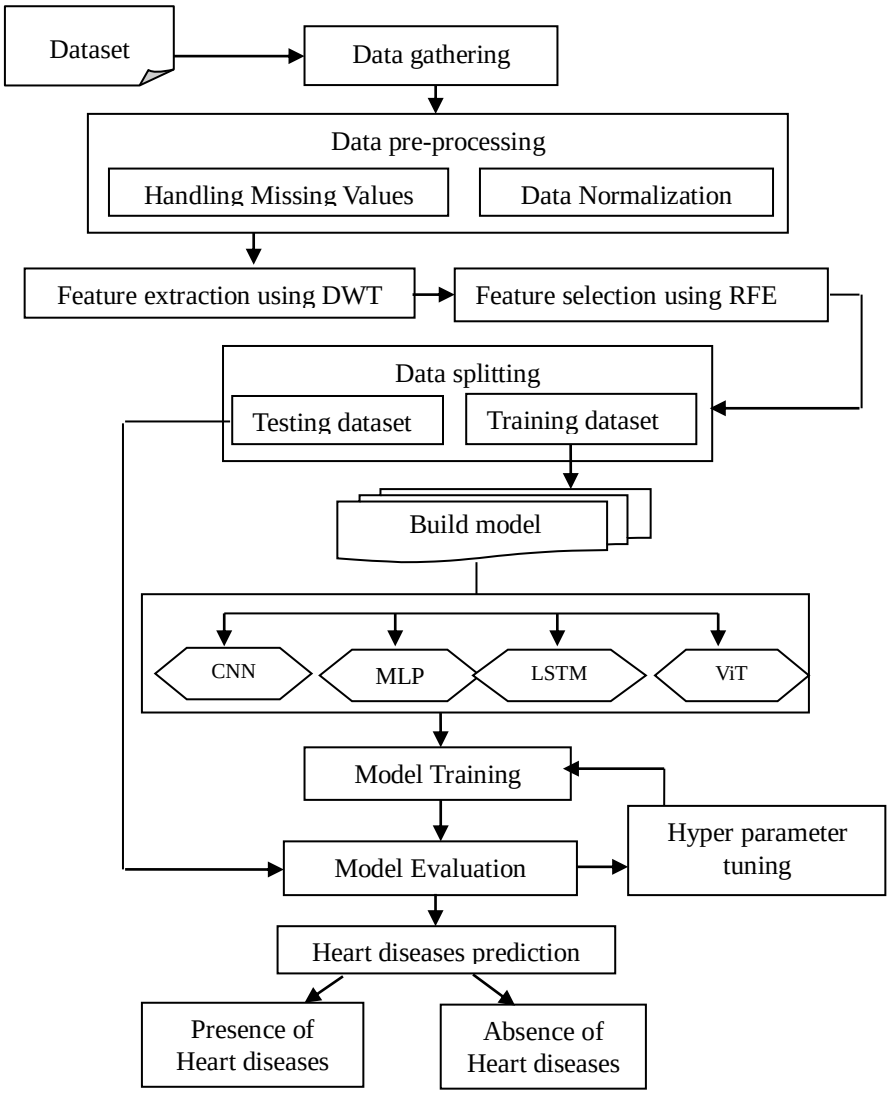


Figure 6. Proposed Methodology

3.4 Proposed Algorithm

Algorithm: Heart Prediction using ML

START

1. Data Collection

Collect dataset D from the patients.

Let $D = \{(x_i, y_i)\}_{i=1}^{n}$

where x_i represents features and y_i Represents labels (heart disease presence/absence).

2. Data Preprocessing

Handling Missing Values:

Identify missing values in D. Replace missing values using imputation methods:

$$x_{i,j} = mean(\{x_{k.j} \mid x_{k.j} \neq missing\})$$

Alternatively, remove instances with missing values.

Data Normalization:

Normalize features to have zero mean and unit variance:

$$x_i = \frac{x_{i,j} - \mu_j}{\sigma_j}$$

Where are μ_j and σ_j the mean and standard deviation of feature j.

3. Feature Extraction

Apply DWT to extract features:

$$W(x) = \sum_{k=0}^{n-1} x_k \,.\, \psi\left(\frac{t - kT}{s}\right)$$

Where ψ is the mother wavelet.

4. Feature Selection

Remove the least important feature and repeat until the desired number of features k is selected.The importance score for feature j:

$$I_j = \sum_{i=1}^{n} |\beta_{i,j}|$$

Where $\beta_{i,j}$ is the weight of feature j, for instance, i.

5. Data Splitting

Split dataset D into the training set D_{train} and testing set D_{test} using a 70-30 ratio:

$$D_{train}, D_{test} = Train\_test\_split(D, test_{size} = 0.3)$$

6. Model Building

Define architectures for CNN, DNN, LSTM, and ViT:

CNN: Convolutional layers followed by pooling layers.

$$Conv(x) = \sigma(W * x + b)$$

MLP: Fully connected layers with activation functions.

$$y = f\left(\sum_{i=1}^{n} w_i\, x_i + b\right)$$

LSTM: LSTM units with gating mechanisms.

$$f_t = \sigma(W_f . [h_{t-1}, x_t] + b_f)$$

ViT: Transformer-based architecture.

$$z_l = MSA\big(LayerNorm(z_l - 1)\big) + z_l - 1$$
$$z_l = MLP\big(LayerNorm(z_l)\big) + z_l$$

7. Model Training

Train models using the training dataset D_{train}:

Minimize loss function L(θ) using optimization algorithms like SGD or Adam.

$$\theta = arg\min_{\theta} L(\theta) = arg\min_{\theta} \frac{1}{n} \sum_{i=1}^{n} \ell\,(f_\theta(x_i), y_i)$$

Where ℓ is the loss function and f_θ is the model.

8. Model Evaluation

Evaluate models using the testing dataset D_{test} on various performance metrics.

9. Hyperparameter Tuning

Define hyperparameter space H. Search for the best hyperparameters θ∗

$$\theta^* = arg\min_{\theta \epsilon H} L_{Val}(\theta)$$

Where L_{Val} is the validation loss.

10. Heart Disease Prediction

Use the best-performing model f_{θ^*} to predict heart disease:

For a new instance x_{new}:

$$\hat{y} = f_{\theta^*}(x_{new})$$

Output the prediction result indicating the presence or absence of heart disease

END

3. Results and Discussion

This section provides the outcomes of the research that are obtained after the implementation of the proposed methodology.

4.1 Evaluation Metrics

The performance assessment of the suggested method is determined by accuracy, precision, recall, and F1 score.

Accuracy: Accuracy is a metric that quantifies the ratio of correct predictions, encompassing both true positives (TP) and true negatives (TN), to the total number of predictions made. It is calculated using the formula:

$$Accuracy = \frac{TP+TN}{FP+FN+TP+TN} \quad (3)$$

where FP represents false positives and FN represents false negatives.

Precision: The ratio of accurately predicted positive cases to the total predicted positive instances is represented as precision.

$$Precision = \frac{TP+FP}{TP} \quad (4)$$

Recall: The ratio of correctly predicted positive instances to all instances that actually belong to the positive class.

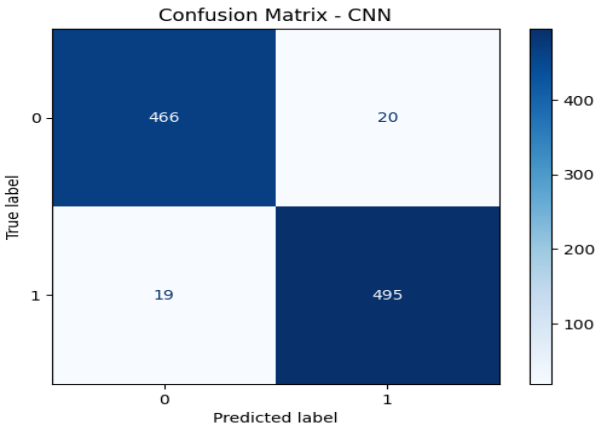
$$Recall = \frac{TP+FN}{TP} \quad (5)$$

F1 Score: The F1 Score is the harmonic mean of precision and recall. It provides a single metric that balances precision and recall concerns.

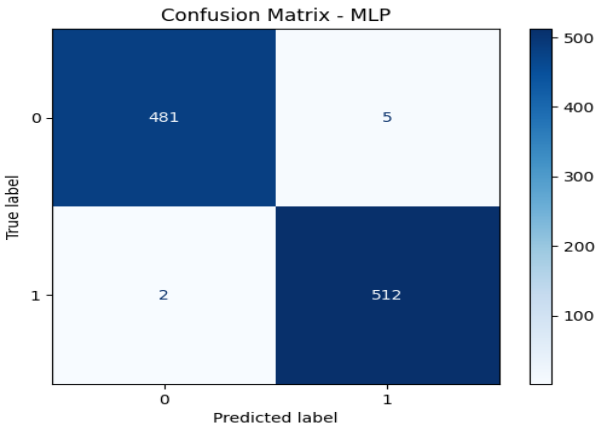
$$F1 - score: = \frac{2*(Precision*Recall)}{(Precision+Recall)} \quad (6)$$

4.2 Confusion Matrix

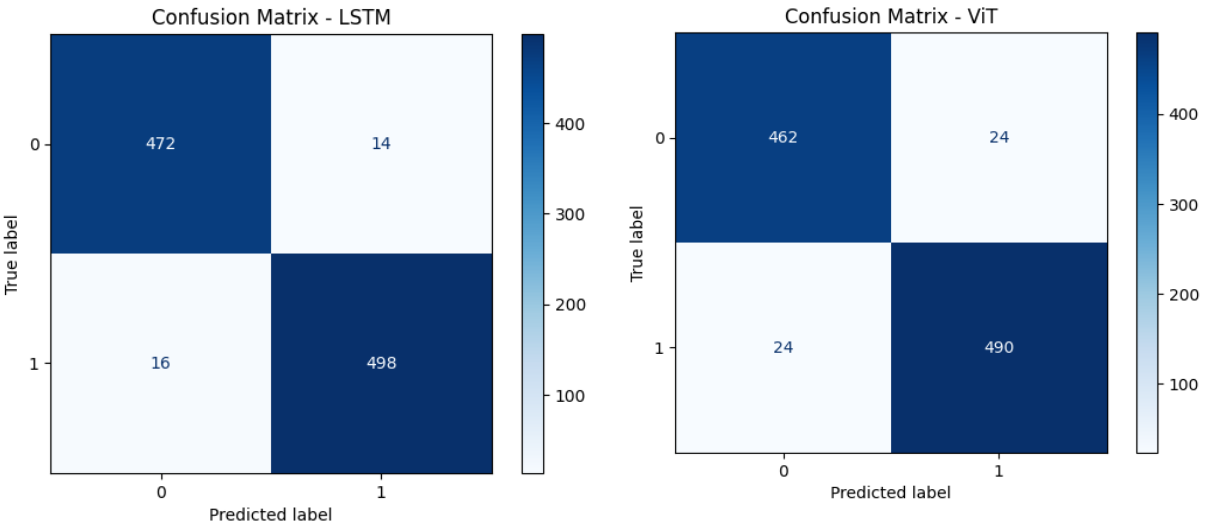
The confusion matrix analysis reveals the performance of various ML models in detecting early heart disease using ECG data. The CNN model effectively identified CVDs, recording 466 TP and 495 TN, but it had 19 FP and 20 FN, suggesting some healthy individuals were misclassified as having heart disease, while a few patients were overlooked as shown in Figure 7(a). The MLP model outperformed others, achieving 481 TP and 512 TN, with only 5 FP and 2 FN, indicating minimal errors and a high accuracy in detecting heart disease as depicted in Figure 7 (b). The LSTM model classified 472 TP and 498 TN, demonstrating a good ability to detect both conditions, but it misclassified 14 FN and 16 FP, reflecting a slightly less accurate performance compared to the MLP as depicted in Figure 7(c). Lastly, the ViT model identified 462 TP and 490 TN, but it had 24 FP and 24 FN, indicating some misclassifications of healthy individuals and missed cases of heart diseaseas shown in Figure 7(d).



(a) Confusion matrix of CNN



(b). Confusion matrix of MLP



(c) Confusion matrix of LSTM (d) Confusion matrixof ViT

Figure 7: Confusion matrix

4.3 Performance Evaluation

This section evaluates the performance of ML models for early heart disease detection, focusing on accuracy, precision, recall and f1-score.

- **Based on the accuracy**

Figure 8 illustrates the accuracy of ML algorithms in the early diagnosis of heart disease. The MLP model attained the maximum accuracy of 99.3%, succeeded by LSTM at 97%, CNN at 96.1%, and ViT at 95.2%.While all models performed well, the MLP model demonstrated superior predictive accuracy in this analysis.

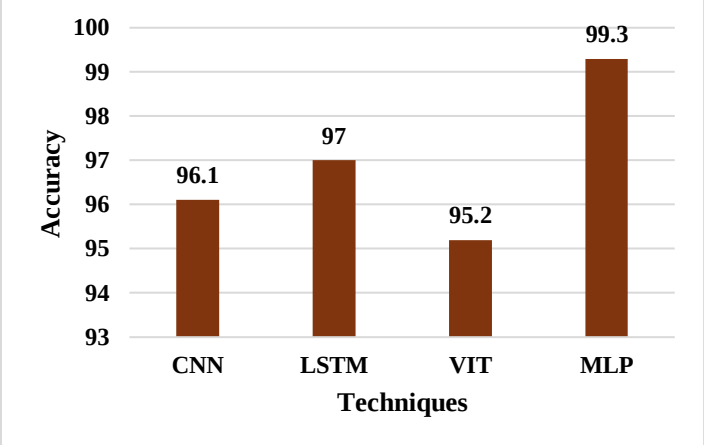


Figure 8: Accuracy

- **Based on the Precision**

Figure 9 depicts the precision of several ML algorithms for the early detection of heart disease. MLP scored the highest precision at 99.03%, which means that it performed the best in terms of accuracy in heart disease detection. The LSTM model scored a high precision at 97.27%, showing the strength of the model in handling time-series data like ECG signals. The CNN scored a precision of 96.12% because it can capture spatial features. The accuracy of the VIT model was marginally lesser by 5.23% precision, hence more suited comparatively to the other experiments.

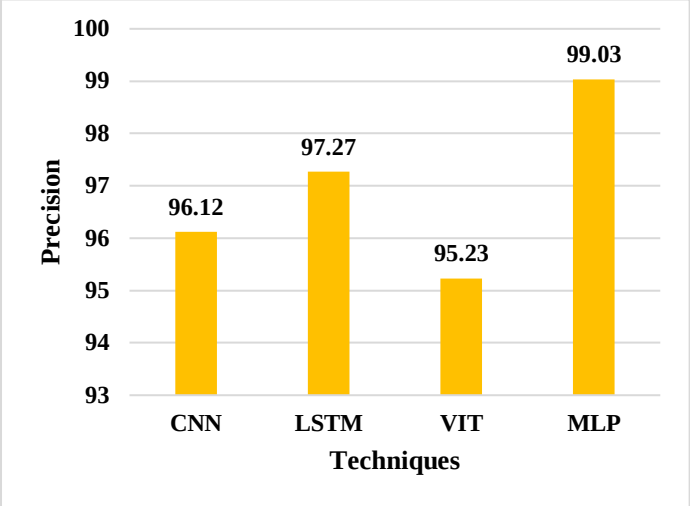


Figure 9: Precision

- **Based on the Recall**

Figure 10 shows the comparison of the recall of ML algorithms in detecting early heart disease with ECG data. MLP showed the highest recall value of 99.61%, thereby showing an excellent capability for positive case detection. The LSTM also came very close, with a recall value of 96.89%. This represents its high ability to detect the heart disease event. The CNN has shown a recall of 96.30% and captures a good amount of the actual positive cases. This was contradictory as the recall of the ViT model was at its lowest, at 93.63%, which means that the model was unable to identify the true positive more effectively compared to other models. In general, the MLP model scored higher in recall for this test.

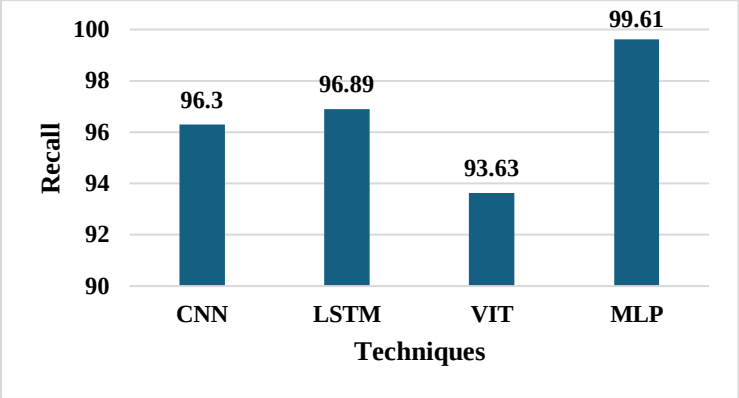


Figure 10: Recall

- **Based on the F1-score**

Figure 11 plots the F1-scores of the ML algorithms designed for early heart disease diagnosis using ECG data. The highest F1-score was 99.83% of the MLP, which proves to be balanced between precision and recall values in predicting heart disease cases. Next, the best F1-score of the ViT model was reported at 99.32%. LSTM scored an F1-score of 97.08%, which gives this approach a good view regarding dealing with false positives and false negatives. CNN came close to the F1-score, that being 96.21% which is much smaller and still represents being really efficient in that case as well. By general analysis, MLP appears to be performing with the high F1 score.

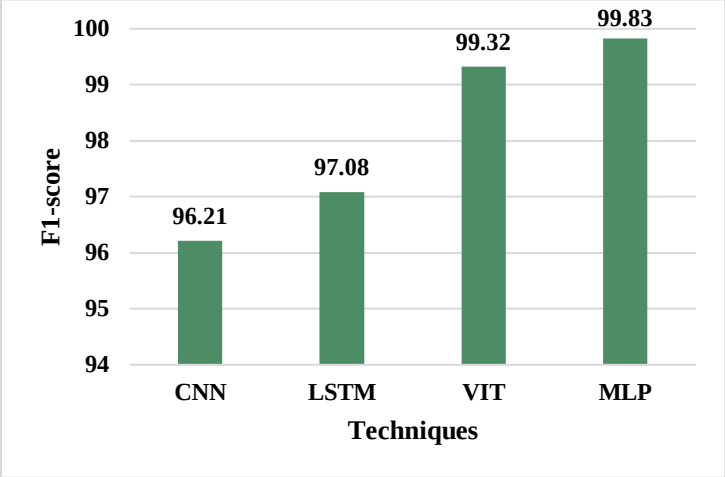


Figure 11: F1-score

4.4 Comparative Analysis

Table 2 shows a comparison of the ML algorithms in the early detection of heart disease using ECG data, showing several models and their accuracy. The highest accuracy was found in the Multi-Layer Perceptron, with 99%. This clearly shows the applicability of this model. With a little difference, the accuracy of the Random Forest was found to be 98%, showing the complexity-handling capacity of the model and providing good predictions in this context.Another powerful algorithm was XGBoost that yields 95.9%, SqueezeNet, was a compact model that still showed high accuracy at 93.75%. Lastly, the Optimized Deep Neural Network (DNN) had the lowest accuracy at 87.7%, yet it still provided competitive results. Such comparison presents the strength behind ensemble methods and deep learning models that can be generated in terms of detecting heart disease based on ECG data; MLP and RF performed much better than those used above.

Table 2: Comparative analysis

Authors	Feature engineering	Technique name	Accuracy
Subba et al., (2024) [68]	Wavelet transforms	RF	98%
Suryani et al., (2022) [69]	Wrapper and filtering	Optimized DNN	87.7%
Nagavelli et al., (2022) [70]	DWT	XGBoost	95.9%
Al Fahoum et al., (2023) [71]	Wavelet transform	SqueezeNet	93.75%
Proposed study	**DWT+RFE**	**MLP**	**99%**

4. Conclusion

This study used ECG data for the purpose of comparing several ML algorithms towards solving the challenging problem of early heart disease detection. The scope of the study was to include the use of DWT for feature extraction and RFE to select the most relevant features, thus guaranteeing a robust model performance. ML models CNN, MLP, LSTM, and ViT were applied in this study and tested on a dataset of 986 patient ECG records. The main findings show that the MLP model achieved a maximum accuracy of 99.3%, which shows good performance in the early diagnosis of heart disease, while LSTM and CNN also have similar metrics. Based on the above results, this paper highlights the potential of ML to enhance the accuracy of diagnostics for cardiovascular diseases.

For future work, a more complex and hybrid architecture of deep learning needs to be explored to further enhance the predictiveness and generalization capabilities of these models. A bigger population dataset could be prepared covering a larger and diverse range of patients. Also, other diagnostic tools and aids could be integrated for gaining a better insight from this model and thus to have real-world applicability.

Reference

[1] Hinton, Robert B., and Katherine E. Yutzey. "Heart valve structure and function in development and disease." Annual review of physiology 73, no. 1 (2011): 29-46.

[2] Jagannathan, Ram, Shivani A. Patel, Mohammed K. Ali, and KM Venkat Narayan. "Global updates on cardiovascular disease mortality trends and attribution of traditional risk factors." *Current diabetes reports* 19 (2019): 1-12.
[3] https://www.who.int/health-topics/cardiovascular-diseases/#tab=tab_1
[4] Cushman, Mary, Christina M. Shay, Virginia J. Howard, Monik C. Jiménez, Jennifer Lewey, Jean C. McSweeney, L. Kristin Newby et al. "Ten-year differences in women's awareness related to coronary heart disease: results of the 2019 American Heart Association National Survey: a special report from the American Heart Association." *Circulation* 143, no. 7 (2021): e239-e248.
[5] Kumar, Dheeraj, and Luv Kumar. "The Pulmonary Artery Diameter Variation among Smokers and Non-smokers." Tuijin Jishu/Journal of Propulsion Technology 45, no. 1: 2024.
[6] https://www.creative-diagnostics.com/Cardiac-Disease.htm
[7] Bully, Paola, Álvaro Sánchez, Edurne Zabaleta-del-Olmo, Haizea Pombo, and Gonzalo Grandes. "Evidence from interventions based on theoretical models for lifestyle modification (physical activity, diet, alcohol and tobacco use) in primary care settings: a systematic review." *Preventive medicine* 76 (2015): S76-S93.
[8] Balwan, WahiedKhawar, and SachdeepKour. "Lifestyle Diseases: The Link between Modern Lifestyle and threat to public health." *Saudi J Med Pharm Sci* 7, no. 4 (2021): 179-84.
[9] Spring, Bonnie, Abby C. King, Sherry L. Pagoto, Linda Van Horn, and Jeffery D. Fisher. "Fostering multiple healthy lifestyle behaviors for primary prevention of cancer." *American Psychologist* 70, no. 2 (2015): 75.
[10] Jaarsma, Tiny, Loreena Hill, Antoni Bayes-Genis, Hans-Peter Brunner La Rocca, Teresa Castiello, JelenaČelutkienė, Elena Marques-Sule et al. "Self-care of heart failure patients: practical management recommendations from the Heart Failure Association of the European Society of Cardiology." *European journal of heart failure* 23, no. 1 (2021): 157-174.
[11] Çolakoğlu, Nurdan, and Berke Akkaya. "Comparison of multi-class classification algorithms on early diagnosis of heart diseases." In *Proc. Int. Conf. Recent Develop. Data Sci. Bus. Anal*, p. 162. 2019.
[12] Kapila, Ramdas, ThirumalaisamyRagunathan, SumalathaSaleti, T. Jaya Lakshmi, and MohdWazih Ahmad. "Heart disease prediction using novel quine McCluskey binary classifier (QMBC)." *IEEE Access* 11 (2023): 64324-64347.
[13] Diller, Gerhard-Paul, Alexandra Arvanitaki, Alexander R. Opotowsky, Kathy Jenkins, Philip Moons, Alexander Kempny, Animesh Tandon et al. "Lifespan perspective on congenital heart disease research: JACC state-of-the-art review." *Journal of the American College of Cardiology* 77, no. 17 (2021): 2219-2235.
[14] Mahalie, Roswitha. "A novel surveilance framework for tracking and predicting health outcomes of cardiovascular diseases risk factors among people living with HIV initiated on art in Khomas region, Namibia." PhD diss., University of Namibia, 2021.
[15] Du, Xuewei, Xujie Su, Wanxue Zhang, Suyan Yi, Ge Zhang, Shan Jiang, Hui Li, Shaoguang Li, and Fan Xia. "Progress, opportunities, and challenges of troponin analysis in the early diagnosis of cardiovascular diseases." *Analytical Chemistry* 94, no. 1 (2021): 442-463.
[16] Saeed, Anum, Kaustubh Dabhadkar, Salim S. Virani, Peter H. Jones, Christie M. Ballantyne, and Vijay Nambi. "Cardiovascular disease prevention: training opportunities, the challenges, and future directions." *Current atherosclerosis reports* 20 (2018): 1-7.
[17] Kazemian, Negin, MortezaMahmoudi, Frank Halperin, Joseph C. Wu, and SepidehPakpour. "Gut microbiota and cardiovascular disease: opportunities and challenges." *Microbiome* 8 (2020): 1-17.
[18] Shafqat, Sarah, SairaKishwer, Raihan Ur Rasool, JunaidQadir, TehminaAmjad, and Hafiz Farooq Ahmad. "Big data analytics enhanced healthcare systems: a review." *The Journal of Supercomputing* 76 (2020): 1754-1799.
[19] National Academies of Sciences, Medicine Division, Board on Global Health, and Committee on Improving the Quality of Health Care Globally. "Crossing the global quality chasm: improving health care worldwide." (2018).

[20] Dash, Sabyasachi, Sushil Kumar Shakyawar, Mohit Sharma, and Sandeep Kaushik. "Big data in healthcare: management, analysis and future prospects." *Journal of big data* 6, no. 1 (2019): 1-25.
[21] Mahalakshmi, K., and P. Sujatha. "Critical Analysis of Feature Selection Methods for Data Preprocessing with Heart Disease Dataset." In *Data Intelligence and Cognitive Informatics: Proceedings of ICDICI 2021*, pp. 667-682. Singapore: Springer Nature Singapore, 2022.
[22] Javid, Irfan, RozaidaGhazali, Muhammad Zulqarnain, and Norlida Hassan. "Data pre-processing for cardiovascular disease classification: A systematic literature review." *Journal of Intelligent & Fuzzy Systems* 44, no. 1 (2023): 1525-1545.
[23] Sami, Osamah, YousefElsheikh, and FadiAlmasalha. "The role of data pre-processing techniques in improving machine learning accuracy for predicting coronary heart disease." *International Journal of Advanced Computer Science and Applications* 12, no. 6 (2021).
[24] Wilson, Peter WF, Ralph B. D'Agostino, Daniel Levy, Albert M. Belanger, HalitSilbershatz, and William B. Kannel. "Prediction of coronary heart disease using risk factor categories." *Circulation* 97, no. 18 (1998): 1837-1847.
[25] Allen, Christopher J., Kaushik Guha, and Rakesh Sharma. "How to improve time to diagnosis in acute heart failure–clinical signs and chest x-ray." *Cardiac failure review* 1, no. 2 (2015): 69.
[26] Waldmann, Victor, Nicolas Combes, Magalie Ladouceur, David S. Celermajer, Laurence Iserin, Michael A. Gatzoulis, Paul Khairy, and Eloi Marijon. "Understanding electrocardiography in adult patients with congenital heart disease: a review." *JAMA cardiology* 5, no. 12 (2020): 1435-1444.
[27] Ripley, D. P., T. A. Musa, L. E. Dobson, S. Plein, and J. P. Greenwood. "Cardiovascular magnetic resonance imaging: what the general cardiologist should know." *Heart* 102, no. 19 (2016): 1589-1603.
[28] Bu, Guilin, Ying Miao, Jingwen Bin, Sheng Deng, Taowen Liu, Hongchun Jiang, and Weiping Chen. "Comparison of 128-slice low-dose prospective ECG-gated CT scanning and trans-thoracic echocardiography for the diagnosis of complex congenital heart disease." *PloS one* 11, no. 10 (2016): e0165617.
[29] Honnashamaiah, Anu. "Dr. RathnakaraS.,"Detection of Arrhythmia in ECG signal using Deep Learning Methods–A exhaustive review & summary of the concepts & techniques"." Tuijin Jishu/Journal of Propulsion Technology, ISSN: 1001-4055.
[30] Krittanawong, Chayakrit, HongJu Zhang, Zhen Wang, Mehmet Aydar, and Takeshi Kitai, "Artificial intelligence in precision cardiovascular medicine," Journal of the American College of Cardiology, vol. 69, no. 21, pp 2657-2664, 2017
[31] Rajkomar, Alvin, Jeffrey Dean, and Isaac Kohane, "Machine learning in medicine," New England Journal of Medicine, vol. 380, no. 14, pp 1347-1358, 2019
[32] Vadranam, Teja Naidu, BV Brahma Rao, and VVR MaheswaraRao. "Detection and Prediction of Comorbities of Diabetes Using Machine Learning Techniques." Tuijin Jishu/Journal of Propulsion Technology 45, no. 2: 2024.
[33] Ketu, Shwet, and Pramod Kumar Mishra. "Hybrid classification model for eye state detection using electroencephalogram signals." *Cognitive Neurodynamics* 16, no. 1 (2022): 73-90.
[34] Ketu, Shwet, and Pramod Kumar Mishra. "Performance analysis of machine learning algorithms for IoT-based human activity recognition." In *Advances in Electrical and Computer Technologies: Select Proceedings of ICAECT 2019*, pp. 579-591. Springer Singapore, 2020.
[35] Ketu, Shwet, and Pramod Kumar Mishra. "Enhanced Gaussian process regression-based forecasting model for COVID-19 outbreak and significance of IoT for its detection." *Applied Intelligence* 51, no. 3 (2021): 1492-1512.
[36] Ketu, Shwet, and Pramod Kumar Mishra. "Empirical analysis of machine learning algorithms on imbalance electrocardiogram based arrhythmia dataset for heart disease detection." *Arabian Journal for Science and Engineering* 47, no. 2 (2022): 1447-1469.
[37] Saxena, K.; Sharma, R.: Efcient heart disease prediction system. ProcediaComput. Sci. 85, 962–969 (2016)

[38] Samuel, Oluwarotimi Williams, Grace MojisolaAsogbon, Arun Kumar Sangaiah, Peng Fang, and Guanglin Li. "An integrated decision support system based on ANN and Fuzzy_AHP for heart failure risk prediction." *Expert systems with applications* 68 (2017): 163-172.
[39] Boon, Khang Hua, Mohamed Khalil-Hani, and M. B. Malarvili. "Paroxysmal atrial fibrillation prediction based on HRV analysis and non-dominated sorting genetic algorithm III." *Computer methods and programs in biomedicine* 153 (2018): 171-184.
[40] Diwakar, Manoj, AmrendraTripathi, Kapil Joshi, MinakshiMemoria, and Prabhishek Singh. "Latest trends on heart disease prediction using machine learning and image fusion." *Materials today: proceedings* 37 (2021): 3213-3218.
[41] Ali, Farman, Shaker El-Sappagh, SM Riazul Islam, Daehan Kwak, Amjad Ali, Muhammad Imran, and Kyung-Sup Kwak. "A smart healthcare monitoring system for heart disease prediction based on ensemble deep learning and feature fusion." *Information Fusion* 63 (2020): 208-222.
[42] Yusuf, Salim, Philip Joseph, SumathyRangarajan, Shofiqul Islam, Andrew Mente, Perry Hystad, Michael Brauer et al. "Modifiable risk factors, cardiovascular disease, and mortality in 155 722 individuals from 21 high-income, middle-income, and low-income countries (PURE): a prospective cohort study." *The Lancet* 395, no. 10226 (2020): 795-808.
[43] Singh, D. P. "An Extensive Examination of Machine Learning Methods for Identifying Diabetes." Tuijin Jishu/Journal of Propulsion Technology 45, no. 2: 2024.
[44] Pachiyannan, Prabu, MuslehAlsulami, DeafallahAlsadie, Abdul KhaderJilaniSaudagar, Mohammed AlKhathami, and Ramesh Chandra Poonia. "A Novel Machine Learning-Based Prediction Method for Early Detection and Diagnosis of Congenital Heart Disease Using ECG Signal Processing." *Technologies* 12, no. 1 (2024): 4.
[45] Alimbayeva, Zhadyra, ChingizAlimbayev, KassymbekOzhikenov, NurlanBayanbay, and AimanOzhikenova. "Wearable ECG Device and Machine Learning for Heart Monitoring." *Sensors* 24, no. 13 (2024): 4201.
[46] Ribeiro, Pedro, Joana Sá, Daniela Paiva, and Pedro Miguel Rodrigues. "Cardiovascular diseases diagnosis using an ECG multi-band non-linear machine learning framework analysis." *Bioengineering* 11, no. 1 (2024): 58.
[47] Utsha, UcchwasTalukder, I. Hua Tsai, and Bashir I. Morshed. "A smart health application for real-time cardiac disease detection and diagnosis using machine learning on ecg data." In *IFIP International Internet of Things Conference*, pp. 135-150. Cham: Springer Nature Switzerland, 2023.
[48] Baghdadi, Nadiah A., Sally Mohammed FarghalyAbdelaliem, AmerMalki, Ibrahim Gad, Ashraf Ewis, and ElsayedAtlam. "Advanced machine learning techniques for cardiovascular disease early detection and diagnosis." *Journal of Big Data* 10, no. 1 (2023): 144.
[49] Yilmaz, Rüstem, and FatmaHilalYağin. "Early detection of coronary heart disease based on machine learning methods." *Medical Records* 4, no. 1 (2022): 1-6.
[50] Hossain, AdibaIbnat, SabitriSikder, Annesha Das, and Ashim Dey. "Applying machine learning classifiers on ECG dataset for predicting heart disease." In *2021 International Conference on Automation, Control and Mechatronics for Industry 4.0 (ACMI)*, pp. 1-6. IEEE, 2021.
[51] Anuar, NayanNazrul, Hamid Hafifah, SubohMohdZubir, Abdullah Noraidatulakma, JaafarRosmina, MhdNurul Ain, Hamid MariatulAkma et al. "Cardiovascular disease prediction from electrocardiogram by using machine learning." (2020): 34-48.
[52] Tyagi, Ankita, and Ritika Mehra. "Intellectual heartbeats classification model for diagnosis of heart disease from ECG signal using hybrid convolutional neural network with GOA." *SN Applied Sciences* 3, no. 2 (2021): 265.
[53] Hammad, Mohamed, Asmaa Maher, Kuanquan Wang, Feng Jiang, and MoussaAmrani. "Detection of abnormal heart conditions based on characteristics of ECG signals." *Measurement* 125 (2018): 634-644.
[54] Sharma, Manish, Ram BilasPachori, and U. Rajendra Acharya. "A new approach to characterize epileptic seizures using analytic time-frequency flexible wavelet transform and fractal dimension." Pattern Recognition Letters 94 (2017): 172-179.

[55] Srivastava, V. K., and Devendra Prasad. "DWT-based feature extraction from ECG signal." *American J. of Eng. Research (AJER)* 2, no. 3 (2013): 44-50.
[56] Tang, J.; Alelyani, S.; Liu, H. Feature selection for classification: A review. Data Class Algor. Appl. 2014, 37, 1–29.
[57] Faysal, Jabed Al, SkTahmidMostafa, Jannatul Sultana Tamanna, KhondokerMirazulMumenin, MdMashrurArifin, Md Abdul Awal, AtanuShome, and Sheikh ShanawazMostafa. "XGB-RF: A hybrid machine learning approach for IoT intrusion detection." In *Telecom*, vol. 3, no. 1, pp. 52-69. MDPI, 2022.
[58] Sharma, Ashish, Dinesh Bhuriya, and Upendra Singh. "Survey of stock market prediction using machine learning approach." In *2017 International conference of electronics, communication and aerospace technology (ICECA)*, vol. 2, pp. 506-509. IEEE, 2017.
[59] Rinish Reddy, R., SadhwikaRachamalla, Mohamed Sirajudeen Yoosuf, and G. R. Anil. "Convolutional Neural Network Based Intrusion Detection System and Predicting the DDoS Attack." In *Data Intelligence and Cognitive Informatics: Proceedings of ICDICI 2022*, pp. 81-94. Singapore: Springer Nature Singapore, 2022.
[60] Khudayer, BaidaaHamza, Mohammed Anbar, Sabri M. Hanshi, and Tat-Chee Wan. "Efficient route discovery and link failure detection mechanisms for source routing protocol in mobile ad-hoc networks." IEEE Access 8 (2020): 24019-24032
[61] Lu, Renjie. "Malware detection with lstm using opcode language." arXiv preprint arXiv:1906.04593 (2019).
[62] Aeinfar, Vahid, HooriehMazdarani, FatemehDeregeh, Mohsen Hayati, and MehrdadPayandeh. "Multilayer Perceptron Neural Network with supervised training method for diagnosis and predicting blood disorder and cancer." In *2009 IEEE International Symposium on Industrial Electronics*, pp. 2075-2080. IEEE, 2009.
[63] Le, Xuan-Hien, Hung Viet Ho, Giha Lee, and Sungho Jung. "Application of long short-term memory (LSTM) neural network for flood forecasting." *Water* 11, no. 7 (2019): 1387.
[64] Sharma, Rohit, Sachin S. Kamble, AngappaGunasekaran, Vikas Kumar, and Anil Kumar. "A systematic literature review on machine learning applications for sustainable agriculture supply chain performance." Computers & Operations Research 119 (2020): 104926
[65] Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805* (2018).
[66] Ashish, Vaswani. "Attention is all you need." *Advances in neural information processing systems* 30 (2017): I.
[67] Borhani, Yasamin, JavadKhoramdel, and EsmaeilNajafi. "A deep learning-based approach for automated plant disease classification using vision transformer." *Scientific Reports* 12, no. 1 (2022): 11554.
[68] Subba, Tanuja, and TejbantaChingtham. "Comparative Analysis of Machine Learning Algorithms With Advanced Feature Extraction for ECG Signal Classification." *IEEE Access* (2024).
[69] Suryani, Esti, SigitSetyawan, and BintangPe Putra. "The cost-based feature selection model for coronary heart disease diagnosis system using deep neural network." *IEEE Access* 10 (2022): 29687-29697.
[70] Nagavelli, Umarani, Debabrata Samanta, and Partha Chakraborty. "Machine Learning Technology-Based Heart Disease Detection Models." *Journal of Healthcare Engineering* 2022, no. 1 (2022): 7351061.
[71] Al Fahoum, A. "Enhanced cardiac arrhythmia detection utilizing deep learning architectures and multi-scale ECG analysis." TuijinJishu/J Propulsion Technol 44, no. 6 (2023): 5539-5554.