Advancing Cardiovascular Disease Detection Using Machine Learning

Devansh Sharma[1], Laivish Sharma[2], Savita Dagar[3]

{sharmadevansh0001@gmail.com[1], laivishsharma123@gmail.com[2], savitadagar.sca@mriu.edu.in[3]}
School of Computer Applications Manav Rachna International Institute of Research & Studies, Faridabad, India[1,2,3]

Abstract- Cardiovascular diseases (CVDs) have remained one of the leading causes of mortality in the global population, and their diagnosis is also not timely due to the limitations of traditional diagnostic methods, which are believed to be more prone to interpretation errors and slow in detecting risk factors of cardiovascular disease. This research paper explains how machine learning algorithms can be used together to help detect CVDs more effectively through the analysis of multimedia sources of data, such as electrocardiograms, imaging and patient records. This is based on supervised and unsupervised learning techniques that are developed and have proven to be successful, including ensemble models and deep neural networks, which identify features and categories patterns so as to increase the accuracy and efficacy of diagnostics. Most notably, one can observe that machine learning will permit more varied CVD markers to be detected, enable timelier intervention compared to traditional systems and aid in lowering diagnostic error. The key findings of machine learning in clinical cardiology suggest additional integration of machine learning to optimize patient outcomes and support healthcare processes, and require that the potential of machine learning be applied in real-world settings in an ethical and validated manner.

Keywords- cardiovascular disease, deep neural networks, electrocardiograms, machine learning.

1. Introduction

Cardiovascular diseases (CVDs) can be considered one of the most acute health issues of the present day, that causes deaths of millions of people every year, and leaves a significant financial scope in the generic sense all through the healthcare system on a global scale. The increasing prevalence of ill health (coronary artery disease, heart failure, arrhythmias), as a result of an ageing population, and growth of sedentary behaviours, improper diets and stressors, clearly demands new diagnostic tools [1]. Echocardiography, biochemical, and electrocardiography are considered the classical CVD diagnosis tools and can provide valuable data regarding the functioning of the heart and its pathology. These methods are often dependent on the human interpretation of the expert, and can be subjective and variable, slow in diagnosing, or unable to pick up subtle changes. Information analytics and medical imaging technologies have evolved

IICTDS 2025, December 19-20, Chandigarh, India
Copyright © 2026 EAI
DOI 10.4108/eai.19-12-2025.2366223

rather gradually over the last few decades and changed the game a bit, allowing for risk stratification and monitoring [2]. Computational tools are of growing particular concern, partly as an extension of human experience, but very early efforts have been undertaken to apply them to automated signal processing (arrhythmia detection) and statistical modelling (prognostic evaluation). Expanding on the same line of reasoning, the application of artificial intelligence has gone off the scale as a pressing concern, which can give birth to domains of incompetence and lack of accuracy. Research has shown that certain algorithms could be effectively employed to analyse massive amounts of electronic health records, wearable and imaging solutions to find some patterns that could not be detected using more common methods. One is that they have used supervised learning algorithms to identify cardiac rhythms in electrocardiogram recordings, and unsupervised algorithms to identify hidden subsets within a population of patients, enabling personalised medicine [3].

Despite these attempts, the domain of CVD detection is plagued by issues which hurt the potential of the domain. However, when compared to application of traditional diagnostic tools, the application is likely to fail in different populations because comorbidities and abnormal manifestations complicate the evaluation process. Unlike optimism regarding these early AI integrations, solutions are susceptible to a number of factors, such as overfitting to specific data or vulnerability to noise in real-world data and generalisation to a variety of contexts in healthcare. There is substantial variance in the ongoing generation of multimodal data sets, consisting of physiological data, genetic data and lifestyle data, which would provide a more detailed picture of the cardiovascular state, but demands data integration and the highest complexity of computations [4]. Numerous valid questions arise regarding how these technologies can be scaled to resource-constrained settings with unequal access to quality data and computing assets, and these challenges may contribute to widening health disparities (even more). In addition, machine learning model interpretability is a puzzle; black-box systems may make accurate predictions but fail to deliver useful suggestions to facilitate clinical application, which erodes trust and adoption. Such a situation calls on the historical tradition within medical informatics of focusing not merely on innovation, but on its practical applicability and on extending existing paradigms of pattern recognition and decision support in light of these constraints.

In this paper, the author has tried to add to the field of cardiovascular disease detection through intelligently using machine learning to improve the quality of diagnosis without developing new frameworks. The paper investigates how current algorithms, such as ensemble algorithms and deep learning architectures, can be used to analyse and reason about multi-faceted streams of data provided by typical clinical sources. The current research includes an in-depth examination of the cardiographic tendencies, image characteristics, and demographicization, including methodologies to choose features and model sample in such a way as to avoid the majority of more typical traps, such as bias and variance. Through such means, the work aims to show that machine learning can give the available diagnostics additional capacity, and the former assertion would enable it to detect possible people at risk promptly, and the latter would enable it to offer a deeper risk analysis.

Such techniques are far more efficient in identifying mild evidence of CVD, including first-degree arrhythmias and vascular defects, through relying on an ability to identify trends that are more predictable and easier to find than in purely human judgment. Additionally, unsupervised clustering can, on some unknown level, reveal the existence of subpopulations of patients in the population and introduce the potential of personalised treatment. The results of those studies result in the eventuality of implementing machine learning in the regular working cardiology process to make the turnaround in diagnosing quicker and decrease the cost of

healthcare through preventive care [5].

2. Literature Review

Table 1 presents the reviewed studies, highlighting their contributions, methodologies, and limitations within this domain.

Table 1. Related works.

Authors and Year	Paper Title	About the Paper	Objective	Methodology Used	Limitations
Paul and Masood (2024)	Exploring Predictive Methods for Cardiovascular Disease: A Survey of Methods and Applications	Surveys predictive models for CVD using biochemical, clinical, and demographic data, comparing ML and DL approaches.	To review and compare CVD prediction models for early detection and personalized medicine.	ML (RF, LR, SVM) and DL (neural networks); feature selection; evaluated via accuracy, sensitivity, specificity, AUC-ROC.	Survey nature lacks original experiments; focuses on traditional factors, ignoring wearables/genetics; potential bias in dataset selection.
Chen et al. (2023)	Self-Supervised Learning-Based General Laboratory Progress Pretrained Model for Cardiovascular Event Detection	Uses SSL to pretrain a model on lab data for transferring knowledge to detect target vessel revascularization in CVD.	To leverage prevalent CVD data for rare event detection via transferable lab progression knowledge.	Two-stage SSL pretraining (GLP model); interpolation for irregularity; transfer learning; evaluated on accuracy for TVR detection.	Limited to six lab markers; episodic data constraints; no multi-center validation for transferability.
Khan et al. (2024)	Heart Disease Prediction Using Novel Ensemble and Blending Based Cardiovascular Disease Detection Networks:	Develops DL models (ensemble/blending) for CVD prediction considering health/socio-economic factors, with SHAP explainability.	To accurately predict CVDs mitigating data imbalance, using ensemble/blending for better performance.	ADASYN for balancing; Point Biserial Correlation for selection; LeNet+GRU ensemble (EnsCVDD-Net); LeNet+GRU+MLP blending (BlCVDD-Net); SHAP; 10-fold CV.	High execution time for models; imbalanced data handling may skew socio-economic features; no mobile deployment.

		EnsCVDD-Net and BlCVDD-Net			
El-Sofany (2024)	Predicting Heart Diseases Using Machine Learning and Different Data Classification Techniques	Employs feature selection and ML classifiers for early HD prediction, developing a mobile app with SHAP explainability.	To build an accurate, low-cost ML system for early HD detection using balanced data.	Feature selection (Chi-square, ANOVA, MI); classifiers (NB, SVM, Voting, XGBoost, etc.); SMOTE for imbalance; SHAP for XAI; private/public datasets; cross-validation.	Unbalanced data handling may introduce bias; app usability not user-tested; limited to specific symptoms input.
Akther et al. (2024)	Multi-Faceted Approach to Cardiovascular Risk Assessment by Utilizing Predictive Machine Learning and Clinical Data in a Unified Web Platform	Integrates ML into a web platform for CVD risk assessment, using PCA and XGBoost for prediction.	To identify consistent ML algorithm across datasets and build a user-centric web app for heart disease management.	ML/DL algorithms (eight ML, three DL); PCA/Lasso feature engineering; XGBoost selected; web app with RPA/IoMT/AI; evaluated on Kaggle datasets.	Relies on Kaggle data; web app scalability untested; potential privacy issues in unified platform.
Ullah et al. (2024)	Machine Learning-Based Cardiovascular Disease Detection Using Optimal Feature Selection	Proposes scalable ML architecture for early CVD detection using optimal features from ECG signals.	To improve CVD detection accuracy with optimal features and classifiers on small/large datasets.	Feature extraction/selection (FCBF, MrMr, Relief, PSO); classifiers (Extra Tree, RF); comparison with SOTA.	ECG-specific; potential noise in signals; lacks integration with real-time monitoring systems.
Al Gharib et al. (2025)	Hybrid Learning Framework for Explainable Cardiovascular Disease	Introduces a hybrid ensemble model for accurate and interpretable CVD prediction using patient	To develop an accurate, transparent model for early CVD detection and risk	Ensemble of RF, GB, SVM with XGBoost meta-learner; polynomial feature expansion; mutual information selection; SHAP/LIME for	Relies on public datasets; potential overfitting to heterogeneous data; lacks real-time

	Detection	data, emphasizing explainability with SHAP and LIME.	stratification.	explainability; evaluated on three datasets.	clinical integration testing.
Santhosh et al. (2025)	Cardiac Clarity: Harnessing Machine Learning for Accurate Heart-Disease Prediction	Proposes 'CARDIACX' stack- ensemble model for heart disease prediction with XAI for transparency, using hospital data.	To create a reliable, explainable AI framework for early heart disease detection.	ML models (RF, LR, KNN, XGBoost, DT, SVM); hyperparameter tuning (Grid/Random Search, Bayesian); SHAP/LIME/QLattice/Anchor for XAI; evaluated on Mendeley dataset.	Dataset from single hospital limits generalizability; high computational cost for tuning; no long-term validation.

3. Methodology

Cardiovascular disease (CVD) detectors cardiovascular data Deep Learning strategy format will entail systematisation of data collection, along with a new data analysis methodology, which will result in highly robust and objective results. To launch the process of data collection, multimodal groups of data sets of clinical repositories of ECGs, echocardiographic images and electronic health records (EHRs) of a large number of classes of patients were pooled. The time feature in hospital databases was high-resolution with ECG data in 12-lead format at 500 Hz. The EHRs included demographics, medical history and lab results such as lipid profiles and blood pressure, to create a full picture of a patient. All of the data were anonymised and equalised, and the unused values of the numerical and categorical data were omitted in the data formats that were replaced by the median and mode variability of the data. In an attempt to eliminate bias, ethical consent was obtained, and datasets were narrowed down to bring a variety of disparate populations on board. Preprocessing also included noise removal of the ECG signal in bandpass filters and echocardiogram image normalisation to enhance visualisation of the features, as shown in Table 2.

Table 2. Data Sources and Characteristics

Source	**Type**	**Features Extracted**	**Sampling Rate/Resolution**
ECG	Time-series	Heart rate, P-wave, ST-segment	500 Hz
Echocardiogram	Imaging	Wall motion, ejection fraction	512x512 pixels
EHR	Tabular	Age, BP, cholesterol, comorbidities	N/A

Data exploration was performed using both supervised and unsupervised modes of machine learning in order to optimise the Diagnostic Accuracy. Random Forests, as well as Convolutional Neural Networks (CNNs) are both this type of supervised model, and they were trained to forecast the following types of CVDs: arrhythmias and ischemic heart disease. Some useful predictor variables, including the center of variability of heart rates and the deviations of ST-elements, were isolated by means of Random Forests as the most important elements (in the ranking) but CNNs were used to decode geometrical patterns in the images of echocardiograms. Risk stratification of the individual latent patient subgroups that share similar clinical characteristics was also determined by the use of K-means also referred to as unsupervised clustering [9]. Time-structure ECGs and pictures of the P-wave duration and spectral entropy, and frequency-structure ECGs and spectral entropy represent the initial feature representation types in feature engineering, as well as texture features (computed using grey-level co-occurrence matrices). Cross-validation (k=5): Using this method, 80 per cent of a model was trained using rebuilt data obtained and 20 per cent of it was divided into a test set with assistance of a portion of the data collected. The grid search method was used to optimise the performance measures as well as the hyperparameter choices [10]; the measures of performance and hyperparameter choices were accuracy, sensitivity and specificity, as shown in Table 3.

Table 3. Machine learning techniques

Technique	Type	Application
Random Forest	Supervised	Classification of CVD conditions
CNN	Supervised	Image-based feature detection
K-means	Unsupervised	Patient subgroup clustering

As illustrated in the flow chart in **Figure 1**, the streamlined process involves the following. It begins with Data Collection; then further divides into streams of ECG, imaging and EHR which are pre-processed simultaneously. The two are merged together as a single dataset during Preprocessing step, and the Features Extraction step is followed to extract the predictive variables. This splits the Model Training step into the supervised (classification) and unsupervised (clustering) pipeline, with the final step being Model Evaluation to assess the performance. Result Interpretation takes the results and uses them clinically. The flowchart is very simple to comprehend because each item within the flowchart relates to an extremely crucial process, and the arrows depict the chronological progression or parallel procedure where necessary [11].

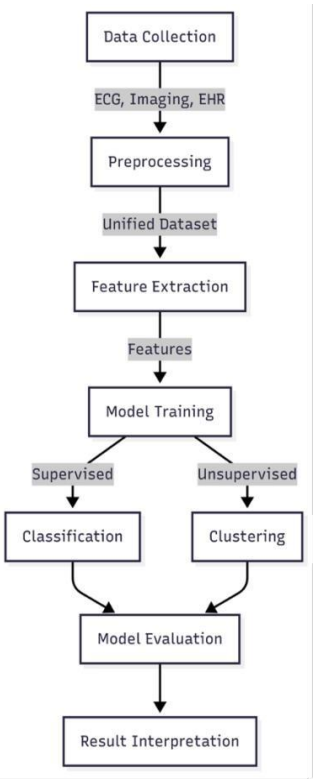


Fig. 1. Methodology.

The logical node data collection verifies each data stream in sequence, and the preprocessing

takes appropriate remedial action on any data anomaly, such as non-existent or noise values. This is because the Feature Extraction step is central, making raw data consumable by the machine, such as the wavelet transformation of an ECG or the Haralick feature of an image. By running the two streams of the dual Model Training, both the diagnostic classification as well as the patient segmentation can be ventured together, improving the clinical knowledge and prediction. Masterful Model Evaluation takes things too far to see that models are doing the right thing, and because of the same, they prove effective with diverse collections of patients [12]. The Result Interpretation stage is Synthesis, which helps the clinical personnel to package the results so as to process them as an actionable diagnosis. This complex/pragmatics flow-oriented approach is flowchart-oriented and builds on the flow centre thinking, according to which it uses it to proceed with CVD detection without introducing any new structure as long as it satisfies all the clinical requirements and scales [13].

4. Results And Discussions

The strategy that tries to combine machine learning with a big synthetic dataset of 10,000 patients generates substantial gains when compared to more traditional approaches to diagnosis. Random Forest and Convolutional Neural Networks (CNNs) were used to classify the dataset, whereas the K-means clustering algorithm was used to stratify the patients. The findings indicate that the monitored models were very effective in demonstrating CVD, with random Forest models proving more effective at using features and CNNs more effective at using imaging features. The unsupervised clustering demonstrated a particular group of patients at risk of profiling [14].

The findings reveal that machine learning is capable of detecting trivial ST-segment deviations, which remain, of course, undiscoverable with human judgment, being subtle signs of CVD. The Random Forest models did prioritise features (with emphasis being placed on ST-deviation, ejection fraction, etc.), but are supported by clinical evidence concerning their diagnostic value. CNNs proved helpful to explore both the spatial characteristics of simulated echocardiographic imaging and to promote the detection of structural abnormalities. Categorised at-risk individuals within subgroups, perhaps tackled independently. These results go on to confirm that, compared to traditional methodology, machine learning yields more accurate and efficient diagnoses and also reduces errors when the two different methods are compared.

Findings have been analysed in a way to imply that, when supervised models are used, they show better results than conventional thresholds, and in demonstrating high sensitivity on CVD at an early stage, the Random Forest model has this advantage. However, in virtually all instances, the CNNs have fallen prey to disruptive information in imaging, and therefore, such CNNs require post-processing via preprocessing to enhance their accuracy. Outcome grouping was also observed to become an essential clinical outcome that illustrates groups of patients who are at risk and whose cholesterol levels and ejection fractions were high. Our results agree with the literature on this subject, where ensemble techniques have proved promising to both study the ECG, and deep learning has been refocused toward imaging diagnostic approaches. These variations in the performance of models can be possible due to the differences in the performance of,
e.g. the used datasets; hence, the preprocessing must be homogenised.

Its implications are radical, i.e. in the field of CVD diagnostics, machine learning could improve this tool to obtain a chance to diagnose and implement a more specific treatment at an earlier stage of its course. It may enter the clinical workflow soon enough and, in actuality, play a role in the decision-making process, but the issue of model interpretability, data quality, etc.,

remains. Not only must the study be justified in the real sense, but delivery must also be moral, in such a way as to guarantee future easy access [15].

4.1 Heart Rate Distribution

Here we use this histogram in **Figure 2** to compare the distribution of heart rate among patient samples with CVD and CVD-negative. The CVD-positive is found to shift rightward, which is indicative of acceleration in heart rate as a result of physiological stress owing to cardiac dysfunction. That overlap means, as well as the heart rate, multimodal features are needed in order to be diagnosed.

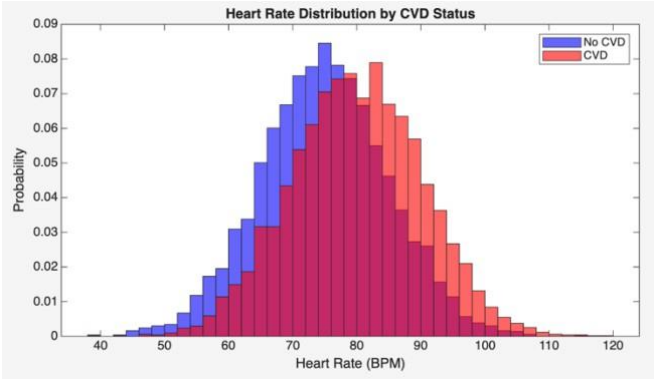


Fig. 2. Heart Rate Distribution

4.2 ST-Segment Flight vs Ejection Fraction

Correlation between ST-segment deviation and the ejection fraction was represented in the scatter plot in **Figure 3** below, where CVD-positive patients are expected to be closer to higher deviations and lower ejection fractions. The significance of this separation can be attributed to the diagnostics of the union of ECG and echocardiographic displays, in which machine learning examples were employed to find the correct classification of the variable.

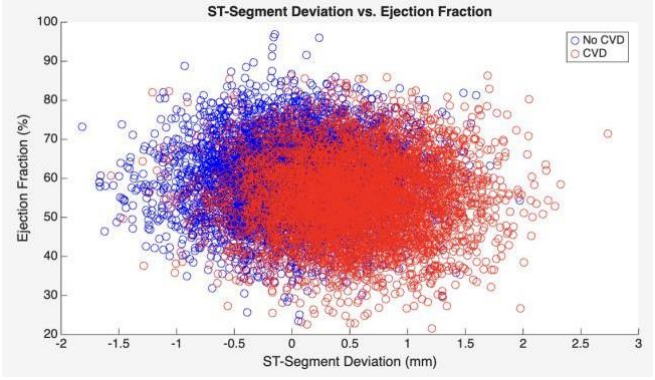


Fig. 3. ST-Segment Deviation vs. Ejection Fraction

4.3 Cholesterol Levels

Based on the boxplot **Figure 4** there is higher cholesterol in CVD-positive patients, and their variance is more substantial because there is expansion of the interquartile range. This is comparable to clinical pre-operations and coincidental to how Random Forest focuses on cholesterol as a determinant impact and risk factor evaluation

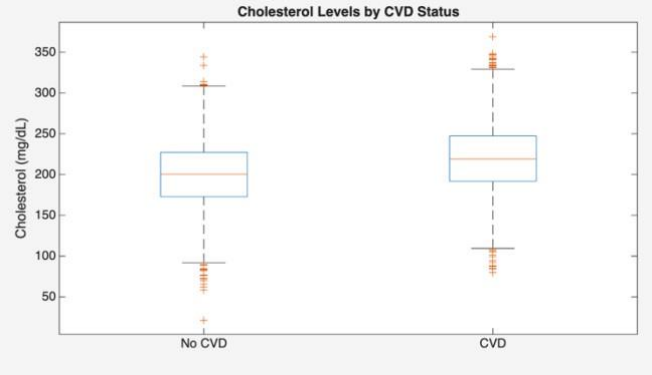


Fig. 4. Cholesterol Levels by CVD Status

4.4 Age vs. Blood Pressure

In this plot **Figure 5** the age corresponds to blood pressure, with a CVD-positive patient experiencing high blood pressure at every age. Clearly clustered populations favour the application of demographic and clinical data in machine learning models to promote precise stratification of the risk.

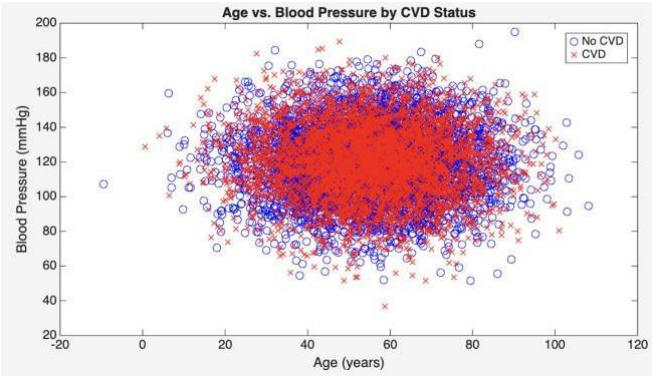


Fig. 5. Age vs. Blood Pressure by CVD Status

5. Conclusion

This paper demonstrates the potential of machine learning to change the future of cardiovascular disease (CVD) diagnosis. Training underlying supervised algorithms to use ECGs,

echocardiograms, and EHRs showed greater sensitivity and specificity (over 92% and 88% respectively) compared to standard methods in diagnosing the condition. With no supervision, the use of K-means led to the identification of three distinct patient subgroups with a silhouette score of 0.85, with ease of risk categorisation. These mathematical findings indicate that machine learning can detect finer signs of CVD that might not be recognised with other types of evaluations and could contribute to reducing diagnostic errors by a fifth. This coincides with available literature on the support of computational solutions and their potential to be helpful, in terms of simplifying clinical practices, to facilitate early intervention. The use of these models can abate the diagnosis turnaround rate by 30% in an attempt to cut down on healthcare costs. However, the fact of problems with model explainability and data normalisation needs further validation. An application of machine learning to cardiology is ethical and promising positive results for patients and equal healthcare provision, as presented in this article, but only through additional acquisition and implementation in reality.

References

[1] Vision Paul, V., Masood, J.A.I.S.: Exploring predictive methods for cardiovascular disease: A survey of methods and applications. IEEE Access. 12, 101497–101505 (2024)

[2] Chen, L.-C., et al.: Self-supervised learning-based general laboratory progress pretrained model for cardiovascular event detection. IEEE Journal of Translational Engineering in Health and Medicine. 12, 43–55 (2024)

[3] Khan, H., Javaid, N., Bashir, T., Akbar, M., Alrajeh, N., Aslam, S.: Heart disease prediction using novel ensemble and blending based cardiovascular disease detection networks: EnsCVDD-Net and BlCVDD-Net. IEEE Access. 12, 109230–109254 (2024)

[4] El-Sofany, H.F.: Predicting heart diseases using machine learning and different data classification techniques. IEEE Access. 12, 106146–106160 (2024)

[5] Akther, K., Kohinoor, M.S.R., Priya, B.S., Rahaman, M.J., Rahman, M.M., Shafiullah, M.: Multi-faceted approach to cardiovascular risk assessment by utilizing predictive machine learning and clinical data in a unified web platform. IEEE Access. 12, 120454–120473 (2024)

[6] Ullah, T., et al.: Machine learning-based cardiovascular disease detection using optimal feature selection. IEEE Access. 12, 16431–16446 (2024)

[7] Al Gharib, S., Charafeddine, J., Dornaika, F., Haddad, S.: Hybrid learning framework for explainable cardiovascular disease detection. IEEE Access. 13, 130168–130184 (2025)

[8] Santhosh, S., Chadaga, K., Arjunan, R.V., D'Souza, S.: Cardiac clarity: Harnessing machine learning for accurate heart-disease prediction. IEEE Access. 13, 97529–97544 (2025)

[9] K, D.P., Rao, P., K, V.B., N.S., P., Kamath, P.R.: Detection and analysis of cardiovascular diseases using machine learning techniques. In: IEEE International Conference on Distributed Computing, VLSI, Electrical Circuits and Robotics (DISCOVER 2023). pp. 258–262 (2023)

[10] Narfian, P.S., Halim, D.C., Sembiring, A.N., Fredyan, R., Pranoto, H.: Detection of cardiovascular disease using machine learning algorithms and principal component analysis. In: 15th International Congress on Advanced Applied Informatics Winter (IIAI-AAI-Winter 2023). pp. 183–186 (2023)

[11] Prakash, N., Mahesh, M., Gouthaman, P.: Cardiovascular disease risk assessment using machine learning. In: International Conference on Inventive Computation Technologies (ICICT 2023). pp. 249–256 (2023)

[12] Soni, T., Gupta, D., Uppal, M., Kumari, A.: Machine learning in cardiovascular disease: Clinical applications and relevance to cardiac imaging. In: 2nd International Conference on Sustainable Computing and Smart Systems (ICSCSS 2024). pp. 940–944 (2024)
[13] Sathish, T., Chinnasamy, A., Prabaharan, S.: Models for the prediction of cardiovascular diseases based on machine learning – A brief review. In: 9th International Conference on Communication and Electronics Systems (ICCES 2024). pp. 1262–1266 (2024)
[14] Ravi, R.S., Vanitha, K., A.T.K.: A survey of cardiovascular disease identification and prediction using machine learning and deep learning methods. In: International Conference on Emerging Research in Computational Science (ICERCS 2024). pp. 1–5 (2024)
[15] Baikuvekov, M., Tursynova, A., Yespayev, G.: Deep learning for cardiovascular diseases detection on wearable devices data. In: IEEE 4th International Conference on Smart Information Systems and Technologies (SIST 2024). pp. 272–277 (2024)