Sarcouncil Journal of Engineering and Computer Sciences

ISSN(Online): 2945-3585

Volume- 02| Issue- 10| 2023

Research Article

Received: 05-09-2023| Accepted: 05-10-2023 | Published: 19-10-2023

Explainable AI in Big Data Health Analytics: Detecting Heart Diseases from High-Dimensional EHRs

Jagan Kurma[1], Jaya Vardhani Mamidala[2], Sunil Jacob Enokkaren[3], Avinash Attipalli[4], Varun Bitkuri[5],and Raghuvaran Kendyala[6]

[1] *Christian Brothers University, Computer Information Systems*
[2] *University of Central Missouri, Department of Computer Science*
[3] *ADP, Solution Architect*
[4] *University of Bridgeport, Department of Computer Science*
[5] *Stratford University ,Software Engineer*
[6] *University of Illinois at Springfield, Department of Computer Science*

Abstract: The integration of Explainable AI into health data analytics of big data could guarantee the detection of heart disease in intricate EHRs, thereby increasing the transparency of the models, clinical decision-making, and individualized care in terms of interpretable predictions and state-of-the-art feature selection methods. This is particularly relevant for wearable ECG devices. Using data from wearable devices that detect the electrocardiogram (ECG) in real-time, the authors of this study provide an explainable AI method for predicting cardiac illness using high-dimensional electronic health records (EHRs). Data cleansing, data normalization, feature selection, and dimensionality reduction are some of the exacting procedures that make up the preprocessing of EHR data in the suggested architecture. Reduces computation overhead and increases efficiency through the use of advanced preprocessing techniques such as Recursive Feature Elimination (RFE) for feature selection and Principal Component Analysis (PCA) for dimensionality reduction. To make accurate and scalable disease predictions using a Recurrent Neural Network (RNN) model, it is crucial to extract time-based patterns from electrocardiogram (ECG) samples. The model's 92% accuracy and strong performance on all robustness indicators precision, recall, and F1-score make it an attractive option. To achieve the goals of explainable AI in healthcare analytics, the suggested methodology would facilitate the development of trustworthy clinical decision support systems by promoting interpretability and transparency.

Keywords: Explainable AI, Big Data, Electronic Health Records (EHR), Recurrent Neural Network (RNNs), feature selection, dimensionality reduction, heart disease detection, Healthcare Analytics.

INTRODUCTION

The widespread adoption of EHR systems has led to a significant increase in healthcare data creation. A wide range of data types is contained in electronic health records (EHRs), including medical records, demographic information, administrative charts, and laboratory test results (Abdullah, S. S. *et al.*, 2020). Although this data holds the potential for improved healthcare delivery, it is difficult to work with due to its complexity and size (Badheka, A. O., et al. 2013). Healthcare providers have a hard time processing and analysing data that is high-dimensional and diverse (Estiri, H. *et al.*, 2019). Big data analytics, data mining, and ML are assisting in the automation of valuable information retrieval. The healthcare industry faces a challenging yet vital task in predicting the incidence of heart disease.

Cardiovascular diseases (CVDs) refer to conditions characterised by narrowed or obstructed blood arteries, which can lead to serious health complications such heart attacks and strokes (Ali, F., *et al.*, 2020). Every year, hundreds of lives are lost due to cardiovascular illnesses (CVDs), which cause more than 18 million fatalities globally, as reported by the WHO. Patient history, physical exams, as well as symptoms are normally used to diagnose patients (Indrakumari, R. *et al.*, 2020), however, ML, and big data tools can increase accuracy and speed of diagnosis (Ismail, A. *et al.*, 2020).

Big Data Analytics facilitates effective management of heterogeneous, high-velocity and large-sized healthcare data. Turning complex EHRs into a more meaningful format can be achieved by using methods such as unsupervised learning and dimensionality reduction, which makes identifying the information involved more reasonable without leading to significant information loss (Badheka, A. O., et al. 2013). Their applications will help avert healthcare expenditure and improve outcomes, as they facilitate the use of data in clinical decision-making.

AI applications are finding new uses in the field of healthcare to automate testing, improve workflows, and support clinical decision-making (Alarsan, F. I., and Younes, M. 2019). Algorithms

*Corresponding Author: Jagan Kurma

DOI- https://doi.org/10.5281/zenodo.17390449

that enable learning and problem-solving allow AI systems to mimic human cognitive functions (Khan, Z. F., and Alotaibi, S. R. 2020). Although AI has revolutionary potential, it raises concerns about data privacy, transparency, and trust.

Explainable AI (XAI), partly clears these issues by making AI-driven decision-making transparent and explainable. XAI can assist clinicians in comprehending and trusting the model outputs used in predicting heart diseases, where the EHR data are high-dimensional (Ebrahimi, Z. *et al.*, 2020). With the increasing use of AI, big data, and management information systems in healthcare, explain ability will play a vital role in clinical acceptance. Integrating AI with big data analytics provides an effective mechanism for identifying heart diseases in challenging EHR data. Explainable AI also promotes trust and transparency, which means that the technology will be more relevant and effective when introduced to the real-world healthcare environment.

Motivation and Contribution of the Study

This research idea is motivated by the increasing interest in high-dimensional (in Electronic Health Records (EHRs), originating in wearable ECG data) detectors of heart disease, which in reality are large-dimensional and interpretable, and also in real-time. Conventional diagnostic processes tend to fail in handling the vast amount and complexity of health data produced in real-life contexts. The increased use of wearable gadgets and big data in healthcare creates a pressing need to find AI-based solutions that provide not only high predictive accuracy but also transparency and explain ability of clinical decisions. As a result, models for detecting cardiac disease are being integrated with deep learning and explainable AI to enhance their reliability, interpretability, and scalability. These are the most important contributions:

- Developed an explainable AI-based framework for heart disease detection using high-dimensional wearable ECG data, enabling interpretable and data-driven clinical insights.
- Improved model performance and reduced noise by implementing a robust data pretreatment pipeline that included cleaning, feature selection, normalisation, and dimensionality reduction.
- Applied an RNN model that could accurately classify sequential ECG data by capturing their temporal relationships.
- To guarantee that the evaluation was fair and reproducible, variables like as reliability, accuracy, precision, recall, F1-score, and an 80:20 train-test split were investigated.
- Proven that the RNN model outperformed more conventional ML and DL techniques, proving that it is well-suited for use in wearable health monitoring systems that operate in real-time.

Justification and Novelty

The innovation of the proposed methodology lies in its integration of explainable AI with a Recurrent Neural Network (RNN) architecture for heart disease detection using high-dimensional data derived from wearable ECG devices, as illustrated in Figure. 1. Unlike conventional models that struggle with temporal dependencies and lack interpretability, the proposed approach leverages sequential learning capabilities of RNNs to accurately capture dynamic ECG signal patterns while maintaining model transparency. The preprocessing pipeline encompassing data cleaning, feature selection, normalization, and dimensionality reduction ensures that only the most relevant and noise-free features are utilized, enhancing both efficiency and accuracy. This methodological design is justified by the growing demand for interpretable, real-time diagnostic systems in big data healthcare environments, where explain ability, reliability, and adaptability to high-dimensional EHRs are essential for clinical acceptance and decision support.

Structure of the Paper

The outline of the paper is as follows: Section II includes the Literature Review, which examines previous methods for forecasting the occurrence of heart disease. Section III covers the technique, which includes data collection, pre-processing, and model design; Section IV covers the results and discussion. The future work and conclusion are presented in Section V.

LITERATURE REVIEW

This literature review offers a comprehensive look at the state-of-the-art in cardiac disease prediction using ML and DL. It showcases models that achieve high accuracies utilising SVM, CNN-LSTM, Logistic Regression, and SHAP-based explain ability tools. These models have found applications in clinical decision support and web-based interfaces.

Ahmed Ismail, Samir Abdlerazek (2020) construct a cloud-based system for the real-time prediction of health problems using large amounts of medical data. The suggested scalable system uses Apache Spark to retrieve attributes from medical data and

Publisher: SARC Publisher

applies the suggested ML algorithm. The goal is to forecast healthcare risks and notify users and healthcare providers of any new suggestions or alarms. Their goal in writing this research is to assess the effectiveness of ML algorithms when applied to EHRs. To provide users with real-time suggestions and alerts based on sensor readings, the proposed study aimed to develop a practical recommendation system that utilises streaming medical data, user profile histories, and a knowledge database. When compared to previous research, the proposed prediction system exhibits a high predictability of 90.6% in the context of cardiovascular disease, indicating that it may yield accurate results (Ismail, A. *et al.,* 2020).

Sujatha and Mahalakshmi (2020) suggest that the use of DT, KNN, SVM, RF, NB, or LR is one method for predicting the existence of cardiovascular disease. They examined the algorithms' Accuracy, Precision, AUC, and F1-score to gauge their efficacy. When compared to other supervised ML algorithms, RF's accuracy rate in predicting the onset of cardiac disease was higher (83.52%), according to the experimental data. The accuracy, AUC, and F1-score for RF classifiers are 88.24% and 84.21%, respectively (Sujatha, P., & Mahalakshmi, K. 2020).

Krittanawong et al. (2020) examine the overall predictive ability of ML models in heart disease and determine its nature. From their creation until March 15, 2019, the MEDLINE, Embase, and Scopus databases were searched using an oversimplified methodology. This led to ML algorithms being able to foretell episodes of heart failure, arrhythmias, coronary artery disease, and stroke. Out of 344 studies that fulfilled their requirements, we were able to include 103 cohorts, totalling 3,377,318 individuals. The AUC for coronary artery disease prediction in the pooling method was 0.88 (95% CI 0.84 to 0.91) with boosting, whereas the AUC for the home-built methods was 0.93 (95% CI 0.85 to 0.97). An AUC of 0.92 (95% CI 0.81979 to 0.97) was achieved for stroke prediction using SVM, an AUC of 0.91 (95% CI 0.81341 to 0.96), and an AUC of 0.90 (95% CI 0.83430 to 90.95) was achieved with CNN (Krittanawong, C., *et al.,* 2020).

Ahmed et al. (2020) provide an approach to real-time cardiac illness prediction that utilises a patient's present health status-defining medical data streams. The system's principal goal is to choose the best ML algorithm for cardiac illness prediction. Relieved and univariate feature selection are the two methods used to extract features from the dataset. One decision tree, one SVM, one LR classifier, and one RF classifier were the four ML approaches under consideration. Each algorithm was taught using data that was either sampled or fully available. To improve ML's accuracy, they used hyperparameter tuning and cross-validation. One area where the suggested system excels is in its ability to process Twitter data streams that contain patient information. To achieve this, Apache Spark and Apache Kafka form the backbone of the system. With a total accuracy of 94.9%, the RF classifier was determined to have achieved the best performance among all models (Ahmed, H. *et al.,* 2019).

Mohan, Thirumalai and Srivastava (2019) The purpose of utilising ML algorithms is to improve the precision of cardiovascular disease prediction by highlighting key traits. The heart disease prediction model was enhanced using a HRFLM, resulting in an accuracy level of 88.7%. The foundation of this model is based on several current classification schemes and attributes. Clinical data analysts face a significant challenge when attempting to forecast cardiovascular disease (Mohan, S. *et al.,* 2019).

Nashif et al. (2018) A ML system that can predict when heart disease would start to develop has been proposed for use in the cloud. A successful ML strategy for cardiac illness diagnosis can be developed by doing separate assessments of many ML algorithms on the open-source data mining platform WEKA, which is based on Java. They utilised two popular open-access datasets to evaluate the proposed method's ability to diagnose cardiac disease using 10-fold cross-validation. Accuracy, sensitivity, and specificity were all 97.50% for the SVM approach (Nashif, S. *et al.*, 2018).

Table I presents a literature review of recent heart disease prediction studies, showcasing various datasets and ML methods like SVM, CNN-LSTM, and SHAP, with accuracies. Limitations include dataset specificity, interpretability, and generalizability, suggesting future improvements in explainable AI and real-time deployment.

Publisher: SARC Publisher

Table 1: Summary Of Previous Work On Big Data Health Analytics Detecting Heart Disease

Authors	Dataset	Methods	Findings	Limitations	Future Work
Ahmed Ismail, et.al. (2020)	Streaming medical data via Apache Spark	Custom ML algorithm on Spark for real-time prediction	Achieved 90.6% accuracy for heart disease prediction; system delivers real-time alerts & recommendations	Specifics of the dataset and algorithm are not detailed; no comparison with baseline models	Improve scalability; extend to other diseases; test across different environments
Sujatha, et.al. (2020)	A Kaggle dataset on cardiovascular disease	RF, DT, KNN, SVM, and LR	The most effective method was Random Forest: Accuracy of 83.52%, F1-score of 84.21%, and AUC of 88.24%	Dataset source not mentioned; only classical ML algorithms considered	Consider DL models and ensemble stacking; analyse on larger datasets
Krittanawong et al. (2020)	103 cohorts from MEDLINE, Embase, Scopus (3.3 million individuals)	SVM, Boosting, CNN, custom-built algorithms (Meta-analysis)	Custom-built models had an AUC of 0.93; SVM for stroke: AUC 0.92; CNN: 0.90	Meta-analysis lacks implementation details; clinical deployment is not explored	Include interpretability/explainability; validate models in clinical trials
Ahmed et. al. (2020)	Medical data streams + Twitter data	Decision Tree, SVM, Random Forest, Logistic Regression with Univariate & Relief feature selection	Random Forest achieved 94.9% accuracy; real-time system using Apache Spark + Kafka	Dependency on streaming infrastructure; Twitter data may have noise	Incorporate more real-time social data; improve semantic understanding of data
Mohan, et.al. (2019)	Cleveland Coronary Heart Disease Registry	Human Regression with a Linear Regression Experiment (HRFLM)	Achieved 88.7% accuracy; improved performance using a hybrid model	Limited to one dataset; feature selection details are minimal	Evaluate on multi-source clinical data; explore other hybrid techniques
Nashif et al. (2018)	Two open-access datasets	SVM in WEKA with 10-fold cross-validation	97.53% accuracy, Sensitivity: 97.5%, Specificity:	Datasets not specified; method restricted to the WEKA	Expand the model to other platforms; investigate real-time deployment options

Publisher: SARC Publisher

			94.94% using SVM	platform	

METHODOLOGY

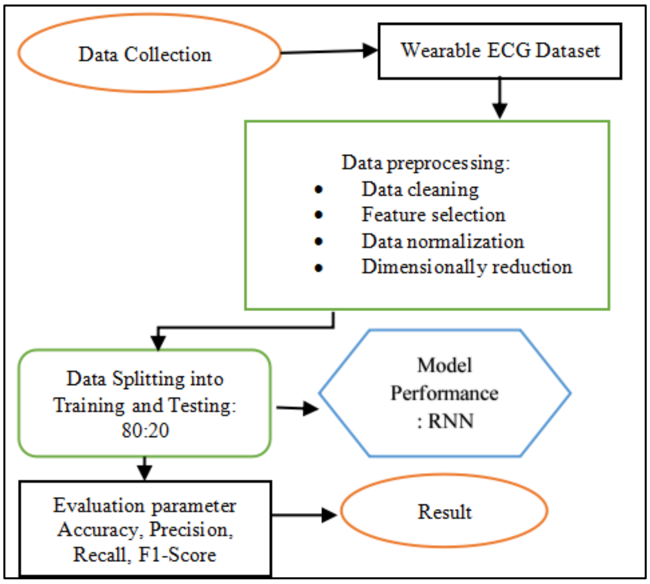

Fig 1: Flowchart Diagram of the Detection Using Heart Disease Detection on Wearable ECG Dataset

The proposed methodology for explainable AI in big data health analytics for heart disease detection from high-dimensional EHRs is illustrated in Figure 1. The process begins with data collection from a wearable ECG dataset, which represents a critical source of real-time physiological signals. Data cleaning, feature selection, normalisation, and dimensionality reduction are all parts of data pre-processing that help with the problems of high dimensionality and noise that are typical in electronic health records. After the data has been pre-processed, it should be divided into two sets: one for training and one for testing. This 80:20 ratio ensures a fair evaluation of the models. As a result of learning the temporal connections within the data, the ECG data is trained using a RNN model, which offers good performance in sequential analysis. Standard measures for evaluating models include F1-score, recall, accuracy, and precision. This methodology enables the interpretable and scalable diagnosis of heart disorders, aligning with the goals of explainable AI in healthcare applications, where reliability and transparency are of the utmost importance.

Data Collection

Reliable cardiovascular disease prognosis using wearable electrocardiograms (ECGs) The information in the dataset comes from a wide range of places, and all of them provide something unique to the overall picture of patients' health. Wearable technology has quickly become an indispensable data source due to its capacity to continuously track metrics including heart rate, physical activity, sleep habits, and blood pressure. The vast amounts of real-time data produced by gadgets such as fitness trackers and smartwatches enable the early identification of irregularities, including arrhythmias. Using these imaging datasets, can gain a better understanding of disease development and risk stratification.

Dataset Preprocessing

An essential step in getting raw heart data ready for ML models is data pre-processing. It all starts with data cleaning, which involves addressing issues such as missing values and inconsistencies. If there are blanks in the data acquired by wearable devices or electronic health records, procedures like imputation can be used to fill them in using median, mean, or predictive values. To harmonise datasets from different sources, they must also ensure that units and phrases are consistent.

Data Cleaning

As part of the data preparation process, purifying the dataset is essential. Filling in blanks when cells are empty or the term null appears in the dataset is crucial to avoid skewed findings.

Feature Selection

Feature selection is a crucial stage that can help reduce noise and enhance the model's performance. Using a technique known as RFE, can choose semantics such as HRV, cholesterol

Publisher: SARC Publisher

levels, and ventricular ejection fraction. Interpretation and overfitting risk are both enhanced by using clinically relevant variables.

Data Splitting

Splitting the data into an 80/20 training set and a 20/20 testing set made it easy to watch the model's performance across the board in great detail while reducing the likelihood of overfitting.

Data Normalisation

Data normalisation involves bringing all characteristics to a single scale. This is done to ensure that the ML model does not become overwhelmed by data with different units or magnitudes, such as blood pressure and heart rate. For more equitable training, this transformation normalizes all features to 0 with a standard deviation of 1 before training the model. It is possible to see this mathematical picture in Equation (1):

$$Normalization(X) = \frac{\text{X}-\text{Xmin}}{Xmix-Xmin} \quad (1)$$

The variable X represents features, where X-min is the minimum and X-max is the maximum.

Dimensionality Reduction

High-dimensional data, such as imaging data, can be effectively handled using dimensionality reduction, which simplifies the data without compromising its essential components. Algorithms, such as Principal Component Analysis (PCA), enable the extraction of key components that represent maximum variability, making computations more complex and enhancing the efficacy of prediction modelling.

Classification with RNN Model

The processing of sequential input is the domain of a subset of NN known as RNNs. Important contextual information included in the data is a temporal ordering. Because RNNs retain their previous inputs, a message can have a lasting effect, unlike feed-forward networks. This is because RNNs contain internal tails (Khanna, D. *et al.,* 2015). This is particularly beneficial in their applications to time series, including their use in monitoring heart rate variations, blood sugar levels, or tracking any variations in the progression of chronic diseases. In healthcare, RNNs can analyse historical patient data, including glucose patterns, drug compliance, and lifestyle habits, to inform predictions of future health risks. In one example, the continuous health data collected with wearable devices can be provided to an RNN in time series to predict when a diabetic patient may get complications. Equation (2) specified as:

$$h_t = \tanh\left(W_{hh}h_{t-1} + W_{xh}x_t + b_h\right) \quad (2)$$

Where, h_t: Presently concealed condition the function h_{t-1}: Before concealed condition the current input is represented by x_t Words with heights W_{hh}, W_{xh}, xh: Current hidden state weight matrices and input state b h: Negative bias term, tanh: Non-linear activation function.

Using the input from the current step and the hidden state from the previous step, as shown in the equation, one can determine the current hidden state. Despite their effectiveness, traditional RNNs face challenges when handling long-term dependencies resulting from diminishing gradients.

The introduction of LSTM networks remedied this. LSTMs are a kind of state-of-the-art RNN that successfully captures data's long-range dependencies through the use of memory cells and gating methods. This makes them highly suitable for predicting long-term health outcomes by preserving relevant patient information over time. LSTMs can differentiate between short-term noise and significant long-term patterns, which is critical in scenarios such as detecting early signs of cardiac deterioration or managing chronic conditions. Equation (3) defined as:

$$C = f_t \odot C_{t-1} + i_t \odot c \quad (3)$$

$\tilde{C}_t$: Potential values for the state update $\odot$:, C_{t-1}: The state of the cell before the update, i_t: The gate that determines what new information to add, and $\tilde{C}_t$: The current state of the cell Multiplication of elements one by one.

This Equation updates the memory cell by selectively forgetting parts of the old cell state and incorporating new information. The design of LSTM enables it to preserve important medical patterns across long time windows, making it an ideal choice for clinical sequence modelling.

Performance Metrics

The datasets utilised for this study include both authentic and fake entries, denoted as 0 and 1, respectively. To compare and contrast the algorithms' results, the Performance Metrics use four metrics (Taha, A. A., & Malebary, S. J. 2020). These metrics, which can take on values between 0 and 1, evaluate the model's ability to thoroughly detect network intrusions. Critical metrics for evaluating accuracy and reliability are TP, TN, FP, and FN; at the same time, understanding the values shown in the confusion matrix is of the utmost

Publisher: SARC Publisher

importance.

- **True Positive (TP):** valid records that are appropriately tagged as valid
- **True Negative (TN):** falsified records that are appropriately marked as such.
- **False Positive (FP):** documents that are not legitimate but are falsely certified as such.
- **False Negative (FN):** official documents that were inadvertently recorded as false documents

Table 2: Performance Matrix

Value	Actual positive	Actual negative
Predicted positive	TP	TP
Predicted negative	FN	FN

Table II presents the performance matrix, where values are calculated separately for each test case during the testing stage, serving as the basis for computing the metric evaluation. These values in the multiclass context lead to the general equations for accuracy, precision, recall, and F-score.

Accuracy

By reducing the sum of all positive and negative instances to a single metric, the accuracy rate of case classification can be determined. The four Equations define the precision according to Equation (4)

$$Accuracy = \frac{\text{TP+TN}}{\text{TP+Fp+TN+FN}} \quad (4)$$

Precision

Equation (5) shows that to find the accuracy, they divide the percentage of positive outcomes properly anticipated by the total number of cases:

$$Precision = \frac{\text{TP}}{\text{TP+FP}} \quad (5)$$

Recall

One important metric is recall, which measures the accuracy of predictions relative to the overall number of true positives. Equation (6) provides a mathematical definition of recall:

$$Recall = \frac{TP}{TP+FN} \quad (6)$$

F1-score

An effective metric for gauging a model's threat detection capabilities is the F1-score, which is the harmonic mean of recall and accuracy.The score takes into account the significant impact of both false positives and false negatives. If the F1-score is high, then the model achieves a satisfactory balance between recall and precision, as seen in Equation (7).

$$F1 - Score = \frac{2(Precision*Recall)}{\text{Precision+Recall}} \quad (7)$$

Loss function

One way to measure how well the model predicts values about the actual value is via the loss function.

These matrices are utilized to determine the machine and DL models.

RESULTS & DISCUSSION

On the Wearable ECG Dataset, the RNN model outperformed the competition, achieving an impressive 92% accuracy rate. It also had great precision (90%), recall (93%), and F1-score (91%), demonstrating its resilience in predicting cardiac illness. Specifically, the high recall demonstrates that the model successfully identifies affirmative cases. Learning and generalization are evident from the training and validation curves, which show a constant decrease in loss and a continuous increase in accuracy. Although there were a few misclassifications, the confusion matrix indicates that the model is highly predictive. To train ML/DL models, such as RNNs, the experiment was conducted on a laptop with the following specifications: an NVIDIA GPU, 512 GB SSD, 16 GB of RAM, and an Intel Core i7 processor. In Table III, and can get an overview of the performance indicators that show how well the suggested model predicts the occurrence of heart disease:

Table 3: ECG dataset used for wearable heart disease detection and RNN model performance outcomes

Measures	RNN Model
Accuracy	92
Precision	90
Recall	93
F1-score	91

Publisher: SARC Publisher

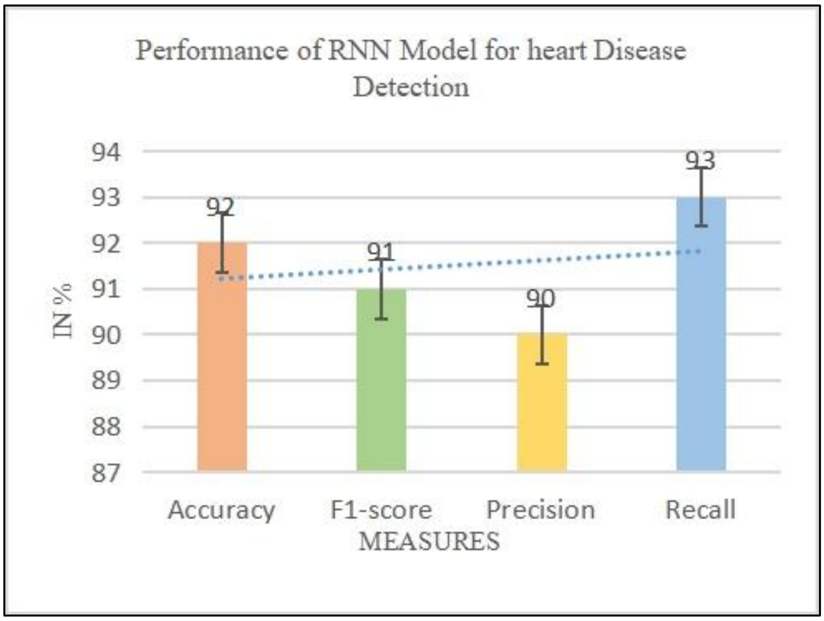

Fig 2: Bar Graph of RNN Model

Table III and Figure 2 present the performance of the RNN model. These measures include F1-score (91%), Accuracy (92%), Precision (90%), and Recall (93%). You may see the values with various colours and error bars to illustrate any potential variation in the bar chart. The graphic highlights the RNN model's strong classification capabilities, particularly its high recall, which indicates accurate identification of positive cases. The F1-score, which verifies the model's stability, shows that recall and precision are in equilibrium. The RNN can be trusted and has been proven effective in the assessed task, consistently achieving high scores on all measures.

Fig 3: The Accuracy Plot on the RNN Model

The accuracy of the RNN model across seven epochs, including training and validation, is shown in Figure 3. Training accuracy has reached 90% and validation accuracy has reached 88%, both of which have been rising consistently. This stable, positive curve depicts the efficient learning and generalisation of the model, with little overfitting, as evidenced by the minimal difference between the validation and training accuracy. The performance curve indicates how well the model can represent the inherent trend in the data, which justifies its reliability and stability in addressing the classification problem in biomedical fields.

Publisher: SARC Publisher

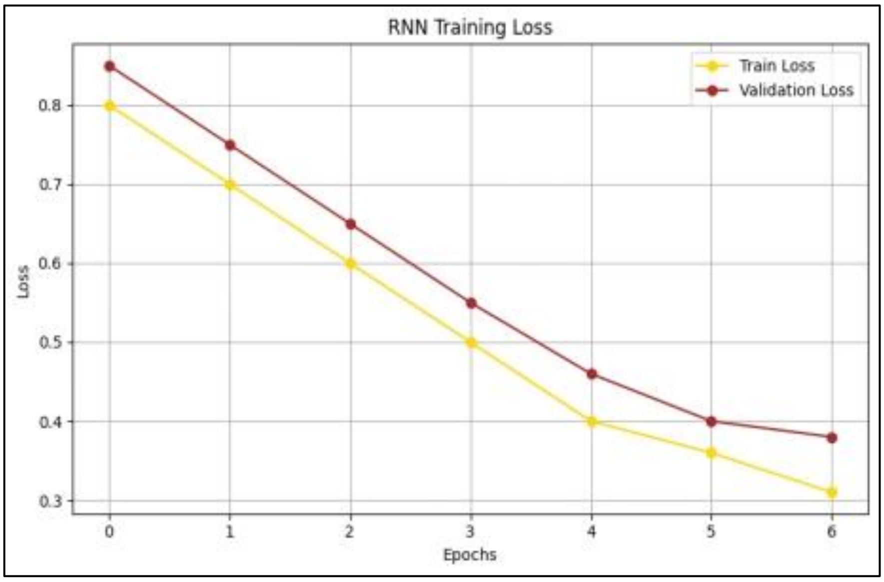

Fig 4: The Loss Function for the RNN Model

Figure 4 illustrates the loss curve of the RNN model across seven epochs, broken down by training and validation loss. Both loss functions exhibit a downward trend, whereas the training loss was scaled down (approximately 0.80 to 0.31) and the validation loss was also scaled down (approximately 0.85 to 0.38). The fact that the values of loss are slowly decreasing in both datasets shows that the learning and convergence conditions of the models are being met properly. Additionally, the similarity between the training and validation losses suggests that there is little overfitting, which confirms that the model can generalise. These facts support the applicability of RNN model in the area of sequence-based predicting activities.

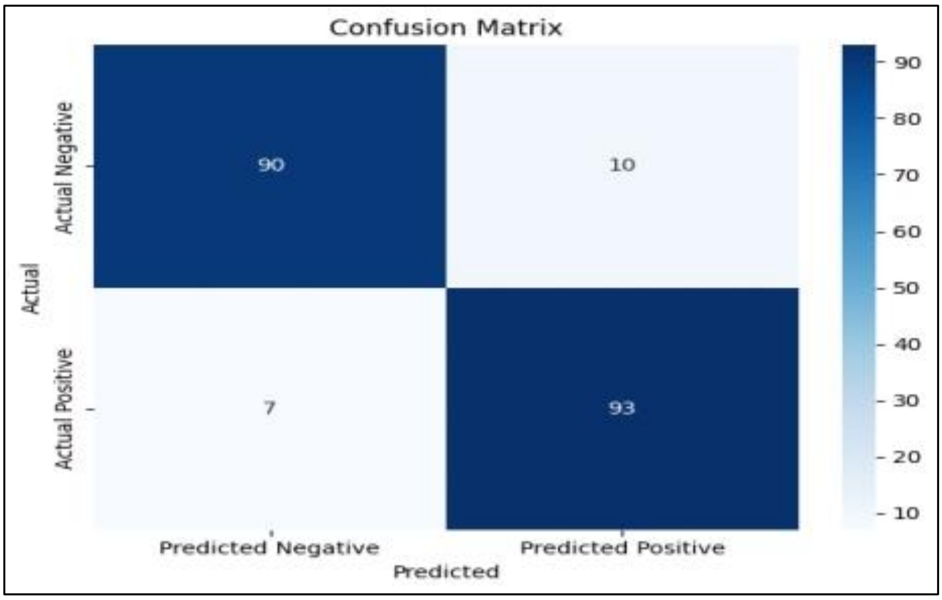

Fig 5: Performance Confusion Matrix of RNN Model

Figure 5 shows the confusion matrix used to assess the RNN model's ability to predict binary outcomes (heart disease). As shown in the matrix, the model incorrectly classified 10 cases as false positives and 7 as false negatives, despite correctly identifying 93 genuine positives and 90 true negatives. Several metrics demonstrate the model's effectiveness in classifying data as either positive or negative, including an accuracy of 92%, precision of 90%, recall of 93%, and an F1-score of 91%. A balanced classification capacity, as demonstrated by excellent results across all criteria, is crucial in medical diagnostics, particularly in accurately identifying genuine positive instances.

Publisher: SARC Publisher

Comparative Analysis

Each model is trained and evaluated on an equal dataset of wearable ECG data because the experimental settings are equivalent. This section offers a comprehensive performance analysis of the proposed model. Table IV presents the comparative findings, demonstrating how the RNN model performs best with respect to the main classification parameters. Among all the models tested, RNN showed the best results for detecting cases of cardiac disease, with an accuracy of 92%, precision of 90%, recall of 93%, and F1-score of 91%. In contrast, the DT model ((Mohan, S. *et al.*, 2019) achieves 98.8% recall and 91.8 F1-score, respectively, while only managing 85 and 86 precision and accuracy, respectively. The LR model (Baccouche, A. *et al.*, 2020) achieves a balanced result, with an accuracy ratio of 85.25% and an F1-score of 86.9%, in contrast to the LSTM model ((Rajdhan, A. *et al.*, 2020), which exhibits an average result with an accuracy of 74% and an F1-score of 79%. When it comes to using wearable ECG records to diagnose cardiac illness, the RNN model appears to be more reliable and predictable, making it a very good algorithm.

TABLE 4: ML AND DL MODELS' COMPARATIVE FOR HEART DISEASE DETECTION ON WEARABLE ECG DATASET

Models	Accuracy	Precision	Recall	F1-SCORE
RNN	92	90	93	91
DT((Mohan, S. *et al.*, 2019)	85	86	98.8	91.8
LSTM(Baccouche, A. *et al.*, 2020)	74	88	74	79
LR(Rajdhan, A. *et al.*, 2020)	85.25	85.7	88.2	86.9

The results show that the RNN version outperforms traditional ML and other DL models in detecting cases of heart disease using wearable ECG information. It demonstrates a balanced performance, effectively capturing all significant classification metrics and time dependencies of the ECG signals. Although Decision Tree (DT) and Logistic Regression (LR) models demonstrate a competitive performance and the LSTM model works fairly well, RNN has the qualities of being both robust and reliable, which is why it can be viewed as the most applicable model to predict the occurrence of the heart disease both correctly and efficiently in wearable healthcare systems and apps.

CONCLUSION & FUTURE WORK

In Conclusion, explainable AI and big data analytics are transforming the healthcare landscape and can be leveraged to enhance diagnostics, treatment planning, and process efficiency. The paper may provide valuable insights into the application of ML algorithms in detecting heart diseases. The potential of ML in healthcare is associated with the possibility of improving the quality of diagnosis, resource management, facilitating the process of personalised therapy, and promoting the development of this essential industry. The given methodology in terms of explainable AI can be used to effectively utilise high-dimensional EHRs, specifically wearable ECGs, to achieve effective detection of heart disease with the help of an RNN model. Having achieved 92% accuracy and good values of recall and F1-score, the method proves to be robust in detecting temporal patterns related to human health and maintains transparency with explainable properties. More sophisticated pre-processing processes, such as RFE with PCA, also help streamline model functioning and reduce computer requirements. In the future, it will aim to enhance the generalizability of the model by using more diverse and multimodal healthcare data, enabling real-time processing, and considering integration into clinical decision support systems to manage preventive and personalised cardiovascular care proactively.

REFERENCES

1. Abdullah, S. S., Rostamzadeh, N., Sedig, K., Garg, A. X., and McArthur, E. "Visual Analytics for Dimension Reduction and Cluster Analysis of High-Dimensional Electronic Health Records." *Informatics* 7.2 (2020). https://doi.org/10.3390/informatics7020017.
2. Badheka, A. O., et al. "ST-T Wave Abnormality in Lead aVR and Reclassification of Cardiovascular Risk (from the National Health and Nutrition Examination Survey-III)." *American Journal of Cardiology* 112.6 (2013): 805–810. https://doi.org/10.1016/j.amjcard.2013.04.058.
3. Estiri, H., Klann, J. G., and Murphy, S. N. "A Clustering Approach for Detecting Implausible Observation Values in Electronic Health Records Data." *BMC Medical Informatics and*

Publisher: SARC Publisher

Decision Making 19.1 (2019): 142. https://doi.org/10.1186/s12911-019-0852-6.
4. Ali, F., et al. "A Smart Healthcare Monitoring System for Heart Disease Prediction Based on Ensemble Deep Learning and Feature Fusion." *Information Fusion* 63 (2020): 208–222. https://doi.org/10.1016/j.inffus.2020.06.008.
5. Indrakumari, R., Poongodi, T., and Jena, S. R. "Heart Disease Prediction Using Exploratory Data Analysis." *Procedia Computer Science* 173 (2020): 130–139. https://doi.org/10.1016/j.procs.2020.06.017.
6. Ismail, A., Abdlerazek, S., and El-Henawy, I. M. "Big Data Analytics in Heart Diseases Prediction." *Journal of Theoretical and Applied Information Technology* 98 (2020): 11.
7. Alarsan, F. I., and Younes, M. "Analysis and Classification of Heart Diseases Using Heartbeat Features and Machine Learning Algorithms." *Journal of Big Data* 6.1 (2019). https://doi.org/10.1186/s40537-019-0244-x.
8. Khan, Z. F., and Alotaibi, S. R. "Applications of Artificial Intelligence and Big Data Analytics in m-Health: A Healthcare System Perspective." *Journal of Healthcare Engineering* 2020 (2020): 1–15. https://doi.org/10.1155/2020/8894694.
9. Ebrahimi, Z., Loni, M., Daneshtalab, M., and Gharehbaghi, A. "A Review on Deep Learning Methods for ECG Arrhythmia Classification." *Expert Systems with Applications X* 7 (2020). https://doi.org/10.1016/j.eswax.2020.100033.
10. Sujatha, P., and Mahalakshmi, K. "Performance Evaluation of Supervised Machine Learning Algorithms in Prediction of Heart Disease." *2020 IEEE International Conference for Innovation in Technology (INOCON)*. IEEE, 2020: 1–7. https://doi.org/10.1109/INOCON50539.2020.9298354.
11. Krittanawong, C., et al. "Machine Learning Prediction in Cardiovascular Diseases: A Meta-Analysis." *Scientific Reports* 10.1 (2020): 16057. https://doi.org/10.1038/s41598-020-72685-1.
12. Ahmed, H., Younis, E. M. G., Hendawi, A., and Ali, A. A. "Heart Disease Identification from Patients' Social Posts: A Machine Learning Solution on Spark." *Future Generation Computer Systems* 111 (2020): 714–722. https://doi.org/10.1016/j.future.2019.09.056.
13. Mohan, S., Thirumalai, C., and Srivastava, G. "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques." *IEEE Access* 7 (2019): 81542–81554. https://doi.org/10.1109/ACCESS.2019.2923707.
14. Nashif, S., Raihan, M. R., Islam, M. R., and Imam, M. H. "Heart Disease Detection by Using Machine Learning Algorithms and a Real-Time Cardiovascular Health Monitoring System." *World Journal of Engineering and Technology* 6.4 (2018): 854–873. https://doi.org/10.4236/wjet.2018.64057.
15. Khanna, D., Sahu, R., Baths, V., and Deshpande, B. "Comparative Study of Classification Techniques (SVM, Logistic Regression and Neural Networks) to Predict the Prevalence of Heart Disease." *International Journal of Machine Learning and Computing* 5.5 (2015): 414–419. https://doi.org/10.7763/IJMLC.2015.V5.544.
16. Taha, A. A., and Malebary, S. J. "An Intelligent Approach to Credit Card Fraud Detection Using an Optimized Light Gradient Boosting Machine." *IEEE Access* 8 (2020): 25579–25587. https://doi.org/10.1109/ACCESS.2020.2971354.
17. Baccouche, A., Garcia-Zapirain, B., Olea, C. C., and Elmaghraby, A. "Ensemble Deep Learning Models for Heart Disease Classification: A Case Study from Mexico." *Information* 11.4 (2020): 207. https://doi.org/10.3390/info11040207.
18. Rajdhan, A., Sai, M., Ravi, D., Agarwa, A., and Ghuli, D. P. "Heart Disease Prediction Using Machine Learning." *International Journal of Engineering Research and Technology* 9.4 (2020). https://doi.org/10.1109/ICAC3N60023.2023.10541622
19. Rajiv, C., Mukund Sai, V. T., Venkataswamy Naidu, G., Sriram, P., & Mitra, P. "Leveraging Big Datasets for Machine Learning-Based Anomaly Detection in Cybersecurity Network Traffic." *J. Contemp. Edu. Theor. Artif. Intel.* JCETAI/102 (2022).
20. Sandeep Kumar, C., Srikanth Reddy, V., Ram Mohan, P., Bhavana, K., & Ajay Babu, K. "Efficient Machine Learning Approaches for Intrusion Identification of DDoS Attacks in Cloud Networks." *J. Contemp. Edu. Theor. Artif. Intel.* JCETAI/101 (2022).
21. Bhumireddy, J. R., Chalasani, R., Tyagadurgam, M. S. V., Gangineni, V. N., Pabbineedi, S., & Penmetsa, M. "Big Data-Driven Time Series Forecasting for Financial

Publisher: SARC Publisher

Market Prediction: Deep Learning Models." *J. Artif. Intell. Big Data* 2.1 (2020): 153–164. DOI: 10.31586/jaibd.2022.1341.
22. Nandiraju, S. K. K., Chundru, S. K., Vangala, S. R., Polam, R. M., Kamarthapu, B., & Kakani, A. B. "Advance of AI-Based Predictive Models for Diagnosis of Alzheimer's Disease (AD) in Healthcare." *J. Artif. Intell. Big Data* 2.1 (2022): 141–152. DOI: 10.31586/jaibd.2022.1340.
23. Tyagadurgam, M. S. V., Gangineni, V. N., Pabbineedi, S., Penmetsa, M., Bhumireddy, J. R., & Chalasani, R. "Designing an Intelligent Cybersecurity Intrusion Identify Framework Using Advanced Machine Learning Models in Cloud Computing." *Univ. Libr. Eng. Technol.* (2022).
24. Vangala, S. R., Polam, R. M., Kamarthapu, B., Kakani, A. B., Nandiraju, S. K. K., & Chundru, S. K. "Leveraging Artificial Intelligence Algorithms for Risk Prediction in Life Insurance Service Industry." *SSRN* 5459694 (2022).
25. Polam, R. M., Kamarthapu, B., Kakani, A. B., Nandiraju, S. K. K., Chundru, S. K., & Vangala, S. R. "Data Security in Cloud Computing: Encryption, Zero Trust, and Homomorphic Encryption." *Int. J. Emerg. Trends Comput. Sci. Inf. Technol.* 2.3 (2021): 70–80.
26. Gangineni, V. N., Pabbineedi, S., Penmetsa, M., Bhumireddy, J. R., Chalasani, R., & Tyagadurgam, M. S. V. "Efficient Framework for Forecasting Auto Insurance Claims Utilizing Machine Learning Based Data-Driven Methodologies." *Int. Res. J. Econ. Manag. Stud.* 1.2 (2021).
27. Vattikonda, N., Gupta, A. K., Polu, A. R., Narra, B., Buddula, D. V. K. R., & Patchipulusu, H. H. S. "Blockchain Technology in Supply Chain and Logistics: A Comprehensive Review of Applications, Challenges, and Innovations." *Int. J. Emerg. Res. Eng. Technol.* 3.3 (2022): 99–107.
28. Narra, B., Vattikonda, N., Gupta, A. K., Buddula, D. V. K. R., Patchipulusu, H. H. S., & Polu, A. R. "Revolutionizing Marketing Analytics: A Data-Driven Machine Learning Framework for Churn Prediction." *Int. J. Artif. Intell. Data Sci. Mach. Learn.* 3.2 (2022): 112–121.
29. Polu, A. R., Narra, B., Buddula, D. V. K. R., Patchipulusu, H. H. S., Vattikonda, N., & Gupta, A. K. "Blockchain Technology as a Tool for Cybersecurity: Strengths, Weaknesses, and Potential Applications." (n.d.).
30. HK, K. "Design of Efficient FSM Based 3D Network on Chip Architecture." *Int. J. Eng.* 68.10 (2020): 67–73.
31. Krutthika, H. K. "Modeling of Data Delivery Modes of Next Generation SOC-NOC Router." *Proc. 2019 Global Conf. Adv. Technol. (GCAT).* IEEE, 2019: 1–6.
32. Ajay, S., Satya Sai Krishna Mohan, G., Rao, S. S., Shaunak, S. B., Krutthika, H. K., Ananda, Y. R., & Jose, J. "Source Hotspot Management in a Mesh Network on Chip." *VDAT* (2018): 619–630.
33. Nair, T. R., & Krutthika, H. K. "An Architectural Approach for Decoding and Distributing Functions in FPUs in a Functional Processor System." *arXiv preprint* arXiv:1001.3781 (2010).
34. Gopalakrishnan Nair, T. R., & Krutthika, H. K. "An Architectural Approach for Decoding and Distributing Functions in FPUs in a Functional Processor System." *arXiv e-prints* arXiv-1001 (2010).
35. Krutthika, H. K., & Aswatha, A. R. "Implementation and Analysis of Congestion Prevention and Fault Tolerance in Network on Chip." *J. Tianjin Univ. Sci. Technol.* 54.11 (2021): 213–231. https://doi.org/10.5281/zenodo.5746712.
36. Krutthika, H. K., & Aswatha, A. R. "FPGA-Based Design and Architecture of Network-on-Chip Router for Efficient Data Propagation." *IIOAB J.* 11(S2) (2020): 7–25.
37. Krutthika, H. K., & Aswatha, A. R. "Design of Efficient FSM-Based 3D Network-on-Chip Architecture." *Int. J. Eng. Trends Technol.* 68.10 (2020): 67–73. https://doi.org/10.14445/22315381/IJETT-V68I10P212.
38. Krutthika, H. K., & Rajashekhara, R. "Network-on-Chip: A Survey on Router Design and Algorithms." *Int. J. Recent Technol. Eng.* 7.6 (2019): 1687–1691. https://doi.org/10.35940/ijrte.F2131.037619.
39. Polam, R. M., Kamarthapu, B., Kakani, A. B., Nandiraju, S. K. K., Chundru, S. K., & Vangala, S. R. "Big Text Data Analysis for Sentiment Classification in Product Reviews Using Advanced Large Language Models." *Int. J. AI, BigData, Comput. Manag. Stud.* 2.2 (2021): 55–65.
40. Gangineni, V. N., Tyagadurgam, M. S. V., Chalasani, R., Bhumireddy, J. R., & Penmetsa, M. "Strengthening Cybersecurity Governance:

Publisher: SARC Publisher

The Impact of Firewalls on Risk Management." *Int. J. AI, BigData, Comput. Manag. Stud.* 2 (2021): 10–63282.

41. Pabbineedi, S., Penmetsa, M., Bhumireddy, J. R., Chalasani, R., Tyagadurgam, M. S. V., & Gangineni, V. N. "An Advanced Machine Learning Models Design for Fraud Identification in Healthcare Insurance." *Int. J. Artif. Intell. Data Sci. Mach. Learn.* 2.1 (2021): 26–34.
42. Kamarthapu, B., Kakani, A. B., Nandiraju, S. K. K., Chundru, S. K., Vangala, S. R., & Polam, R. M. "Advanced Machine Learning Models for Detecting and Classifying Financial Fraud in Big Data-Driven Systems." *Int. J. Artif. Intell. Data Sci. Mach. Learn.* 2.3 (2021): 39–46.
43. Tyagadurgam, M. S. V., Gangineni, V. N., Pabbineedi, S., Penmetsa, M., Bhumireddy, J. R., & Chalasani, R. "Enhancing IoT Security Through Intelligent Intrusion Detection Using ML Models." *Int. J. Emerg. Res. Eng. Technol.* 2.1 (2021): 27–36.
44. Vangala, S. R., Polam, R. M., Kamarthapu, B., Kakani, A. B., Nandiraju, S. K. K., & Chundru, S. K. "Smart Healthcare: Machine Learning-Based Classification of Epileptic Seizure Disease Using EEG Signal Analysis." *Int. J. Emerg. Res. Eng. Technol.* 2.3 (2021): 61–70.
45. Kakani, A. B., Nandiraju, S. K. K., Chundru, S. K., Vangala, S. R., Polam, R. M., & Kamarthapu, B. "Big Data and Predictive Analytics for Customer Retention: Exploring the Role of Machine Learning in E-Commerce." *Int. J. Emerg. Trends Comput. Sci. Inf. Technol.* 2.2 (2021): 26–34.
46. Penmetsa, M., Bhumireddy, J. R., Chalasani, R., Tyagadurgam, M. S. V., Gangineni, V. N., & Pabbineedi, S. "Next-Generation Cybersecurity: The Role of AI and Quantum Computing in Threat Detection." *Int. J. Emerg. Trends Comput. Sci. Inf. Technol.* 2.4 (2021): 54–61.
47. Polu, A. R., Vattikonda, N., Gupta, A., Patchipulusu, H., Buddula, D. V. K. R., & Narra, B. "Enhancing Marketing Analytics in Online Retailing through Machine Learning Classification Techniques." *SSRN* 5297803 (2021).
48. Kalla, D. "AI-Powered Driver Behavior Analysis and Accident Prevention Systems for Advanced Driver Assistance." *Int. J. Sci. Res. Mod. Technol.* 1 (2022).
49. Dinesh, K. "Navigating the Link Between Internet User Attitudes and Cybersecurity Awareness in the Era of Phishing Challenges." *Int. Adv. Res. J. Sci. Eng. Technol.* (2022).
50. Kalla, D., Kuraku, D. S., & Samaah, F. "Enhancing Cybersecurity by Predicting Malwares Using Supervised Machine Learning Models." *Int. J. Comput. Artif. Intell.* 2.2 (2021): 55–62.
51. Katari, A., & Kalla, D. "Cost Optimization in Cloud-Based Financial Data Lakes: Techniques and Case Studies." *ESP J. Eng. Technol. Adv.* 1.1 (2021): 150–157.

Source of support: Nil; **Conflict of interest:** Nil.

Cite this article as:

Kurma, J., Mamidala, J. V., Enokkaren, S. J., Attipalli, A., Bitkuri, V. and Kendyala, R. " Explainable AI in Big Data Health Analytics: Detecting Heart Diseases from High-Dimensional EHRs." *Sarcouncil Journal of Engineering and Computer Sciences* 2.10 (2023): pp 1-13.

Publisher: SARC Publisher