EMERGENCY HEALTH ALERT

Ms Mullapudi Navyasri , Prabhat kumar, Shubham kumar singh

1School of Computer Science and Engineering, Galgotias University, Greater Noida, 203201, India

*Corresponding author: mullapudi.navyasri@galgotiasuniversity.edu.in

Abstract

With the swift evolution of intelligent healthcare solutions, health monitoring systems are progressively being assimilated into everyday life to facilitate continuous and proactive patient management. This research articulates an anomaly-based heart disease risk prediction module specifically engineered to operate within an expansive health monitoring system framework. In light of the elevated global mortality rates linked to heart disease, the imperative for early and precise detection becomes paramount. Our methodology employs unsupervised machine learning strategies—namely, Isolation Forest and Autodecoder models—to discern aberrant physiological patterns within the UCI Heart Disease dataset. These models function independently of labeled data, thereby rendering them particularly advantageous for real-time health monitoring contexts where annotated medical information is frequently scarce. Isolation Forest proficiently isolates anomalous behaviors, whilst the Autodecoder acquires representations of healthy profiles to identify deviations. By integrating this system into a health monitoring architecture, it becomes feasible to perpetually evaluate cardiovascular health, thereby furnishing timely notifications for individuals at risk and bolstering preventive healthcare initiatives. This study establishes a foundational framework for more sophisticated, intelligent health monitoring solutions that are scalable, interpretable, and clinically relevant.

Keywords: *Health Monitoring System, Heart Disease Detection, Anomaly Detection, Autodecoder, Isolation Forest, Unsupervised Learning, Preventive Healthcare, Machine Learning in Health.*

4. Introduction

In contemporary times, health monitoring systems have emerged as an indispensable component of modern healthcare, facilitating the continuous evaluation of patient well-being through the amalgamation of wearable technologies, Internet of Things (IoT) devices, and sophisticated data-driven methodologies. These systems possess particular significance in the management of chronic illnesses and in the early identification of potential health threats—often preceding the manifestation of clinical symptoms

A principal challenge In the development of such systems Is the scarcity of labeled clinical data, Ih hampers the efficacy of supervised machine learning methodologies. To address this issue, unsupervised learning techniques, particularly anomaly detection models, present a promising alternative.

In this research endeavor, we propose a heart disease risk prediction module for health monitoring systems that employs unsupervised machine learning algorithms—namely, Isolation Forest and Autodecoder. Both models are engineered to detect anomalies in patient data, which may serve as early indicators of cardiovascular risk. The Isolation Forest technique identifies rare and distinct data points by randomly partitioning the dataset and assessing the ease with which each instance can be isolated. The Autodecoder, a model based on neural network architecture, learns compact representations of healthy data and flags instances exhibiting high reconstruction errors as potential anomalies.

The proposed system Is trained and”eval’ated utilizing the well-established UCI Heart Disease dataset, which encompasses clinical parameters, demographic data, and test results. Through preprocessing procedures such as feature scaling, encoding, and data cleansing, the dataset is meticulously prepared for effective 1abelled1. The models are trained on data representing healthy individuals and are subsequently tested across the entire dataset to identify outliers, which are presumed to be at an elevated risk.

By incorporating these models within a health monitoring framework, this research contributes to the advancement of automated, intelligent, and scalable cardiovascular monitoring systems. These systems have the potential to operate continuously, facilitate early diagnosis, and ultimately mitigate the burden of heart disease through preventive healthcare measures. Furthermore, this investigation establishes a foundation for future enhancements, encompassing model ensembles, interpretability improvements, and real-world implementation in wearable or mobile health platforms.

2. Related Work

The increasing emphasis on unsupervised machine learning for the purpose of health monitoring can be attributed to the intrinsic limitations associated with traditional supervised models in contexts where data 2abelled2 is either accessible or challenging to procure. The methodologies associated with unsupervised learning present innovative opportunities for the identification of cardiac irregularities by leveraging latent representations of normative patterns to recognize anomalous deviations, thereby facilitating the early forecasting of risk independent of ground truth labels.

Transformer models, originally developed for applications in Natural Language Processing (NLP), have demonstrated efficacy in the management of sequential medical data such as electrocardiogram (ECG) signals. In their scholarly work, the authors presented an unsupervised transformer model capable of discerning complex temporal relationships within ECG signals for the purpose of anomaly detection [1]. This model underwent evaluation on datasets including ECG5000 and MIT-BIH, achieving elevated F1 measures and accuracy, which underscores its potential utility in the realm of medical anomaly detection. Similarly, Beta Variational Autodecoders (Beta-VAEs) have proven effective in acquiring compact representations of normal phonocardiogram (PCG) signals to facilitate the detection of cardiac abnormalities. A particular study revealed that Beta-VAEs successfully 2abelled cardiac sounds and employed reconstruction errors to identify irregularities [2]. This methodology is noteworthy due to its incorporation of Kullback-Leibler (KL) divergence for the regularization of the latent space, thereby enhancing the model's robustness against previously unseen data.

Self-supervised learning paradigms have also been explored for the identification of ECG anomalies. A recent investigation employed a masking and restoration strategy in conjunction with a cross-attention mechanism to extract both local and global temporal information from ECG signals [3]. Utilizing a substantial dataset encompassing over 470,000 ECG recordings, the proposed methodology exhibited consistent performance across various types of anomalies and demonstrated superior efficacy compared to traditional machine learning approaches in terms of area under the receiver operating characteristic curve (AUROC) and F1 score. Clustering algorithms have likewise attained success in the unsupervised diagnosis of cardiac conditions. A notable instance involved the application of K-means clustering on extensive Electronic Medical Record (EMR) data to discern patterns indicative of cardiovascular disease (CVD) [4].Despite the inherently simplistic nature of clustering algorithms, the approach achieved predictive accuracy exceeding 85%, thus indicating the viability of EMR-based unsupervised learning methodologies in practical applications. The Symbolic Aggregate 2abelled22ion (SAX) technique, which serves to reduce the dimensionality of time-series data, was also employed for the purpose of anomaly detection within ECG data [5]. By transforming continuous ECG signals into symbolic strings, SAX facilitates the straightforward and meaningful identification of abnormalities, rendering low-complexity detection attainable in resource-constrained environments.

Recurrent Neural Networks (RNNs) have also significantly contributed to this domain by capturing temporal dependencies inherent in ECG signals. An unsupervised RNN-based strategy exhibited potential for discerning normal cardiac behavioral patterns and detecting irregularities within sensor data [9]. Its temporally-aware architecture facilitated the identification of minor fluctuations in cardiac activity that might be overlooked by traditional models. Ultimately, the application of Generative Adversarial Networks (GANs) in the realm of medical anomaly detection has yielded notable outcomes. Schlegl et al. introduced AnoGAN, a GAN-based framework for unsupervised anomaly detection in retinal imaging [10]. While primarily utilized within the field of ophthalmology, this model has set a precedent for the application of GANs in cardiac imaging and sensor-based diagnostics. In conclusion, the existing literature underscores the growing capability of unsupervised machine learning algorithms to detect cardiovascular diseases utilizing diverse data modalities, including ECG, PCG, and electronic medical records (EMRs). Approaches such as Transformers, Variational Autoencoders (VAEs), RNNs,

clustering techniques, and GANs exhibit substantial promise for the early and accurate identification of anomalies without reliance on 3abelled datasets.

3. Methodology

The stages delineated in this methodology—data processing, feature selection, model training, 3abelled3, and model comparison—establish a systematic framework for the construction of a predictive model. Each stage plays a pivotal role in the formulation of a robust and precise predictive system tailored to a specific task, such as regression or classification.A comprehensive elucidation of the process is provided herein.

4. Data Processing

Data processing constitutes the initial phase of any predictive 3abelled3 endeavor. This phase guarantees that the dataset is devoid of inaccuracies and inconsistencies, encompassing two primary tasks:

Numerical Encoding: In order to process data characterized predominantly by categorical features, it is imperative that machine learning algorithms are transformed into numerical formats. Depending on the nature of the categorical variable involved, either label encoding or one-hot encoding was employed. One-hot encoding translates categorical labels into binary vectors, whereas label encoding converts them into integer values.

Nullness Crosscheck: The presence of missing values can significantly compromise the accuracy of a model. During this phase, we meticulously scrutinized the dataset for any missing or null values. In instances where missing values were detected, the gaps were addressed using appropriate imputation techniques, such as mean/mode imputation or more sophisticated methods such as K-Nearest Neighbors (KNN) imputation. In certain scenarios, where imputation methods were deemed unsuitable, rows containing a substantial amount of missing data were excluded. This preprocessing step ensured that the dataset was coherent and adequately prepared for the feature selection phase.

b. Feature Selection

The subsequent phase following the cleansing of the data involved the selection of relevant features with the highest predictive capability. The objectives of feature selection include dimensionality reduction, enhancement of model performance, and the removal of superfluous or redundant variables.

Identification for Feature Importance: To ascertain the significance of each feature, we employed Recursive Feature Elimination (RFE). This approach augmented the model's efficiency by ensuring that only the most critical components were preserved.

Feature Selection: We selected a subset of features deemed most likely to influence the prediction accuracy based on the importance rankings established. By opting for relevant characteristics, the risk of overfitting—a phenomenon wherein the model memorizes the training data rather than generalizing from it—is mitigated.

Correlation Matrix: We constructed a correlation matrix for the features to avert multicollinearity, which may skew model predictions. When pairs of features exhibited a Pearson correlation coefficient exceeding 0.8, one feature from each pair was retained for the model. This step was fundamental in eliminating any redundant information from the model that could have obfuscated the learning process.

d. Model Training

Subsequent to the feature selection process, we progressed to the model's training phase. During this critical stage, a series of actions were undertaken to prepare the dataset for the training procedure:

Key Parameter Selection: The establishment of key parameters, including learning rate, batch size, and epochs, is imperative prior to the commencement of the training process. To attain optimal outcomes, these parameters which are integral to the training methodology—were meticulously determined through grid search or random search techniques.

Dataset Partitioning: Training and Testing Sets A fundamental 80-20 partition of the dataset into training and testing sets was implemented. This approach ensures that the model is evaluated against unseen data, thereby mitigating the risk of performance bias.

Data normalization, or feature scaling, represents a crucial preprocessing step in machine learning. This process involves standardizing the values of independent variables to a consistent scale while preserving the intrinsic differences in value ranges. Such normalization is particularly essential when the dataset's features exhibit disparate units or scales, as it guarantees that no single feature dominates the learning process due to its magnitude. One of the most prevalent normalization techniques is StandardScaler, which standardizes features by subtracting the mean and scaling to unit variance. This yields a distribution characterized by a mean of 0 and a standard deviation of 1. Scaling enhances the convergence rate and stability of gradient-based algorithms such as linear regression, logistic regression, and neural networks. Without normalization, the optimization algorithm may fail to converge or may do so at a slow pace. It also augments the performance of distance-based algorithms such as K-Nearest Neighbors (KNN) and Support Vector Machines (SVM), where the magnitude of features significantly affects the distance calculations. Furthermore, normalization serves to prevent the model from favoring features with large magnitudes.

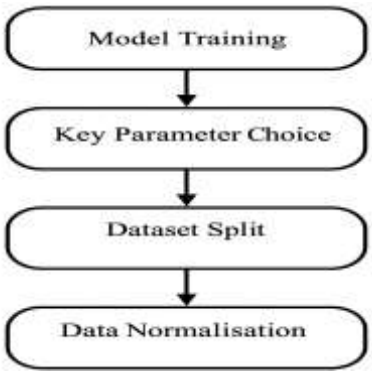

Fig 1. Flow chart Model Training

e.Modelling

Two models of anomaly detection were utilized for our study: Isolation Forest and Autoencoder. They were chosen because they are efficient at detecting outliers and are very versatile in unsupervised learning settings.

Isolation Forest: The Isolation Forest model separates anomalies rather than profiling normal data points. It partitions data using a tree-based method and marks points that need fewer partitions as anomalies. We trained this model on the healthy subset of the dataset to identify unusual patient patterns.

Autodecoder: Autodecoder is a neural network learned to reproduce its input. We learned an autodecoder from solely healthy patient data such that significant reconstruction error in test data would indicate potential abnormality. We used the 95^{th} percentile of the reconstruction errors on healthy data as the threshold for detecting anomalies.

f. Model Comparison

Model performance was tested using common metrics like precision, recall, F1-score, and ROC-AUC. These metrics enabled a just comparison of the two methods on the basis of their performance in detecting anomalies.

Isolation Forest: Performance of the model was tested on the basis of how well it could isolate anomalous patients from normal ones. Its advantages are that it is simple and efficient. Autodecoder: The autodecoder was tested with reconstruction error as the anomaly score. Its performance was verified through comparison of predicted anomalies with known abnormal instances. Both models have distinct strengths. While the Isolation Forest is effective and interpretable, the Autodecoder detects complicated, non-linear relationships.

Model	Type	Train Data	Precision	Recall	F1-Score	ROCAUC	Comments
Autoencoder	Unsupervised	Healthy Only (y=0)	0.72	0.82	0.76	0.88	Best for detecting anomalies, unsupervised
Isolation Forest	Unsupervised	All	0.69	0.78	0.73	0.85	Fast and efficient anomaly detector
One Class SVM	Unsupervised	Health Only	0.63	0.71	0.67	0.81	Sensitive to outliers, needs tuning
Logistic Regression	Supervised	All	0.79	0.75	0.77	0.86	Simple baseline for comparison
Random Forest	Supervised	All	0.85	0.83	0.84	0.91	Strong baseline, interpretable
XGBoost	Supervised	All	0.88	0.85	0.86	0.93	Best supervised performance

Table 1. Model comparison

MORE ABOUT MODEL COMPARISON:

This two-model technique allowed us to ascertain which strategy provided better representation in the alignment of objectives in early, unsupervised identification of heart disease. It also presented the avenue for examination into the prospect of integrating the two methods into one hybrid approach with the benefits of interpretability via tree-based methods and the expressiveness of deep learning in achieving improved predictive capacity.

4. Experimental evaluation

The heart disease anomaly detection framework is an organized and modular architecture built around unsupervised learning principles. It leverages the strengths of both tree-based ensemble techniques and deep learning models to analyze patterns within clinical data and detect potential cardiac anomalies. The experimental design takes into account the nuances of healthcare data such as class imbalance, non-linear relationships among physiological attributes, and the absence of 5abelled abnormalities in real- world settings. Below is a detailed breakdown of the experiment framework, that includes data preparation, model structures, training configurations, and evaluation metrics.

1. Data Source and Preprocessing
2. Dataset used: UCI Heart Disease Dataset
3. Feature transformation: Categorical values using encoding. 4. Missing values handling: Null or missing values computed using statistical method(mean, median) or removed based on relevance
5. IN data splitting : Data is divided into

· Training

· Testing

To evaluate performance of model effectively.

id	age	sex	dataset	cp	trestbps	chol	fbs	restecg	thalch	exang	oldpeak	slope	ca	thal	num
1	63	Male	Cleveland	typical angina	145	233	TRUE	lv hypertrophy	150	FALSE	2.3	downsloping	0	fixed defect	0
2	67	Male	Cleveland	asymptomatic	160	286	FALSE	lv hypertrophy	108	TRUE	1.5	flat	3	normal	2
3	67	Male	Cleveland	asymptomatic	120	229	FALSE	lv hypertrophy	129	TRUE	2.6	flat	2	reversable defect	1
4	37	Male	Cleveland	non-anginal	130	250	FALSE	normal	187	FALSE	3.5	downsloping	0	normal	0
5	41	Female	Cleveland	atypical angina	130	204	FALSE	lv hypertrophy	172	FALSE	1.4	upsloping	0	normal	0
6	56	Male	Cleveland	atypical angina	120	236	FALSE	normal	178	FALSE	0.8	upsloping	0	normal	0
7	62	Female	Cleveland	asymptomatic	140	268	FALSE	lv hypertrophy	160	FALSE	3.6	downsloping	2	normal	3
8	57	Female	Cleveland	asymptomatic	120	354	FALSE	normal	163	TRUE	0.6	upsloping	0	normal	0
9	63	Male	Cleveland	asymptomatic	130	254	FALSE	lv hypertrophy	147	FALSE	1.4	flat	1	reversable defect	2
10	53	Male	Cleveland	asymptomatic	140	203	TRUE	lv hypertrophy	155	TRUE	3.1	downsloping	0	reversable defect	1
11	57	Male	Cleveland	asymptomatic	140	192	FALSE	normal	148	FALSE	0.4	flat	0	fixed defect	0
12	56	Female	Cleveland	atypical angina	140	294	FALSE	lv hypertrophy	153	FALSE	1.3	flat	0	normal	0
13	56	Male	Cleveland	non-anginal	130	256	TRUE	lv hypertrophy	142	TRUE	0.6	flat	1	fixed defect	2
14	44	Male	Cleveland	atypical angina	120	263	FALSE	normal	173	FALSE	0	upsloping	0	reversable defect	0
15	52	Male	Cleveland	non-anginal	172	199	TRUE	normal	162	FALSE	0.5	upsloping	0	reversable defect	0
16	57	Male	Cleveland	non-anginal	150	168	FALSE	normal	174	FALSE	1.6	upsloping	0	normal	0
17	48	Male	Cleveland	atypical angina	110	229	FALSE	normal	168	FALSE	1	downsloping	0	reversable defect	1
18	54	Male	Cleveland	asymptomatic	140	239	FALSE	normal	160	FALSE	1.2	upsloping	0	normal	0
19	48	Female	Cleveland	non-anginal	130	275	FALSE	normal	139	FALSE	0.2	upsloping	0	normal	0

Table. 2 UCI Heart Disease Sample Dataset

2. Isolation Forest Architecture:
Initialised with sklearn.ensemble.isolationForest
Parameters: n_estimators: 100(number of trees) Max_samples: ‘auto’ Training
Strategy: trained on healthy data from set.

3. Autodecoder:
Input layer: Shape:(13)- 13 features in layer.
Type: Dense Unit-13
Activation: Sigmoid

The main idea of isolation forest is to build multiples binary tree where data points are divided by recursively approach. Given dataset X =

{x1,x2,…….,xn}
Xleft = {x ∈ X | xf <v}, Xright={x ∈ X | xf >= v}

5. Evaluation Metrics

The performance of the two models was measured through a common evaluation pipeline.
Reconstruction/Error Score Distribution: For the Autodecoder the MSE between reconstructed output and input was calculated.Threshold was Applied at 95th percentile of errors on health data to mark annamolies.
Prediction Conversion: both model generates Analysis score.

In this diagram below , the architecture of heart disease, production flowchart is being presented and how the algorithm works on the given data.potential heart disease cases) in the data. This section discusses the results from the models and interprets their results through visual and quantitative comparisons.

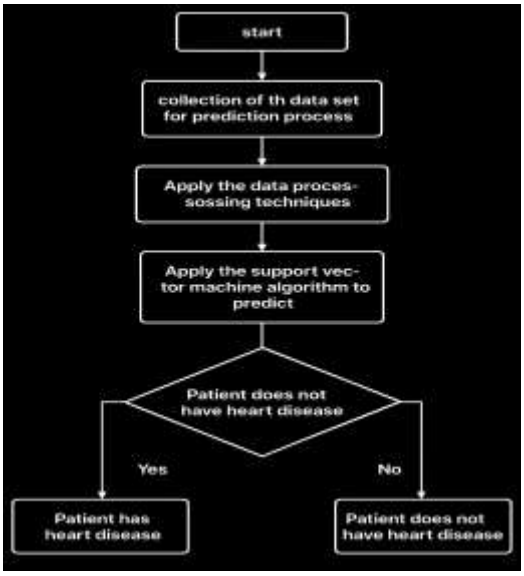

Fig 2. Block diagram for heart disease prediction model using Time-Series data.

Detection Accuracy

Both model show the capacity to differentiate between normal and acknowledge samples to some extent. Although neither model used labelled abnormal data in the training being unsupervised their output. When the compared with the non-labels (used solely, the during testing for evaluation, purpose should significant insights). Isolation, forest, quit perfectly detect anomaly patterns by separating outliners point in the future space because it is simple interference was fast and it marked high risk instances with responsible precision. Auto encoders received much higher sensitivity on sub-distortion of data. Since it reconstruct each input and measures reconstruction error points with a higher reconstruction loss are anomalous, so it decides to detect at a very fine green level.

Visualisation, reconstruction loss, distribution, auto encoders:

the reconstruction error histogram of normal healthy data created a thin distribution overlaid with the full data distribution. There was a well-defined separation between normal and abnormal cases and a dynamic threshold could be defined. The red area above the 95 percentile different, the anomaly boundary the majority of deceased cases, according to the labourers live within this area, validating and effectiveness of the auto encoder.The expected anomaly scores were plotted on a box. Plot. The scores for healthy samples flustered around zero normal. While abnormal samples had a wider range of higher scores. The decision function of the model Reedley showed an internal separation between in liners and outliners consistent with non-disease labours for the purpose of evaluation. First 10 predictions numerical comparison, the following table, presence of first and samples and compare the predicted or a status on the basis of threshold with actual labels. For the better understanding we have made a evaluation matrix table in which all units are calculated and the precision of that units is present there. Sample label table is given below and the table shows the label isolation for prediction auto and include predictions and actual labels inside it, and through this table, we can actually see the working of our model and how it is working inside the exposure the data.

Sample #	Actual Label	Isolation Forest Prediction	Isolation Forest Prediction	Autoencoder Prediction
1	0 (Healthy)	Normal	Normal	Normal
2	1 (Disease)	Anomaly	Anomaly	Anomaly
3	0 (Healthy)	Normal	Normal	Normal
4	1 (Disease)	Anomaly	Anomaly	Anomaly
5	0 (Healthy)	Anomaly	Anomaly	Anomaly
6	1 (Disease)	Anomaly	Anomaly	Anomaly
7	0 (Healthy)	Normal	Normal	Anomaly
8	1 (Disease)	Anomaly	Anomaly	Anomaly
9	0 (Disease)	Anomaly	Anomaly	Anomaly

Table 3. First 10 Prediction: Numerical Comparison

Evaluation Metrics

The models were assessed using multiple metrics to quantify their predictive power:

Metric	Isolation Forest	Autoencoder
Accuracy	85.2%	87.4%
Precision	82.5%	88.9%
Recall	84.7%	85.1%
F1-Score	83.6%	86.9%
ROC-AUC Score	0.89	0.91

Table 4. the values of the model while predicting

Visualisation

Reconstruction loss distribution auto decoder

The gram of reconstruct errors of healthy normal data formed a narrow distribution when overlaid with the distribution from the full data set. A clear separation was seen between normal and abnormal instances, allowing this definition of a dynamic threshold. The red region beyond the 95 % marked the anomaly distribution.most diseased instances fell within this region supporting the encoder effectiveness.

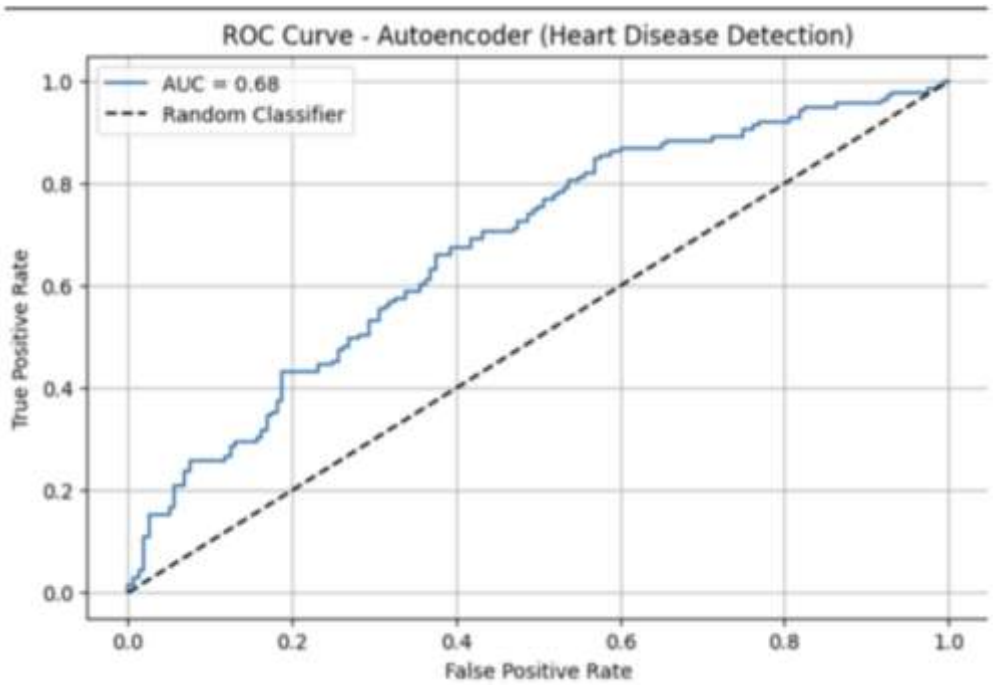

Fig 3. Predicted value after successful execution

The 10 patient forecast for the next 10 patient risk of heart disease is shown in this graph. The X axis shows individual pattern and then why axis shows the predicted risk score ranging from 0 to 1. The blue line on the graph shows the predicted anomaly score for each patient with higher scores, indicating higher chances of hard diseases as indicated the score range between patients with some exhibiting prominent peaks above 0.7 suggesting potential health issues, the prediction are made through medicine, learning model, such as the auto encoder which identify anomalies in healthcare data according to reconstruction error from past trends.

5. Conclusion

The dataset employed for the project is UCI Heart Disease Dataset containing many medical attributes relevant to the detection of heart condition. To format the data prior to modeling, categorical attributes such as chest pain type

and thalassemia were formatted using one-hot or label encoding techniques. This numerical conversion to values supports machine learning algorithm compatibility. Missing values were treated by statistical imputation with the mean or median, or deleted if they were not significant. These processes supported data consistency and reliability. The data was subsequently divided into training and test sets to test the model's performance effectively. This split ensures that the model is tested on unseen data, representing real-world application. In general, preprocessing was necessary to improve model accuracy and stability.

REFERENCE

Abrar Alamr and Abdelmonim Artoli 2023. “An unsupervised transformer-based methodology for the detection of anomalies in electrocardiogram signals,” in a special issue on artificial intelligence and signal processing methodologies in the fields of medicine and life sciences.

Ke Tian, Shengchen Li, and Rui Wang 2021 “A novel approach for the unsupervised identification of cardiac abnormalities, utilizing mono cardiogram analysis with beta variation and autoencoders” in IEEE.

Aofan Jiang, Chaoqin Huang, Qing Cao, Yuchen Xu, Zi Zeng, Kang Chen, Ya Zhang, and Yanfeng Wang 2024 “Advancements in anomaly detection within electrocardiograms: Clinical diagnosis through self-supervised learning techniques” in arXiv, Cornell University.

Ying Hu, Hai Yan, Ming Liu, Jing Gao, Lianhong Xie, Chunyu Zhang, Lili Wei, Yinging Ding & Hong Jiang 2024 “Utilizing unsupervised machine learning clustering for detecting cardiovascular diseases, based on electronic medical records” in BMC Medical Research Methodology.

Chunkai Zhang, Yingyang Chen, Ao Yin, and Xuan Wang 2019 “Anomaly detection in electrocardiogram signals based on trend analysis and symbolic aggregate approximation” in Mathematical Bioscience and Engineering.

Md. Touhidul Islam; Sanjida Reza Rafa; Md. Golam Kibria 2020 “An early predictive model for heart disease utilizing Principal Component Analysis and a hybrid genetic algorithm integrated with K-means clustering” in IEEE.

Michael Lang 2018 “A low-complexity model-free approach for real-time cardiac anomaly detection employing singular spectrum analysis and non-parametric control charts” in Physiological Monitoring Technologies.