Heart Disease Prediction Using Deep Learning Techniques

V. Adlin beenu [1], G. Jini mol [2]
[1, 2] Dept of CSE
[1, 2] Arunachala College Of Engineering

Abstract- *The term "heart disease" refers to several types of heart conditions. The most common type of heart disease in the United States is coronary artery disease (CAD), which affects the blood flow to the heart. Decreased blood flow can cause a heart attack. Sometimes heart disease may be "silent" and not diagnosed until a person experiences signs or symptoms of a heart attack, heart failure, or an arrhythmia, but there are ways to prevent and manage many types of heart disease. This project proposes a deep learning approach for heart disease prediction. The nature of heart disease is complex, and hence, the disease must be handled carefully. This project provides accurate and effective deep learning techniques. The K-means clustering technique is used to identify groups of data objects in a dataset. Particle swarm optimization (PSO) is a method for estimating solutions to numerical maximization and minimization problems that are very difficult or impossible. A deep learning technique (LSTM classifier) is used to classify the attacks in several conditions. This project significantly improves the healthcare system and serves as a valuable diagnostic tool for healthcare professionals. This project is implemented using Python Jupyter software.*

Keywords- Coronary artery disease, deep learning, heart disease prediction, K-means clustering, Particle swarm optimization, LSTM classifier.

I. INTRODUCTION

According to WHO, Heart Diseases are a leading cause of death worldwide. It is quite difficult to identify the cardiovascular disease (CVD) because of some contributory factors which contribute to CVD like high blood pressure, cholesterol level, diabetics, abnormal pulse rate, and many other factors. Sometimes CVD symptoms may vary for different genders chest pain while a female patient has some other symptoms with chest pain like chest discomfort: such as nausea, extreme fatigue, and shortness of breath. Researchers have been exploring a wide range of techniques to predict heart diseases but the disease prediction at an early stage is not very efficient due to many factors, including but not limited to complexity, execution time, and accuracy of the approach. As such, proper treatment and diagnosis can save many lives.

One American dies every 36 seconds due to CVD. More than 665 million people die due to heart disease which one in every 4 deaths. Cardiovascular disease costs a lot to the US healthcare system. In the years 2014 and 2015, it cost about $219 billion per year in terms of healthcare services, medicine, and lost productivity due to death. Early diagnosis can also help to prevent heart failure which can lead to the death of a person. Angiography is considered as the most precise and accurate method for the prediction of cardiac artery disease (CAD), but it is very costly which makes it less accessible to low-income families. A number of factors such as blood pressure, cholesterol, creatine, etc., affect heart health, so it makes it difficult to diagnose. The authors in analyzed different factors that cause heart disease and identified controllable factors such as alcohol usage, smoking, diabetics, high cholesterol, and limited physical activity. In the modern era, electronic health records (EHRs) are also helpful for clinical and research purposes. The physical examination might have some errors and in the case of heart disease, these minor errors can cost a life in the future. Machine learning-based expert systems effectively diagnose CVD and as a result death ratio is reduced.

Heart failure disease (HFD), which has the highest global prevalence, has a multifactorial pathology. In some cases, cardiac output is reduced due to changes in mechanical properties and altered cardiac electrical activity. A number of neurohormonal regulatory mechanisms are activated in the early stages of HFD. Although these compensatory mechanisms can compensate for the effects of HFD in the short term, they can cause dyspnea on exertion, pulmonary and peripheral edema, heart remodeling, and exacerbate ventricular dysfunction, which can cause permanent changes in preload and after load. Many treatment options are offered to patient with HFD such as life style changes, medications or implantable devices such as a pacemaker or defibrillator. Following up with this population is a major concern, given that hospitalization due to acute HFD decompensation is the leading cause of health care expenditure. According to studies and statistics, heart disease is the most serious problem that people face, particularly HFD. Early detection and diagnosis of cardiac disease, like other diseases, is the first step in treatment and care.

Heart disease/ Coronary illness is a popular term that implies that the heart isn't used regularly. Children can be brought into the world with heart-related problems. This is called inherent heart disorder. Many of the heart/coronary related diseases can be acquired. Coronary illness, which incorporates ischemic coronary illness, coronary heart disappointment and different sicknesses of the heart, is the main source of all around. It represents the greater part of all s because of cardiovascular illness. Coronary illness can cause angina (chest torment), heart assaults (myocardial localized necrosis) and unexpected heart failure. A heart assault is typically the principal indication of coronary illness. Peoples/Individuals with a cardiovascular ailment or who are at high cardiovascular hazard (because of the nearness of at least one hazard factors, for example, hypertension, diabetes, hyperlipidemia or officially settled malady) need early identification and the board utilizing directing and prescriptions, as fitting. The extent of deaths and incapacity from coronary illness was essentially higher in men than in ladies, however, was comparable among people for stroke. Death due to cardiovascular diseases increased day by day.

So, diagnoses & prediction of heart disease earlier is compulsory and treatments for individuals/peoples who are probably going to have a coronary illness and help them have a longer life. A favored technique to determine the issues of precise finding and the conveyance of focused treatments is the continuous execution of complete physical assessments. Be that as it may, total and continuous physical assessments would prompt information over-burden. Develops a technique approach in this work by combining an LSTM model with classification methods. The main purpose of this algorithm to produce accurate results for finding out heart disease prediction.

II. LITERATURE SURVEY

Developed a Cardiovascular disease is a major cause of death and morbidity worldwide. Predicting heart disease survivors is a significant challenge in clinical data analytics. This study examines the heart failure survivors from a dataset of 299 hospitalized patients. The goal is to identify significant features and effective data mining techniques that can improve the predictability of cardiovascular patient survival. To address the problem of class imbalance, the Synthetic Minority Oversampling Technique (SMOTE) is used. Various optimization techniques have been used to improve a variety of metrics such as accuracy, precision, and recall. The primary goal of this study is to highlight a comparison of various machine learning techniques in order to select the best method for predicting heart disease survival. However multiple combinations of machine learning models to benefit from their combined advantages improving the Prediction of Heart Failure Patients' Survival is not considered in this paper.

Described identification method of heart failure using improved vector machine based on duality optimization technique. The goal of this research is to create a new automatic method for diagnosing HFD from ECG signals using an improved support vector machine The proposed algorithm uses a hybrid approach of dual SVM and nonparametric algorithms to detect HFD in ECG signals, resulting in increased reliability and accuracy of identification and diagnosis of heart failure classes in the early stages. The advantage of this method is applicable to ECG signals with several amplitude and duration of the P-QRS-T waves, differentiating between normal and abnormal ECG signals based on heart rate duration from the time interval from one PQRST wave to the next PQRST wave, predicting and diagnosing the HF from an ECG signal only. Drawback of this analysis is chronic disease classification in biomedical signals is not available in this paper.

Presented Detection of Atrial fibrillation (AF) from premature atrial contraction (PAC) and premature ventricular contraction (PVC). In this paper demonstrate a new density Poincare plot-based' machine learning method for detecting AF from PAC/PVCs using electrocardiogram (ECG) recordings. Methods- First propose generating this new density Poincare plot, which is derived 'from the difference of the heart rate (DHR) and provides the DHR's overlapping phase-space trajectory information. In this paper, demonstrate a novel image domain-based method for detecting PAC/PVCs from A. However, the binary Poincare plot only shows the triangular kite shape and not how many times the triangular patterns were formed in the Poincare phase space.

Recommended heart disease prediction using hybrid machine learning techniques. Machine learning (ML) has been shown to be effective in assisting in decision-making and prediction from the large amount of data generated by the healthcare industry. It which will improve the accuracy of cardiovascular disease prediction. The prediction model is introduced with various feature combinations and several well-known classification techniques. The primary goal of this study is to improve the performance accuracy of heart disease prediction. However diverse mixtures of machine learning techniques is not carried out this paper.

Discussed the optimized random forest for improved heart disease detection. Diagnosis of heart failure is a difficult task, especially in underdeveloped and developing countries where human experts and equipment are scarce. For feature selection, the proposed diagnostic system employs the random search algorithm (RSA) and the random forest model for heart

failure prediction. The grid search algorithm is used to optimize the proposed diagnostic system. To assess the precision of the proposed method, two types of experiments are carried out. This method achieved classification accuracy while also improving training accuracy. Furthermore, HF disease is expected to account for more than 80% of the predicted increase in mortality in developing countries.

Examined to find the key features of cardiovascular disease prediction using machine learning techniques. The prediction model incorporates various feature combinations and established classification methods. It achieves a higher level of precision through the heart disease prediction model. The factors that contribute to cardiovascular disease can be identified through this research. For data analysis, a comparison of important variables was shown using the Internet of Medical Things (IOMT) platform. The proposed RFRF-ILM method is used to combine the characteristics of the linear model and the random forest. RFRF-ILM predicts heart disease with high accuracy. However, if the disease is detected at an early stage and prevention measures are implemented as soon as possible, the mortality rate can be drastically reduced.

Suggested the method for detecting and localizing ischemic heart disease that is both quick and accurate. T waves were segmented from averaged MCG recordings, and 164 features were extracted. These characteristics were divided into three categories: time domain characteristics, frequency domain characteristics, and information theory characteristics. The goal of this study is to develop a reliable multi-channel MCG based IHD early detection and localization system. The proposed machine learning method gives clinicians a quick and accurate diagnosis tool for interpreting MCG data, which increases its acceptance in clinics. However, due to its lower sensitivity and specificity, the majority of suspected IHD patients are forced to undergo cardiac computed tomography (CT) or coronary angiography (CAG), putting them at risk of excessive radiation exposure, allergic reaction to contrast agent, and other complications.

Proposed the Heart sounds are used to detect CHF. The approach combines traditional Machine Learning (ML) with end-to-end Deep Learning (DL). The classic ML learns from expert features, whereas the DL learns from a signal's spectro-temporal representation. The method was tested using recordings from 947 subjects from six publicly available datasets and one CHF dataset collected specifically for this study. This technique investigate the differences in heart sounds during the transition from the decompensate to the decompensate states of CHF in order to develop personalized monitoring models. The study went beyond the standard healthy. Patient classification and investigated personalized models for detecting different stages of CHF. Capable of distinguishing between the decompensate and the decompensate phases with an accuracy calculated using a LOSO evaluation. While we are aware that there is a risk of over fitting in these.

Explained recently, wearable devices have become popular with wide applications in the health monitoring system which has stimulated the growth of the Internet of Medical Things. The IOMT has an important role to play in reducing the mortality rate by the early detection of disease. The proposed study aims to identify key characteristics of heart disease prediction using machine learning techniques. Many studies on heart disease diagnosis have been conducted, but the accuracy of the findings is low. Using the Levy flight algorithm, the proposed MSSO-ANFIS improves search capability. The regular learning process in ANFIS is gradient-based and has a tendency to become stuck in local minima. The MSSO-ANFIS prediction model achieves a higher accuracy with a precision than the other approaches. The performance of the presented LCSA technique for optimal feature selection is compared with existing CSA with respect to the fitness value against the number of iterations.

Suggested a Cardio vascular autonomic neuropathy (CAN) is one of the most commonly overlooked diabetes complications. It is distinguished by damage to the autonomic nerves that regulate heart rate and vascular compliance. A novel approach is proposed here to investigate the feasibility of using heart rate variability (HRV) features collected over 24 hours and embedded within machine learning algorithms to provide a comprehensive screening for CAN patients. Proposed models could help primary care centers stratify the risk of CAN, leading to early treatment and preventing sudden cardiac death due to silent myocardial infarction. The method is thought to be simple and effective, especially in clinical settings with limited resources. Forecasting the existence of CAN in non-CAN diabetic patients on a long-term is not carried out in this work.

III. PROPOSED SYSTEM

This proposed work uses a deep learning method to predict the heart attacks. A SDN Dataset is used, and the initial stage of data pre-processing involves an analysis of the input dataset. The datasets are pre-processed using Standard Scaler. The dataset's useless data is handled using data pre-processing techniques. Following data preparation, the stage of the data visualization process is produced. In order to detect characteristics and correlations in the dataset, data visualization is used. After visualization, this dataset is sent to

the K-means cluster algorithm. The K-means cluster algorithm is used to identify clusters of data objects in a dataset. For the data, the following deep learning technique (LSTM classifier) is allowable. In order to detect difficulties in that region, the LSTM classifier is used to classify the attacks. Finally, the data are evaluated to achieve a prediction output.

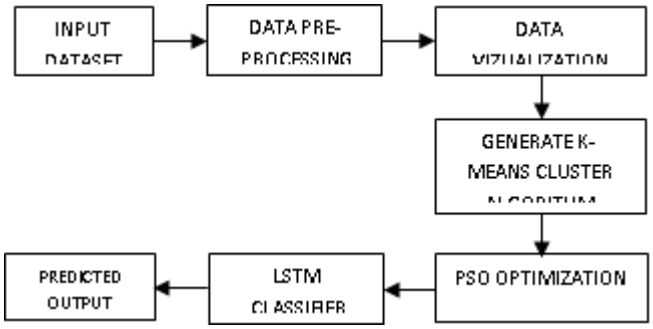

Fig.3.1 Proposed Block Diagram

Data Preparation

Data preparation is the process of preparing raw data so that it is suitable for further processing and analysis. Key steps include collecting, cleaning, and labeling raw data into a form suitable for machine learning (ML) algorithms and then exploring and visualizing the data. The components of data preparation include data preprocessing, profiling, cleansing, validation and transformation; it often also involves pulling together data from different internal systems and external sources.

Steps in the Data Preparation Process

Data preparation is done in a series of steps. There's some variation in the data preparation steps listed by different data professionals and software vendors, but the process typically involves the following tasks:

Data Collection - Relevant data is gathered from operational systems, data warehouses, data lakes and other data sources. During this step, data scientists, members of the BI team, other data professionals and end users who collect data should confirm that it's a good fit for the objectives of the planned analytics applications.

Data Discovery and Profiling - The next step is to explore the collected data to better understand what it contains and what needs to be done to prepare it for the intended uses. To help with that, data profiling identifies patterns, relationships and other attributes in the data, as well as inconsistencies, anomalies, missing values and other issues so they can be addressed.

Data Cleansing - Next, the identified data errors and issues are corrected to create complete and accurate data sets. For example, as part of cleansing data sets, faulty data is removed or fixed, missing values are filled in and inconsistent entries are harmonized.

Data Structuring - At this point, the data needs to be modeled and organized to meet the analytics requirements. For example, data stored in comma-separated values (CSV) files or other file formats has to be converted into tables to make it accessible to BI and analytics tools.

Data Transformation and Enrichment - In addition to being structured, the data typically must be transformed into a unified and usable format. For example, data transformation may involve creating new fields or columns that aggregate values from existing ones. Data enrichment further enhances and optimizes data sets as needed, through measures such as augmenting and adding data.

Data Validation and Publishing - In this last step, automated routines are run against the data to validate its consistency, completeness and accuracy. The prepared data is then stored in a data warehouse, a data lake or another repository and either used directly by whoever prepared it or made available for other users to access.

Data Visualization

Data visualization is a field in data analysis that deals with visual representation of data. It graphically plots data and is an effective way to communicate inferences from data. Using data visualization, it can get a visual summary of our data. With pictures, maps and graphs, the human mind has an easier time processing and understanding any given data. Data visualization plays a significant role in the representation of both small and large data sets, but it is especially useful when we have large data sets, in which it is impossible to see all of our data, let alone process and understand it manually.

Segmentation

Image segmentation is a commonly used technique in digital image processing and analysis to partition an image into multiple parts or regions, often based on the characteristics of the pixels in the image. Image Segmentation involves converting an image into a collection of regions of pixels that are represented by a mask or a labeled image. A common technique is to look for abrupt discontinuities in pixel values, which typically indicate edges that define a region. Another common approach is to detect similarities in the regions of an image. Some techniques that follow this

approach are region growing, clustering, and Thresholding. A variety of other approaches to perform image segmentation have been developed over the years using domain-specific knowledge to effectively solve segmentation problems in specific application areas. So let us start with one of the clustering-based approaches in Image Segmentation which is K-Means segmentation.

K-Means Cluster

K-Means segmentation algorithm is an iterative algorithm that tries to partition the dataset into Kpre-defined distinct non-overlapping subgroups (clusters) where each data point belongs to only one group. It tries to make the intra-cluster data points as similar as possible while also keeping the clusters as different (far) as possible. It assigns data points to a cluster such that the sum of the squared distance between the data points and the cluster's centroid (arithmetic mean of all the data points that belong to that cluster) is at the minimum. The less variation we have within clusters, the more homogeneous (similar) the data points are within the same cluster.

PSO Optimization

PSO is a stochastic optimization technique based on the movement and intelligence of swarms. In PSO, the concept of social interaction is used for solving a problem. It uses a number of particles (agents) that constitute a swarm moving around in the search space, looking for the best solution. Each particle in the swarm looks for its positional coordinates in the solution space, which are associated with the best solution that has been achieved so far by that particle. It is known as pbest or personal best. Another best value known as gbest or global best is tracked by the PSO. This is the best possible value obtained so far by any particle in the neighborhood of that particle.PSO is best used to find the maximum or minimum of a function defined on a multidimensional vector space.

Classification

The classification concept that serves as a mechanism to show interfaces, classes, data types and components. A classifier describes a set of instances that have common behavioural and structural features. The LSTM classification is used in this project.

Long Short-Term Memory Networks Algorithm

Long short-term memory networks are a special kind of recurrent neural networks. LSTM is an unsupervised learning method, although technically, they are trained using supervised learning methods, referred to as self-supervised. All RNNs, including the LSTM, consist of units. These are ANNs with multiple hidden layers that contain cycles from subsequent neurons to preceding ones. These cycles create a special recurrent layer within the network, called hidden state, which acts like a memory and allows the RNN unit to handle time sequenced data. Since the unit is recurrent, it handles the input by time steps. It is a variety of recurrent neural networks (RNNs) that are capable of learning long-term dependencies, especially in sequence prediction problems. Using LSTM, time series forecasting models can predict future values based on previous, sequential data. This provides greater accuracy for demand forecasters, which results in better decision-making for the business.

IV. RESULT AND DISCUSSION

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	52	1	0	125	212	0	1	168	0	1.0	2	2	3	0
1	53	1	0	140	203	1	0	155	1	3.1	0	0	3	0
2	70	1	0	145	174	0	1	125	1	2.6	0	0	3	0
3	61	1	0	148	203	0	1	161	0	0.0	2	1	3	0
4	62	0	0	138	294	1	1	106	0	1.9	1	3	2	0
5	58	0	0	100	248	0	0	122	0	1.0	1	0	2	1
6	58	1	0	114	318	0	2	140	0	4.4	0	3	1	0
7	55	1	0	160	289	0	0	145	1	0.8	1	1	3	0
8	46	1	0	120	249	0	0	144	0	0.8	2	0	3	0
9	54	1	0	122	286	0	0	116	1	3.2	1	2	2	0
10	71	0	0	112	149	0	1	125	0	1.6	1	0	2	1

Figure 4.1Input dataset

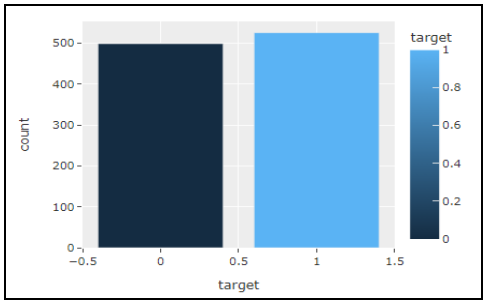

Figure 4.2 Target Distribution

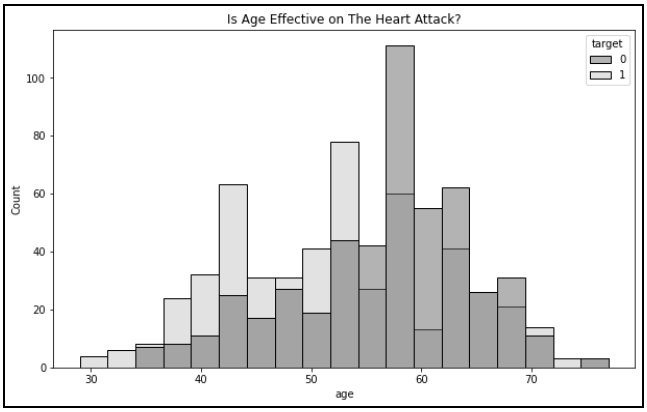

Figure 4.3 Age Estimation.

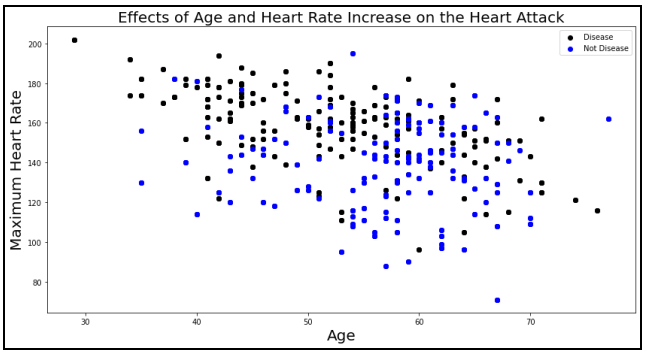

Figure 4.4Comparison of Age and Heart Attack

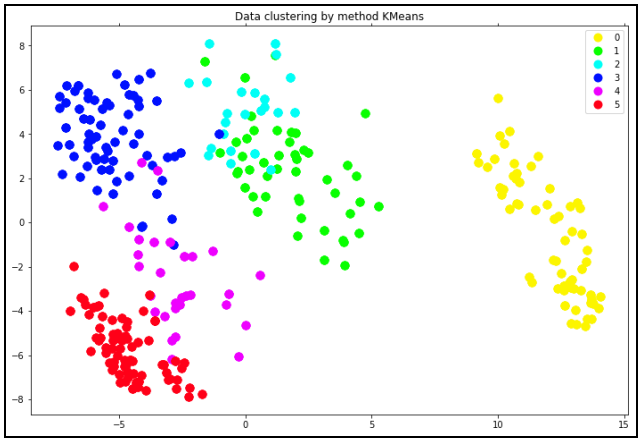

Figure 4.5 Data Clustering by method K-means

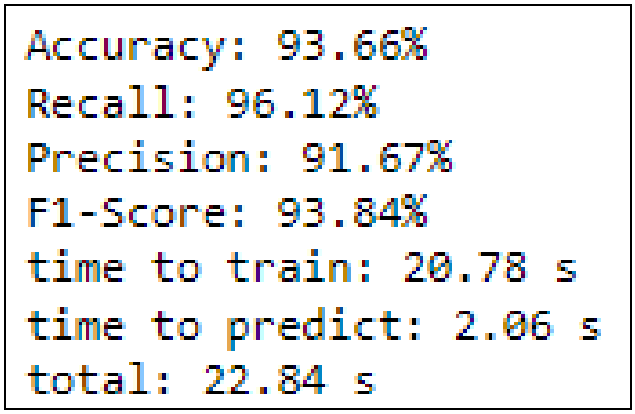

Figure 4.6Performance matrixfor LSTM

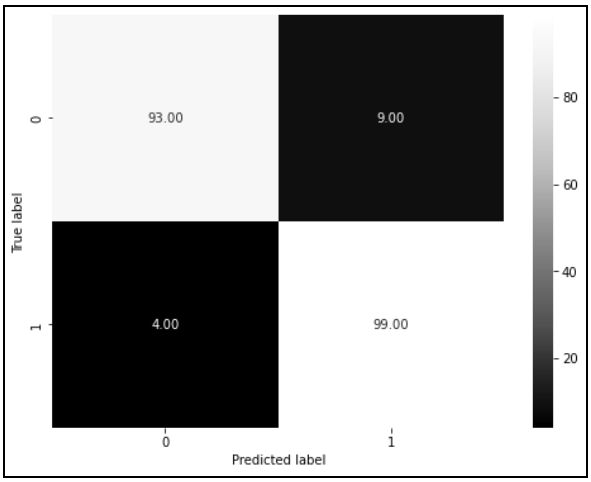

Figure 4.7Confusion Matrix for LSTM

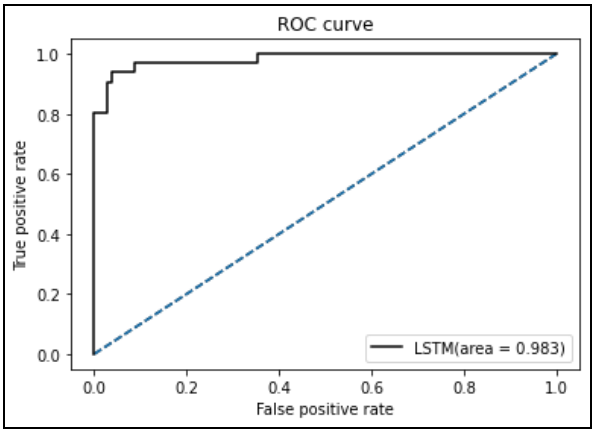

Figure 4.8ROC Curve for LSTM

V. CONCLUSION

This project proposes a deep learning approach for heart disease prediction. The methods that are used for comparison are the confusion matrix, precision, specificity, sensitivity, and F1 score. The classification algorithms of LSTM are well performed in this project. Lastly, evaluate the comparison to see how well the classification techniques succeeded. The results indicate that deep learning is another option for detecting attacks that may cause disruptions in the future.

In future work, our research can be further expanded to include the genetic algorithm. A more effective approach for the genetic algorithms will be explored to further improve accuracy.

REFERENCES

[1] A. Ishaq et al., "Improving the Prediction of Heart Failure Patients' Survival Using SMOTE and Effective Data

Mining Techniques," in IEEE Access, vol. 9, pp. 39707-39716, 2021.

[2] G. G. N. Geweid and M. A. Abdallah, "A New Automatic Identification Method of Heart Failure Using Improved Support Vector Machine Based on Duality Optimization Technique," in IEEE Access, vol. 7, pp. 149595-149611, 2019.

[3] S. K. Bashar et al., "Novel Density Poincaré Plot Based Machine Learning Method to Detect Atrial Fibrillation From Premature Atrial/Ventricular Contractions," in IEEE Transactions on Biomedical Engineering, vol. 68, no. 2, pp. 448-460, Feb. 2021.

[4] S. Mohan, C. Thirumalai and G. Srivastava, "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques," in IEEE Access, vol. 7, pp. 81542-81554, 2019.

[5] A. Javeed, S. Zhou, L. Yongjian, I. Qasim, A. Noor and R. Nour, "An Intelligent Learning System Based on Random Search Algorithm and Optimized Random Forest Model for Improved Heart Disease Detection," in IEEE Access, vol. 7, pp. 180235-180243, 2019.

[6] C. Guo, J. Zhang, Y. Liu, Y. Xie, Z. Han and J. Yu, "Recursion Enhanced Random Forest With an Improved Linear Model (RERF-ILM) for Heart Disease Detection on the Internet of Medical Things Platform," in IEEE Access, vol. 8, pp. 59247-59256, 2020.

[7] R. Tao et al., "Magnetocardiography-Based Ischemic Heart Disease Detection and Localization Using Machine Learning Methods," in IEEE Access, vol. 66, pp. 1658-1667, 2019.

[8] M. Gjoreski, A. Gradišek, B. Budna, M. Gams and G. Poglajen, "Machine Learning and End-to-End Deep Learning for the Detection of Chronic Heart Failure from Heart Sounds," in IEEE Access, vol. 8, pp. 20313-20324, 2020.

[9] M. A. Khan and F. Algarni, "A Healthcare Monitoring System for the Diagnosis of Heart Disease in the IoMT Cloud Environment Using MSSO-ANFIS," in IEEE Access, vol. 8, pp. 122259-122269, 2020.

[10] M. Alkhodari et al., "Screening Cardiovascular Autonomic Neuropathy in Diabetic Patients With Microvascular Complications Using Machine Learning: A 24-Hour Heart Rate Variability Study," in IEEE Access, vol. 9, pp. 119171-119187, 2021.

[11] N. L. Fitriyani, M. Syafrudin, G. Alfian and J. Rhee, "HDPM: An Effective Heart Disease Prediction Model for a Clinical Decision Support System," in IEEE Access, vol. 8, pp. 133034-133050, 2020.

[12] C. Guo, J. Zhang, Y. Liu, Y. Xie, Z. Han and J. Yu, "Recursion Enhanced Random Forest With an Improved Linear Model (RERF-ILM) for Heart Disease Detection on the Internet of Medical Things Platform," in IEEE Access, vol. 8, pp. 59247-59256, 2020.

[13] A. I. Humayun, S. Ghaffarzadegan, M. I. Ansari, Z. Feng and T. Hasan, "Towards Domain Invariant Heart Sound Abnormality Detection Using Learnable Filterbanks," in IEEE Access, vol. 24, no. 8, pp. 2189-2198, 2020.

[14] A. Javeed, S. Zhou, L. Yongjian, I. Qasim, A. Noor and R. Nour, "An Intelligent Learning System Based on Random Search Algorithm and Optimized Random Forest Model for Improved Heart Disease Detection," in IEEE Access, vol. 7, pp. 180235-180243, 2019.

[15] L. Zhao, C. Liu, S. Wei, C. Liu and J. Li, "Enhancing Detection Accuracy for Clinical Heart Failure Utilizing Pulse Transit Time Variability and Machine Learning," in IEEE Access, vol. 7, pp. 17716-17724, 2019.