ISSN (Online): 2583-7052

International Journal of Advanced Trends in Engineering and Management (IJATEM)
Vol. 02 (10), October 2023, pp. 24-35

Article Title: Efficient Heart Disease Prediction Using PSO Optimized LSTM Classifier

Efficient Heart Disease Prediction Using PSO Optimized LSTM Classifier

A. Tamizharasi, S. Mohanalakshmi

Assistant Professor, Department of Computer Science and Engineering, R.M.D. Engineering College, kavarapettai, Thiruvallur.

Professor, Department of Electronics and Communication Engineering, Rohini College of Engineering and Technology, 629401, Tamil Nadu, India.

ABSTRACT

Accurate diagnosis and prognosis of cardiovascular disease are essential in the medical field, enabling cardiologists to administer appropriate treatments. Machine learning has gained prominence in the healthcare sector due to its ability to identify patterns in data. Leveraging machine learning for classifying cardiovascular disease occurrences, significantly minimize misdiagnosis. This research focuses on developing a robust model capable of accurately predicting cardiovascular diseases, ultimately aiming to decrease the mortality rates associated with these conditions. To find clusters of data objects in a dataset, K-means clustering algorithm is implemented. The technique known as particle swarm optimization (PSO) is used to estimate solutions to extremely challenging or unsolvable numerical minimization and maximization problems. LSTM classifier, a deep learning technology is employed to categorize the attacks under various scenarios. In addition to greatly enhancing the healthcare system, this research gives medical practitioners access to an invaluable diagnostic tool. The outcomes of proposed work is examined using programs written in Python Jupyter software tool.

Keywords: Electronic Health Records (EHR), Heart Failure Disease (HFD), Particle Swarm Optimization (PSO), Long-Short Term Memory (LSTM), K-Means Cluster, Cardio Vascular Disease (CVD).

1 Introduction

Heart disease is any condition that impairs the heart's capacity to works normally. Heart failure is most commonly caused by reduction of the coronary arteries, .Globally, it is the major cause of death [1]. The greatest cause of death globally is cardiovascular disease (CVD), a gathering of situations related to the heart. Researchers have studied a wide variety of methods to expect heart diseases, however early detection is not very effective due to a number of factors, with but not limited to the technique's accuracy, complexity, and execution time [2]. The primary technique for identifying coronary artery disease is angiography, a non-invasive imaging modality that is utilized for both the heart and the coronary arteries [3]. Risk factors make it extremely difficult to diagnose heart disorders; as a result, current technologies for detecting the presence of cardiac diseases have evolved [4]. Heart health is complicated by a multitude of factors, including blood pressure, cholesterol, creative, and more, making diagnosis challenging. The scientists examined many risk factors for heart disease and found that

International Journal of Advanced Trends in Engineering and Management (IJATEM)
Vol. 02 (10), October 2023, pp. 24-35

smoking, alcohol consumption, diabetes, high cholesterol, and insufficient physical exercise are among the ones to be controlled [5]. These benefits were offered by Electronic Health Record (EHR) The development of an efficient system that would transform the healthcare sector and solve the problems with paper-based medical records was the goal of EHR systems [6]. Heart failure is a condition in which many neurohormonal regulating mechanisms are activated in the early stages. These compensatory mechanisms results in the consequences of a high-fat diet (HFD), including increased ventricular dysfunction, exertional dyspnea, peripheral edema, pulmonary remodelling, and persistent changes to preload and afterload [7]. Typically, K-means clustering is employed as an unsupervised machine learning approach for clustering and sorting data into discrete clusters based on criteria other than heart-related outcomes or diseases [8]. Particle Swarm Optimization (PSO) is a critical method for selecting features. In terms of speed and memory, it is computationally inexpensive and in order to identify cardiac illness early, PSO and ensemble learning methodologies and sensors are used to generate medical data [9]. A dataset that can be accurately represented as a sequence and has a temporal or sequential aspect to the data is required if you are committed to using LSTM classifiers for the prediction of cardiac disease. Time series data from the electrocardiogram (ECG), heart rate variability, and other physiological markers may be included in this. In this instance, the input data would be handled as a time series, with each time step denoting a distinct data point or data window [10]. Considering that hospitalization for acute HFD decompensation is the primary cause of health care costs, following up with this population is a serious problem. The list below demonstrates how the paper is organised: Section I gives a description of the introduction. The recent works are described in Section II. The proposed model and description are explained in Part III. The results are presented in Section IV. In Section V, the conclusion is presented.

2 Recent Works

Chunyan Guo, *et al* [2020] [11] analysed to identify the salient characteristics of machine learning-based cardiovascular disease prediction. The prediction model combines a number of feature combinations with tried-and-true classification techniques. With the use of the heart disease prediction model, it attains a better degree of precision. This research is used to identify the elements that lead to cardiovascular disease. IOMT platform was used for data analysis to compare significant variables. The features of the random forest and the linear model are combined using the proposed RFRF-ILM approach. The accuracy of RFRF-ILM's heart disease prediction is great. However, death rate is significantly decreased if the illness is recognized early and preventative measures are put in place as soon as feasible.

Rong Tao, *et al* [2019] [12] have introduced an accurate and speedy way to identify and locate ischemic heart disease. From averaged MCG recordings, T waves were separated, and 164 characteristics were obtained. The time field features, frequency domain characteristics, and information theory characteristics of three groups into which these traits were separated. This

International Journal of Advanced Trends in Engineering and Management (IJATEM)
Vol. 02 (10), October 2023, pp. 24-35

work aims to provide a dependable multi-channel MCG-based IHD early detection and localization solution. Clinicians' acceptance of the suggested machine learning approach is increased since it provides them with a rapid and precise diagnostic tool for analysing MCG data. However, because cardiac computed tomography (CT) or coronary angiography (CAG) has lesser sensitivity and specificity, most suspected IHD patients are required to undergo these procedures, which puts them at risk for excessive radiation exposure, allergic response to complications and contrast agent.

Martin Gjoreski, *et al* [2020] [13] have proposed heart sounds by means of detecting CHF. It blends end-to-end Deep Learning (DL) with conventional Machine Learning (ML). In contrast to the DL, which learns from a signal's Spectro-temporal representation, conventional ML learns from expert characteristics. The approach was tested applying recordings from 947 people across six datasets that were made accessible to the public and one CHF that is specifically gathered for this purpose. To create customized monitoring models, this method examines the variations in cardiac sounds as CHF moves from the decompensated to the recompensated phases. The study looked into customized models for identifying various stages of CHF, going beyond the conventional healthy patient classification. An accuracy determined by a LOSO evaluation, is able to discriminate between the recompensated and decompensated phases.

Mohamad ayob khan, *et al* [2020][14] As recently explained, wearable technology has gained popularity and many uses in health monitoring systems, have fuelled the expansion of the Internet of Medical Things. Through early disease diagnosis, the IOMT plays a significant part in lowering the death rate. The study aims to identify key aspects of heart disease prediction by machine learning techniques. Numerous research has been done on the diagnosis of cardiac disease, however, the results are not very accurate. The MSSO-ANFIS enhances search performance by utilizing the Levy flight algorithm. The gradient-based regular learning process in ANFIS is prone to becoming trapped in local minima. Compared to the other methods, with precision, the prediction model achieves a higher accuracy. In terms of the fitness value against the number of iterations, the proposed LCSA technique for optimal feature selection is demonstrated and contrasted with the current CSA.

Mohanad Alkhodar *et al* [2021] [15] have introduced Cardiovascular autonomic neuropathy (CAN) is a frequently disregarded consequence of diabetes. Damage to the autonomic neurons, which control vascular compliance and heart rate, sets it apart. Here, a novel method is put forth to examine the viability of providing a thorough screening for patients with chronic illness (CAN) by utilizing heart rate variability (HRV) data that are gathered over the course of a day and integrated into machine learning algorithms. The suggested models have the potential to assist primary care facilities in stratifying patients' risk of CAN, which could result in early treatment and save sudden cardiac death from silent myocardial infarction. It is believed that

International Journal of Advanced Trends in Engineering and Management (IJATEM)
Vol. 02 (10), October 2023, pp. 24-35

this approach is easy to use and efficient, particularly in low-resource clinical settings. It is not done to predict the long-term presence of CAN in diabetic patients who do not have CAN.

3 Proposed Work Explanation

The proposed work predicts cardiac attacks using a deep learning technique, shown in Figure 1. An analysis of the input dataset is the first step in the data pre-processing process, which uses an SDN dataset. StandardScaler is used for pre-processing the datasets. Data pre-processing methods are used to deal with the unnecessary data in the dataset. The stage of the data visualization process is created after the data has been prepared. Data visualization is used to find features and correlations within the dataset. This dataset is forwarded to the K-means cluster algorithm after visualization. To find clusters of data objects in a dataset, apply the K-means cluster algorithm. The following deep learning method (LSTM classifier) is acceptable for the data. The LSTM classifier is used to categorize the attacks in order to identify problems in that area. In order to produce a prediction result, the data are finally evaluated.

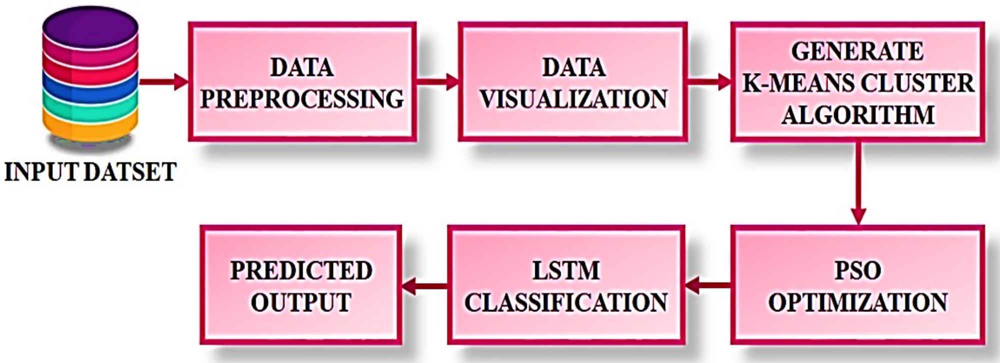

Figure 1: *Proposed Block diagram*

3.1 Input Dataset

The following dataset given in the table are defined as fallows.11 dataset have been taken with people of age above 52 to 71.

3.2 Data Preparation

The method of getting rare data prepared for additional processing and analysis is known as data preparation. In order to prepare raw data for machine learning (ML) algorithms, it is important to first gather, clean, and label the data before analysing and displaying it. Among the components of data preparation are transformation, validation, cleansing, profiling, data preprocessing. Another frequent activity is compiling data from numerous external and internal sources.

International Journal of Advanced Trends in Engineering and Management (IJATEM)
Vol. 02 (10), October 2023, pp. 24-35

3.3 Steps in data preparation process

Data collection- Relevant data is gathered from operational systems, data lakes, data warehouses, and other data sources. In this stage, data scientists, BI team associates, end users who collect data, and other data specialists should confirm that the data matches the purposes of the intentional analytics applications.
Data Discovery and profiling- The next stage is to examine the collected data to find out more about what it contains and what has to be done to make it for its future uses.
Data cleansing- To build comprehensive accurate data sets, the identified data flaws and issues are then addressed. For instance, incorrect data is corrected or eliminated, missing ideals are filled in, and unreliable entries are coordinated as part of data set cleansing.
Data structuring- It is now essential to model and organize the data in order to meet the analytics necessities. For example, data must be converted into tables before BI and analytics tools may access it from comma-separated values (CSV) files or other file formats.
Data Transformation and Enrichment- In order to enhance presence organized, the data typically needs to be converted into a standardized and practical format. For example, introducing new columns or fields that combine values from existing ones could be a requirement of data conversion.
Data Validation and Publishing- Automated processes are useful to the data in this final step to verify its accuracy, constancy, and extensiveness. After that, the set data is kept in a data lake, warehouse or some other repository, and is either utilized by the creator directly or made available to other users.

3.4 Data Preprocessing

Through the extraction of patient medical history—such as blood pressure, blood sugar levels, and chest pain—from a dataset including patient medical history, this initiative predicts individuals who will develop cardiovascular disease.

3.5 Data Visualization

The field of data analysis that deals with the visual display of data is called data visualization. It depicts data graphically and is a useful tool for communicating conclusions drawn from data. The human mind processes and comprehends material more easily when it is presented in the form of images, maps, and graphs. Both tiny and huge data sets is represented effectively with the help of data visualization.

3.6 k-means clustering

The reiterative K-Means segmentation technique pursues to divide the dataset into Kpre-defined unique, non-overlapping subgroups (clusters), with each data point belonging to a single group. The goal is to maintain as much distance between the clusters and as much similarity between the intra-cluster data points as feasible. The process involves grouping data

International Journal of Advanced Trends in Engineering and Management (IJATEM)
Vol. 02 (10), October 2023, pp. 24-35

Article Title: Efficient Heart Disease Prediction Using PSO Optimized LSTM Classifier

points into clusters so that the total squared distance between the cluster's centroid—the arithmetic means of all the data points in that cluster—is as small as possible. The data points inside a cluster are more homogenous (similar) when there is less diversity within them.

3.7 PSO optimization

PSO is a stochastic optimization method inspired by swarm intelligence and mobility. The idea of social interaction is used to problem solving in PSO. It employs a swarm of particles, or agents, to seek for the optimal solution by moving around in the search space. In the solution space, every particle in the swarm searches for its positional coordinates, which correspond to the best.

3.8 Long short-term memory networks algorithm

Although they are officially learned using supervised learning techniques, often known as self-supervised techniques, LSTM is an unsupervised learning method.

3.9 Prediction performance matrix

The number of properly forecast attack flows is represented by TP the number of acceptably predicted normal flows is represented by TN the number of incorrectly prophesied attack flows is represented by FP and the number of incorrectly predicted normal flows is represented by FN in the experimental evaluation that follows.

Accuracy

It's the ratio of accurate forecasts to total predictions; a higher accuracy means the learning model made a more accurate prediction, which is indicated by

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \quad (1)$$

Precision

The ability to prevent incorrectly classifying negative records as positive; a low rate of false positives is associated with a high precision rate. It is the proportion of accurately anticipated attacks to all predicted attacks, represented by

$$\text{precision} = \frac{TP}{TP + FP} \quad (2)$$

Recall

Recall, also known as True Positive Rate, is the proportion of accurately anticipated to actual attacks, represented by

$$\text{Recall} == \frac{TP}{TP + FN} \quad (3)$$

F1 score

International Journal of Advanced Trends in Engineering and Management (IJATEM)
Vol. 02 (10), October 2023, pp. 24-35

The weighted average of Precision and Recall, or F1 score, is represented by

$$F1 = \frac{2 \cdot (\text{Recall} \cdot \text{Precision})}{\text{Recall} + \text{Precision}} \tag{4}$$

AUC

The ROC Curve plots Recall against Incorrect Constructive Amount at several classification levels. AUC stands for Extent under the Curve.

$$FPR = \frac{FP}{FP+TN} \tag{5}$$

4 Results

The validation of proposed work is examined using Python software and the corresponding outcomes for prediction of heart disease is as follows.

	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 2: *Input Dataset*

The dataset utilized by the proposed work consisting of certain measures is illustrated in Figure 2.

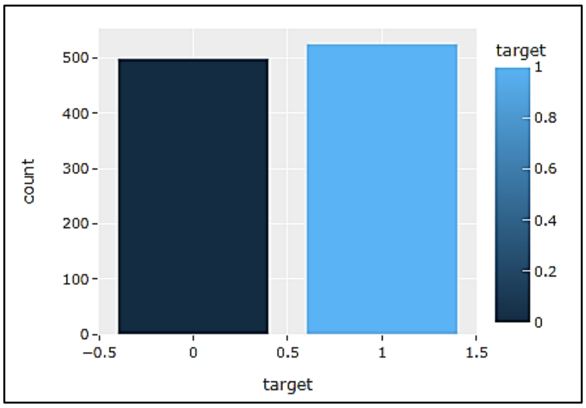

Figure 3: *Target Distribution*

International Journal of Advanced Trends in Engineering and Management (IJATEM)
Vol. 02 (10), October 2023, pp. 24-35

Figure 3 illustrates the target distribution, which is followed by preprocessing the data that is crucial to enhance its quality. During preprocessing, the computer receives the gathered HEART datasets as input and is trained appropriately.

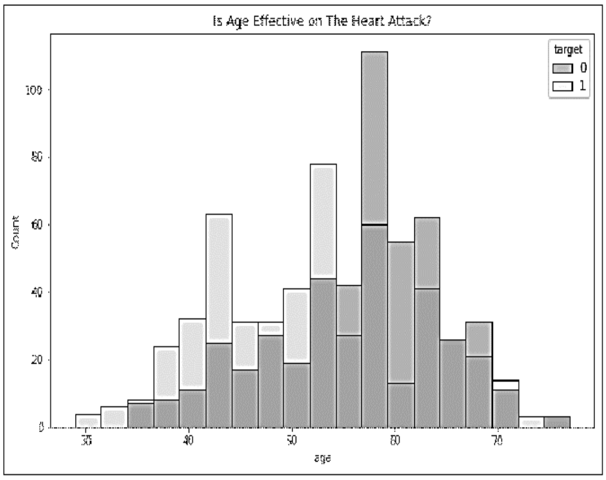

Figure 4: *Age Estimation*

Figure 4 indicates the age at which the heart attack began. Data visualization is the process of visually portraying the data to identify patterns and relationships.

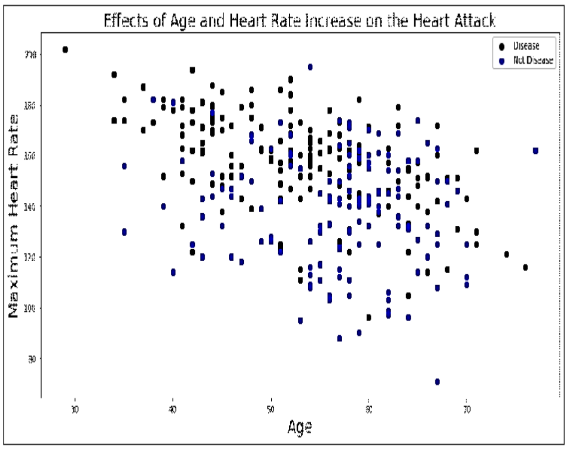

Figure 5: *Comparison of age and heart attack*

Figure 5 demonstrates how age and elevated heart rate relate to the repercussions of a heart attack. Characterizing the clusters is the key to deciphering the significance of k-means clusters, shown in Figure 6. It may view the location of each unique data point across all variables by using a parallel coordinates plot. It is determine each cluster stands for by comparing the values of each variable across clusters.

ISSN (Online): 2583-7052

International Journal of Advanced Trends in Engineering and Management (IJATEM)
Vol. 02 (10), October 2023, pp. 24-35

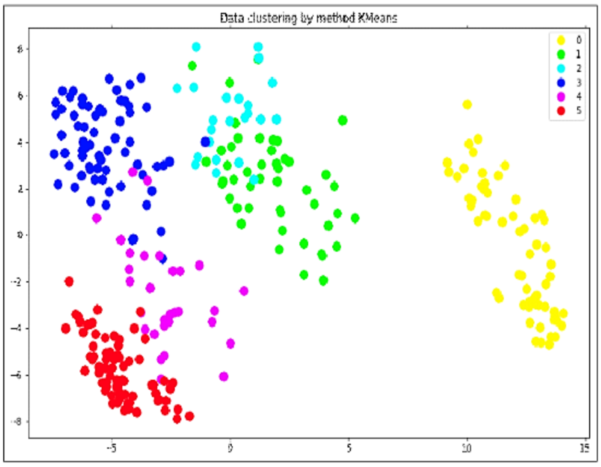

Figure 6: *Data Clustering by methods K-means*

The values of recall (RE), F1 score, and precision (PR) are determined using the logistic regression classification method is shown in Figure 7.

```
Accuracy: 93.66%
Recall: 96.12%
Precision: 91.67%
F1-Score: 93.84%
time to train: 20.78 s
time to predict: 2.06 s
total: 22.84 s
```

Figure 7: *Performance matrix for LSTM*

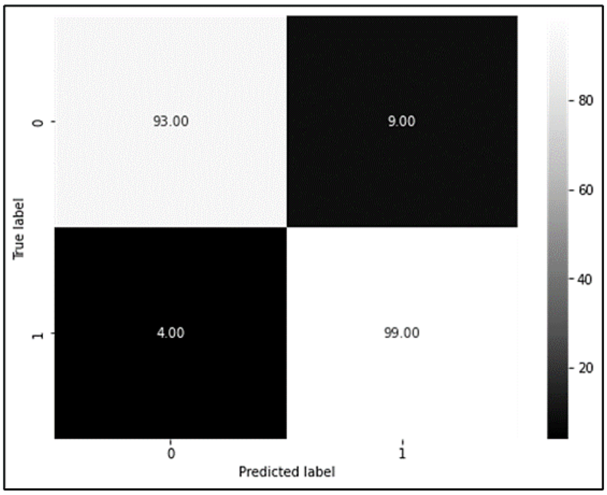

Figure 8: *Confusion matrix for LSTM*

International Journal of Advanced Trends in Engineering and Management (IJATEM)
Vol. 02 (10), October 2023, pp. 24-35

The LSTM confusion matrix is displayed in Figure 8. A confusion matrix, or 2x2 table that contrasts the model's predicted values with the test dataset's actual values, is a popular tool for assessing the quality of a logistic regression model.

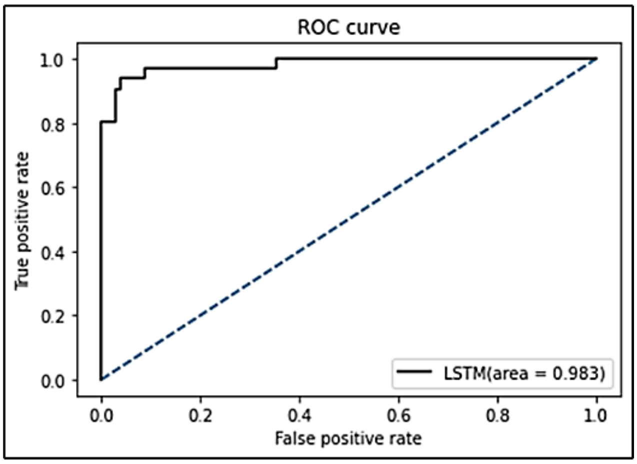

Figure 9: *ROC curve for LSTM*

In Figure 9, the ROC curve for LSTM is shown. A graph that shows a classification model's performance over all categorization levels is called a receiver operating characteristic curve, or ROC curve. It shows a curve representing two limitations.

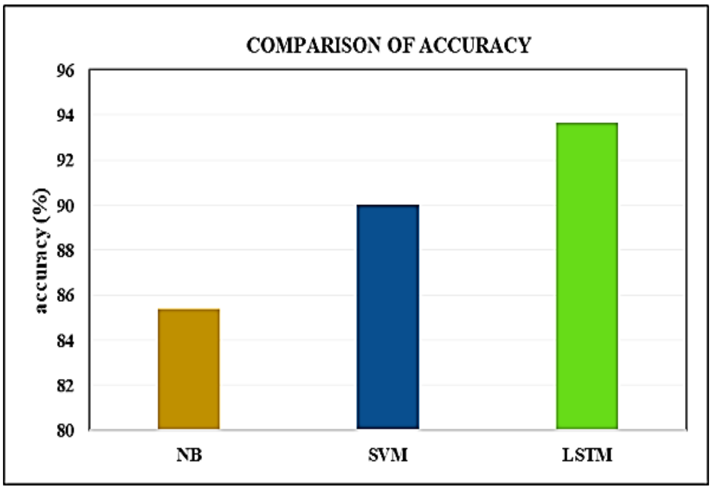

Figure 10: *Comparison of Accuracy Graph*

The following graph 10 explains about the difference between the existing proposed part in which it gives best accuracy of 93.6% when compared to the NB, SVM.

International Journal of Advanced Trends in Engineering and Management (IJATEM)
Vol. 02 (10), October 2023, pp. 24-35

Article Title: **Efficient Heart Disease Prediction Using PSO Optimized LSTM Classifier**

5 Conclusion

A deep learning method for predicting heart disease is proposed in this study. The initial process of preprocessing performed for achieving optimal data for further operation. The implementation of K-means clustering segments the pre-processed output and finally classification of heart disease is done using LSTM classifier. The performance of classifier is validated and the outcomes obtained includes F1 score of 93.84%, precision value of 91.67, recall measure of 96.12% and accuracy of 93.66% respectively. The outcomes foreshows that that proposed LSTM classification algorithm work effectively. According to the findings, deep learning represents an additional avenue for identifying potential future disruption-causing threats. As a future work, to increase accuracy even more, a more efficient method will be investigated.

Reference

1. Liaqat Ali; Atiqur Rahman; Aurangzeb Khan; Mingyi Zhou; Ashir Javeed; Javed Ali Khan, Year: 2019, "An automated diagnostic system for heart disease prediction based on ${\chi^{2}} $ statistical model and optimally configured deep neural network", IEEE Access, Vol: 7, pp. 34938 – 34945.

2. Fatima Zohra; Ali Javed; Hussain Dawood, Year: 2020, "Survival Rate Prediction Model of Cardio Vascular Disease Patients by Quantifying the Risk Profile using SVM", In 2020 2nd International Conference on Computer and Information Sciences (ICCIS), pp. 1 – 4. IEEE.

3. Cheng-Jun Zhang; Denghui Xia; Chao Zheng; Jianyong Wei; Yu Cui; Yanzhen Qu; Fangzhou Liao, Year: 2020, "Automatic identification of coronary arteries in coronary computed tomographic angiography", IEEE Access, Vol: 8, pp. 65566 – 65572.

4. P. Ramprakash; R. Sarumathi; R. Mowriya; S. Nithyavishnupriya, Year: 2020, "Heart disease prediction using deep neural network", In 2020 International Conference on Inventive Computation Technologies (ICICT), pp. 666 – 670. IEEE.

5. Abid Ishaq; Saima Sadiq; Muhammad Umer; Saleem Ullah; Seyedali Mirjalili; Vaibhav Rupapara; Michele Nappi, Year: 2021, "Improving the Prediction of Heart Failure Patients' Survival Using SMOTE and Effective Data Mining Techniques", in IEEE Access, Vol: 9, pp. 39707 – 39716, 2021.

6. Ayesha Shahnaz; Usman Qamar; Ayesha Khalid, Year: 2019, "Using blockchain for electronic health records", IEEE access, Vol: 7, pp. 147782 – 147795.

7. Umarani Nagavelli; Debabrata Samanta; Partha Chakraborty, Year: 2022, "Machine learning technology-based heart disease detection models", Journal of Healthcare Engineering.

International Journal of Advanced Trends in Engineering and Management (IJATEM)
Vol. 02 (10), October 2023, pp. 24-35

8. Amita Malav; Kalyani Kadam; Pooja Kamat, Year: 2017, "Prediction of heart disease using k-means and artificial neural network as hybrid approach to improve accuracy", International Journal of Engineering and Technology, Vol: 9, no: 4, pp. 3081 – 3085.

9. Indu Yekkala; Sunanda Dixit; M. A. Jabbar, Year: 2017, "Prediction of heart disease using ensemble learning and Particle Swarm Optimization", In 2017 International Conference on Smart Technologies for Smart Nation (SmartTechCon), pp. 691 – 698. IEEE.

10. Kuang Junwei; Hangzhou Yang; Liu Junjiang; Yan Zhijun, Year: 2019, "Dynamic prediction of cardiovascular disease using improved LSTM", International Journal of Crowd Science, Vol: 3, no: 1, pp. 14 – 25.

11. Chunyan Guo; Jiabing Zhang; Yang Liu; Yaying Xie; Zhiqiang Han; Jianshe Yu, Year: 2020, "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.

12. Rong Tao; Shulin Zhang; Xiao Huang; Minfang Tao; Jian Ma; Shixin Ma; Chaoxiang Zhang; Tongxin Zhang; Fakuan Tang; Jianping Lu; Chenxing Shen, Year: 2019, "Magnetocardiography-Based Ischemic Heart Disease Detection and Localization Using Machine Learning Methods", in IEEE Access, Vol: 66, pp. 1658 – 1667.

13. Martin Gjoreski; Anton Gradišek; Borut Budna; Matjaž Gams; Gregor Poglajen, Year: 2020, "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.

14. Mohammad Ayoub Khan; Fahad Algarni, Year: 2020, "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.

15. Norma Latif Fitriyani; Muhammad Syafrudin; Ganjar Alfian; Jongtae Rhee, Year: 2020, "HDPM: An Effective Heart Disease Prediction Model for a Clinical Decision Support System", in IEEE Access, Vol: 8, pp. 133034 – 133050.