Neural Computing and Applications (2023) 35:7253–7266
https://doi.org/10.1007/s00521-022-07064-0

S.I. : NEURO, FUZZY AND THEIR HYBRIDIZATION

An automatic heart disease prediction using cluster-based bi-directional LSTM (C-BiLSTM) algorithm

P. Dileep[1] · Kunjam Nageswara Rao[2] · Prajna Bodapati[2] · Sitaratnam Gokuruboyina[3] · Revathy Peddi[4] · Amit Grover[5] · Anu Sheetal[6]

Received: 13 January 2021 / Accepted: 5 February 2022 / Published online: 22 March 2022

Abstract
Heart disease involves many diseases like block blood vessels, heart attack, chest pain or stroke. Heart disease will affect the muscles, valves or heart rate, and bypass surgery or coronary artery surgery will be used to treat these problems. In this paper, UCI heart disease dataset and real time dataset are used to test the deep learning techniques which are compared with the traditional methods. To improve the accuracy of the traditional methods, cluster-based bi-directional long-short term memory (C-BiLSTM) has been proposed. The UCI and real time heart disease dataset are used for experimental results, and both the datasets are used as inputs through the K-Means clustering algorithm for the removal of duplicate data, and then, the heart disease has been predicted using C-BiLSTM approach. The conventional classifier methods such as Regression Tree, SVM, Logistic Regression, KNN, Gated Recurrent Unit and Ensemble are compared with C-BiLSTM, and the efficiency of the system is demonstrated in terms of accuracy, sensitivity and F1 score. The results show that the C-BiLSTM proves to be the best with 94.78% accuracy of UCI dataset and 92.84% of real time dataset compared to the six conventional methods for providing better prediction of heart disease.

Keywords Heart disease · K- means clustering · BiLSTM · Prediction · Performance metrics

1 Introduction

Heart disease (HD) has been an emerging problem of human beings, and it has been considered as the major disease globally, and the census shows that many people die annually due to this disease. Consequently, an effective diagnosis model is necessary for the patients who are suffering with heart disease, and much attention is required to provide basic care [1]. Many research works have been carried out so far to find the risk factors that lead to heart

✉ P. Dileep
p.dileep@mrcet.ac.in

Kunjam Nageswara Rao
kunjamnag@gmail.com

Prajna Bodapati
prajna.mail@gmail.com

Sitaratnam Gokuruboyina
myphdreserach@outlook.com

Revathy Peddi
revathy5813@gmail.com

Amit Grover
amitgrover321@gmail.com

Anu Sheetal
anusheetal2013@gmail.com

[1] Department of Computer Science and Engineering, Malla Reddy College of Engineering and Technology, Hyderabad, Telangana 500100, India

[2] Department of CS and SE, AUCE(A) at Andhra University, Visakhapatnam, Andhra Pradesh, India

[3] Department of CSE, LENDI Institute of Engineering and Technology, Vizianagaram, Andhra Pradesh, India

[4] Department of CSE, ACE Engineering College, Hyderabad, Telangana, India

[5] Department of Electronics and Communication Engineering, Shaheed Bhagat Singh State University, Ferozepur, Punjab, India

[6] Department of Engineering and Technology, Guru Nanak Dev University, Regional Campus, Gurdaspur, Punjab, India

disease. Providing a promising solution for identifying the heart risk factors is still a research challenge. The main risk factors identified in developing heart diseases are high blood cholesterol with hypertension and diabetes. Other risk factors related to heart disease are life style factors such as eating, drinking alcohol, smoking, obesity, etc.[1]. The disease prediction techniques are commonly used to find the relationship among various features related to the disease and the hidden information inherent in data [2].

Various machine learning and deep learning algorithms are carried out to achieve significant performance in the prediction of heart disease. ML is a method in which large data are divided into subset data, and subsequently, the patterns in the data are discovered as well as used for decision making [3]. The state of the art ML methodologies have been utilized to determine and measure the severity of heart failure and to predict the adverse events that are performed for the detection of heart failure [4–7]. The existing classifiers of regression tree (RT), SVM, Logistic Regression (Logistic), KNN, Gated Recurrent Unit (GRU), CNN, Ensemble (EL), DBN and DNN are used to predict the heart disease.

Convolutional neural networks (CNN) classification schemes have gained its popularity, due to its power to discover mid and high level image representations [8]. Convolution Neural Network receives 91 feature maps and 2 convolution training levels, and each level consists of 4 fully connected levels as well as each level comprises 1024 hidden units. The performance of the systems is evaluated with different numbers of convolutional layers. Then, the same numbers of map features are used to increase the build level from 2 to 5 in order to observe the effect of the build level. The level of fully connected does not change. The model has evaluated the Cleveland dataset [9, 10]. The model has 13 clinical features derived from the established UCI data as input. The proposed model is trained using an enhanced back propagation algorithm. The best performance of the three convolutional layers is 95% in terms of accuracy.

Deep learning is used to solve complex problems in the real world. Deep neural network classification algorithm has been utilized to determine the exact heart disease of patients. In this article, a framework with a systematic approach for the prediction of heart disease using deep neural networks (DNNs) has been proposed [11]. A layered method with many hidden convolutional layers as well a database method that uses a large amount of data in a deep neural network (DNN) has been utilized. Multiple hidden convolutional layers are also used to simplify the learning process. Then, this model is used to predict the heart disease. Empirical results show that the suggested DNN has achieved 91.246% accuracy, and it is better than the traditional NN model and group model.

Deep belief network (DBN) is one of the proficient classification algorithms, and it utilizes deep learning approach. The deep belief network is a graphical model that learns how to obtain detailed hierarchical representations of training data and to generate hidden variables on multiple levels. In the graphic model, there are connections between levels, rather than connections between units in each level [12]. The proposed architecture has two levels of structures and three levels of complete connectivity. The proposed architecture can avoid overfitting. The last three layers are fully connected, and the properties obtained after the convolutional layer are classified. The proposed study uses 16 medical characteristics that are important for predicting heart disease. As a result, by using DBN Sorter, the heart disease prediction rate achieved is 90% [13]. Prediction accuracy is an important issues in conventional algorithms. To overcome this issue, the proposed heart disease prediction system model introduces in-depth analyses methods in the data mining process to predict the heart disease incidence by using enhanced cluster-based BiLSTM classification algorithm.

1.1 Author contribution

- A cluster based on neural network models has been introduced for classifying different types of heart diseases. The main objective of this research is to enhance the accuracy prediction of an unbalanced heart disease dataset.
- Furthermore, in recent years, many studies have restricted the use of feature selection methods for the model. Therefore, the proposed cluster-based model that works on the mechanism of parallel computation allows using all the features without any restriction of feature selection method.
- This model has been used to realize automatic recognition and for the extraction of entities in unstructured medical texts, a model by combining cluster and Bi-directional, has been proposed.
- The advantage of BiLSTM method is that it analyzes the data bi-directionally and also analyzes the linear relationship between the characteristics. These processes help to increase the prognostic performance of heart disease.
- BiLSTM classifier is applied to estimate the performance of the system. When the classification results are compared with theexisting results. The better improvement is shown. Our experimental outcomes by using a real-time data set exhibit an improved diagnosis prediction performance strategy.

In heart diseases prediction, the proposed cluster-based BiLSTM renders more accuracy to detect heart disease.

The rest of the paper has been organized as follows: Literature survey is presented in Sect. 2, and the proposed method is explained in Sect. 3. The results are discussed in Sect. 4, and the conclusion is provided in Sect. 5.

1.2 Literature survey

Krishnan et al. [14] proposed a new hybrid deep learning model for heart disease prediction using recurrent neural networks (RNNs) in combination with a multiple gated recurrent units (GRU), long short-term memory (LSTM) and Adam Optimizer. The proposed model showed excellent accuracy of 98.6876%, which is the highest among the conventional RNN models.

Nguyen et al. [15] proposed high-performance architecture for optimizing decision making. Since reliability and stability are paramount factors in medical support systems, improving predictive performance and maintaining model stability are always top priorities. Experimental results show that the proposed approach achieves more powerful and reliable performance than existing ML and DL techniques in statistical, image based and sequential benchmark datasets.

Baccouche et al. [16] proposed an ensemble learning frameworks for various neural network models and how to aggregate random undersampling. Perform a data preprocessing step by characteristic selection to improve the performance of the classification algorithm. Experiments with unidirectional and bi-directional neural network models show that ensemble classifiers using BiGRU models using CNN models have the highest accuracy and F1 score of 91–96% for different types of hearts.

Shorewala et al. [17] proposed the prediction of coronary heart disease (CHD) using a risk factor approach. Predictive techniques such as K-nearest neighbors, Binary Logistic classification and Naive Bayes are evaluated on the basis of metrics such as accuracy, recall and ROC curves. These base classifiers are compared against ensemble modeling techniques such as bagging, boosting, and stacking.

Zhang et al. [18] proposed a heart disease prediction algorithm that is combining Embedded feature selection with deep neural networks. This built-in feature selection algorithm is based on the linear SVC algorithm and uses the L1 norm as a penalty item to select a subset of features that are highly relevant to heart disease. Network weights are initialized with the He initializer to prevent gradient varnishing and explosions, allowing predictors to get better performance.

Krishnan et al. [19] proposed a model aims to run a deep learning advanced recurrent neural network (RNN) model to improve the accuracy of existing predictive models that must be 98.23% or higher. In this paper, we discussed the deep learning method, derive a performance comparison between existing systems, and proposed an improved RNN model for better provision in terms of accuracy and feasibility. The presence of a multi-gated recurrent unit has improvised the performance of RNN models with an accuracy of 98.4%.

Javid et al. [20] used the UCI heart disease dataset tests ML techniques with traditional methods (random forests, support vector machines, K-nearest neighbors, etc.) and deep learning models (long-term, short-term memory, gate repeat unit neural networks, etc.). Combine multiple classifiers to explore voting-based models to improve the accuracy of weak algorithms. Interim persuasive approaches are being used to coordinate the way that ensemble techniques are implemented to improve the accuracy of heart disease prediction.

Zhou et al. [21] have suggested a technique to develop neural systems and genetic algorithms in order to maximize the weights as well as to improve ANN runtime performance. Distinguished crucial hazard factors are applied by the system for the forecasting of heart illnesses, and it is not expensive in clinical analyses. The system has been designed with genetic algorithm optimization, and it is applied to neural network which works better compared to back propagation.

Luo et al. [22] have analyzed a heart disease prediction system which has applied deep neural network algorithm to determine the likely options of heart associated diseases. Deep belief network, which is one of the proficient classification algorithms, utilizes deep learning approach in DNN, and this method renders more accuracy compared to CNN algorithm.

For faster training speed, a full sequence has been prepared by Zhou et al. [23] for various DNN structures for better execution performance, and it also conceives a fractional parallel execution of sequence for applying numerous Graphical Processing Units (GPUs). It is demonstrated that series training could be effortlessly covered with various DNNs by somewhat aligning fault signals in the output layer.

In deep learning, an approach has been reported by Gonging Wang et al. [18] to segment right ventricular (RV) endocardium and epicardium. This method possesses two tasks like limiting the Region of Interest (ROI), the area comprises significant features, and also, it will mitigate RV segmentation.

The two tasks are coordinated into a training system, and each task is resolved via two multilayer convolutional neural systems. A full automatic prediction heart disease has been discovered by Ali et al. [24] in 3-dimensional echocardio graphy (3DE). For efficient volume estimation, left ventricular volume evaluation is performed by applying regression and random forests. A Left Ventricular (LV)

volume prediction method without segmentation has been structured by deep learning technology using large scale cardiac magnetic resonance imaging (CMRI) datasets.

Yogita Solanki et al. [25] have evaluated a system in which blood alcohol content (BAC) detection has been carried out using pixel wise patch-based steps. To evaluate the system performance, ground truth samples are used, and they are compared with 840 mammograms from 210 cases. The deep learning approach has a huge impact on human experts, and the models can be used in similar fields to detect BAC on mammograms to differentiate cardiovascular risks.

Nandhini et al. [26] have proposed automatic diagnostic system which is used with χ^2 statistical models to refine the characteristics and to classify heart disease. The proposed model has been assessed using six different assessment indicators including accuracy, sensitivity, specificity, MCC, AUC and ROC graphs. This work has predicted the heart disease in real time, and it can be used for coronary artery disease. Unlike many other systems, it is ready for monitoring and forecasting. The system's diagnostic system is ready to use ML algorithms to predict disease center and then, the prediction results are supported by the disease center data set. The experiment has been excluded from the control study, and hence, the accuracy of the proposed system with the random forest classifier has reached 89%.

Ramprakash et al. [27] have demonstrated various issues associated with healthcare and different machine learning algorithms. The target of this research is to form a model for heart condition detection, and it deals with nonlinear features, an enormous amount of knowledge and supplying more accurate results than the existing researches. This deep learning with keras package has increased the speed of prediction. This model performs better than the prevailing approaches with 85.72% precision, 88.24% recall, 12.1% loss and 89.15% accuracy.

Sumit et al. [28] have organized a logical order which generally follows the methodological section. Compared to the alternative algorithms and developments, this approach tests the results of intelligent forecasts. Talos development lies in DNN's new development technique. Talos offers higher accuracy of (90.76%) for alternative optimizations.

Repaka et al. [29] have focused on the definition of SHDP (Smart cardiovascular disease prediction), which takes into account the NB classification approach as well as the advanced encryption standard algorithm to solve the prediction problem. It is found that the dominant technique outperforms Naive Bayes in relevance accuracy, as it provides 89.77% accuracy despite the decrease in attributes. Mohan et al. [30] have implicated hybrid random forest linear model, to increase the accuracy of heart disease prediction. Many disease characteristics have been used in standard classifiers for live prediction performance.

1.2.1 Dataset features information

The UCI Cleveland repository [31] with cardiac data set has been used in the experiment. Based on the extensive experimentation, 14 out of the 76 most effective properties have been found. The Cleveland dataset consists of 14 most dominant attributes and 303 samples with 8 absolute functions and 6 numeric functions. Table 1 shows the description of the data set.

Patients selected from this data set are between 29 and 77-years-old. A value of 0 is used to represent a female patient, and a value of 1 is used to represent a male patient. There are three types of chest pains that can indicate heart disease. Typical angina type 1 is due to blockage of the heart arteries caused by decreased blood flow to the heart muscle. The main cause of type 1 angina is mental or emotional stress. The second type occurs for a number of reasons, but sometimes, it can be an angina rather than the actual cause of heart disease called chest pain. The next resource is trestbps, which represent dormant blood pressure readings. Cholesterol levels are expressed in Chol. Fasting blood sugar levels are expressed in Fbs. If Fbs is greater than 120 mg/dL, a value of 0 is assigned, and a value of 1 indicates whether Fbs is less than 120 mg/dL. The results of the dormant electrocardiogram are expressed as Restecg.

The maximum heart rate is represented by dropout, the preferred exercise for angina pectoris 0 is no pain, 1 is represented by exang, and ST depression stimulated by exercise is indicated by oldpeak, maximum exercise slope. The ST segment is plotted as a slope, and the number of main vessels colored by fluoroscopy is plotted as an approximation. The stress test period is displayed like this, and the final goal is class properties. The value of the class attribute is used to differentiate between patients with heart disease and patients without heart disease. A value of 1 indicates patients with heart disease, and a value of 0 indicates normal.

1.3 Proposed cluster-based BilLSTM

Although many machine learning methods have been used to predict heart disease, these existing methods are still ineffective. The aim of the present study is to improve the effectiveness of heart disease prognosis with the help of the proposed BiLSTM. The BiLSTM method analyzes data in two directions, and it generally improves predictive performance. These techniques are used to ensure a linear distribution of the starting structure, allowing LSTM to improve accuracy. LSTM is a deeplearning method that effectively analyzes data and discovers key features required for prediction.

Table 1 Dataset description

No	Attributes	Descriptions	Value range
1	Age	Patients age (in Year)	29–79
2	Sex	0: female and 1: male	0 and 1
3	Cp	Type of chest pain Type 0: Typical Angina Type 1: Atypical angina Type 2: Non-anginal pain Type 3: Asymptomatic	0–3
4	Trestbps	Resting blood sugar (in mm Hg on admission to the hospital)	94–200
5	Chol	Serum cholesterols in mg/dl	126–564
6	Fbs	Fasting blood suger > 120 mg/dl (1 = true; 0 = false)	0 and 1
7	Restecg	Resting ECG result	0–2
8	Thalach	Maximum heart rates achieved	71–202
9	Exang	Exercise induced angina	0,1
10	Oldpeak	ST depression induced by exercise relative to rest	1–3
11	Slope	Slope or peak exercise ST Segment Value 1: Upsloping Value 2: Flat Value 3: Downsloping	1–3
12	Ca	number of major vessels (0–3) colored by flourosop	0–3
13	Thal	3 = Normal; 6 = Fixed defect; 7 = Reversable defect	3,6,7
14	Target	The predicted attribute 0: Yes; 1: No	0 and 1

The architecture of the BiLSTM model is shown in Fig. 1. The model uses BIO labeling scheme (Begin, Inside and Outside). The $Q = (q_1, q_2, \ldots q_k)$ represents the context information carried out by the character embedding trained by Skip-Gram. HyC manifests as hypertension, and its temporal properties are continuous. Meanwhile, HyD manifests itself as hypertension and its temporal properties of data. First, the sentences appear as a series of vectors through the key layer $X = (x_1, x_2, \ldots x_t, \ldots x_n)$ where n is the length of the sentence.

A tanh layer on the top of BiLSTM is then used to predict the reliability score of the character for which all the possible tags are listed as network exit scores.

$$e_t = \tanh(W_e h_t)$$

where W_e = weight matrix of training model.

The BiLSTM learning model performs data partitioning by applying k-means clustering algorithm and trains each of the partitioned data. In a highly parallel manner, the advantage of suggested BiLSTM is that its training process hassmaller model size and LSTM is discovered from less training data. The C-BiLSTM model for heart disease prediction is shown in Fig. 2.

1.4 K-means clustering for data partitioning

A clustering algorithm is applied to cluster the objects which are more similar (in some sense or another) to each other than to those in other groups (clusters). There are broad ranges of clustering algorithms available, and they can be separated into three main groups: partitioning, hierarchical and density-based algorithms. In this work, a partitioning-based algorithm is applied, and it is often applied for partitioning large datasets into groups of similar data. Selection of number of clusters at the beginning is a major disadvantage of this algorithm, however; the manual selection of the number of clusters doesn't affect the performance of partitioning datasets.

1.4.1 Cluster centroid steps

Let the input data points be $A_x = \{A_1, A_2, ..A_z\}$, where $A_x|\ \mathrm{x} = (\{\mathrm{x} = 1, 2, \ldots X\}$ represents the X number of input data with z -dimensional vectors. The goal of k-means clustering method is to group the input data in to N clusters.

Thus, a subset with cluster centers, $G(G = G_1, G_2, \ldots G_n)$ is found for the A_x by reducing the objective function, $F = \sum_{p=1}^{q} \sum_{A_q \in G_p}^{q} m_{p,q}(A_q, G_p)$, where A_q represents the data with z-dimensional vector, and $m_{p,q}$ (A_q, G_p)depicts the distance among the data A_q and the cluster center G_p.

The distance between the data A_q and the cluster center G_p is calculated in Eq. (1) as,

$$m_{p,q}(A_q, G_p) = \left\|A_q - G_p\right\|^2 \tag{1}$$

Thus, the objective function can be written in Eq. (2) as,

$$F = \sum_{p=1}^{q} \sum_{A_q \in G_p}^{q} \left\|A_q - G_p\right\|^2 \tag{2}$$

The cluster centers can be determined when the values of the objective functions are known. The minimization of the objective function is calculated in Eqs. (3, 4),

$$\frac{\partial F}{\partial G_p} = \frac{\partial}{\partial G_p}\left[\sum_{q=1}^{z_p} \left(A_q - G_p\right)^2\right] \tag{3}$$

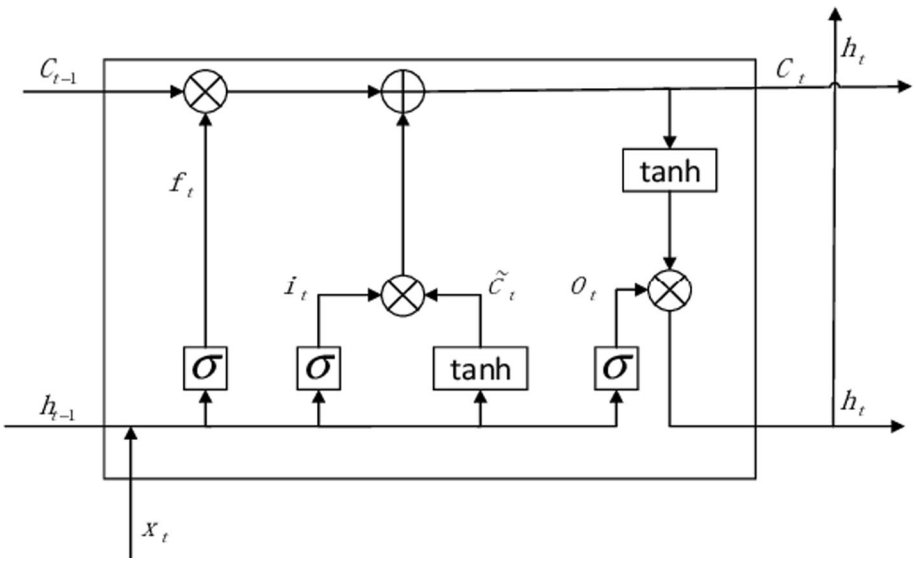

Fig. 1 The proposed cluster-based BiLSTM block diagram of heart disease prediction

$$\frac{\partial F}{\partial G_p} = -2\sum_{q=1}^{z_p}\left(A_q - G_p\right) = 0 \tag{4}$$

Now, the cluster centers can be found in Eq. (5) as,

$$G_p = \frac{1}{z_p}\left(\sum_{q=1}^{z_p} A_q, p = (1, 2, \ldots N)\right) \tag{5}$$

In the above equation, z_p represents the number of data points in cluster 'p'.

Major steps involved in k-means clustering are:

Step 1 Determine the number of groups "N", where "N" is the number of partitions to be created for the data set.
Step 2 Start with N cluster centroids by randomly separating all objects into *N* clusters by calculating the distance. This step is used to verify or to confirm that the centroids are different from each other.
Step 3 In this stage, calculating the distances of all clusters of centroids is processed and then based on the nearest centroid, each object to the cluster will be processed.
Step 4 Calculate the centrifuges of the improved clusters with the new updated cluster center as the average of the creation instances.
Step 5 Re-iterate Step 3 until the centroids do not alter any more.

1.5 Cluster-based BiLSTM

A neural network has been designed to be a bilinear LSTM (BiLSTM) with many levels of nodes between the input and the output along with a series of steps to perform the identification and processing of resources. However, the processing time required for BiLSTM training is high, and the cluster-based BiLSTM is used to increase the training speed.

The BiLSTM cluster-based learning model performs data partitioning with a k-means clustering algorithm and trains each partitioned data on multiple BiLSTM simultaneously. In a highly parallel way, the advantage of the suggested cluster-based BiLSTM is that its training process has smaller model dimensions, and each BiLSTM is discovered by less training data. The proposed model for the heart disease prediction method is shown in Fig. 2.

1.6 4 Block diagram

Classification of heart disease is the final section in which type of heart disease cluster-based BiLSTM classifier. One of the advantages of this method is that the learned features from raw training data are demonstrated automatically, and they are quite different from the traditional method as shown Fig. 3.

1.7 Cluster-based BiLSTM architecture

Cluster-based BiLSTM has been utilized in the model for predicting heart diseases, and at the same time, the approach aids to reduce the computation cost and time. The BiLSTM principle of operation recognizes the current output and applies it as an input to the next hidden layer. The number of tests available during the training phase is huge, but deep learning techniques are very effective. The proposed cluster-based BiLSTM has been developed and for every group of data, the cluster-based BiLSTM applies multiple BiLSTM structures. C-BiLSTM architecture is depicted in Fig. 4. Five function layers are included: Input layer, embedding layer, BiLSTM layer, attention layer and

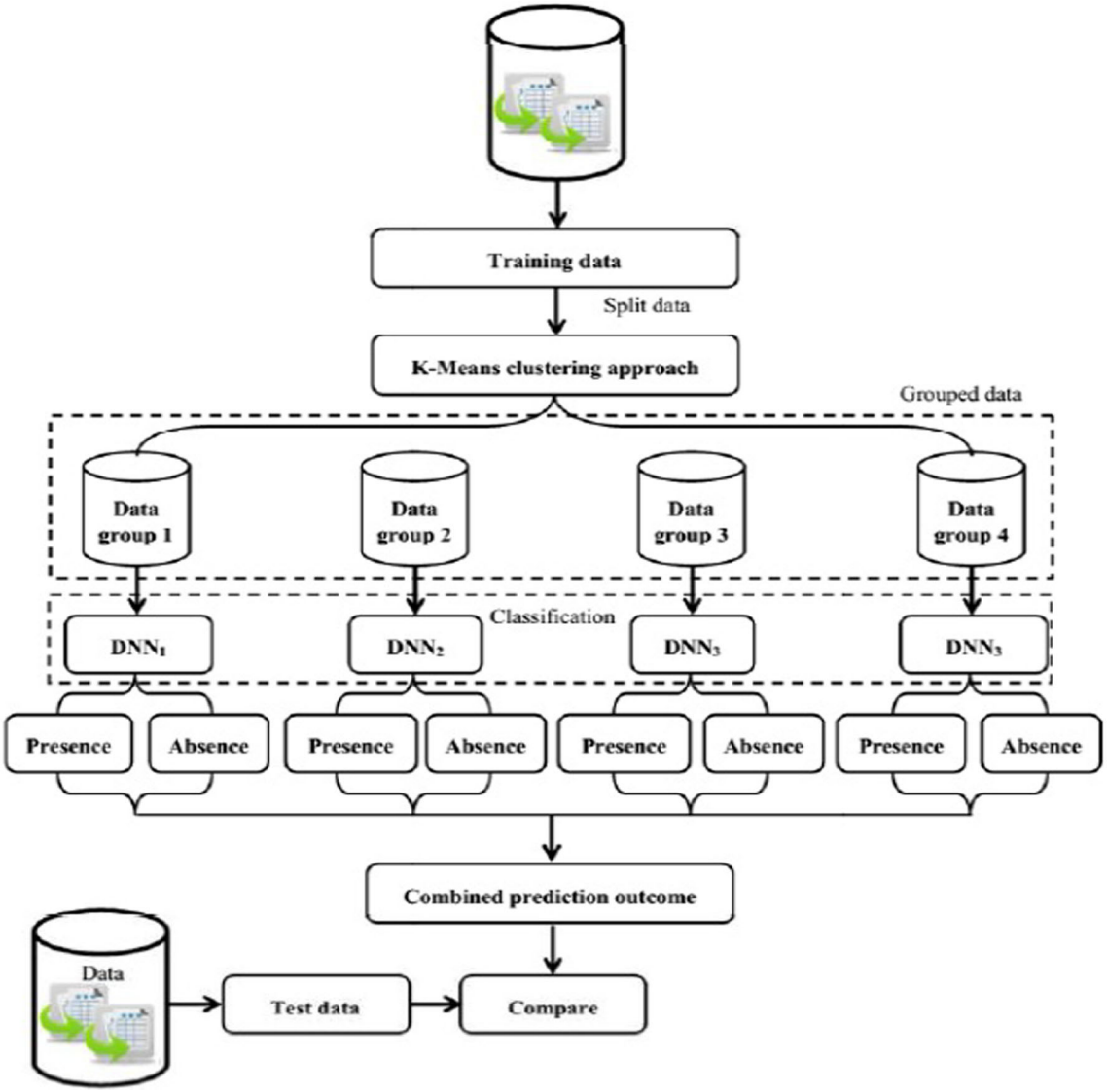

Fig. 2 C-BiLSTM heart disease prediction framework

output layer. In the forward propagation process, the output of each layer is used as the input of the next layer.

To evaluate the error of entire network, the cost function is applied. The minimization criterion of cost function helps in altering proper weights and biases of BiLSTM, during pre-training so as to decrease the error. At each iteration, the cost function is reduced and in which, the weights and biases are found to maximize the network output. Once the network gets trained, the trained BiLSTM can be applied to anticipate the heart diseases. Likewise, multiple BiLSTMs are trained with the cluster groups. At last, the network efficiency is formalized by applying the test data (Fig. 5).

The test data are also grouped into number of clusters as adequate to the number of groups in the training data for validation. Each group of test data is aligned to the trained BiLSTM, so that each BiLSTM network is able to predict heart diseases, and finally, the outcomes from each BiLSTM are aggregated.

2 Result and discussion

2.1 Data processing

The proposed algorithm has been modeled in MATLAB platform. Cleveland heart disease data from the california irvine (UCI) data warehouse have been used for analysis. Each segmented data are individually trained using k-means grouping. This dataset individually predicts the results based on the presence or absence of heart disease. Heart disease data collected in real time from Hospital

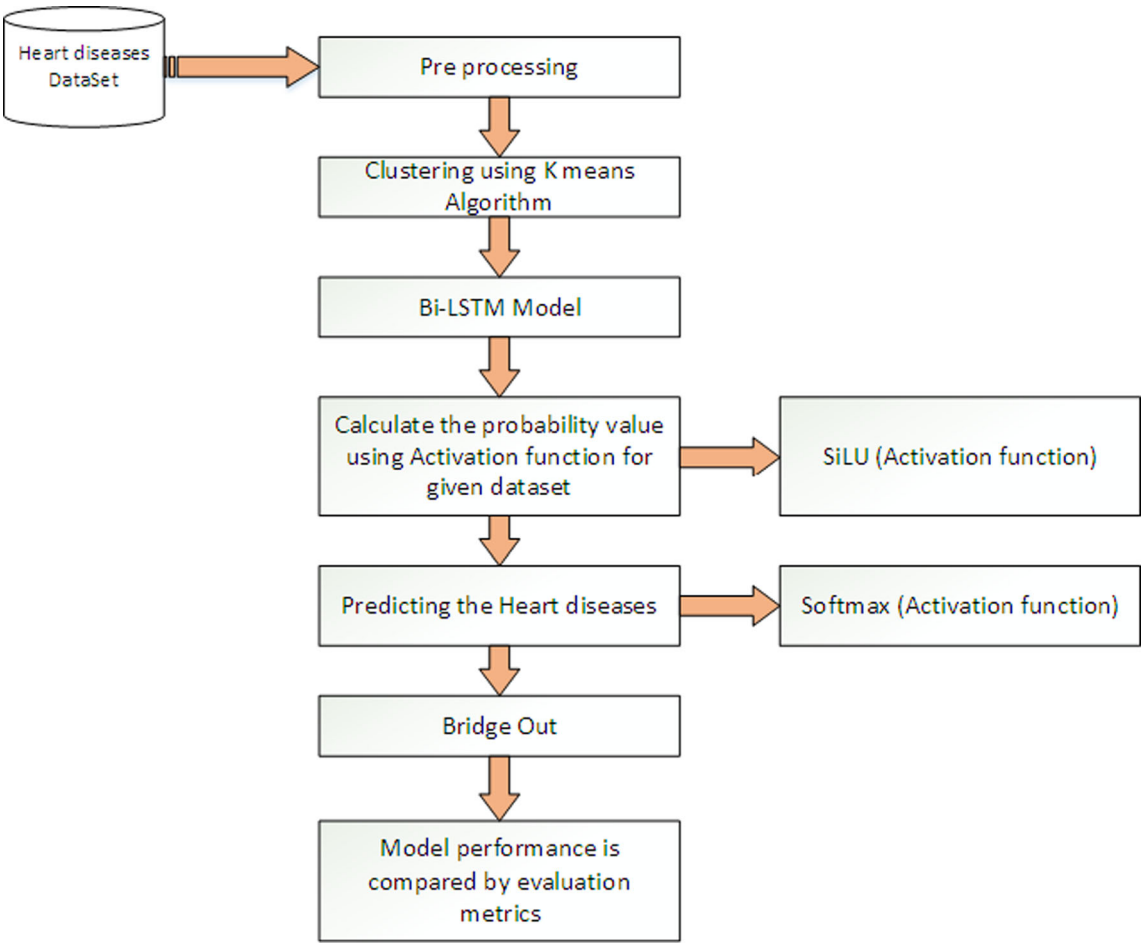

Fig. 3 Proposed Cluster-based BiLSTM

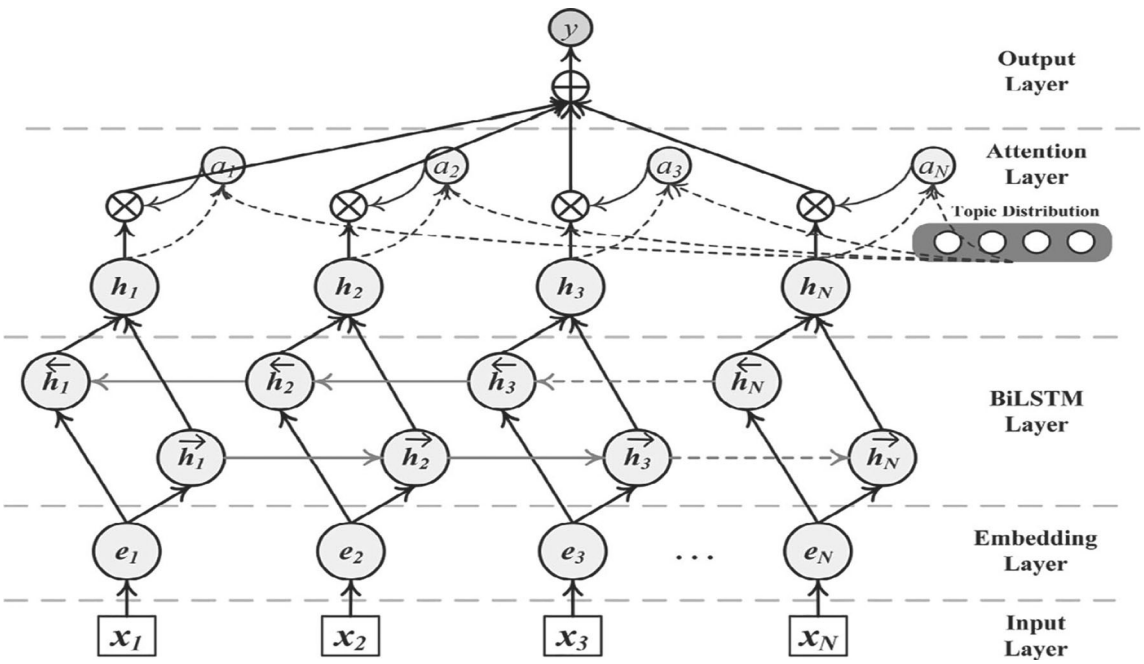

Fig. 4 Cluster-BiLSTM architecture

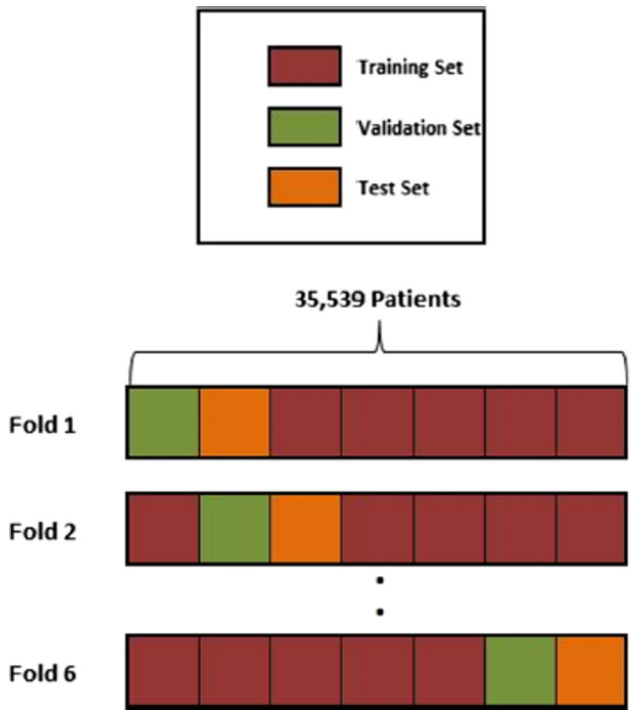

Fig. 5 Sixfold cross-validation

Medica Norte are retrieved, and they contain 300 patient records with various biological indicators including age, sex, systolic and diastolic blood pressure and heart rate. For this test, Python software has been used for simulation with an Intel i3 processor, 4 GB RAM and 1 TB hard drive.

2.2 Training method

All the developed models are trained with datasets generated from 300 patient record registers. That is, the data are divided into training dataset and testing dataset with a ratio of 6:1:1. The performance is reported using the test data for the area under the AUC curve. The Graphical User Interface (GUI) of the proposed approach is presented in Fig. 6. With the test sample provided as in Fig. 3, the model has predicted the correct output as 'abnormal'.

A heat map is also used to clearly analyze the correlation of all the features in Fig. 7.

Regarding the age distribution attribute, Fig. 8 illustrates people with and without heart disease. It can be seen that the ages are in between 40 and 52 years. It can be understood when the age is associated with heart disease, people aged from 50 to 52 and 40 to 41 are primarily consolidated by heart disease. Additionally, histograms are designed to display discrete property data that focus on the distribution of marginal features related to disease and not disease, as shown in Figs. from 9, 10, 11.

According to the age distribution characteristics, there are usually people with heart diseases and people without heart disease. The patient ages are varied from 40 to 52 years. It is found if the age is associated with disease, people aged from 50 to 52 and 40 to 41 suffer from heart disease and dominant associations. It is also designed to be age-adjacent to reflect the possibility of a highly correlated stopping characteristic (thalach). It is generally observed that people with heart disease have higher heart rates than people without heart disease. In addition, as age increases, the maximum heart rate decreases to a negative value of −0.3 as depicted in Fig. 12.

2.3 Evaluation metrics

The model has been evaluated with the evaluation metrics namely Accuracy, Prediction Time, Precision, Recall and F-measure.

2.3.1 Precision

Precision is the determination of true positives for the ratio of the sum of true positives to false positives.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{6}$$

2.3.2 Recall

Recall is defined as the true positive versus the false negative sum ratio.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{7}$$

2.3.3 Accuracy

Precision is calculated using sensitivity and specificity by the following equation,

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} * 100 \tag{8}$$

2.3.4 F-measure

The measure F is an assessment of test precision that is, the weighted harmonic mean of the test precision and recovery.

2.4 Performance analysis

The values obtained for the suggested model are tabulated in Tables 2 and 3

Table 2 and Fig. 13 depict the comparison of different performance metrics of the proposed (C-BiLSTM) approach and the existing classifiers of Regression Tree

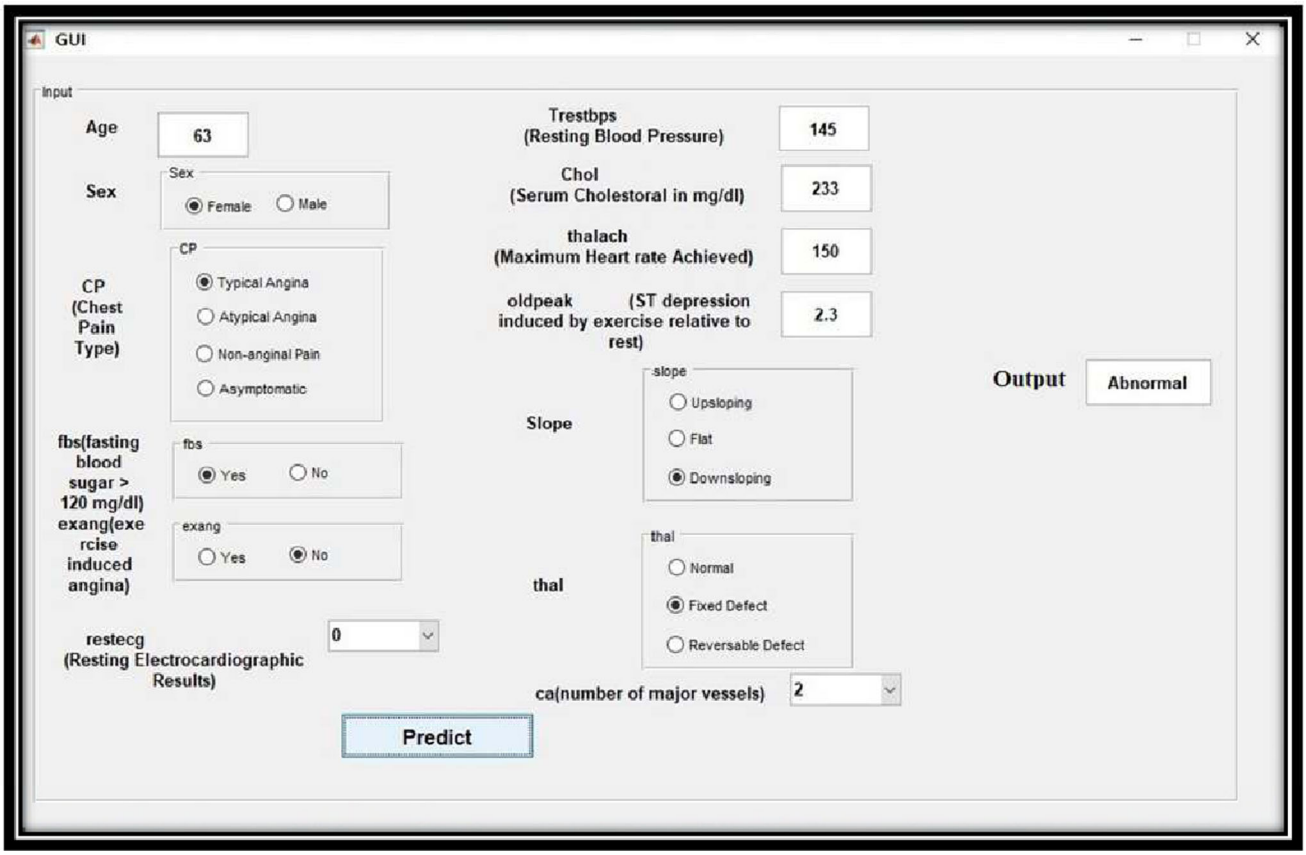

Fig. 6 GUI interface of heart disease prediction

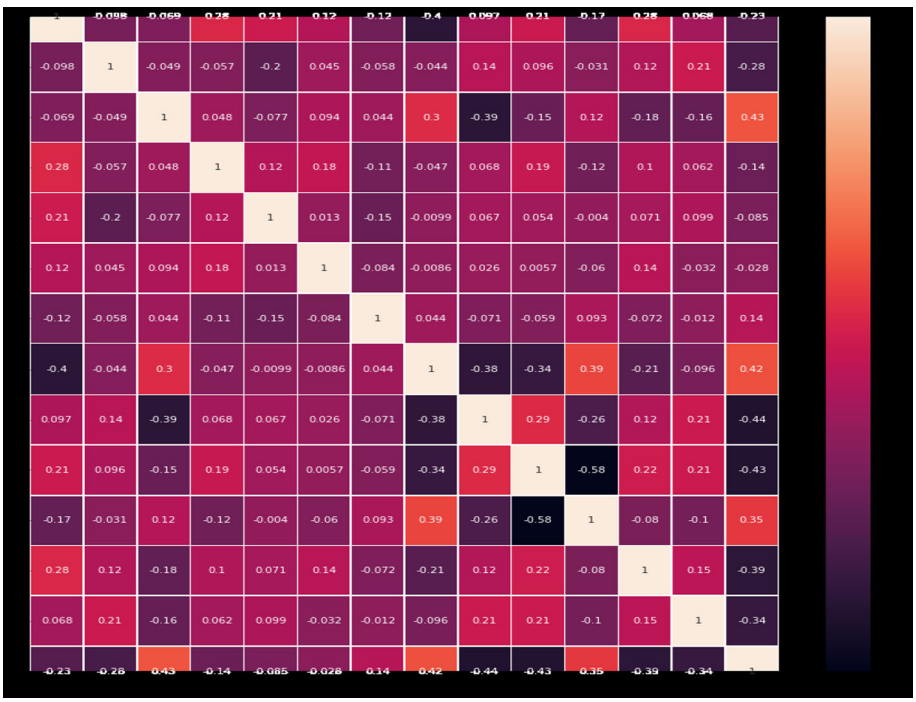

Fig. 7 Correlation-cross values heat map

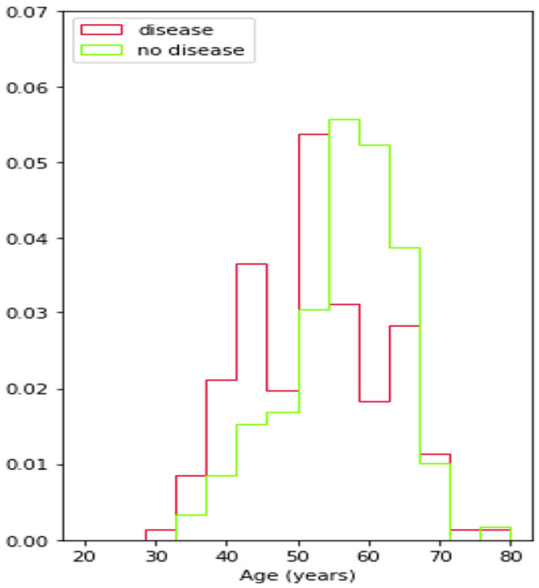

Fig. 8 Age distribution

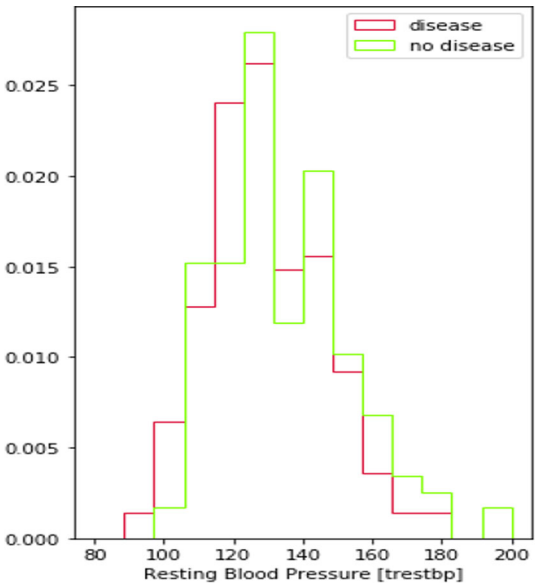

Fig. 10 Blood pressure distribution

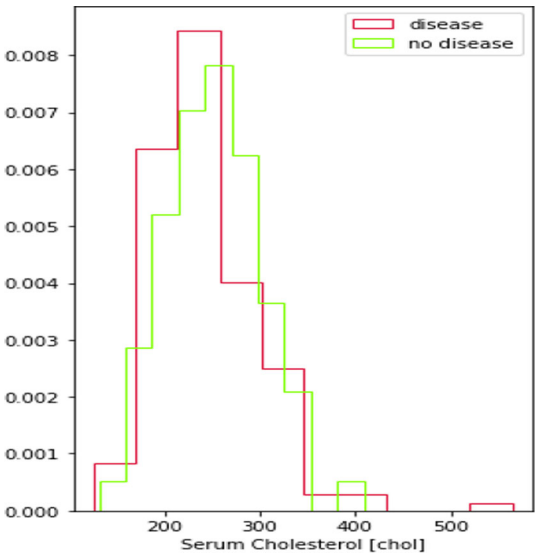

Fig. 9 Serum cholesterol distribution

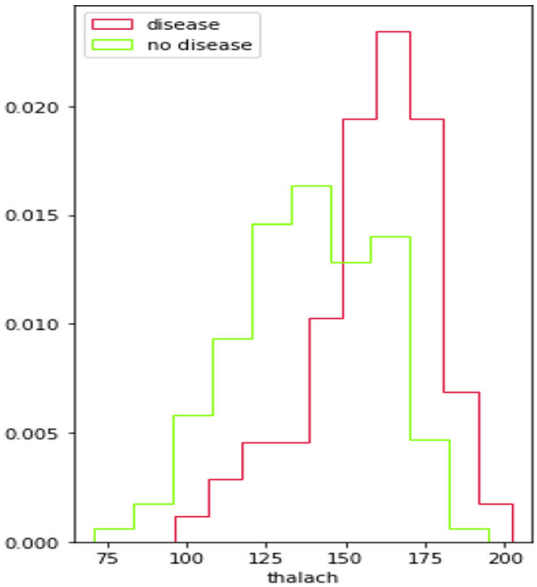

Fig. 11 Maximal heart rate acquire

(RT), SVM, Logistic Regression(Logistic), KNN, Gated Recurrent Unit(GRU), CNN, Ensemble(EL), DBN and DNN. The proposed method has obtained high classification accuracy of 94.78%, sensitivity of 99.56% and specificity of 94.18% compared to the existing classifiers. The accuracy of the existing classifier is 91.24% for DNN, 90% for DBN, 88.54% for Ensemble, 85% for CNN,81.46% for Gated Recurrent Unit, 82.80% for KNN, 81.38% for Logistic Regression, 78.79% for SVM and 75.40% for Regression Tree. From these analyses, it is evident that the proposed C-BiLSTM approach has greatly improved in the context of all performance metrics compared to other classifiers for the online UCI dataset. The F1-score value of the C-BiLSTM is 92.55% which is the highest value among all other conventional methods.

Table 3 and Fig. 14 illustrate the comparison of different performance metrics of the proposed (C-BiLSTM) approach and the existing classifiers of Regression Tree (RT), SVM, Logistic Regression(Logistic), KNN, Gated Recurrent Unit(GRU), CNN, Ensemble(EL), DBN and DNN. The proposed method has obtained high classification accuracy of 92.84%, sensitivity of 96.20%, specificity of 93.36% compared to the existing classifiers. The accuracy of the existing classifier is 90.90% for DNN, 90.17% for DBN, 90.21% for Ensemble, 85.90% for CNN, 85.34%

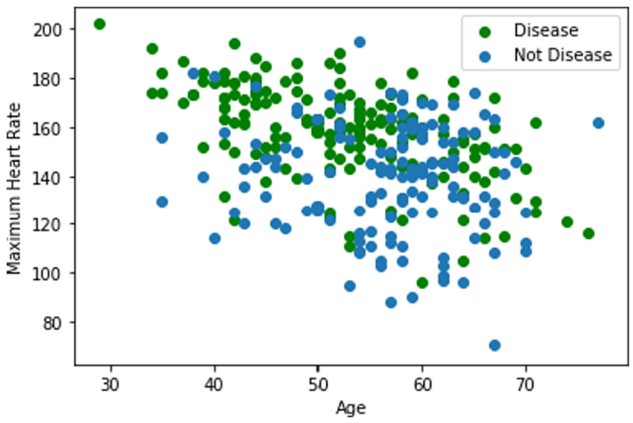

Fig. 12 Age vs. Thalach

for Gated Recurrent Unit, 89.23% for KNN, 86.74% for Logistic Regression, 85.17% for SVM and 81.73% for Regression Tree. From these analyses, it can be revealed that the proposed C-BiLSTM approach has greatly improved in terms of all performance metrics compared to other classifiers for the real time dataset. The F1-score value of the C-BiLSTM is 93.41% which is the highest value among all other conventional methods.

3 Discussion

The proposed cluster-based BiLSTM method has analyzed the data in two directional ways and provided the linear relationship between the attributes. Hence, the proposed method has attained higher performance compared to the existing methods. The target of this work is to classify four types of heart diseases. However, each type has very unbalanced distribution due to the large proportion of collected records by indicating that the patients have not been diagnosed with any type of disease. Subsequently, the proposed framework has demonstrated a technique of averaging and random under-balancing of the records, in order to equally generate proportional dataset for building the classification models.

Table 2 Experimental results of the proposed models of UCI data set

Classifier/performance metrics	Dataset	Accuracy (%)	Sensitivity (%)	Recall (%)	F1 (%)
Regression tree	Cleveland heart disease dataset	75.40	64.34	56.78	55.90
SVM		78.79	78.62	76.56	75.32
Logistic regression		81.38	86.36	80.85	81.68
KNN		82.80	88.26	82.34	82.32
Gated recurrent unit		81.46	84.43	84.45	84.82
CNN		85	86.90	86.12	87.34
Ensemble		88.54	89.67	92.78	90.91
DBN		90	91.21	91.78	92.14
DNN		91.24	91.67	91.91	92.07
C -BiLSTM		94.78	95.56	94.18	92.55

Table 3 Experimental results of the proposed models of Real time data set

Classifier/performance metrics	Dataset	Accuracy (%)	Sensitivity (%)	Recall (%)	F1 (%)
Regression tree	Real time heart disease dataset	81.73	90.54	81.85	81.76
SVM		85.17	92.91	85.24	85.18
Logistic regression		86.74	93.78	86.88	86.92
KNN		89.23	94.85	90.23	90.36
Gated recurrent unit		85.34	82.43	81.45	81.82
CNN		85.90	85.70	86.34	87.10
Ensemble		90.21	91.32	90.14	90.05
DBN		90.17	91.0	91.78	91.45
DNN		90.90	93.50	92.32	92.21
C—BiLSTM		92.84	96.20	93.36	93.41

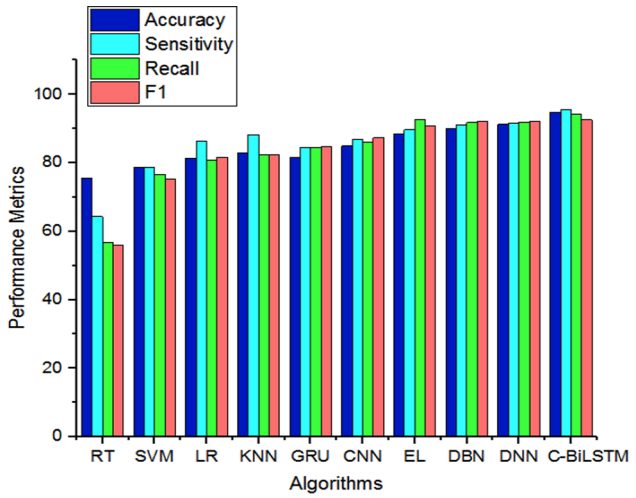

Fig. 13 Comparison of C-BiLSTM with the existing algorithm of UCI Dataset

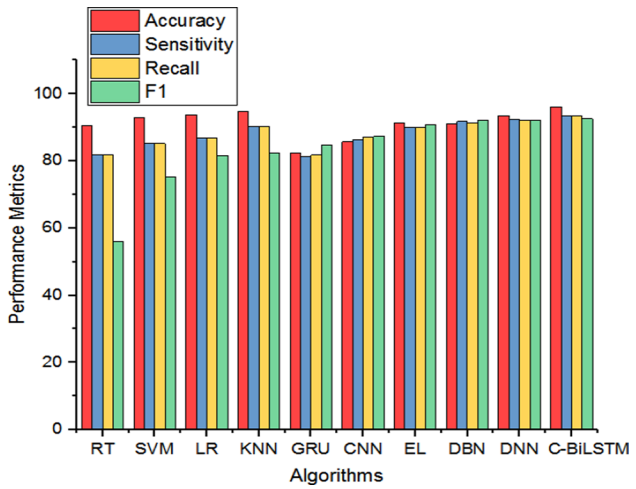

Fig. 14 Comparison of C-BiLSTM with existing algorithm of real time dataset

The proposed method has obtained high classification accuracy of 94.78%, sensitivity of 99.56% and specificity of 94.18% compared to the existing classifiers. The accuracy of the existing classifier is 91.24% for DNN, 90% for DBN, 88.54% for Ensemble, 85% for CNN, 81.46% for Gated Recurrent Unit, 82.80% for KNN, 81.38% for Logistic Regression, 78.79% for SVM and 75.40% for Regression Tree. From these analyses, it is realized that the proposed C-BiLSTM approach has greatly improved in view of all performance metrics compared to other classifiers for the online UCI dataset. The proposed method has achieved high classification accuracy of 92.84%, sensitivity of 96.20%, and specificity is 93.36% compared to the existing classifiers. The accuracy of the existing classifier is 90.90% for DNN, 90.17% for DBN, 90.21% for Ensemble, 85.90% for CNN, 85.34% for Gated Recurrent Unit, 89.23% for KNN, 86.74% for Logistic Regression, 85.17% for SVM and 81.73% for Regression Tree. From these analyses, it is proved that the proposed C-BiLSTM approach has greatly improved all performance metrics compared to other classifiers for the Real Time heart disease dataset.

4 Conclusion

In this paper, an effective method of classifying heart disease has been proposed by using cluster-based BiLSTM. The heart disease prediction experiment has been carried out through cluster-based BiLSTM algorithm by using structured data. The proposed algorithm is used to identify the risk of cardiovascular disease and its corresponding risk factors. The dataset has been chosen from online repositories and real time dataset. The techniques of preprocessing applied are filled in missing values and removing correlated columns. Next, the classifier is applied to the preprocessed dataset, and then, BiLSTM is constructed. Finally, the accuracy of the models is calculated, and the analyses are performed based on efficiency calculations. The proposed methods are compared with the conventional classifiers such as Regression tree, SVM, Logistic Regression, KNN, Gated Recurrent Unit, CNN, Ensemble, DBN and DNN. High classification accuracy obtained through the proposed method is 94.78%, sensitivity is 99.56%, and specificity is 94.18% for online UCI dataset whereas the accuracy is 94.78% for real time dataset. In future, a further study will be carried out with other feature selection and optimization techniques to improve the effectiveness of predictive classifiers for the diagnosis of cardiovascular disease.

Acknowledgements The authors would like to thank the editors and the reviewers for their insightful comments and suggestions.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Human and animal rights There is no animal involvement in this research.

Informed consent The authors declare that they have no consent.

References

1. Mozaffarian D, Wilson PW, Kannel WB (2008) Beyond established and novel risk factors: lifestyle risk factors for cardiovascular disease. Circulation 117(23):3031–3038

2. Poirier P (2008) Healthy lifestyle: even if you are doing everything right, extra weight carries an excess risk of acute coronary events. Circulation 117:3057–3059
3. Murphy KP (2012) Machine learning: a probabilistic perspective. MIT press, Cambridge
4. Kanimozhi VA, Karthikeyan T (2016) A survey on machine learning algorithms in data mining for prediction of heart disease. Int J Adv Res Compu Commun Eng 5(4):552–557
5. Tripoliti EE, Papadopoulos TG, Karanasiou GS, Naka KK, Fotiadis DI (2017) Heart failure: diagnosis, severity estimation and prediction of adverse events through machine learning techniques. Comput Struct Biotechnol J 15:26–47
6. Anbarasi M, Anupriya E, Iyengar NCHSN (2010) Enhanced prediction of heart disease with feature subset selection using genetic algorithm. Int J Eng Sci Technol 2(10):5370–5376
7. Florence S, BhuvaneswariAmma NG, Annapoorani G, Malathi K (2014) Predicting the risk of heart attacks using neural network and decision tree. Int J Innov Res Comput Commun Eng 2(11):2320–9798
8. HD Masethe, MA Masethe (2014) Prediction of heart disease using classification algorithms. In: Proceedings of the world congress on engineering and computer Science 2:22-24
9. Zhou P, Liu C, Liu Q, Dai L, Jiang H, (2013) A cluster-based multiple deep neural networks method for large vocabulary continuous speech recognition. In: IEEE international conference on acoustics, speech and signal processing, pp. 6650–6654
10. Lee J-G, Jun S, Young-Won Cho MS, Lee H, Kim GB, Joon-BeomSeo MD, Kim N (2017) Deep learning in medical imaging: general overview. Korean J Radiol 18(4):570–584
11. Shankar V, Kumar V, Devagade U (2020) Heart disease prediction using CNN algorithm. SN COMPUT SCI 1:170. https://doi.org/10.1007/s42979-020-0097-6
12. Karthikeyan T, Kanimozhi VA (2017) Deep learning approach for prediction of heart disease using data mining classification algorithm deep belief network. Int J Adv Res Sci Eng Technol 4(1):3194–3201
13. Dileep P, Rao KN (2019) An efficient feature selection based heart disease prediction model. Int J Adv Sci Technol 28(9):309–323
14. Krishnan S, Magalingam P, Ibrahim R (2021) Hybrid deep learning model using recurrent neural network and gated recurrent unit for heart disease prediction. Int J Electr Comput Eng 11(6):5467–5476
15. Nguyen DK, Lan CH, Chan CL (2021) Deep ensemble learning approaches in healthcare to enhance the prediction and diagnosing performance: the workflows, deployments, and surveys on the statistical, image-based, and sequential datasets. Int J Environ Res Public Health 18(20):10811. https://doi.org/10.3390/ijerph182010811
16. Baccouche A, Garcia-Zapirain B, Castillo Olea C, Elmaghraby A (2020) Ensemble deep learning models for heart disease classification: a case study from Mexico. Information 11:207. https://doi.org/10.3390/info11040207
17. Shorewala V (2021) Early detection of coronary heart disease using ensemble techniques. Inform Med Unlocked 26:100655. https://doi.org/10.1016/j.imu.2021.100655
18. Zhang D, Chen Y, Chen Y, Ye S, Cai W, Jiang J, Xu Y, Zheng G, Chen M (2021) Heart disease prediction based on the embedded feature selection method and deep neural network. J Healthc Eng 2021:6260022. https://doi.org/10.1155/2021/6260022
19. Krishnan S, Magalingam P, Ibrahim RB (2020) Advanced recurrent neural network with tensor flow for heart disease prediction. Int J Adv Sci Technol 29(5):966–977
20. Javid I, Alsaedi AKZ, Ghazali R (2020) Enhanced accuracy of heart disease prediction using machine learning and recurrent neural networks ensemble majority voting method. Int J Adv Comput Sci Appl (IJACSA) 11(3):540–551. https://doi.org/10.14569/ijacsa.2020.0110369
21. G Luo, R An, K Wang, S Dong H Zhang (2016) A deep learning network for right ventricle segmentation in short-axis MRI Computing in Cardiology Conference (CinC)
22. S Dong, G Luo, G Sun, K Wang H Zhang (2016) A combined multi-scale deep learning and random forests approach for direct left ventricular volumes estimation in 3D echocardiography.Computing in Cardiology Conference (CinC)
23. G Luo, G Sun, K Wang, S Dong H Zhang (2016) A novel left ventricular volumes prediction method based on deep learning network in cardiac MRI Computing in Cardiology Conference (CinC), Vancouver, BC, Canada
24. Wang J et al (2017) Detecting cardiovascular disease from mammograms with deep learning. IEEE Trans Med Imaging 5:1172–1181. https://doi.org/10.1109/TMI.2017.2655486
25. Ali L, Rahman A, Khan A, Zhou M, Javeed A, Khan JA (2019) An automated diagnostic system for heart disease prediction based on x^2 statistical model and optimally configured deep neural network. IEEE Access 7:34938–34945
26. Solanki Y, Sharma S (2019) Analysis and prediction of heart health using deep learning approach. Int J Comput Sci Eng 7(8):2347–2693
27. Nandhini S, Debnath M, Sharma A, Pushkar (2018) Heart disease prediction using machine learning. Int J Recent Eng Res Dev 3(10):39–46
28. Ramprakash P, Sarumathi R, Mowriya R, Nithyavishnupriya S (2020) Heart disease prediction using deep neural network. Int Conf Invent Comput Technol. https://doi.org/10.1109/ICICT48043.2020.9112443
29. Sharma S, Parmar M (2020) Heart diseases prediction using deep learning neural network model. Int J Innov Technol Explor Eng 9(3):2244–2248
30. AN Repaka, SD Ravikanti RG Franklin 2019 Design and implementing heart disease prediction using naives Bayesian. In: 3rd International conference on trends in electronics and informatics (ICOEI), Tirunelveli, India
31. Newman D, Hettich S, Blake C, Merz C (1998) UCI Repository of machine learning databases. http://www.ics.uci.edu/∼mlearn/MLRepository.html