Received: October 16, 2019. Revised: November 14, 2019. 31

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

A Prediction Technique for Heart Disease Based on Long Short Term Memory Recurrent Neural Network

Manohar Manur[1]* Alok Kumar Pani[1] Pankaj Kumar[2]

[1]*Department of Computer Science and Engineering, CHRIST (Deemed to be University), Bengaluru, India*
[2]*Department of Computer Science and Technology, Motihari College of Engineering, Bihar, India*
* Corresponding author's Email: mm6219311@gmail.com

Abstract: In recent years, heart disease is one of the leading cause of death for both women and men. So, heart disease prediction is considered as a significant part in the clinical data analysis. Standard data mining techniques like Support Vector Machine (SVM), Naïve Bayes and other machine learning techniques used in the earlier research for heart disease prediction. These methods are not sufficient for effective heart disease prediction due to insufficient test data. In this research, Bi-directional Long Short Term Memory with Conditional Random Field (BiLSTM-CRF) has been proposed to increase the efficiency of heart disease prediction. The input medical data were analyzed in a bi-directional manner for effective analysis, and CRF provided the linear relationship between the features. The BiLSTM-CRF method has been tested on the Cleveland dataset to analyze the performance and compared with existing methods. The results showed that the proposed BiLSTM-CRF outperformed the existing methods in heart disease prediction. The average accuracy of the proposed BiLSTM-CRF is 90.04%, which is higher than the existing methods.

Keywords: Bi-directional long short term memory, Conditional random field, Heart disease prediction, Naïve Bayes, Support vector machine.

1. Introduction

In the present decades, heart disease is the major cause of death for both women and men. Hence, an efficient heart disease prediction model creates awareness among patients to provide essential care [1]. Generally, the data mining techniques are used to find the relationship between the various factors and hidden information in the data [2]. Several machine learning techniques also applied to obtain considerable performance of heart disease prediction. Feature selection is an important part of heart disease prediction [3]. Methods with high precision and reliability provide more support to data in identifying prospective patients through accurate prediction. The SVM, neural networks, logistic regression, Naïve Bayes and clustering algorithm provide only considerable performance in disease prediction [4]. Additionally, the presence of uncertainty and missing data affects the performance of the prediction model [5].

Moreover, the huge volume of healthcare data increases the complexity, so method with high scalability is required to process the collected data [6]. Deep learning is a new technique in artificial intelligence and it has been applied in various fields for effective analysis. The deep learning techniques effectively speed up the process in handling the huge volume of data, that is higher than the other traditional machine learning algorithms like SVM, random forest, and naïve Bayes [7]. Besides, the optimization techniques such as Particle Swarm Optimization (PSO), Fruit Fly Optimization (FFO) etc. are applied in the medical data classification for feature optimization. These techniques do not provide the global optimum solution for the data classification [8-10]. To overcome this issue, a new prediction model named as BiLSTM-CRF is proposed in this research paper. The proposed BiLSTM-CRF method has the advantage of analysing the data in bi-directional and also analyse the linear relationship between the features that helps

to increase the performance of heart disease prediction.

This paper is arranged as follows, the Introduction is given in section 1, the Literature survey on recent heart disease prediction method is presented in section 2, the Proposed methodology is explained in section 3, the Experimental result is discussed in section 4 and the Conclusion is done in section 5.

2. Literature survey

Data mining techniques help to predict heart disease in patients using medical records. The latest researches on heart disease prediction using machine learning techniques are surveyed in this section.

J. Nahar, T. Imam, K.S. Tickle, and Y.P.P. Chen [11] tested the computational intelligent methods in the heart disease prediction using Cleveland dataset. This system analysed the capacity of expert judgment based on feature selection techniques and compared with the standard classification algorithms. The experimental outcome showed that the feature selection techniques effectively improve the performance of disease prediction. The feature selection techniques were combined with different classifiers to investigate the performance of computational intelligent methods, in that Naïve Bayes classifier showed promising results. The classifier cannot handle the conditional independence data that affect the posterior probability estimation.

M.Z. Alam, M.S. Rahman, and M.S. Rahman [12] developed a new feature ranking based method for medical data classification. In this literature, a new ranker algorithm was used to rank the features in the dataset and then the random forest classifier was applied in higher ranked feature for the heart disease prediction. This experiment was verified on the 10 different benchmark datasets including Cleveland. The simulation result showed that the developed system achieved better performance compared to the existing systems. In contrast, the developed feature ranking method doesn't analyse the non-linear relationship among the features that affect the performance of prediction.

M.S. Amin, Y.K. Chiam, and K.D. Varathan [13] reviewed the existing methodologies for a significant feature selection and data mining techniques for heart disease prediction. The hybrid features were used to develop the predictive model and tested with seven classification methods including SVM, and Naïve Bayes. The heart disease data sets were collected from UCI Machine Learning Repository. The Cleveland dataset is the commonly used dataset for heart disease prediction. The nine significant features and the top three data mining techniques were identified by performing the experiment. In this research work, the classification was carried out with three limited features that may decrease the performance of heart disease prediction.

C.B.C. Latha, and S.C. Jeeva, [14] developed a new collaborative classification to increase the accuracy of prediction. The heart disease dataset was used to test the performance of the ensemble method and compared it with other methodologies. The developed method was also involved in analysing the stages of the disease. The result of the developed system showed that this method increased the performance of heart disease detection. In addition, the feature selection techniques were also utilized to increase the performance of prediction. But, this method has low interpretability and a linear relationship between the features that reduce the prediction accuracy.

S. Mohan, C. Thirumalai, and G. Srivastava, [15] involved in finding the significant features using machine learning techniques to increase the accuracy of heart disease prediction. The combination of features was applied in the standard classifiers to measure the performance of the prediction. The hybrid random forest with a linear model (HRFLM) provided higher accuracy in the heart disease prediction. In pre-processing, the multiclass variables and binary classification were applied in the attributes of the dataset to check the presence or absence of heart disease. The age is an important factor for heart disease prediction, which has been neglected in the paper.

From this literature survey, it is clear that the existing methods have some limitations. To address these limitations, a new deep learning-based approach is proposed to improve the performance of heart disease prediction.

3. Proposed method

The heart disease prediction has been carried out by many existing methods which are based on machine learning and clustering methods, but still those existing methods lack in efficiency. The objective of this research is to improve the efficiency of heart disease prediction using deep learning method (BiLSTM-CRF). The BiLSTM method analyses the data in two directional ways that tend to improve the performance of prediction. The CRF techniques have been used to provide the linear distribution of the output structure that helps LSTM to improve the accuracy. The LSTM is a deep learning method that analyses the data effectively and

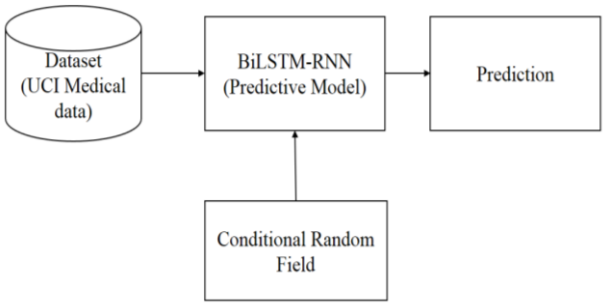

Figure.1 Block diagram of the proposed method

retrieves the key feature which are required for prediction. Fig. 1 shows the block diagram of the proposed method.

3.1 Bi-direction long short term memory - conditional random field

The Bi-LSTM is a deep learning model, which analyse the data in the forward and backward manner to reduce the error in classification. The CRF is applied in the Bi-LSTM for the sequence learning. The Bi-LSTM consists of a forward and backward layer to analyse the information from the data as well as tokens. LSTM is a type of Recurrent Neural Network (RNN), which has the capacity to store the most relevant information. The explanation about RNN, LSTM, BiLSTM and Bi-LSTM-CRF is given in the below section.

RNN is a class of artificial neural model that has the connections between the units to create a directed cycle. The arbitrary embedding sequences $x = (x_1, \ldots, x_T)$ is used as input and exhibit the dynamic temporal behaviour based on the internal memory network. This network consists of hidden unit h, output layer y and the last time step T and it also represents the length of input data in the sequence learning model. For every time step t, RNN computes the hidden states h_t based on previous step h_{t-1} and the input data of current step x_t is mathematically shown in Eq. (1).

$$h_t = g(Ux_t + Wh_{t-1}) \quad (1)$$

Where, U and W denote weight matrices of the network; $g(.)$ is a non-linear activation function or element-wise logistic sigmoid function [16]. The output time step t is computed as $y_t = softmax(Vh_t)$, where V is another weighted parameter. LSTM is a type of RNN, which is developed to solve the gradient vanishing problem. The LSTM is developed with two gates i.e., an input gate i_t and output gate o_t with the cell activation vector c_t.

BiLSTM consists of two LSTM layer to analyse the data from the past and future token context, sequentially. One-layer process the data from left to right and another vice versa. In each time step t, a hidden forward layer contains hidden unit function $\vec{h}$, which is processed based on the previous state $\vec{h}_{t-1}$ with input data at current time step $\vec{h}_t$. The backward layer process the same hidden unit function based on future content $\overleftarrow{h}_{t+1}$ with current data. The forward and backward data representation combined into long vector.

Another widely used model in disease prediction is CRF, which is a discriminative graphical model that has a single log-linear distribution over structured output as a function of input data. The observed variable is denoted as X and the random variable is denoted as Y. G is an undirected graph developed based on Y to show dependency between the vertices. CRF analyses the conditional probability of the output values $y \in Y$ with input values $x \in X$ that is proportional to the product of potential functions on cliques of the graphs, as shown in Eq. (2).

$$p(y|x) = \frac{1}{Z_x} \prod_{s \in S(y,x)} \Phi_s(y_s, x_s) \quad (2)$$

Where Z_x is denoted as a normalization factor of output values, $S(y,x)$ is cliques set of G, and $s(y_s, x_s)$ is clique potential on clique s. The BiLSTM-CRF model has a softmax with overall possible tag sequences, that delivers a probability for the sequence y. The computation of the output sequence prediction is mathematically shown in Eq. (3).

$$y* = argmax_{y \in Y} \sigma(X, y) \quad (3)$$

Where, $\sigma(X, y)$ is indicated as score function, which is defined in Eq. (4).

$$\sigma(X, y) = \sum_{i=0}^{n} A_{y_i, y_{i+1}} + \sum_{i=1}^{n} P_i, y_i \quad (4)$$

Where, A is denoted as matrix of transition scores, $A_{yi,yi} + 1$ is indicated as the score of a transition from the tag of y_i to y_{i+1}. The length of the sentence n with matrix scores P by the BiLSTM network, P_i, y_i is the score of the y_i^{th} tag of the i^{th} data.

In BiLSTM-CRF, the dropout technique is used after the input layer, in order to reduce the overfitting of the training data. This technique helps to prevent complex coadaptation on the training data. The experimental result of the proposed BiLSTM-CRF is presented in the next section.

4. Experimental results

Usually, the heart disease prediction model analyses various factors of the patient to predict the risk of heart disease. The BiLSTM method analyses various features in the collected dataset to predict heart disease. Generally, the deep learning techniques are applied in several research fields to increase accuracy of prediction and classification. In this work, a new deep learning method (BiLSTM–CRF) has been proposed for the heart disease prediction. Here, the BiLSTM–CRF method has been compared with a few standard data mining techniques. This section provides a detailed information about the experimental results of the BiLSTM–CRF approach. In this work, python software is used for simulation with Intel i7 processer, 8GB of RAM, and 500 GB hard disk.

4.1 Dataset

In this research, Cleveland dataset (UCI machine repository) is utilized to evaluate the performance of the BiLSTM–CRF method for heart disease prediction. This dataset contains 303 medical records with 76 attributes [17]. The Cleveland dataset has an attribute named "num" to analyze the heart disease diagnosis in patients on different scales, which ranges from 0 to 4. The value 0 represents that there is no heart disease present in the patients and the value from 1 to 4 states the severity of heart disease. In recent years, many studies have been carried out on the Cleveland dataset for heart disease prediction [11 –15]. The BiLSTM–CRF method is compared with previous researches in order to show the superiority of the proposed work. Addithinally, the BiLSTM-CRF method is analysed with other datasets like Cleveland, Hungary, Switzerland, etc.

4.2 Parameter setting

In this research, 10-fold cross validation is used to test the performance of the proposed and existing methods. The dataset is divided into 10 sets in that nine sets are used for training and one set is applied for testing, and the iteration is set at 10.

4.3 Performance metric

In this research study, five performance metrics such as, accuracy, precision, f-measure, sensitivity and specificity have been used to evaluate the performance of the proposed and existing methods. Precision is defined as the fraction of correct instance between selected instances. The F-measure provides the mean value between precision and recall, where recall is the measure of selected correct instance among total instances. Accuracy is the measure of trueness in the prediction instance, which provides the correctness of the prediction. The formulas of precision, recall, and accuracy are given in the Eq. (5), (6), and (7).

$$Precision = \frac{TP}{TP+FP} \tag{5}$$

$$F-measure = \frac{2TP}{2TP+FP+FN} \tag{6}$$

$$Accuracy = \frac{TP+TN}{n} \tag{7}$$

In addition, sensitivity measures the proportion of actual positive and the specificity measures the proportion of actual negative. The formula of sensitivity and specificity is shown in Eq. (8) & (9) respectively.

$$Sensitivity = \frac{TP}{TP+FN} \tag{8}$$

$$Specificity = \frac{TN}{TN+FP} \tag{9}$$

Where, TP represents true positive, TN is true negative, n is the total number of instances, FP is denoted as false positive and FN indicates as false negative.

4.4 Performance analysis

The proposed method (BiLSTM–CRF) tested on Cleveland dataset in light of precision, f-measure, and accuracy. In addition, the BiLSTM–CRF method is compared with the standard data mining techniques [13-15] in heart disease prediction for evaluating the effectiveness of the proposed work.

In the table 1, the performance of the proposed method (BiLSTM-CRF) compared to existing data mining techniques by means of accuracy. It shows

Table 1. Performance evaluation in light of accuracy

Technique	Average Accuracy (%)
Vote [13]	78.2
Naïve Bayes	78.2
Support Vector Machine (SVM)	78.15
Logistic Regression (LR)	78.03
Neural Network	75.18
k-NN	63.5
Decision Tree	63.5
Bagging Naïve Bayes [14]	84.16
Boosting in Naïve Bayes [14]	84.16
HRFLM [15]	88.4
BiLSTM-CRF	90.04

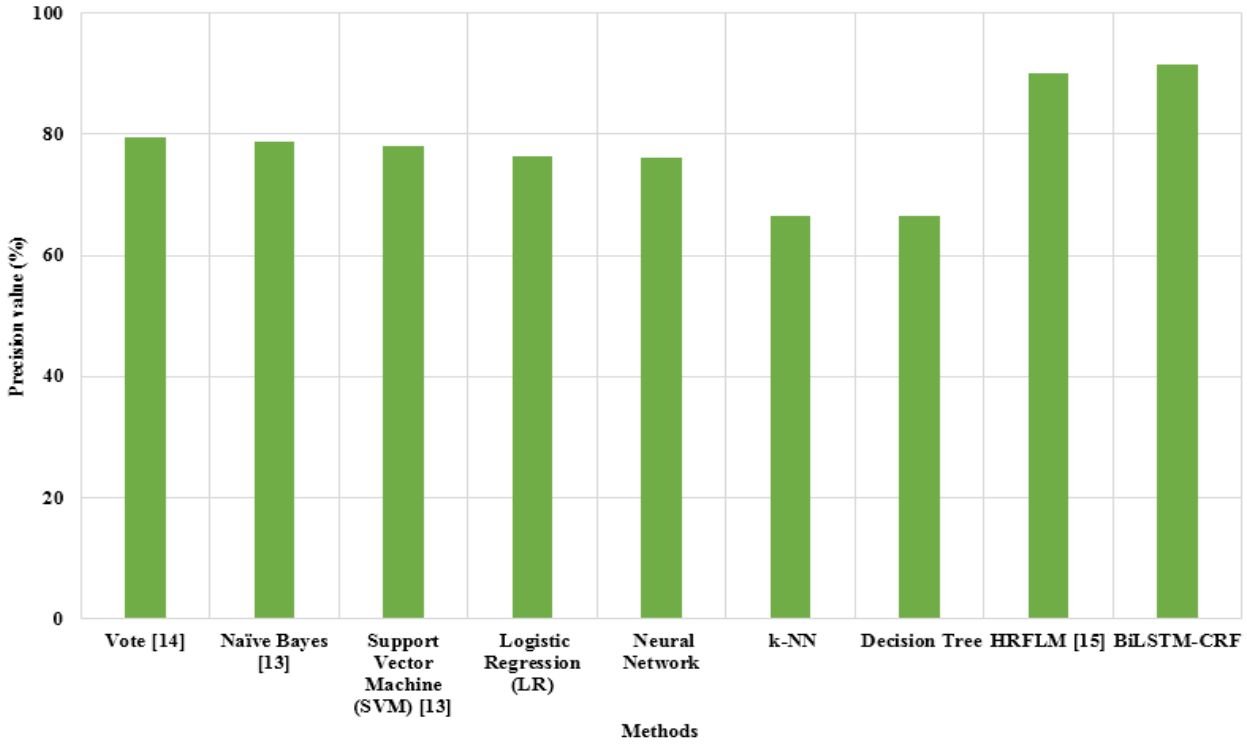

Figure.2 Precision value of various data mining techniques in Cleveland dataset

that the proposed method (BiLSTM-CRF) achieved high accuracy compared to other techniques. The Vote [13] and HRFLM [15] attained high performance related to standard data mining schemes. The boosting and bagging methods applied in the Naïve Bayes classifier that showed considerable performance [14].

The proposed BiLSTM-CRF method has an average accuracy of 90.04%, while the existing method HRFLM [15] has an average accuracy of 88.4%. The relationship between the features is limited at certain aspects, which is considered as the drawback of HRFLM. In HRFLM, the age factor is neglected at the feature selection that is essential in heart disease prediction. The proposed method uses an age factor as features until the hidden relationship is analysed at both directions. The BiLSTM – CRF can effectively analysis features in both directions. In addition, CRF is applied to represent the linear relationship of the outcome that helps to increase the performance of the proposed method.

The precision value of the existing and proposed method in the heart disease prediction is shown in Fig. 2. The proposed BiLSTM–CRF method shows a higher precision value compared to other existing methods. The proposed BiLSTM-CRF method has the advantage of bi-directional data analysis and the CRF provides the linear relationship between the attributes. This tends to improve the performance of the proposed BiLSTM-CRF method. The proposed BiLSTM-CRF method has a precision value of 91.6 %, while existing HRFLM [15] method has a precision value of 90.1 %. From the standard data mining techniques, it is shown that the Naïve Bayes has higher performance in the prediction.

The f-measure value of the existing and proposed method in the heart disease prediction is shown in Fig. 3. The proposed BiLSTM-CRF method attained a higher f-measure compared to other existing methods. The f-measure value of the BiLSTM-CRF is 91.78%, while the existing methodology (HRFLM) has the f-measure value of 90%. Compared to the standard data mining techniques, the Naïve Bayes method has a higher F-measure value of 80.25%. The deep learning methods can effectively analysis the features in the dataset as the proposed method analyzes the data in a bidirectional manner. The proposed BiLSTM–CRF method effectively analyses the relationship between the data attributes.

Fig. 4 represents the comparison of the accuracy achieved by each technique. The proposed BiLSTM-CRF method attained the highest accuracy compared to other existing methods. The proposed method has an accuracy of 92.46 %, while existing Vote method has an accuracy of 86.2 %. The proposed method attained high performance, because the data are analyzed in two directions that effectively provides the information between the data. The CRF method provides a linear relationship between the aspects and output.

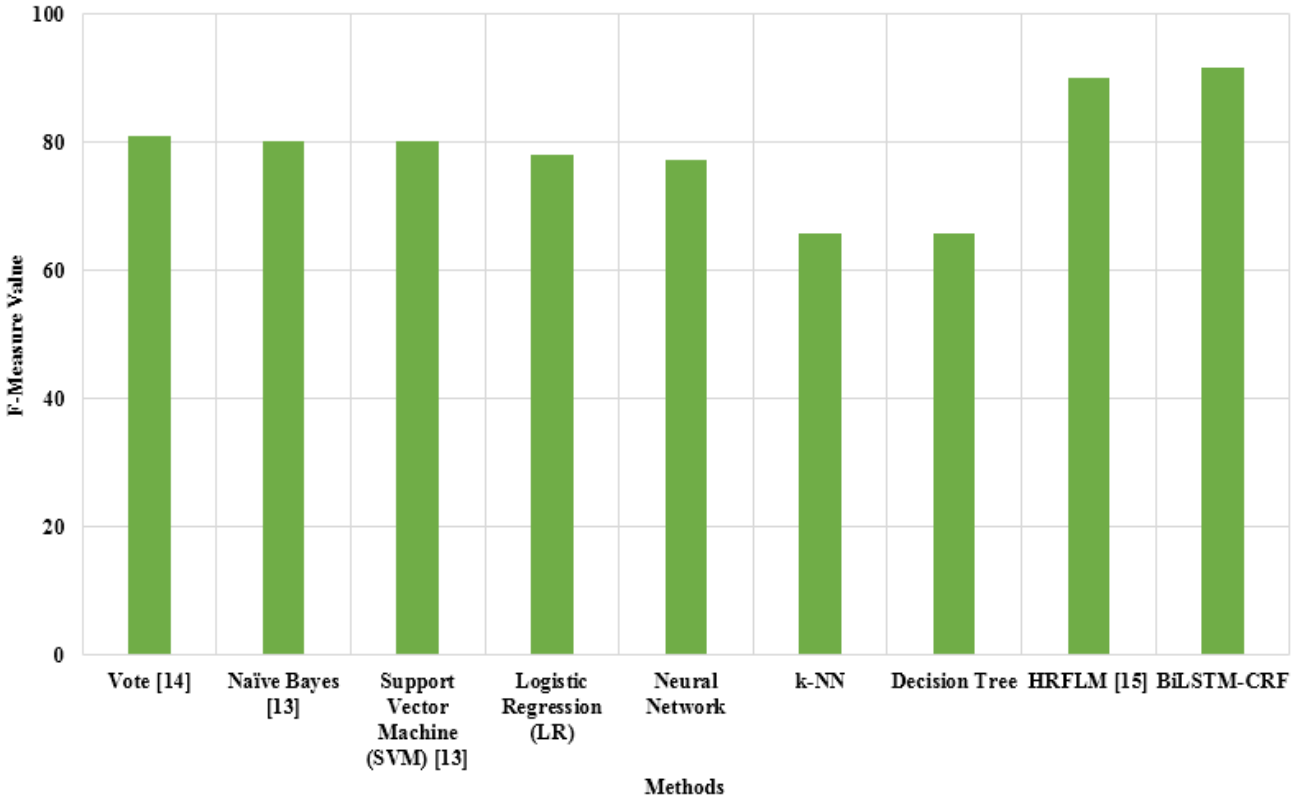

Figure.3 F-Measure value of various data mining techniques in Cleveland dataset

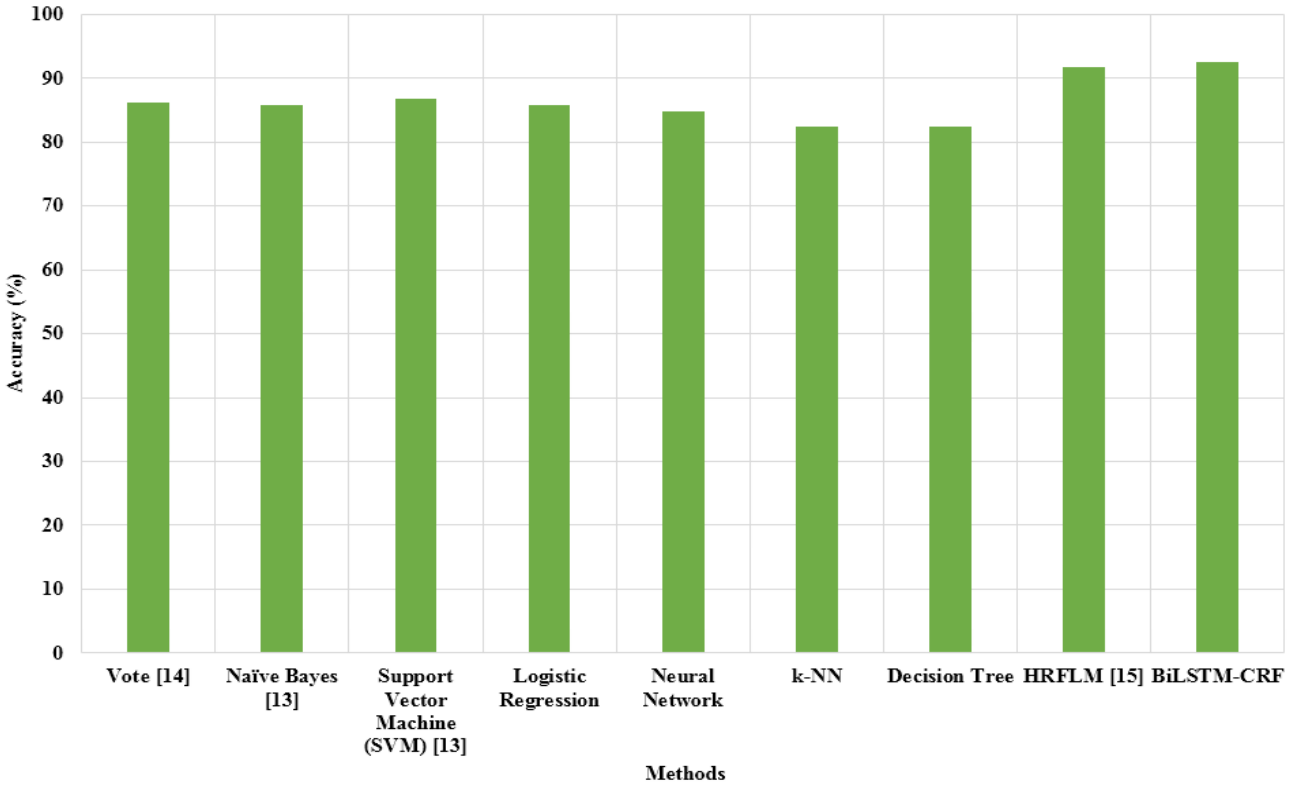

Figure.4 Highest accuracy achieved by data mining techniques

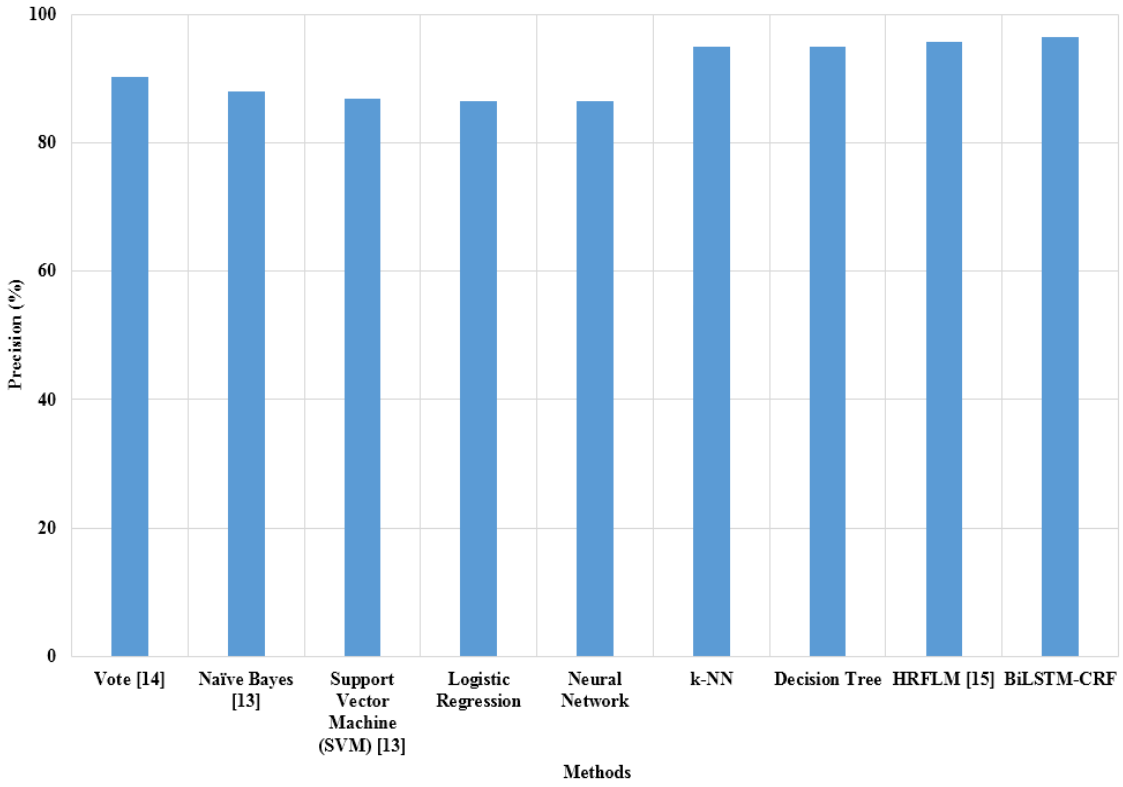

Figure.5 Highest precision value of data mining techniques in heart disease prediction

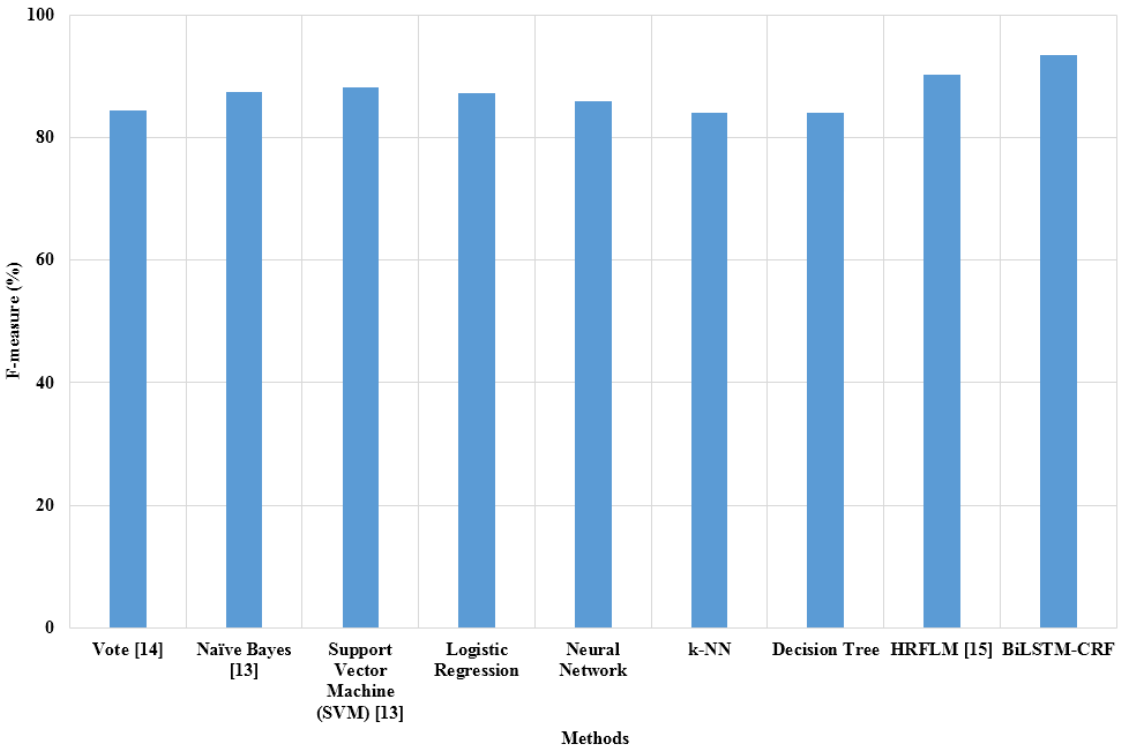

Figure.6 Highest F-measure value of data mining techniques in heart disease prediction

Fig. 5 represents a graphical depiction of proposed and existing methodologies in terms of precision. The proposed method has the highest precision of 96.4% and the second-highest precision is achieved by Decision tree and k-NN of 95%. Though, the average precision value of the decision tree and k-NN classifier are minimum compared to other existing methods. The proposed BiLSTM-CRF method has a higher precision value than other existing methods.

Table 2. Average result of proposed BiLSTM-CRF method on other datasets

Models	Accuracy (%)	Sensitivity (%)	Specificity (%)
Decision Tree	85	98.8	0
Random Forest	86.1	98.8	10
Navie Bayes [13]	75.8	79.8	60
Support Vector Machine [13]	86.1	100	0
VOTE [14]	87.41	-	-
HRFLM [15]	88.4	92.8	82.6
BiLSTM-CRF	89.3	93.1	80.2

The highest f-measure value of existing data mining techniques and the proposed method is shown in Fig. 6. The BiLSTM-CRF method has achieved the highest f-measure value related to other methods. The proposed BiLSTM-CRF method has the f-measure of 93.47%, while the existing method has the f-measure of 84.41%. The proposed BiLSTM-CRF method analyzes the data in two directional ways and provides the linear relationship between the attributes. So, the proposed BiLSTM-CRF attained higher performance compared to existing methods.

4.5 Performance Analysis on the various datasets

In addition, the proposed BiLSTM-CRF method is tested on other datasets like Cleveland, Hungary, Switzerland and VA Long Beach. The average result of the proposed BiLSTM-CRF method on other datasets is shown in Table 2. The proposed BiLSTM-CRF method attained better performance compared to other methods, because it can analyse the data in two directions and also represents the linear relationship between the extracted features.

5. Conclusion

Heart disease is the leading cause of death for both women and men, so it requires an effective prediction technique for early diagnosis. Several data mining and machine learning techniques have been developed to improve the performance of the prediction. This research aimed to improve the performance of heart disease prediction using the BiLSTM-CRF model. The developed model attained highest precision value compared to other data mining techniques, because the proposed model can analyse data in a bi-directional manner and also investigates the linear relationship between the extracted features. In this work, the Cleveland dataset is used to test the performance of the proposed and other existing techniques. The comparison result showed that the proposed BiLSTM–CRF method attained high performance than the existing methods. The proposed method achieved the classification accuracy of 90.04%, which is high compared to other techniques. In future optimization techniques can be included for further improvement in the performance of prediction.

Acknowledgments

I thank CHRIST (Deemed To Be University) Bangalore, for providing a favourable environment to carry out my research work.

References

[1] J. Rodríguez, S. Prieto, and L.J.R. López, “A novel heart rate attractor for the prediction of cardiovascular disease”, *Informatics in Medicine Unlocked*, Vol.15, pp.100174, 2019.

[2] L. Zhang, Z. Chen, J. Su, and J. Li, “Data mining new energy materials from structure databases”, *Renewable and Sustainable Energy Reviews*, Vol.107, pp.554-567, 2019.

[3] I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, and I. Chouvarda, “Machine learning and data mining methods in diabetes research”, *Computational and structural biotechnology journal*, Vol.15, pp.104-116, 2017.

[4] S. Yang, J.Z. Guo, and J.W. Jin, “An improved Id3 algorithm for medical data classification”, *Computers & Electrical Engineering*, Vol.65, pp.474-487, 2018.

[5] M. Seera, and C.P. Lim, “A hybrid intelligent system for medical data classification”, *Expert Systems with Applications*, Vol.41, No.5, pp.2239-2249, 2014.

[6] T. Nguyen, A. Khosravi, D. Creighton, and S. Nahavandi, “Medical data classification using interval type-2 fuzzy logic system and wavelets”, *Applied Soft Computing*, Vol.30, pp.812-822, 2015.

[7] A.M. Karim, M.S. Güzel, M.R. Tolun, H. Kaya, and F.V. Çelebi, “A new framework using deep auto-encoder and energy spectral density for medical waveform data classification and processing”, *Biocybernetics and Biomedical Engineering*, Vol.39, No.1, pp.148-159, 2019.

[8] L. Shen, H. Chen, Z. Yu, W. Kang, B. Zhang, H. Li, B. Yang, and D. Liu, “Evolving support vector machines using fruit fly optimization for medical data classification”, *Knowledge-Based Systems*, Vol.96, pp.61-75, 2016.

[9] Y. Tian, K. Zhang, J. Li, X. Lin, and B. Yang, “LSTM-based traffic flow prediction with missing data”, *Neurocomputing*, Vol.318, pp.297-305, 2018.

[10] C. Tian, C. Li, G. Zhang, and Y. Lv, "Data driven parallel prediction of building energy consumption using generative adversarial nets", *Energy and Buildings*, Vol.186, pp.230-243, 2019.
[11] J. Nahar, T. Imam, K.S. Tickle, and Y.P.P. Chen, "Computational intelligence for heart disease diagnosis: A medical knowledge driven approach", *Expert Systems with Applications*, Vol.40, No. 1, pp. 96-104.
[12] M.Z. Alam, M.S. Rahman, and M.S. Rahman, "A Random Forest based predictor for medical data classification using feature ranking", *Informatics in Medicine Unlocked*, Vol.15, pp.100180, 2019.
[13] M.S. Amin, Y.K. Chiam, and K.D. Varathan, "Identification of significant features and data mining techniques in predicting heart disease", *Telematics and Informatics*, Vol.36, pp.82-93, 2019.
[14] C.B.C. Latha, and S.C. Jeeva, "Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques", *Informatics in Medicine Unlocked*, pp. 100203, 2019.
[15] S. Mohan, C. Thirumalai, and G. Srivastava, "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques", *IEEE Access*, Vol.7, pp. 81542-81554, 2019.
[16] T. Chen, R. Xu, Y. He, and X. Wang, "Improving sentiment analysis via sentence type classification using BiLSTM-CRF and CNN. *Expert Systems with Applications*", Vol.72, pp.221-230, 2017.
[17] V.A. Medical Center, Long Beach and Cleveland Clinic Foundation:Robert Detrano, M.D., Ph.D.