Journal of Theoretical and Applied Information Technology
30th April 2025. Vol.103. No.8

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

DISEASE RISK PREDICTION USING ELECTRONIC HEALTH RECORD DATA BASED ON FLYING SQUIRREL SEARCH OPTIMIZATION WITH BIDIRECTIONAL – LONG SHORT TERM MEMORY

PRASANTHI YAVANAMANDHA[1] , D. S. RAO[2]

[1]Research Scholar, Koneru Lakshmaiah Education Foundation, Department of Computer Science and Engineering, Hyderabad, Telangana, India

[2]Professor, Koneru Lakshmaiah Education Foundation, Department of Computer Science and Engineering, Hyderabad, Telangana, India

E-mail: [1]prasanthi.yavanamandha@klh.edu.in, [2]dsrao@klh.edu.in

ABSTRACT

Nowadays, the attention towards effective disease risk prediction has increased, due to its importance in monitoring the future health status of patients. This prediction helps to provide the right treatment for patients to prevent severe stages of diseases. However, the existing risk prediction models based on Machine Learning (ML) have limitations in learning temporal information from the Electronic Health Record (EHR) data. To overcome this, a Flying Squirrel Search Optimization (FSSO) algorithm for feature selection and Bi-directional Long Short Term Memory (Bi-LSTM) is proposed to enhance the accurate disease risk prediction using EHR data. The proposed FSSO based feature selection method efficiently reduces the dimensionality of EHR data and helps to identify the most predictive features for disease risk accurately. By utilizing the bidirectional layers, Bi-LSTM model learns the dependencies in both past and future directions, which makes the model suitable for capturing comprehensive temporal patterns that influence disease risk. Initially, the EHR data is acquired and preprocessed by solving the missing values in the dataset to enhance the risk prediction process. Experimental results of the proposed method achieved an accuracy of 0.938 for MIMIC-IV dataset when compared to existing methods such as XGBoost and RDF.

Keywords: *Bi-Directional Long Short Term Memory, Disease Risk Prediction, Electronic Health Record, Flying Squirrel Search Optimization, Machine Learning.*

1. INTRODUCTION

The recent development of Electronic Health Records (EHR) has been crucial in the health sector which is an essential way that keep and maintain health records of all patients. Moreover, this EHR enables these health records as valuable data which is used for scientific research of public health. An EHR consists of various types of information including laboratory tests, demographic details, medical procedures, and diagnoses-related information, which is used to extract patterns of medical events evolution that change over time [1-3]. The disease risk is predicted according to the investigation of historical health records that help to improve the future health status of patients [4].

Since EHR data with high-dimensional sparsity, multi-source heterogeneous information and irregular time intervals make the model difficult they are used for disease risk prediction [5,6]. Risk prediction plays a vital role in monitoring and diagnosing diseases which helps to reduce mortality risks and to treat the patients at the right time [7,8]. Accurately evaluating the future disease at the beginning stage is essential for informing the development of personalized prevention strategies [9].

The EHRs, are one of the rich sources of patient data that involves temporal information on medical encounters regarding the disease, which is a suitable learning risk model [10]. In this research, EHR data is obtained from the benchmark datasets such as MIMIC-III and MIMIC_IV databases. Both

datasets have issue of missing data values that greatly impact in disease risk prediction. Thus, in this research a missing value imputation technique is utilized to enhance risk prediction. In the last decade, Machine Learning (ML) based techniques have been extensively used in the healthcare field for distinct purposes such as disease detection, and classification also for risk prediction [11-12]. Using machine learning in medical diagnosis provides more accurate results reducing the errors caused by humans [13]. Machine learning which is termed as a subset of artificial intelligence, aims to predict the desired results of a task by providing a well-preprocessed dataset based on historical results without human interference [14]. Various classification algorithms, such as logistic regression, random forests, and support vector machines, have been employed to predict diseases, including heart disease, diabetes, and tumors. These methods offer improved prediction capabilities and enable early diagnosis, which is critical for effective treatment and better outcomes in healthcare [15,16]. The ML algorithms-based risk prediction significantly enhanced the decision-making process for clinicians and helped to diagnose correctly, which increases patient's lifetime [17]. Furthermore, this prediction analysis has contributed to the early detection of serious diseases, that reduce risks and identify high-risk patients quickly [18].

However, ML models have limitations such as that they heavily depend on the data, overfitting and fail to learn the temporal information that affects the effective risk prediction of diseases using the EHR data. On the other hand, Deep learning (DL) is advancing in various domains such as language processing, signal processing, image recognition and classification leveraging large scale datasets to achieve state-of-the-art performance in numerous tasks [19]. Thus, in this research DL-based prediction method for assessing the risk level of diseases using the patient data obtained from EHR data [19]. However, existing risk prediction model based on DL approaches have limitations such as that neural network models learn spatial features than features with temporal information which are crucial in disease risk prediction. To overcome this limitation faced by ML and existing DL-based prediction methods, a Flying Squirrel Search optimization (FSSO-Bi-LSTM) model is proposed for the risk prediction of diseases using the EHR data efficiently. The proposed Bi-LSTM model learns the temporal dependencies effectively with help of FSSO algorithm. The main contributions of this research are:

- The FSSO-Bi-LSTM method is proposed to select significant features and learn the temporal information about the progressive diseases for precise disease risk prediction using EHR data.
- The proposed Bi-LSTM approach-based risk prediction model learns the temporal information which process the data both forward and backward directions, that makes the prediction model to capture temporal patterns efficiently that influence disease risk in health record data.
- The proposed FSSO based feature selection method with an adaptive dynamic mechanism efficiently helps to identify the most significant predictive features for efficient disease risk based on the fitness function.

The literature survey section deals with a comprehensive review of related works. The methodology section speaks about a detailed explanation of the proposed risk prediction model. This ensures a durable foundation for understanding the research framework and the innovative approach that was introduced. The proposed model has been examined against the existing machine learning models like XGBoost, RF, CNN, and LSTM, which depict exceptional performance. These outcomes are systematically detailed in the results section, supported by specific evaluation metrics. Key findings and significance of the proposed method are summarized and the scope for further findings is mentioned in the conclusion.

2. LITERATURE REVIEW

Fan Yang et al. [20] presented a Mortality Risk Prediction Model based on Deep Learning (Deep MPM) considers the 2 levels of the mechanism to analyze risk prediction from the patient's multiple longitudinal medical records. The presented DeepMPM model based on two-attention mechanism integrated with LSTM model was employed to predict the risk levels of diseases using the EHR data. The main advantage of the presented method was that the analyzed DeepMPM model efficiently learns the temporal information that enhances the prediction performance. However, the presented DeepMPM model has performance degradation while predicting risks due to missing values present in the EHR data. Since the presented LSTM mode have limitations of that rely on constant hidden states that make the prediction difficult to capture long term information which impact in inaccurate prediction for certain diseases like diabetic retinopathy, Alzheimer and so on.

Shuai Niu et al. [21] suggested a non-parametric predictive clustering-based risk prediction model integrates the Dirichlet Process Mixture Model (DPMM) with a cluster-based neural network involved in the EHR risk prediction. This captures the underlying distribution of data leading to improved clustering and more accurate risk prediction, especially in the case of HER, which achieved better accuracy. The combination of DPMM and cluster-based neural networks leads to maximum training time and difficult to interpret the various risk predictions it challenges the analysis of the risk prediction. Also, the unstable clustering of patient's data impact in inaccurate prediction results and increase risk.

Yu-wen Chen et al. [22] suggested an attention-based Temporal Convolution Network (TCN) model achieved better performance in the prediction of mortality risk with time series data handling hospital mortality risk prediction. This integration of attention mechanisms allows the model to focus on the most relevant part of time series data and then efficiently identify the mortality risk improving prediction accuracy. The struggle to analyze the non-time series in a combination of TCN and attention mechanism was unable to process the dynamic risk prediction. Also, the TCN model has drawback to learn uneven spaced data which means irregular time gaps of EHR data that lead to incorrect risk assessments.

Xuping Lin et al. [23] developed an Extreme Gradient Boost and Random Decision Forest (XGBoost-RDF) model for predicting mortality risks of myocardial infarction using the EHR data. This captures the underlying distribution of data leading to improved clustering and more accurate risk prediction, especially in the case of EHR, which achieved better accuracy. The integration of XGBoost and RDF leads to maximum training time and difficulty learning the disease information effectively challenging to analysis of the accurate disease risk prediction. However, the developed XGBoost –RDF model failed to capture temporal dependencies in the health record data. Thus, this developed model process the features independently which make inappropriate to predict risk for progressive disease.

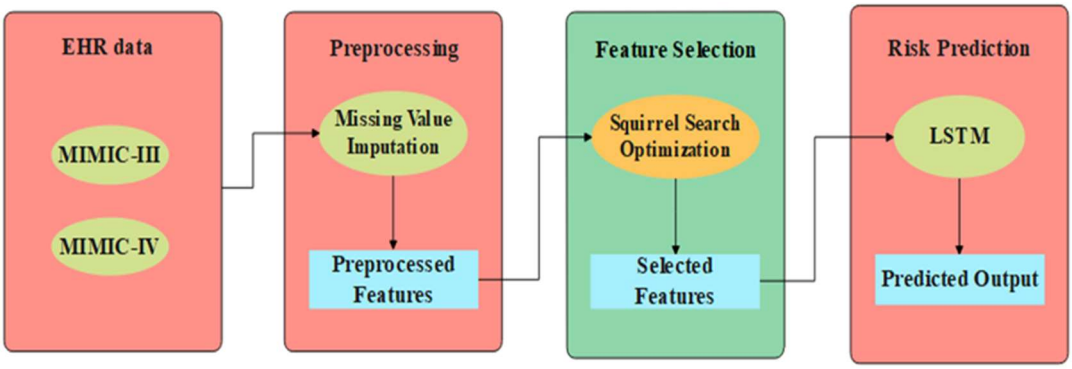
Shengwei Lin et al. [24] designed an ensemble ML model based on random forest, support vector machine, naïve bayes, and k-nearest neighbor for acute kidney injury using EHR data. The designed ensemble model was used to predict the mortality risk of acute kidney injury with the patient's health record data. The main advantage of the designed ensemble model was that the random forest algorithm was employed for feature selection to remove redundant information and improve risk prediction performance. However, the designed ensemble ML model struggles to analyze the temporal information and is also unable to process the dynamic risk prediction. However, the designed ensemble ML model struggles to analyze the temporal information and is also unable to process the dynamic risk prediction. This limitation impact in risk prediction that failed to analyze long term data and provide inaccurate prediction results.

Figure 1. Block diagram for the proposed disease risk prediction model

From the above literature survey of existing works performed for disease risk prediction using EHR data some of the research limitations are described as follows: the two attention LSTM [20] and TCN [22] models have the ability to learn temporal information but they failed to capture long term dependencies and uneven spaced patient's data that affects the accurate risk prediction. Moreover, the DPMM and cluster model [21], XGBoost and RDF model [23] and Ensemble ML models [24] were struggled to capture the temporal dependencies a process the features independently lead to reduce the performance of the model. To address these limitations, a FSSO-Bi-LSTM model [25] is proposed in this research to predict the disease risks effectively by utilizing the EHR data. The proposed Bi-LSTM model process the sequential data both forward and backward that allow it to maintain the contextual information which enhance the prediction accuracy.

3.METHODOLOGY

The proposed disease risk prediction model using EHR data includes four phases: Dataset, preprocessing, feature selection and proposed risk prediction. The block diagram for the proposed disease risk prediction model is illustrated in Figure 1. Firstly, EHR data is obtained from MIMIC-III [26] and MIMIC-IV [27] datasets and preprocessed by the missing value imputation technique. Then an optimization based feature selection is proposed to remove redundant features and to improve the risk prediction process. Finally, a DL-based algorithm is used to predict the risk levels of disease accurately.



3.1. Dataset

For effective disease risk prediction using EHR data, MIMIC-III and MIMIC-IV are the two benchmark datasets used in this research.

3.1.1. MIMIC- III dataset

This dataset is one of the largest open-access EHR datasets [26], that contains both structured and unstructured health data which is collected during several visits of patients to the hospital. The MIMIC-III dataset consists of 22,220 EHRs of patient visit data that include both auxiliary information and medical notes obtained from 19,017 different individuals across several single or multiple hospital visits.

3.1.2. MIMIC-IV dataset

This dataset contains comprehensive health records of 73,181 patients where they are admitted to several Intensive Care Units (ICU) at BIDMC in Boston, Massachusetts, from 2008 to 2019 [27]. MIMIC-IV dataset includes demographic indicators, well-documented events, laboratory results, important sign readings, patient survival status and fluid balance assessments. Moreover, the database consists of International Classification of Diseases and Revision (ICD-9 and ICD-10) codes, that offer a standardized framework for systematic classification. These data are fed as input to the preprocessing process to enhance the EHR data for risk prediction.

3.2. Preprocessing

The acquired EHR data is fed as input to the preprocessing stage that converts raw data into a beneficial format for improving risk prediction. The EHR data consists of missing values which impact the accurate disease risk level prediction, the missing value imputation technique is utilized to address this issue. The KNN algorithm is utilized for the missing value imputation process, which imputes the missing data based on the distance between sample vectors. For every missing feature, it considers the k closest samples which have observed this feature and averages their values about numerical data. If the missing value is a categorical data, the result obtained from KNN is the frequent class of the k nearest neighbors.

3.3. Feature Selection

The preprocessed data is fed as input to select the most significant features by utilizing squirrel search optimization with an adaptive dynamic mechanism technique. The flying squirrel search algorithm mimics the hunting or food-searching behavior of squirrels which integrates the exploration capability of the cuckoo search method with the proposed FSSO algorithm. In this optimization, squirrels are assumed to fly in several directions to identify an optimal hickory tree together nutrient-rich food resource. The agents are initially split into smaller sets, and within every subset and then adjusted adaptively to enhance balance between exploration and exploitation.

In total population of squirrels is split into 70% and 30% which belong to exploration and exploitation groups respectively. Flying Squirrel agents in the exploration group explore for potential regions that are near to their present location in search space through the iteration process, the best solution is selected among the available options in the surrounding area by the fitness value. The velocity and location of the squirrels are mathematically given by matrices in the equation. (1) and (2):

$$FS_{i,j} = \begin{bmatrix} FS_{1,1} & FS_{1,2} & FS_{1,3} & \cdots & FS_{1,d} \\ FS_{1,1} & FS_{1,1} & FS_{1,1} & \cdots & FS_{2,d} \\ FS_{3,1} & FS_{3,2} & FS_{3,3} & \cdots & FS_{3,d} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ FS_{n,1} & FS_{n,2} & FS_{n,3} & \cdots & FS_{n,d} \end{bmatrix} \quad (1)$$

$$V_{i,j} = \begin{bmatrix} V_{1,1} & V_{1,2} & V_{1,3} & \cdots & V_{1,d} \\ V_{1,1} & V_{1,1} & V_{1,1} & \cdots & V_{2,d} \\ V_{3,1} & V_{3,2} & V_{3,3} & \cdots & V_{3,d} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ V_{n,1} & V_{n,2} & V_{n,3} & \cdots & V_{n,d} \end{bmatrix} \quad (2)$$

where, $FS_{i,j}$ and $V_{i,j}$ denotes position i^{th} flying squirrel in the j^{th} dimension; The initial positions of the squirrels $FS_{i,j}$ is determined by a uniform distribution within defined lower and upper limits. In the SSO algorithm, the position of each squirrel is updated according to the following cases 1 to 6 which is determined by a randomly generated value of p and is mathematically expressed in equations (3) to (5). If p is greater than or equal to 0.5, then the following cases will be implemented and among these cases, the most appropriate or suitable case is selected.

Case 1: Position of FS_{at} and moving to the hickory nut tree:

$$FS_{at}^{t+1} = \begin{cases} FS_{at}^{t} + d_g \times G_c(FS_{ht}^{t} - FS_{at}^{t}) & if\ R_1 \geq P_{dp} \\ Random\ location & otherwise \end{cases} \quad (3)$$

Case 2: Position of FS_{nt} and moving to the acorn nut trees:

$$FS_{nt}^{t+1} = \begin{cases} FS_{nt}^{t} + d_g \times G_c(FS_{at}^{t} - FS_{nt}^{t}) & if\ R_2 \geq P_{dp} \\ Random\ location & otherwise \end{cases} \quad (4)$$

Case 3: Position of FS_{nt} and moving to the hickory nut trees:

$$FS_{nt}^{t+1} = \begin{cases} FS_{nt}^{t} + d_g \times G_c(FS_{ht}^{t} - FS_{nt}^{t}) & if\ R_3 \geq P_{dp} \\ Random\ location & otherwise \end{cases} \quad (5)$$

where, FS_{at}^{t}, FS_{ht}^{t}, and FS_{nt}^{t} represents the position of squirrels on acorn nut tree, hickory nut tree, and normal tree, R_1, R_2, R_3 denotes the random number, d_g random distance for gliding, t denotes present iteration. If p is less than 0.5, then the following cases will be implemented which are mathematically formulated in equations (6) to (8):

Case 4: Position of FS_{nt} and moving diagonally:

$$FS_{nt}^{t} = \begin{cases} FS_{nt}^{t} + V_{nt}^{t} + c_1 r(FS_{ht}^{t} - FS_{nt}^{t}) + c_2 r(FS_{at}^{t} - FS_{nt}^{t}) & if p_a < a \\ Random\ FS_{rand}^{t} \in FS_{nt}^{t} & otherwise \end{cases} \quad (6)$$

Case 5: Position of FS_{nt} and moving vertically or horizontally based on the fitness value $F_n(FS_{nt}^{t})$:

$$FS_{nt}^{t} = \begin{cases} FS_{nt}^{t} + V_{nt}^{t} + c_3 r(FS_{rand}^{t} - FS_{nt}^{t}) & if\ F_n(FS_{rand}^{t}) < F_n(FS_{nt}^{t}) \\ FS_{nt}^{t} + V_{nt}^{t} + c_1 r(FS_{rand}^{t} - FS_{nt}^{t}) & otherwise \end{cases} \quad (7)$$

Case 6: Position of FS_{nt} and moving will be exponential:

$$FS_{nt}^{t+1} = FS_{nt}^{t+1} + |(FS_{rand}^{t} - FS_{nt}^{t}) \exp(bt) \cos(2\pi t)| \quad (8)$$

3.3.1. Fitness function

This function is utilized to select high-quality solutions which are obtained by the FSSO algorithm. For the best solution, the number of selected features and classification error rate are considered in this research where the optimal solution must have a lower error rate and a smaller set of features. To evaluate the solution quality, the fitness function for optimal feature selection is represented in equation (9):

$$F_n = h_1\, Err(O) + h_2 \frac{|s|}{|f|} \quad (9)$$

Where, F_n denotes fitness function; $Err(O)$ indicates optimized error rate; s represents selected; f indicates a total number of features; h_1 and h_2 specifies the importance of classification error rate and number of selected features. Based on this fitness function, the most important feature with relevant information is selected and fed to the proposed risk prediction model.

3.4. Proposed Prediction

The selected important features are fed to the proposed disease risk prediction model for prediction of the disease's risk level. Since the EHR is time series-related data recurrent neural network-based methods are utilized in this research which are the most suitable prediction related works. The Bi-LSTM model is a type of LSTM network that updates and utilizes those vectors through the forget, input and output gates to learn long-term-dependent and short-term-dependent information about the features.

There are three important gates present in the Bi-LSTM model that have the same input but different functions to provide accurate results. The forgetting gate is used to accept long-term memory, and output from the previous unit module, and decides which portion to retain or forget; the input gate searches the new information from input, and stores it in the cell state; the output gate processes that information to obtain the final output. These processes can be formulated as in eq. (10) to eq. (15):

$$i_t = \sigma\left(W_i.\left[h_{t-1,}x_t\right] + b_i\right) \quad (10)$$

$$f_t = \sigma\left(W_f.\left[h_{t-1,}x_t\right] + b_f\right) \quad (11)$$

$$o_t = \sigma\left(W_o.\left[h_{t-1,}x_t\right] + b_o\right) \quad (12)$$

$$\bar{C}_t = tanh(W_C.[h_{t-1,}x_t] + b_C) \quad (13)$$

$$C_t = f_t \otimes C_{t-1} + i_t \otimes C_t \quad (14)$$

$$h_t = o_t \otimes tanh\,(C_t) \quad (15)$$

Where, i_t, f_t, o_t, and c_t represents input, forget, output gates, and memory cells, W_i, W_f, W_o b_i, b_f, b_o indicates weights and bias of input, forget, and output gates; σ and $tanh$ denotes activation functions. The input, forget and output gates in Bi-LSTM model can process the input in sequence and fuse the text feature of history, current and future, greatly improving the efficiency of network training compared with LSTM and recurrent neural networks by the addition of long-distance information. At last, the prediction results are obtained softmax layer with the temporal information learned from selected features.

4.RESULTS AND DISCUSSION

The performance analysis of the proposed FSSO-Bi-LSTM method for disease risk prediction using EHR data of MIMIC-III and IV datasets is depicted in this section. The proposed prediction method is simulated using Python 3.9 with a system configuration of i7 processor, 16 GB RAM and Windows 10 OS. Table 1 represents the parameter settings of proposed FSSO-Bi-LSTM model.

Table 1:Parameter settings of proposed method

Parameter	Values
Bi-LSTM	
Dropout rate	0.2
Epochs	50
Learning rate	0.0001
Activation function	Softmax
Optimizer	Adam
Batch Size	32
FSSO algorithm	
No. of iterations	100
No. of population	30
Fitness function	Accuracy and less number of features

Performance measures used for evaluation are Precision, Sensitivity (Recall), Specificity and F1-score. The mathematical expression of the performance metrics is illustrated in eq. (16) to (20):

$$Accuracy = \frac{TP+TN}{TN+TP+FN+FP} \times 100 \quad (16)$$

$$Precision = \frac{TP}{TP+FP} \quad (17)$$

$$Sensitivity = \frac{TP}{TP+FN} \quad (18)$$

$$Specificity = \frac{TN}{TN+FN} \quad (19)$$

$$F1-Score = 2 \times \frac{(precision \times sensitivity)}{(precision+sensitivity)} \quad (20)$$

where, TN is True Negative, TP is True Positive, FN is False Negative, and FP is False Positive respectively.

4.1. Quantitative and Qualitative Analysis

The performance analysis of the proposed method for EHR data based disease risk prediction using MIMIC-III and MIMIC-IV datasets is represented in Tables 2 to 5. Performance evaluation of FSSO-Bi-LSTM using MIMIC-III dataset is illustrated in Table 2. The proposed method utilized for disease risk prediction is evaluated and compared with existing prediction approaches such as the eXtreame Gradient Boost (XGBoost) algorithm, Random Forest (RF) Convolutional Neural Network (CNN), and LSTM, utilized in disease risk prediction.

Table 2: Performance Analysis of the Proposed FSSO-LSTM Method in the MIMIC-III Dataset

Methods	Precision	Sensitivity	Specificity	F1-score
XGBoost	0.7869	0.7964	0.7992	0.7916
RF	0.7987	0.8167	0.8057	0.8076
CNN	0.8074	0.8257	0.8149	0.8164
LSTM	0.8176	0.8362	0.8283	0.8267
Proposed FSSO-LSTM method	0.8964	0.8871	0.8796	0.8917

The performance analysis of the proposed FSSO-Bi-LSTM method using MIMIC-IV dataset is illustrated in Table 3. The proposed FSSO-Bi-LSTM method algorithm is evaluated and compared with existing approaches such as XGBoost, RF, CNN, and LSTM, utilized in disease risk prediction.

Table 3: Performance Analysis of the Proposed FSSO-LSTM method in the MIMIC-IV Dataset

Methods	Accuracy	Precision	Sensitivity	Specificity	F1-score
XGBoost	0.862	0.837	0.822	0.816	0.829
RF	0.887	0.846	0.873	0.875	0.859
CNN	0.901	0.885	0.896	0.884	0.890
LSTM	0.914	0.894	0.907	0.891	0.900
Proposed FSSO-Bi-LSTM method	0.938	0.927	0.916	0.902	0.921

The performance analysis of the proposed FSSO algorithm-based feature selection method utilizing MIMIC-III dataset is illustrated in Table 4. The proposed method is evaluated and compared with existing approaches such as Particle Swarm Optimization (PSO), Harris Hawks Optimization (HHO), Whale Optimization Algorithm (WOA), and Coati Optimization Algorithm (COA) utilized in disease risk prediction.

Table 4: Performance analysis of the FSSO-based feature selection method in the MIMIC-III dataset

Methods	Precision	Sensitivity	Specificity	F1-score
PSO	0.7791	0.7617	0.7633	0.7703
HHO	0.7843	0.7934	0.7711	0.7888
WOA	0.7982	0.8049	0.7864	0.8015
COA	0.8026	0.8126	0.7959	0.8075
Proposed FSSO	0.8964	0.8871	0.8796	0.8917

The performance analysis of the proposed FSSO algorithm based feature selection method utilizing the MIMIC-IV dataset is illustrated in Table 5. The proposed method algorithm is evaluated and compared with existing approaches such as PSO, HHO, WOA, and COA utilized in disease risk prediction using EHR data.

Table 5: Performance analysis of FSSO based feature selection method in MIMIC-IV dataset

Methods	Accuracy	Precision	Sensitivity	Specificity	F1-score
PSO	0.892	0.886	0.862	0.837	0.873
HHO	0.903	0.895	0.887	0.846	0.890
WOA	0.916	0.924	0.891	0.885	0.907
COA	0.927	0.931	0.904	0.894	0.917
Proposed FSSO	0.938	0.927	0.916	0.902	0.921

4.2. Comparative Analysis

The comparative analysis of the proposed model with the existing disease risk prediction model using EHR data with two different datasets is depicted in this section. The comparative analysis with existing methods for two datasets MIMIC-III and MIMIC-IV datasets is illustrated in Figures 2 and 3. The existing approaches used for comparison are Deep MPM [20], DIRPRED [21], XGBoost and RDF [23] and GBM [24] respectively. Performance metrics namely accuracy, precision, sensitivity, specificity and F1-score are used for comparative analysis of proposed FSSO-Bi-LSTM in disease risk prediction.

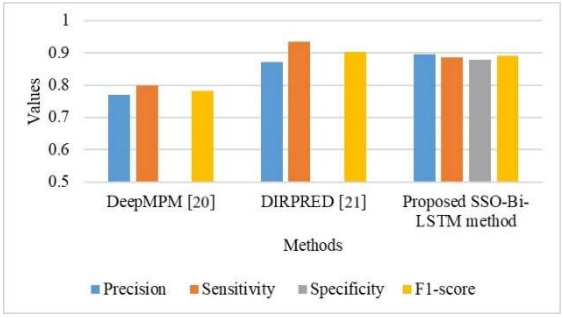


Figure 2: Comparative analysis of proposed method for MIMIC-III dataset

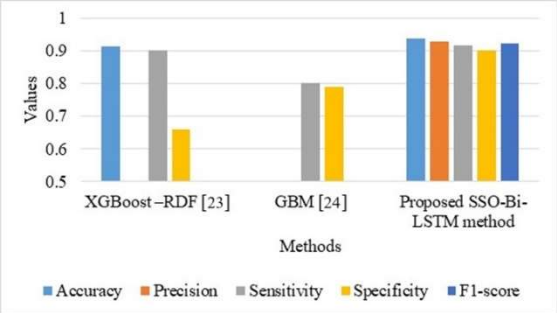


Figure 3: Comparative analysis of proposed method for MIMIC-IV Dataset

4.3. Discussion

The proposed FSSO-Bi-LSTM method achieved better results when compared to the existing disease risk prediction method using the EHR data. The existing method has drawbacks such as DeepMPM [20] model has performance degradation while predicting risks due to missing values present in the EHR data. The integrated DPMM and cluster-based neural network [21] model consume high training time and also face challenges in understanding various risks for accurate prediction. The struggle to analyse the non-time series in combination with TCN and attention mechanism [22] was unable to process the dynamic risk prediction. XGBoost and RDF [23] lead to maximum training time and difficulty learning the disease information effectively challenges to analysis the accurate disease risk prediction. Ensemble ML model [24] struggles to analysis the temporal information and is unable to process the dynamic risk prediction.

To overcome this, an SSO-Bi-LSTM method is used for an accurate disease risk prediction method. The proposed Bi-LSTM approach-based risk prediction model learns the dependencies of both forward and backward directions, which makes the prediction model more suitable to capture temporal patterns or information that influence disease risk. The proposed FSSO algorithm-based feature selection with an adaptive dynamic mechanism is employed to select the most significant features that have best fitness values for efficient disease risk prediction.

5.CONCLUSION

The FSSO-Bi-LSTM method is proposed to enhance the accurate disease risk prediction using EHR data. The proposed FSSO algorithm with an adaptive dynamic mechanism-based feature selection method is used to identify the most predictive features for identifying disease risk effectively using EHR data. However, the existing risk prediction models based on ML and DL methods failed to capture temporal features and long term dependencies which are crucial in disease risk prediction. Since, certain progressive disease like diabetic retinopathy, Alzheimer diseases require history of data for accurate prediction. Thus, in this research a DL model Bi-LSTM which is more suitable for sequential data and an optimization algorithm for eliminating redundant features to enhance the prediction results. By utilizing the bidirectional layers, the Bi-LSTM model learns the dependencies in both past and future directions, which makes the model suitable for capturing comprehensive temporal patterns that influence disease risk.

Initially, the EHR data is acquired and preprocessed by solving the missing values in the dataset to enhance the risk prediction process. For missing values, the imputation KNN model is used for balancing the missing values in EHR data. Then, the significant features are selected by the proposed SSO algorithm, and the risk levels of disease are predicted by the Bi-LSTM model effectively. Experimental results of the proposed method achieved a precision of 0.8264 for the MIMIC-III dataset and an accuracy of 0.938 for the MIMIC-IV dataset when compared to existing methods such as DeepMPM and XGBoost -RDF approaches. However, the proposed FSSO-Bi-LSTM model face challenges to learn long term sequence that affect the accurate disease risk prediction in certain conditions. Hence as a future work advanced DL methods such as Transformers-based methods and attention mechanisms will be used for effective and precise disease risk prediction using EHR data.

REFERENCES:

[1] A. S. M. Mosa, C. Thongmotai, H. Islam, T. Paul, K. S. M. T. Hossain, and V. Mandhadi, “Evaluation of machine learning applications

using real-world EHR data for predicting diabetes-related long-term complications", *Journal of Business Analytics*, Vol. 5, No. 2, 2021, pp. 141–151.

[2] V.M. Ruiz, M.P. Goldsmith, L. Shi, A.F. Simpao, J.A. Gálvez, M.Y. Naim, V. Nadkarni, J.W. Gaynor, and F.R. Tsui, "Early prediction of clinical deterioration using data-driven machine-learning modeling of electronic health records", *The Journal of Thoracic and Cardiovascular Surgery*, Vol. 164, No. 1, 2022, pp. 211-222.e3

[3] F. Li, P. Wu, H. H. Ong, J. F. Peterson, W.-Q. Wei, and J. Zhao, "Evaluating and mitigating bias in machine learning models for cardiovascular disease prediction", *Journal of Biomedical Informatics*, Vol. 138, 2023, p. 104294.

[4] J. H. Hurst, C. Zhao, H. P. Hostetler, M. Ghiasi Gorveh, J. E. Lang, and B. A. Goldstein, "Environmental and clinical data utility in pediatric asthma exacerbation risk prediction models", *BMC Medical Informatics and Decision Making*, Vol. 22, No. 1, 2022.

[5] C.-C. Chiu, C.-M. Wu, T.-N. Chien, L.-J. Kao, C. Li, and C.-M. Chu, "Integrating Structured and Unstructured EHR Data for Predicting Mortality by Machine Learning and Latent Dirichlet Allocation Method", *International Journal of Environmental Research and Public Health*, Vol. 20, No. 5, 2023, p. 4340.

[6] Y. Li, M. Mamouei, G. Salimi-Khorshidi, S. Rao, A. Hassaine, D. Canoy, T.Lukasiewicz, and K. Rahimi, "Hi-BEHRT: Hierarchical Transformer-Based Model for Accurate Prediction of Clinical Events Using Multimodal Longitudinal Electronic Health Records", *IEEE Journal of Biomedical and Health Informatics*, Vol. 27, No. 2, 2023, pp. 1106–1117.

[7] I.E. Nogues, J. Wen, Y. Zhao, C.L. Bonzel, V.M. Castro, Y. Lin, S. Xu, J. Hou, and T. Cai, "Semi-supervised Double Deep Learning Temporal Risk Prediction (SeDDLeR) with Electronic Health Records", *Journal of Biomedical Informatics*, Vol. 157, 2024, p. 104685.

[8] Y. Xu, H. Ying, S. Qian, F. Zhuang, X. Zhang, D. Wang, J. Wu, and H. Xiong, "Time-aware Context-Gated Graph Attention Network for Clinical Risk Prediction", *IEEE Transactions on Knowledge and Data Engineering*, 2022, pp. 1–12.

[9] B. Li, Y. Jin, X. Yu, L. Song, J. Zhang, H. Sun, H. Liu, Y. Shi, and F. Kong, "MVIRA: A model based on Missing Value Imputation and Reliability Assessment for mortality risk prediction", *International Journal of Medical Informatics*, Vol. 178, 2023, p. 105191.

[10] O. Ben-Assuli, T. Heart, R. Klempfner, and R. Padman, "Human-machine collaboration for feature selection and integration to improve congestive Heart failure risk prediction", *Decision Support Systems*, Vol. 172, 2023, p. 113982.

[11] T. You, Q. Dang, Q. Li, P. Zhang, G. Wu, and W. Huang, "TransLSTD: Augmenting hierarchical disease risk prediction model with time and context awareness via disease clustering," *Information Systems*, Vol. 124, 2024, p. 102390.

[12] E. K. Oikonomou and R. Khera, "Machine learning in precision diabetes care and cardiovascular risk prediction", *Cardiovascular Diabetology*, Vol. 22, No. 1, 2023.

[13] D. D. Himabindu and S. P. Kumar, "A Survey on Computer Vision Architectures for Large Scale Image Classification using Deep Learning", *International Journal of Advanced Computer Science and Applications*, Vol. 12, No. 10, 2021.

[14] D. H. Damineni, P. K. Sekharamantry, and R. Badugu, "An Adaptable Model for Medical Image Classification Using the Streamlined Attention Mechanism", *International Journal of Online and Biomedical Engineering (iJOE)*, Vol. 19, No. 16, 2023, pp. 93–110.

[15] N. Narisetty, A. Kalidindi, M. V. Bujaranpally, N. Arigela, and V. V. Ch, "Ameliorating Heart Diseases Prediction using Machine Learning Technique for Optimal Solution", *International Journal of Online and Biomedical Engineering (iJOE)*, Vol. 19, No. 16, 2023, pp. 156–165.

[16] P. N. Srinivasu, N. Sandhya, R. H. Jhaveri, and R. Raut, "From Blackbox to Explainable AI in Healthcare: Existing Tools and Case Studies", *Mobile Information Systems*, Vol. 2022, 2022, pp. 1–20.

[17] Q. Li, X. Yang, J. Xu, Y. Guo, X. He, H. Hu, T. Lyu, D. Marra, A. Miller, G. Smith, and S. DeKosky, "Early prediction of Alzheimer's disease and related dementias using real-world electronic health records", *Alzheimer's & Dementia*, Vol. 19, No. 8, 2023, pp. 3506–3518.

[18] I. Boudali, S. Chebaane, and Y. Zitouni, "A predictive approach for myocardial infarction risk assessment using machine learning and big clinical data," *Healthcare Analytics*, Vol. 5, 2024, p. 100319.

[19] R. AlSaad, Q. Malluhi, A. Abd-alrazaq, and S. Boughorbel, "Temporal self-attention for risk prediction from electronic health records using non-stationary kernel approximation," *Artificial Intelligence in Medicine*, Vol. 149, 2024, p. 102802.

[20] F. Yang, J. Zhang, W. Chen, Y. Lai, Y. Wang, and Q. Zou, "DeepMPM: a mortality risk prediction model using longitudinal EHR data", *BMC Bioinformatics*, Vol. 23, No. 1, 2022.

[21] S. Niu, Q. Yin, J. Ma, Y. Song, Y. Xu, L. Bai, W. Pan, and X. Yang, "Enhancing healthcare decision support through explainable AI models for risk prediction", *Decision Support Systems*, Vol. 181, 2024, p. 114228.

[22] Y.W. Chen, Y.J. Li, P. Deng, Z.Y. Yang, K.H. Zhong, L.G. Zhang, Y. Chen, H.Y. Zhi, X.Y. Hu, J.T. Gu and J.L. Ning, "Learning to predict in-hospital mortality risk in the intensive care unit with attention-based temporal convolution network", *BMC Anesthesiology*, Vol. 22, No. 1, 2022.

[23] X. Lin, X. Pan, Y. Yang, W. Yang, X. Wang, K. Zou, Y. Wang, J. Xiu, P. Yu, J. Lu, and Y. Zhao, "Machine learning models to predict 30-day mortality for critical patients with myocardial infarction: a retrospective analysis from MIMIC-IV database", *Frontiers in Cardiovascular Medicine*, Vol. 11, 2024.

[24] S. Lin, W. Lu, T. Wang, Y. Wang, X. Leng, L. Chi, P. Jin, and J. Bian, "Predictive model of acute kidney injury in critically ill patients with acute pancreatitis: a machine learning approach using the MIMIC-IV database", *Renal Failure*, Vol. 46, No. 1, 2024.

[25] D. S. Rao, F. J. Shaikh, "Prediction of Cancer Disease using Machine learning Approach," *Materials Today: Proceedings,* Vol. 50, No. 1, 2022, pp. 40-47.

[26] MIMIC-III dataset: https://paperswithcode.com/dataset/mimic-iii (Accessed in December 2024)

[27] MIMIC-IV dataset: https://paperswithcode.com/dataset/mimic-iv (Accessed in December 2024)