IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

SPECIAL SECTION ON EMERGING DEEP LEARNING THEORIES AND METHODS FOR BIOMEDICAL ENGINEERING

Received July 13, 2020, accepted July 19, 2020, date of publication July 24, 2020, date of current version August 5, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3011721

A Clinical Prediction Model in Health Time Series Data Based on Long Short-Term Memory Network Optimized by Fruit Fly Optimization Algorithm

WEIJIA LU[1], (Member, IEEE), LIANG MA[2], HAO CHEN[2], XIAOJUAN JIANG[2], AND MING GONG[2]

[1]Science and Technology Department, Affiliated Hospital of Nantong University, Nantong 226001, China
[2]Information Department, Affiliated Hospital of Nantong University, Nantong 226001, China

Corresponding author: Weijia Lu (luweijia.good@163.com)

ABSTRACT Aiming the problems that the clinical data of different patients is difficult for reasonable representation and the time interval between medical events is different, which lead to the difficulty of clinical prediction, a clinical prediction model based on the long short-term memory (LSTM) network optimized by fruit fly optimization algorithm in health time series data is proposed. First, FastText method is used to represent the interpretable vector of medical events, which can extract the concept relationship rich in medical information more effectively. Then, considering the strong dependence of clinical data on time stamp, LSTM network is used to model clinical events for better extraction of long-term and short-term information, so as to improve the prediction performance of the model. Finally, the fruit fly optimization algorithm is used to find the optimal super parameters of LSTM network, which can improve the training efficiency and prediction precision of the network. Experimental results on MIMIC datasets show that the prediction precision, Recall@k and MAP@k of the proposed model are better than those of other models. The validity of the model is proved.

INDEX TERMS Fruit fly optimization algorithm, LSTM network, FastText method, clinical prediction, health time series data, MIMIC dataset.

I. INTRODUCTION

Healthy physique is the biggest wealth and core competitiveness of people. With the continuous improvement of material living standards, people has paid more and more attention to their own health and the requirements of medical service level are constantly improved [1]. The medical requirement has gradually changed from ''getting sick first and then treating'' to ''early screening and prevention of diseases, early detection and early treatment, and personalized diagnosis''. It requires that medical services can timely detect, analyze and evaluate individual's physical health status, provide personalized health consultation and guidance, and prevent disease in the bud. Nowadays, the development of medical information system is rapid. With the wide application of electronic medical record (EMR), a large number of methods for mining and predicting clinical time series data have emerged [2]. The massive electronic medical record data of a large number of patients needs further mining and research, and the information contained in it can bring a lot of convenience for the future auxiliary diagnosis. Based on the electronic health record system, using information technology to quickly process the massive medical data, an intelligent clinical decision support system (CDSS) is built to provide clinicians with clinical supports such as advice, reminder, alarm, prediction and so on [3], [4]. Electronic health record (EHR) is the digital and text information generated by medical personnel using medical information system in the medical activities. It records various physiological indexes related to health in the whole life process of human beings, and contains a large amount of valuable medical knowledge and health information. Therefore, EHR is a valuable research resource in clinical medical prediction [5].

However, the original medical data exists many problems, such as multi-dimensional, sparsity, irregularity, bias ratio (BIAS) and so on. The clinical data of different patients are often expressed in the form of medical data sequences with different time intervals, and they are strongly dependent on

The associate editor coordinating the review of this manuscript and approving it for publication was Yizhang Jiang.

time stamps [6]. Therefore, it is difficult to directly apply traditional machine learning or statistical models to predict the clinical endpoint of patients. Obtaining information from complex, high-dimensional and heterogeneous biomedical data is a key challenge for prediction. In modern biomedical research, various types of data, such as electronic health records, medical images and texts, have emerged. These data are complex, heterogeneous, lack of annotation and usually unstructured [7]. Among them, time series is an important feature. The analysis of time series data is an active research topic in recent decades. It is considered as one of the ten difficult problems in data mining because of its unique property [8].

The electronic health record data in the clinical medical prediction is very similar to the user interaction data in the recommendation system. Therefore, the application of the recommendation system in the clinical medical prediction can be explored [9]. First, their data has features of sparse and default [10]. Second, their data are potentially mutagenic [11]. At the same time, the clinical medical prediction and recommendation system also face the problem of personalization. Because the electronic health record contains all kinds of physiological index information collected from each patient in the treatment process, it has particularity and complexity [12]: On the one hand, due to the unique physical condition of each patient, the slight difference of information contained in their electronic health records may be caused by completely different diseases, so it is necessary to conduct a more in-depth analysis of the patients' electronic health records and make more targeted and personalized prediction; on the other hand, because each physiological index is not completely independent, they are interrelated and influence each other, so different combinations of abnormal index information may correspond to completely different diseases [13]. Although various algorithms have been widely used in the auxiliary diagnosis based on electronic medical record data, and have achieved better results, most of them are only focused on the diagnosis and prediction of a single disease. It is very unrealistic to develop a specific model for each disease. Therefore, there is still a lack of effective methods for a wide range of clinical event prediction problems [14].

The rest of this paper is arranged as follows: Chapter 2 introduces the related research, introduces the research status in this field; Chapter 3 defines the problems to be solved, and then explains the overall framework of the algorithm in this paper. Chapter 4 is the core part of this paper. It systematically introduces the proposed LSTM network based on Drosophila optimization algorithm; Chapter 5 verifies the proposed method by experiment; Chapter 6 summarizes the algorithm in this paper and points out the future research direction.

II. RELATED WORKS

In recent years, most of the techniques for analyzing electronic health record data are based on traditional machine learning and statistical techniques, such as logistic regression, support vector machines (SVM) and random forest. In the recommendation system, the algorithms based on deep neural network have been widely used to solve the problem of personalization [15]. Deep neural network is a kind of deep computing structure. The number of hidden layers is usually large. This kind of multi-layer calculation can continuously mine and discover the deeper data features. The deep neural network can automatically extract the fusion features of different index information, and obtain more accurate prediction results. Using deep neural network to automatically extract fusion features fully considers the combination of physiological indexes of patients. It can predict the current physical condition of patients more accurately and provide personalized diagnosis and prediction service [16]. In clinical medical prediction, the main four prediction problems are: disease deterioration prediction, in-hospital mortality prediction, length of hospital stay prediction, disease classification prediction [17].

Disease deterioration prediction: During hospitalization, according to the physiological detection index, the patient's physical condition is diagnosed, and the rapid changes of the patient's condition are found in time, which automatically gives early warning and triggers the alarm to inform the medical staff. There are various traditional methods for disease deterioration prediction. Many hospitals widely use the ''tracking-trigger'' mechanism. For example, based on the biological characteristics in time domain, reference [18] proposed a resource optimization model using utility function for clinical information transmission of IoMT. Once the early warning score of patients is too low or some physiological detection indexes exceed the threshold value, the early warning alarm will be triggered, and special nursing care will be provided for the patients with deteriorating condition to improve their condition [19]. Reference [20] proposed a prototype solution for public health monitoring data of dashboard based on real use cases, which aims to enhance clinical and policy decision-making. This solution can collect health monitoring data, store and visualize it to inform the patients of the predicted deterioration. However, although the physiological indexes of some patients did not exceed the threshold, the patients' condition deteriorate, resulting in the failure to give patients effective treatment in time, that is, unable to achieve personalized diagnosis.

In-hospital mortality prediction: The high-risk patients can be determined by predicting the mortality rate in the early stage of hospitalization. The research on predicting the mortality rate of patients can be traced back to more than half a century ago. The main method is to calculate Apgar score according to the physiological index data within 24 hours after admission [21]. As the early warning scoring mechanism for predicting the deterioration of patients' condition, the model focuses on the physiological detection index data within 24 hours after admission and the abnormal physiological detection index data of patients. Reference [4] proposed a time-based mortality prediction model. It only predicted the mortality at the final time, but could not predict the

real-time mortality of patients. Many researchers used artificial neural network to predict mortality rate, and achieved better prediction results, but the convergence speed of artificial neural network was slow and the training time was long. Reference [22] proposed a transboundary model for clinical decision support. This model took the concept of work practice into account. As a designed function, it could realize context-aware information sharing and safe health data management in transboundary clinical decision support. At present, neural network has not been widely used to detect the deterioration of disease or the continuous prediction of mortality.

Length of hospital stay prediction: The hospital measures the severity of patients' illness and reasonably allocates hospital resources by predicting the length of hospital stay. Most of the studies on length of stay focus on determining the factors that affect the length of stay, rather than establishing prediction models [23]. Both disease severity score and early warning score were used to predict the length of hospital stay and achieved good results. Reference [24] provided an open data integration platform for clinical medical data of patients across multiple health information systems, which could accommodate and integrate other heterogeneous data sources, and was conducive to the centralization of data assets. This centralization enabled every stakeholder in a patient-centered care environment to participate actively in decision-making. The neural network model used to predict the length of hospital stay of patients is usually a two-classification model, which aims to find out the patients who are hospitalized for a long time. But it cannot accurately predict the length of hospital stay of patients, so it cannot timely and reasonably schedule medical resources.

Disease classification prediction: Disease classification is a relatively new medical informatics problem, which has aroused the interest of machine learning researchers. Upton defined disease classification as a multi-output classification problem. Aiming at the clinical medical time series data with variable length, recurrent neural network (RNN) realizes the classification and combination of more than 100 different diseases by mining the commonness of diseases in the hidden layer [25]. Reference [26] simulated the diagnosis behavior of doctors to classify diseases based on the long short-term memory network (LSTM).

At present, most of the studies are carried out on specific datasets, which are lack of generality, and they have not solved the above four clinical medical prediction problems at the same time [27]. Recommendation systems are mainly divided into collaborative-filtering based recommendation system, content-based recommendation system and hybrid recommendation system. However, these algorithms have their own limitations in dealing with the problem of data sparsity and weighing the recommendation quality on different evaluation criteria. Reference [28] explored and constructed a two-way long short-term memory neural network to predict the results of blood culture test. Based on the time calculation model, this method used nine clinical parameters measured over time to predict the clinical health condition. Compared with the traditional logistic regression model, the prediction effect is significantly improved. This method effectively solves the above four clinical medical prediction problems and makes it a recognized benchmark evaluation model.

LSTM is a kind of recurrent neural network. Considering that the time series data of electronic health record are modeled in LSTM, it can accurately predict the trend of the patient's physical condition with time, and effectively improve the prediction precision. At the same time, the traditional LSTM model can extract long short-term information, so it is often used to deal with the problem of sequence prediction. However, for the data with different time intervals, the model still has some limitations, and it does not fully consider the cross combination of physiological index characteristics of patients, which needs further improvement. In many works, the information of diseases and drugs is represented by a one-hot vector, which loses the rich medical meaning in the vector space and needs further representation learning [29].

Based on the above analysis, in order to solve the problem that the clinical data of different patients are difficult to be represented reasonably and the time interval between clinical events is different, a clinical prediction model on health time series data based on long short-term memory network optimized by fruit fly optimization algorithm is proposed [30]. The basic ideas are as follows: ① FastText method is adopted for interpretable vector representation of medical events to capture the concept relationship rich in medical information more effectively; ② In view of the strong dependence of clinical data on time stamp, LSTM network is used to model clinical events. Moreover, LSTM can better capture long short-term information to improve the prediction performance of the model; ③ Fruit fly optimization algorithm is used to search the optimal super parameters of LSTM network and improve the training efficiency and prediction precision of the network. The innovations of the proposed method are as follows:

(1) In order to improve the prediction precision of the model, the proposed model uses FastText model to express medical concepts. Through vector representation, the distance between the concepts with similar semantics at the spatial level is smaller, so as to obtain more accurate prediction of the diseases with fewer occurrences.

(2) Considering the strong dependence of clinical data on time stamp, the proposed model uses LSTM network to model clinical events, which can better capture long short-term information and improve the prediction performance of the model.

(3) Because the number of hidden layers and learning rate of LSTM have the greatest impact on precision of the network, the fruit fly optimization algorithm is used to find the optimal LSTM network super parameters for improvement of the training efficiency and prediction precision of the network.

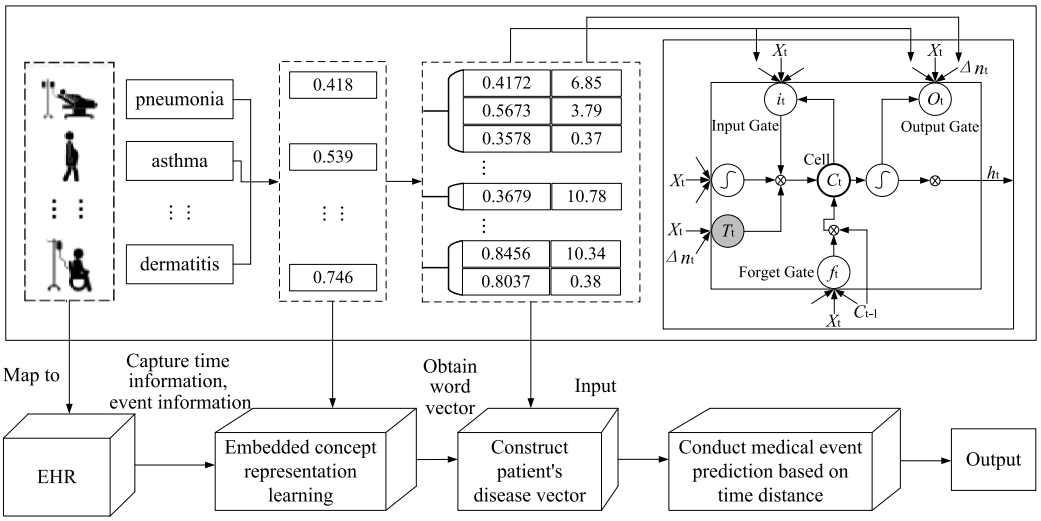

FIGURE 1. Clinical event representation and prediction of the proposed model.

III. PROBLEM DEFINITION AND GENERAL FRAMEWORK

The general framework of the proposed model is shown in FIGURE. 1. First, the embedded representation of medical concepts is trained to make the data has certain medical semantic information, and then the trained vector is used to construct the patient representation. The constructed vectors are fed into the proposed model for prediction.

Definition 1: Define $P = \{p_1, p_2, p_3, \ldots, p_k\}$ as the set of all patients in the record, $D = \{d_1, d_2, d_3 \ldots d_j\}$ as the clinical records in the dataset. For each patient, the diagnosis record is given by a tuple $r_i = (d_i, t_s)$, where d_i is the medical event recorded at time t_s. Because the probability of current events is closely related to the interval of prior events, the time interval of each event is calculated and recorded as E_t. The tuple r_i is rewritten as $r_i = (d_i, E_{t_i})$.

Definition 2: Define E_j as the set of events that occurred in patient p on the jth hospitalization, namely, $E_j = [(d_1, E_{t1}), (d_2, E_{t2}) \cdots (d_f, E_{tf})]$. The hospitalization record of each patient is given by $H_p = [E_1, E_2 \ldots E_j \ldots]$. If the patient is diagnosed with multiple diseases at the same time, the same time stamp is given.

In summary, the task is to predict the possible medical events of the patient p_m in time t_s through the historical information in the electronic medical record [31].

IV. THE PROPOSED PREDICTION MODEL

A. WORD VECTOR REPRESENTATION

In order to make the proposed model better deal with events and their time relationships, inspired by FastText method, medical events are preprocessed with embedded word vectors, and different real value vectors are assigned to each event, so as to obtain low dimensional and rich semantic expression. Finally, a multi-dimensional matrix is obtained to effectively capture the potential relationships between events.

FastText model is a natural language processing model based on Skip-gram, which considers the internal structure of words. Since similar concepts of diseases often have similar lexical structures, such as lymphadenitis and lymphoma, the FastText model with subword information is more suitable for medical event representation.

1) SKIP-GRAM MODEL

Skip-gram model predicts the words that appear in the context through the selected target words. That is, the Skip-gram model selects the target words and tries to predict its neighbors. Moreover, it continuously selects the target and predicts the context by sliding the context window [32]. The task is to optimize the logarithmic likelihood function:

$$L = \sum_{i=1}^{N} \sum_{c \in W_i} \log p\left(w_c \mid w_i\right) \tag{1}$$

where N is the length of medical concept vector to be trained, c is the size of sliding window, w_i is central concept word, W_i is the context word set of w_i, w_c is the context concept word of w_i. If the function that defines the similarity of two words is s, Softmax function can be used to define $p\left(w_c \mid w_i\right)$:

$$p\left(w_c \mid w_i\right) = \frac{e^{s(w_i, w_c)}}{\sum_{j=1}^{M} e^{s(w_i, w_j)}} \tag{2}$$

where M is total number of medical concepts, s is word similarity.

The similarity function is defined by inner product of word vector, as follows:

$$s\left(w_c \mid w_i\right) = \mathbf{u}_{w_i}^{N} \mathbf{v}_{w_c} \tag{3}$$

where $\mathbf{v}$ and $\mathbf{u}$ is the vector representation of w_c and ω_t, c_i is the context word set of ω_i.

Therefore, the task of predicting context words can be decomposed into a group of independent binary classification tasks. The goal of binary classification task is to predict whether a context word exists [33]. For the word at position l,

all the context words are regarded as positive examples, and negative examples are randomly selected from the vocabulary for training. For the selected context location c, the objective function is solved to obtain the optimal solution, which is the final vector representation. The mathematical expression is as follows:

$$J=\sum_{i=1}^{N}\left[\sum_{c\in C_t}\log\left(1+e^{-s(w_i,w_c)}\right)+\sum_{n\in N_{i,c}}\log\left(1+e^{s(w_i,w_c)}\right)\right] \tag{4}$$

2) FASTTEXT MODEL

In medical vocabulary, many words have the same or similar prefixes and suffixes. Considering the internal structure of words in the expression of word vector, an intuitive method is to divide each word into character level n-gram to express. The words are marked with "<" or ">", and the word vector are calculated with the vector sum of n-gram. Taking "lymphadenitis" and "lymphoma" as examples, if $n = 3$, the two words can be expressed as follows:

Lymphadenitis: <1y,1ym, ymp, mph, pha, had, ade, den, eni, nit, iti, tis, is> <lymphadenitis>

Lymphoma: <1y, 1ym,ymp, mph, pho, hom, oma, ma><1ymphoma >

Many subwords in above two words have the same information. This representation can capture the similar relationship more effectively and further capture the similar medical concept information. At the same time, because some rare concepts may have the same substructure as common concepts, the expression of subwords further increases the precision of training and improves the prediction performance [34].

Supposing that the size of subword dictionary is G, $G_w \subset \{1, \ldots, G\}$ represents the n-gram that appears in the given word w, and vector z_g contains each g in n-gram. The word is then represented by the sum of the vector z_g. The function is redefined as:

$$s\left(w_c, w_i\right) = \sum_{g\in G_{w_t}} z_g^N \mathbf{v}_{w_c} \tag{5}$$

FastText allows the shared representation of words, so that the diseases with low incidence in electronic medical record data can also be reliably represented to obtain accurate vector expression.

B. LSTM NETWORK BASED ON FRUIT FLY OPTIMIZATION ALGORITHM

1) LONG SHORT-TERM MEMORY NETWORK PREDICTION ALGORITHM

LSTM is an improved algorithm of recurrent neural network. It uses the memory gate to effectively solve the problem of gradient disappearing, and can learn the features of nonlinear long-term data. The basic network unit is shown in FIGURE. 2.

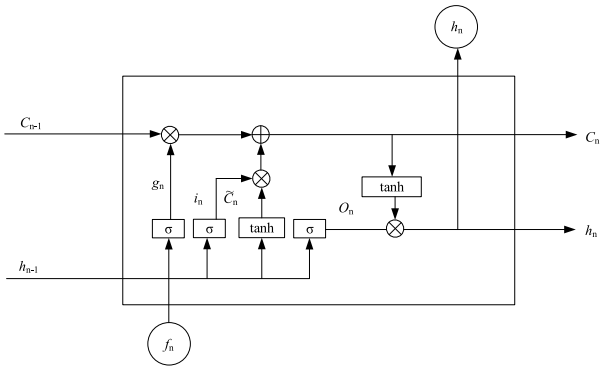

FIGURE 2. Structure of LSTM network unit.

The expression of feature mapping in the basic unit is:

$$\begin{cases} g_n = \sigma\left(\vartheta_f \cdot [h_{n-1}, d_n]\right) + b_f \\ i_n = \sigma\left(\vartheta_i \cdot [h_{n-1}, d_n]\right) + b_i \\ C_n = \tan\left(\vartheta_c \cdot [h_{n-1}, d_n] + b_c\right) \end{cases} \tag{6}$$

where $\vartheta_f, b_f, \vartheta_i, b_i, \vartheta_c, b_c$ are training parameters of network; σ is Sigmoid function; i, g, c respectively represent input gate, output gate and memory unit. A "forget gate" is designed in the memory unit to solve the problem of gradient disappearance due to data dependence by changing the weights of forget and input. The details are as follows:

$$C_n = g_n \cdot C_{n-1} + i_n \cdot \tilde{C}_n \tag{7}$$

The data in this algorithm is generally divided into training value and observed value. The training value is used for training the network, and the observed value is used to compare with the predicted value. For the data with simple change, the prediction precision of this method is relatively high. But for complex data, the prediction precision will be greatly reduced. In order to improve the adaptability of the algorithm, the observed values should be used to update the network in real time [35]. In other words, the current observations are always used to train the network at each time, and the predicted values of the next time are obtained under the new network.

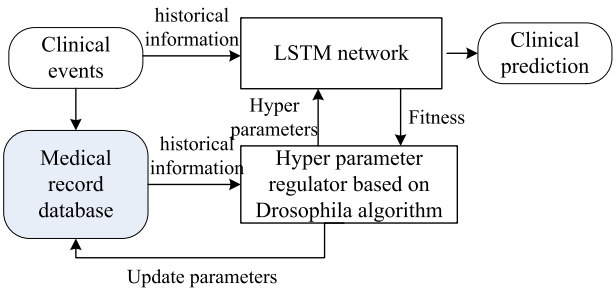

FIGURE 3. The proposed prediction flowchart.

2) CLINICAL PREDICTION ALGORITHM BASED ON LSTM NETWORK OPTIMIZED BY FRUIT FLY OPTIMIZATION ALGORITHM

FIGURE. 3 shows the proposed prediction flowchart. The database is updated in real time by the health time series data

in the electronic medical record and trained offline to obtain the optimal LSTM network parameters.

The health time series data are sent to the LSTM network optimized by fruit fly optimization algorithm for training. The medical record information is input to update the network at the moment, and the clinical prediction value at the next moment is obtained.

C. SUPER PARAMETER SELECTION

The selection of super parameters is very important for the training efficiency and prediction precision of the network. The number of units in the hidden layer and the learning rate have the greatest impact on LSTM network [36]. There are two steps in the selection of super parameters: 1) determine the variation range of the parameters according to the clinical types and prediction requirements; 2) use the fruit fly optimization algorithm to find the optimal number of hidden units and learning rate within a given range.

1) FRUIT FLY OPTIMIZATION ALGORITHM

Fruit fly optimization algorithm is a kind of heuristic algorithm that imitates fruit fly foraging behavior to obtain the global optimal value. Due to the advantages of swarm intelligence optimization algorithms in optimization speed and parameters, it is suitable for adjusting super parameters.

The idea of fruit fly optimization algorithm is to control the fruit fly population step by step in the solution space according to the fitness function (smell concentration decision function) [37]. It usually contains four steps:

(1) Initialization: set the initial parameters of fruit fly population, including the size of fruit fly population, the maximum number of iterations, the initial position and the step length of individual fruit fly to find the target, namely the random direction and distance of fruit fly flight:

$$\begin{aligned} X\ (i) &= X_0 + Step \\ Y\ (i) &= Y_0 + Step \end{aligned} \tag{8}$$

where X_0 and Y_0 are the initial position of fruit fly.

(2) Judgment: according to the fitness function, calculate the smell concentration (Smell) of fruit fly position

$$\begin{aligned} Smell\ (i) &= \text{Function}\ (S\ (i)) \\ S\ (i) &= \frac{1}{Sqrt\ (X(i)\hat{}2 + Y(i)\hat{}2)} \end{aligned} \tag{9}$$

(3) Movement: select the individual with the highest concentration in the fruit fly population, record the position of the individual as optimal position, and command the remaining fruit flies to move toward the optimal position according to the initial step length.

(4) Iteration: repeat the step (2) and (3) until the smell concentration meets the set conditions or reaches the maximum number of iterations. The fitness function selects root mean square error (RMSE), which is defined as:

$$\min \delta_R = \sqrt{\sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2 / n} \tag{10}$$

where δ_R is root mean square; y_i is discrete position data to be processed; $\hat{y}_i$ is predicted value of position; n is number of data.

2) OPTIMAL SELECTION OF LSTM SUPER PARAMETERS BASED ON FRUIT FLY OPTIMIZATION ALGORITHM

The super parameters are determined according to the historical data, that is, the optimal value of the super parameters is selected offline [38], [39]. The pseudo code of the LSTM super parameter selection algorithm based on fruit fly optimization is shown in Algorithm 1.

Algorithm 1 Pseudo Code of Optimal Selection of LSTM Super Parameters Based on Fruit Fly Optimization

Begin

1. According to the recognition results of health time series data in electronic medical record, determine the basic search range of unit number and learning rate by referring to historical data.
2. **for** $i = i + 1$ **do**
3. **if** $i \le n$ or $\delta_R \ge \delta_{R\min}$, **then**
4. Initialize the population of fruit fly and assign different super parameters to fruit fly individuals.
5. Run LSTM offline and calculate the prediction error:

$$\min \delta_R = \sqrt{\sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2 / n}$$

6. Evaluate the super parameter fitness represented by fruit fly individual, and calculate the smell concentration of each individual:

$$\begin{aligned} Smell\ (i) &= \text{Function}\ (S\ (i)) \\ S\ (i) &= \frac{1}{Sqrt\ (X(i)\hat{}2 + Y(i)\hat{}2)} \end{aligned}$$

7. Select the individual with the best fitness function as the moving target of the group.
8. Other fruit fly individuals adjust their positions according to the step length.
9. **end if**
10. Update the super parameters for this kind of clinical prediction in the database.

End

V. EXPERIMENTAL RESULTS AND ANALYSIS

Experimental environment is based on the Keras library implemented on NVIDIA GeForce Titan X GPU with Python. The experiments uses RMSprop mini-batch method, each mini batch includes 128 training samples. Dataset is divided into three groups, including training set, validation set and test set. 80% of samples in the dataset are used as the training set, 10% as the validation set to adjust the super parameters, and the rest as the test set.

A. EXPERIMENTAL SETUP

Medical information mart for intensive care (MIMIC) is an openly available dataset developed by the MIT Lab. This dataset records the clinical data of 46520 patients and 58976 hospital records in the intensive care unit of Beth Israel Deaconess Medical Center. It covers clinical diagnosis, vital signs, laboratory tests and other information.

On this dataset, the patients with more than two admission records and the code of their diagnosis records are extracted, thus obtaining the admission records of 7537 patients, with an average of 2.65 hospitalized times per person. The distribution is shown in FIGURE. 4.

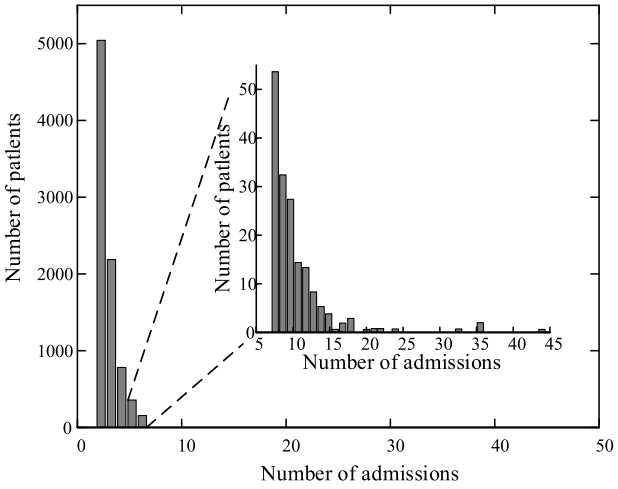

FIGURE 4. Statistic of hospitalized times of patients.

The above data constitute the initial dataset, and a multi-level nested list is constructed to predict the diagnosis category in the next visit. The nested list is as follows:

$$H_p^o = \begin{bmatrix} \left[\left(d_1^1, \Delta t_1^1\right), \left(d_2^1, \Delta t_2^1\right) \ldots \left(d_f^1, \Delta t_f^1\right)\right], \\ \left[\left(d_1^2, \Delta t_1^2\right), \left(d_2^2, \Delta t_2^2\right) \ldots \left(d_g^2, \Delta t_g^2\right)\right] \\ \ldots, \left[\left(d_1^p, \Delta t_1^p\right), \left(d_2^p, \Delta t_2^p\right) \ldots \left(d_h^p, \Delta t_h^p\right)\right] \end{bmatrix} \tag{11}$$

where p and o are number of patients and hospitalized times, $p = 7537$, $o = 19993$. d and t are clinical diagnostic events and time interval.

In the fruit fly optimization algorithm, the population size of fruit fly is 20 and the maximum iterations are 50. After the optimization parameters are normalized, the fixed moving step length is [0.01, 0.01]. In the LSTM network, in order to ensure the speed and precision of training, the maximum number of training is 200. The learning rate and the number of hidden layer units are determined by the optimization algorithm, where the learning rate range is within [0.001, 0.5]and hidden layer number is within [50, 500].

The diagnosis records of each patient are extracted and sent to the FastText model in chronological order. In order to represent the concept of disease more accurately and reduce the calculation time, the window size is set to 5, and a 128-dimensional vector is selected for vector representation. T-distributed stochastic neighbor embedding (T-sne) is a very popular algorithm to reduce the dataset from multi-dimensional to two-dimensional or three-dimensional, which is used for dimension reduction of word vector.

B. EVALUATION INDEX

The precision rate Pr and recall rate Re are used to evaluate the performance of the clinical prediction model. For target user u_i, Pr and Re are defined as:

$$\begin{aligned} Pr &= \frac{TP}{TP + FP} \\ Re &= \frac{TP}{TP + FN} \end{aligned} \tag{12}$$

where TP represents the number of positive samples predicted to be positive, FP represents the number of negative samples predicted to be positive and FN represents the number of positive samples predicted to be negative.

Since the clinical prediction problem is essentially a sequence prediction problem, the Recall@k method and MAP@k (mean average precision) method are adopted to evaluate the performance.

Recall@k: For a sequence predicted by the model, if there are m_i correct diagnosis records in the first k records, and n_i positive records in the original diagnostic sequence of a patient, define $Recall@k = \frac{m_i}{n_i}$. The Recall@k of each patient in the dataset is calculated and averaged. The average value is the recall rate of the model.

MAP@k (mean average precision): This index reflects the prediction performance of the model in the sequence. If the correct diagnosis predicted by the model is in the more front position, the MAP@k value is the higher. If the model cannot predict the correct diagnosis, the MAP@k is 0.

Besides, considering the imbalance of data, area under the receiver operation characteristic curve (AUROC) is adopted as the performance evaluation standard of the model. AUROC is widely used to measure the performance of binary classifiers. For each prediction model, cross validation AUROC is calculated, and its average value reflects the performance and confidence interval of the model.

C. PARAMETER PERFORMANCE ANALYSIS

In the proposed method, the learning rate ϑ used to adjust LSTM training has a great influence on performance, so it needs to be analyzed in depth. The value of ϑ varies from 0 to 1. The precision, Recall@5 and MAP@5 of the proposed method are shown in FIGURE. 5, where the original LSTM is compared with the LSTM network optimized by fruit fly optimization algorithm.

It can be seen from FIGURE. 5 that when $\vartheta = 0$, the network lacks learning, so the precision is the lowest. Similarly, when $\vartheta = 1$, the precision is reduced due to the frequent variation in the network training. Therefore, when $\vartheta = 0.5$,

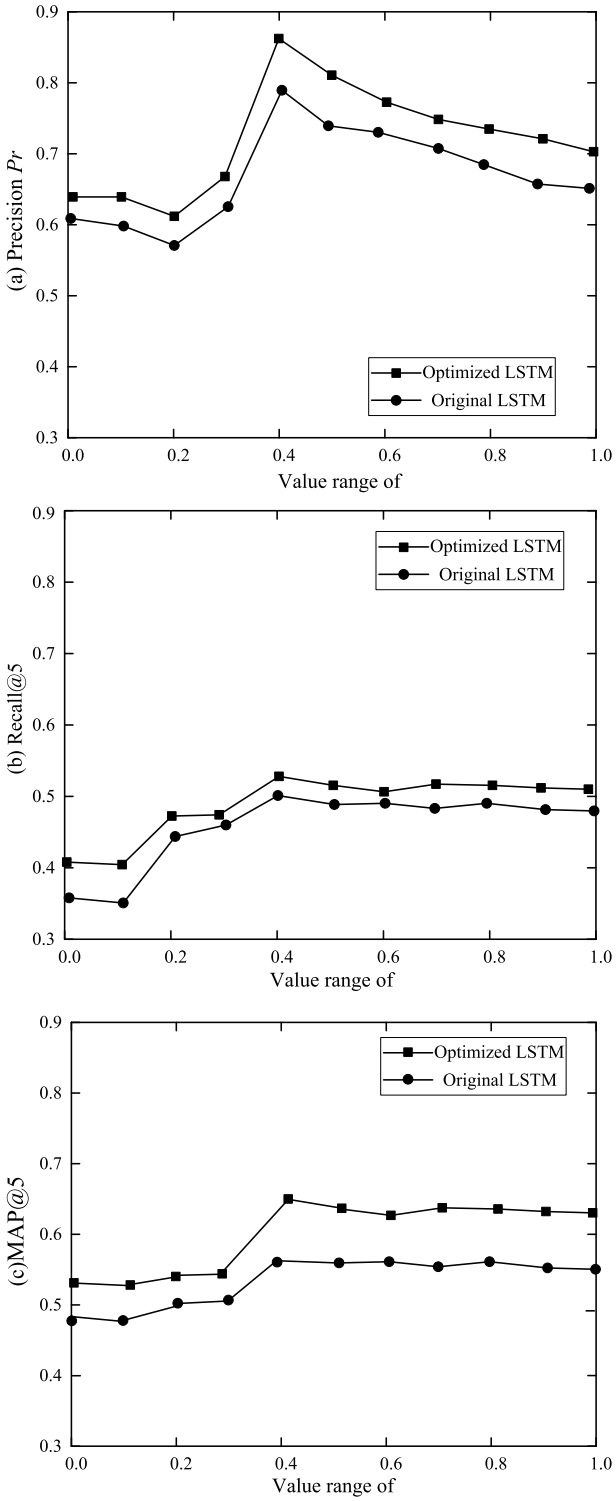

FIGURE 5. The influence of parameter ϑ on the performance of the proposed method.

the precision of the proposed method is the best. Taking the patient history information and other factors into account, the proposed method has high precision and low recall rate, and the performance is ideal. In addition, the performance of the proposed method after the super parameters optimization is better than that of the original LSTM.

D. RESULTS COMPARISON AND ANALYSIS

1) COMPARISON OF RESULTS IN DIFFERENT TIME WINDOWS

Since the proposed model needs to be combined with the current data information for prediction, the data at different times will have a certain impact on the prediction performance. Therefore, the proposed method is compared with the model in reference [20], [22] and [28] in different time windows. The results are shown in Table 1, in which the time windows are set as 1h, 6h and 12h respectively.

TABLE 1. Indexes of different time windows.

Index	time windows	Ref. [20]	Ref. [23]	Ref. [29]	Proposed method
Precision	1h	0.683	0.732	0.801	0.871
	6h	0.541	0.658	0.685	0.805
	12h	0.423	0.530	0.602	0.746
Recall@5	1h	0.721	0.739	0.812	0.875
	6h	0.607	0.638	0.741	0.817
	12h	0.526	0.589	0.653	0.756
MAP@5	1h	0.693	0.752	0.827	0.883
	6h	0.556	0.671	0.695	0.814
	12h	0.437	0.548	0.619	0.756

As can be seen from the above Tab. 1, compared with other models, all indexes of the proposed model are the highest, so it has better performance. In addition, with the increase of time window, the prediction performance of each model decreases slightly. That maybe because after a long period of time, the medical conditions may change in different directions, either worse or better, so it is difficult to accurately predict. With the passage of time, the prediction performance of the model decreases, but the decrease of the proposed model is the smallest, which demonstrates the effectiveness of the proposed model and it can be used in clinical prediction of health time series data.

2) RESULTS COMPARISON OF RECALL@K AND MAP@K

The proposed model was compared with the model in reference [20], [22] and [28] on health time series data for many times to calculate Recall@k and MAP@k. The results are shown in FIGURE. 6.

As can be seen from the FIGURE. 6, when k = 20, the Recall@20 and MAP@20 of the proposed model are the highest. The proposed model uses FastText method for data preprocessing and fruit fly algorithm to optimize LSTM network parameters, so the validity of the model can be proved. The performance of reference [22] is slightly poor at the beginning, but with the increase of K value, satisfactory results are obtained. Compared with reference [28] used the original LSTM model, Recall@20 increases by 17.06%. In reference [22], three-layer automatic encoder was used to generate patient vectors, which was combined with logistic regression classification to predict disease diagnosis

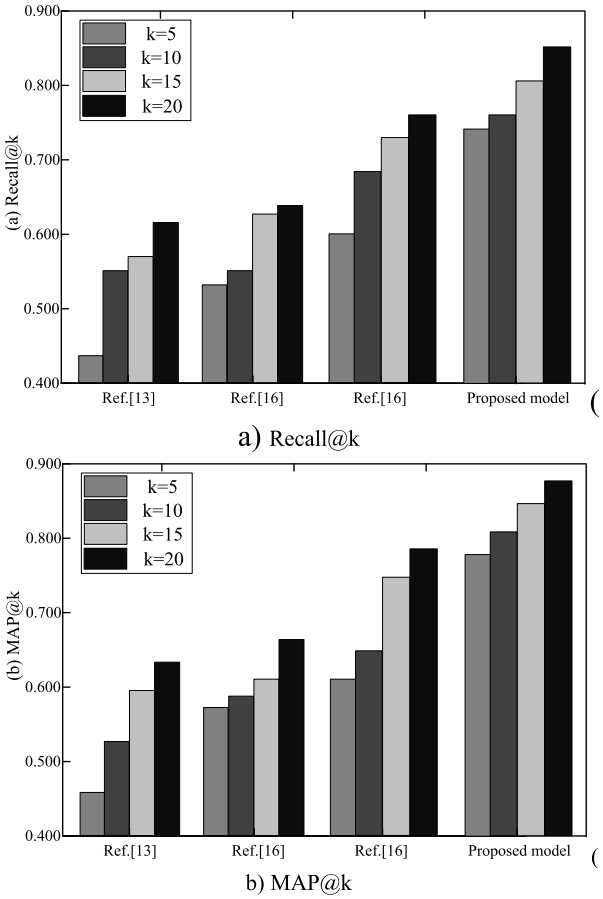

FIGURE 6. Comparison of model performance under different k values.

in a window based on ICD9. Compared with reference [1], the MAP@15 of [22] increases by 3.7%. It can be further demonstrated that the proposed model has better prediction performance at any k value.

3) AUROC RESULTS COMPARISON

Including 64 features for feature selection learning, the AUROC values (mean value, 95% confidence interval) of different prediction models were shown in FIGURE. 7. Among them, 32 features were included in the model in reference [20], 35 features were selected from the model in reference [22], and 38 features were learned in the model in reference [28].

As can be seen from the FIGURE. 7, the performance of traditional LSTM network in reference [28] is poor, and the AUROC is only 0.72 [0.64, 0.81]. Reference [22] uses three-layer automatic encoder combined with various factors, so the prediction performance is improved, with the AUROC 0.84 [0.80, 0.88]. Similarly, the performance of optimized LSTM in reference [20] is better than that in reference [28]. The proposed model uses fruit fly algorithm to optimize LSTM network parameters, and uses FastText for data preprocessing, which further improves the prediction precision. Therefore, the AUROC is the highest, up to 0.87 [0.83, 0.90]. The results of AUROC further demonstrate that the proposed

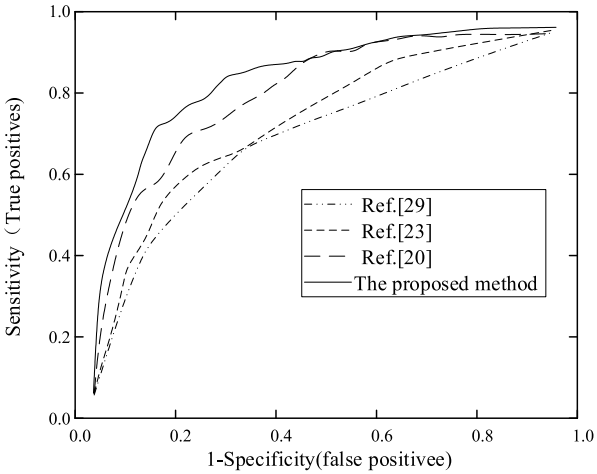

FIGURE 7. Comparison of AUROC values of different prediction models.

model can guarantee better prediction performance when there is a certain imbalance in the data.

VI. CONCLUSION

With the development and application of medical information technology, varieties of physiological indicators related to patient's health are recorded in electronic medical records. By mining electronic health records, the patient's health status can be auxiliary analyzed to provide further clinical prediction. Therefore, this paper proposes a clinical prediction model based on long short-term memory network optimized by fruit fly algorithm in health time series data. The interpretable vector representations of medical events are obtained by FastText method, and they are input into the LSTM network for learning and prediction. The fruit fly optimization algorithm is used to find the optimal super parameters of LSTM network to improve the training efficiency and prediction precision of the network. In addition, the experimental results on the MIMIC dataset show that prediction precision, Recall@k and MAP@k of the proposed model are all higher than other models. With the passage of time, the prediction performance of the model decreases, but the decrease of the proposed model is the smallest, which demonstrates the effectiveness of the proposed model and it can be used in clinical prediction of health time series data.

In the future work, the sequences with different lengths on different datasets will be evaluated. In order to extract event information and real value information at the same time, the model will be further optimized by using more abundant experimental data to process heterogeneous data effectively. In this way, different types of time series events are simulated at the same time in a variety of scenarios, such as drug dosage and measured value, which is conducive to efficient prediction.

REFERENCES

[1] E. Ghaleb, M. Popa, and S. Asteriadis, "Multimodal and temporal perception of audio-visual cues for emotion recognition," in *Proc. 8th Int. Conf. Affect. Comput. Intell. Interact. (ACII)*, Cambridge, U.K., Sep. 2019, pp. 552–558.

[2] K. Shinohara, S. Tanaka, H. Imai, H. Noma, K. Maruo, A. Cipriani, S. Yamawaki, and T. A. Furukawa, "Development and validation of a prediction model for the probability of responding to placebo in antidepressant trials: A pooled analysis of individual patient data," *Evidence Based Mental Health*, vol. 22, no. 1, pp. 10–16, Feb. 2019.

[3] M. B. Abdallah, M. Blonski, S. Wantz-Mezieres, Y. Gaudeau, L. Taillandier, J.-M. Moureaux, A. Darlix, N. M. de Champfleur, and H. Duffau, "Data-driven predictive models of diffuse low-grade gliomas under chemotherapy," *IEEE J. Biomed. Health Inform.*, vol. 23, no. 1, pp. 38–46, Jan. 2019.

[4] P. Tiwari, K. L. Colborn, D. E. Smith, F. Xing, D. Ghosh, and M. A. Rosenberg, "Assessment of a machine learning model applied to harmonized electronic health record data for the prediction of incident atrial fibrillation," *JAMA Netw. Open*, vol. 3, no. 1, pp. 284–288, 2020.

[5] M. Sperrin, D. J. Webb, P. Patel, K. J. Davis, S. Collier, A. Pate, D. A. Leather, and J. M. Pimenta, "Chronic obstructive pulmonary disease exacerbation episodes derived from electronic health record data validated using clinical trial data," *Pharmacoepidemiol. Drug Saf.*, vol. 28, no. 10, pp. 1369–1376, Oct. 2019.

[6] S. Padmanabhan, L. Carty, E. Cameron, R. E. Ghosh, R. Williams, and H. Strongman, "Approach to record linkage of primary care data from clinical practice research datalink to other health-related patient data: Overview and implications," *Eur. J. Epidemiol.*, vol. 34, no. 1, pp. 91–99, Jan. 2019.

[7] Y. F. Zhou and N. Chen, "The LAP under facility disruptions during early post-earthquake rescue using PSO-GA hybrid algorithm," *Fresenius Environ. Bull.*, vol. 28, no. 12, pp. 9906–9914, 2019.

[8] K. R. Jadhav and N. N. Patil, "Clinical document architecture (CDA) generation and integration for health data exchange based on cloud computing a survey," *Int. J. Comput. Sci. Eng.*, vol. 7, no. 1, pp. 801–805, Jan. 2019.

[9] T. Shaw, A. Janssen, R. Crampton, F. O'Leary, P. Hoyle, A. Jones, A. Shetty, N. Gunja, A. G. Ritchie, H. Spallek, A. Solman, J. Kay, M. A. Makeham, and P. Harnett, "Attitudes of health professionals to using routinely collected clinical data for performance feedback and personalised professional development," *Med. J. Aust.*, vol. 210, no. 6, pp. 17–21, Apr. 2019.

[10] P. G. Bambekova, W. Liaw, R. L. Phillips, and A. Bazemore, "Integrating community and clinical data to assess patient risks with a population health assessment engine (PHATE)," *J. Amer. Board Family Med.*, vol. 33, no. 3, pp. 463–467, May 2020.

[11] D. J. Feller, O. J. Bear Don't Walk IV, J. Zucker, M. T. Yin, P. Gordon, and N. Elhadad, "Detecting social and behavioral determinants of health with structured and free-text clinical data," *Appl. Clin. Informat.*, vol. 11, no. 1, pp. 172–181, Jan. 2020.

[12] H. Boshnak, S. AbdelGaberb, and E. Y. AmanyAbdoc, "Ontology-based knowledge modelling for clinical data representation in electronic health records," *Int. J. Comput. Sci. Inf. Secur.*, vol. 16, no. 10, pp. 68–86, 2019.

[13] N. Coppersmith, I. Sarkar, and E. Chen, "Quality informatics: The convergence of healthcare data, analytics, and clinical excellence," *Appl. Clin. Informat.*, vol. 10, no. 2, pp. 272–277, Mar. 2019.

[14] Y. Zhou, H. Yu, Z. Li, J. Su, and C. Liu, "Robust optimization of a distribution network location-routing problem under carbon trading policies," *IEEE Access*, vol. 8, pp. 46288–46306, 2020.

[15] D. McGraw and C. Petersen, "From commercialization to accountability: Responsible health data collection, use, and disclosure for the 21st century," *Appl. Clin. Informat.*, vol. 11, no. 2, pp. 366–373, Mar. 2020.

[16] H. Habibzadeh, K. Dinesh, O. Rajabi Shishvan, A. Boggio-Dandry, G. Sharma, and T. Soyata, "A survey of healthcare Internet of Things (HIoT): A clinical perspective," *IEEE Internet Things J.*, vol. 7, no. 1, pp. 53–71, Jan. 2020.

[17] N. Shah, G. Martin, S. Archer, S. Arora, D. King, and A. Darzi, "Exploring mobile working in healthcare: Clinical perspectives on transitioning to a mobile first culture of work," *Int. J. Med. Informat.*, vol. 125, pp. 96–101, May 2019.

[18] K. Betts, S. Kisely, and R. Alati, "Predicting common maternal postpartum complications: Leveraging health administrative data and machine learning," *BJOG, Int. J. Obstetrics Gynaecol.*, vol. 126, no. 6, pp. 702–709, May 2019.

[19] S. Pirbhulal, O. W. Samuel, W. Wu, A. K. Sangaiah, and G. Li, "A joint resource-aware and medical data security framework for wearable healthcare systems," *Future Gener. Comput. Syst.*, vol. 95, pp. 382–391, Jun. 2019.

[20] V. Strotbaum, M. Pobiruchin, B. Schreiweis, M. Wiesner, and B. Strahwald, "Your data is gold–data donation for better healthcare?" *Inf. Technol.*, vol. 61, nos. 5–6, pp. 219–229, Oct. 2019.

[21] M. M. Wahi and N. Dukach, "Visualizing infection surveillance data for policymaking using open source dashboarding," *Appl. Clin. Informat.*, vol. 10, no. 3, pp. 534–542, May 2019.

[22] K. Sato, T. Ohno, T. Ishii, C. Ito, and T. Kaise, "The prevalence, characteristics, and patient burden of severe asthma determined by using a japan health care claims database," *Clin. Therapeutics*, vol. 41, no. 11, pp. 2239–2251, Nov. 2019.

[23] O. Anya, H. Tawfik, M. M. Alani, and J. Hu, "Cybersecurity design considerations for cross-boundary clinical decision support," *J. Reliable Intell. Environ.*, vol. 5, no. 2, pp. 91–103, Jul. 2019.

[24] R. Ahmed, T. Toscos, R. R. Ghahari, R. J. Holden, E. Martin, S. Wagner, C. Daley, A. Coupe, and M. Mirro, "Visualization of cardiac implantable electronic device data for older adults using participatory design," *Nephron Clin. Pract.*, vol. 10, no. 4, pp. 707–718, 2019.

[25] M. Jayaratne, D. Nallaperuma, D. De Silva, D. Alahakoon, B. Devitt, K. E. Webster, and N. Chilamkurti, "A data integration platform for patient-centered e-healthcare and clinical decision support," *Future Gener. Comput. Syst.*, vol. 92, pp. 996–1008, Mar. 2019.

[26] A. Bhattacharjee, "Bayesian competing risks analysis without data stratification," *Clin. Epidemiol. Global Health*, vol. 8, no. 1, pp. 265–270, Mar. 2020.

[27] L. Li, G. Liu, L. Zhang, and Q. Li, "Deep learning-based sensor fault detection using S-long short term memory networks," *Struct. Monit. Maintenance*, vol. 5, no. 1, pp. 51–65, 2018.

[28] V. A. Rudrapatna and A. J. Butte, "Opportunities and challenges in using real-world data for health care," *J. Clin. Invest.*, vol. 130, no. 2, pp. 565–574, Feb. 2020.

[29] T. Van Steenkiste, J. Ruyssinck, L. De Baets, J. Decruyenaere, F. De Turck, F. Ongenae, and T. Dhaene, "Accurate prediction of blood culture outcome in the intensive care unit using long short-term memory neural networks," *Artif. Intell. Med.*, vol. 97, pp. 38–43, Jun. 2019.

[30] A. Bablani, D. R. Edla, V. Kuppili, and D. Ramesh, "A multi stage EEG data classification using k-means and feed forward neural network," *Clin. Epidemiology Global Health*, vol. 8, no. 3, pp. 718–724, Sep. 2020.

[31] F. E. Shamout, T. Zhu, P. Sharma, P. J. Watkinson, and D. A. Clifton, "Deep interpretable early warning system for the detection of clinical deterioration," *IEEE J. Biomed. Health Inform.*, vol. 24, no. 2, pp. 437–446, Feb. 2020.

[32] R. Blitz and M. Dugas, "Conceptual design, implementation, and evaluation of generic and standard-compliant data transfer into electronic health records," *Appl. Clin. Informat.*, vol. 11, no. 3, pp. 374–386, May 2020.

[33] L. Kobayashi, A. Oyalowo, U. Agrawal, S.-L. Chen, W. Asaad, X. Hu, K. A. Loparo, G. D. Jay, and D. L. Merck, "Development and deployment of an open, modular, near-real-time patient monitor datastream conduit toolkit to enable healthcare multimodal data fusion in a live emergency department setting for experimental bedside clinical informatics research," *IEEE Sensors Lett.*, vol. 3, no. 1, pp. 1–4, Jan. 2019.

[34] C. M. Rey, "Wearable data revolution: Digital biomarkers are transforming research, promising a revolution in healthcare," *Clin. Omics*, vol. 6, no. 2, pp. 10–13, 2019.

[35] J. Nzinga, G. McGivern, and M. English, "Hybrid clinical-managers in Kenyan hospitals: Navigating between professional, official and practical norms," *J. Health Org. Manage.*, vol. 33, no. 2, pp. 173–187, Mar. 2019.

[36] D. F. Garway-Heath, H. Zhu, Q. Cheng, K. Morgan, C. Frost, D. P. Crabb, T.-A. Ho, and Y. Agiomyrgiannakis, "Combining optical coherence tomography with visual field data to rapidly detect disease progression in glaucoma: A diagnostic accuracy study," *Health Technol. Assessment*, vol. 22, no. 4, pp. 1–106, Jan. 2018.

[37] W. Tang, J. Y. Gao, X. Y. Ma, C. H. Zhang, L. T. Ma, and Y. S. Wang, "Application of recurrent neural network in prognosis of peritoneal dialysis," *J. Peking Univ. Health Sci.*, vol. 51, no. 3, pp. 602–608, 2019.

[38] S. Lin, Q. Zhang, F. Chen, L. Luo, L. Chen, and W. Zhang, "Smooth Bayesian network model for the prediction of future high-cost patients with COPD," *Int. J. Med. Informat.*, vol. 126, pp. 147–155, Jun. 2019.

[39] A. Patra and J. A. Noble, "Hierarchical class incremental learning of anatomical structures in fetal echocardiography videos," *IEEE J. Biomed. Health Inform.*, vol. 24, no. 4, pp. 1046–1058, Apr. 2020.

[40] R. Xie, I. Khalil, S. Badsha, and M. Atiquzzaman, "Collaborative extreme learning machine with a confidence interval for P2P learning in healthcare," *Comput. Netw.*, vol. 149, pp. 127–143, Feb. 2019.