International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Heart Disease Prediction using Long Short-Term Memory (LSTM) Deep Learning Methodology

K. Divya Vani

Department of Electronics and Communication Engineering
St.Martin's Engineering College, Dhulapalli, Secunderabad, India

ABSTRACT

Heart disease is one of the deadly diseases. A large population in the world is suffering from this problem. As we consider death rate and a large number of people who are suffering from heart disease, it is revealed how important is early diagnosis of heart disease. There are many traditional methods of prediction for such illness but they are not looking sufficient. Previous dynamic prediction models rarely handle multi-period data with different intervals, and the large-scale patient hospital records are not effectively used to improve the prediction performance. This paper aims to focus on the prediction of cardiovascular disease using the improved long short-term memory (LSTM) model.Based on the traditional LSTM, this paper proposed a new model by improving the internalforgetting gate input. First, the irregular time interval is smoothed to obtain the timeparameter vector, and then it is used as the input of the forgetting gate to overcome theprediction obstacle caused by the irregular time interval.

Keywords: Cardiovascular disease, dynamic prediction, LSTM.

Introduction

The heart is the organ that pumps blood, with its life-giving oxygen and nutrients, to all the tissues of the body.If the pumping action of the heart becomes inefficient,vital organslike brain and kidneys suffers, if the heart stopsworking altogether,death occurs within minutes. The heart diseasehasbeen considered asone of the complexes and life deadliest human diseases in the world. Life itself is completely dependent on the efficient operation of heart. Symptoms of heartdisease includeshortness of breath, weakness of physical body, swollen feet and fatigue[1].

The heart disease diagnosis and treatment are very complex, especially in the developing countries, due to the rare availability of diagnostic apparatus and otherresources which affect properprediction andtreatment of heart patients [2].This makes heart disease a major concern to be dealt with. But it is difficult to identify heart disease because of several contributory risk factors such as diabetes, high blood pressure, high cholesterol, abnormal pulse rate, and many other factors.

The invasive–based techniques to the diagnoseof heart disease are based on theanalysis of thepatient's medical history, physical examination report and analysis of concerned symptoms bymedical experts [3]. Often there is a delay in the diagnosis due to human errors. Due to such constraints, scientists have turned towards modern approaches like Data Mining and Machine Learning for predicting the disease.

Data mining plays an important role in building intelligent model for medical system to detect the heart disease[4]using the available dataset of patients, which involves riskfactor associated with the disease.Medical practitioners may provide helpfor the detection. Several software tools and variousalgorithms have been proposed by researchers for developing effective medicaldecision support system.

Machine learning helps computers to learn and act accordingly. It helps the computer to learn thecomplex model and predict the data and also has the ability to calculate complex mathematics on big data. The machine learning based heart disease predicting systems will be precise and will reduce the risk [5].The value of machine learning technology is recognized wellin health care industry whichhaslargepool of data. It helps medical experts to predict the disease and lead to improvise the treatment.Machinelearningpredictive models such asdecisiontree,k-nearestneighbour, logistic regression, random forest, support vector machine is utilized to predict whether a person is having heart disease or not.

However, medical data are often constricted by smaller sets of observations than whatis usuallypreferred to allow for sufficient training and testing of models built using machinelearning algorithms. Without sufficiently sized data sets, it is very difficult to determine if a model isgeneralizable to previously unseen sets of data. Using synthetic data to overcome constraints inherent in small medical research data sets could be a solution to protect patient privacy and allow forapplication of machine learning algorithms. The larger data sets allow for sufficientlysized training and testing partitions which enable the machine learning algorithm to learn from experience by exposure to a large set of observations, and then to be tested upon another large set of observations that have not previously been introduced tothe model.Using the synthetic data, we train and validate the Machine Learning Models then compare theprediction outcome accuracy to that using the original observations [6]. Once satisfied with theconsistency ofclassification prediction between the original data set and the surrogate data set, we generate an expanded surrogate data set in stage three. While based on the Cleveland data set, this expanded set contains previously unstudied attributes. This expanded data set is used to test and train aneural network model,

having partitioned the syntheticdatainto large testing and training subsets. We then compare the outcome of the prediction accuracy of the deep learningmodel to the traditionalmachine learningmodels. We find that using the expanded surrogate data set to build adeep learningmodel results in the best classificationprediction accuracy and stability.

The rest of the paper is organized asfollows;Section 2 discusses survey on machine learning techniques for predicting heartdisease. Section 3 provides proposed system design, algorithms and methods used for heart disease prediction. Section 4 discusses performance the results that. Finally, Section 5 ends with a conclusion of current work.

Literature

Guo et al., (2020) [7] the proposed Recursion enhanced random forest with an improved linear model (RFRF-ILM) to detect heart disease. This paper aims to find the key features of the prediction of cardiovascular diseases through the use of machine learning techniques. The prediction model is adding various combinations of features and various established methods of classification. it produces a better level of performance with precision through the heart disease prediction model. In this study, the factors leading to cardiovascular disease can be diagnosed. A comparison of important variables showed with the Internet of Medical Things (IoMT) platform, for data analysis.

Latha et al., (2019) [8] This author investigates a method termed ensemble classification, which is used for improving the accuracy of weak algorithms by combining multiple classifiers. Experiments with this tool were performed using a heart disease dataset. A comparative analytical approach was done to determine how the ensemble technique can be applied for improving prediction accuracy in heart disease. The focus of this paper is not only on increasing the accuracy of weak classification algorithms, but also on the implementation of the algorithm with a medical dataset, to show its utility to predict disease at an early stage.

Tao et al., (2018) [9]This author focused on developing a fast and accurate automatic ischemic heart disease detection/localization methodology. Methods: T wave was segmented from averaged MCG recordings and 164 features were subsequently extracted. These features were categorized into three groups: time domain features, frequency domain features, and information theory features. Next, we compared different machine learning classifiers including: KNN, DT, SVM and XGBoost. To identify IHD case, we selected three classifiers with best performance and applied model ensemble to average results.

Arabasadi, et al., (2017) [10] Much research has, therefore, been conducted using machine learning and data mining so as to seek alternative modalities. Accordingly, we herein propose a highly accurate hybrid method for the diagnosis of coronary artery disease. As a matter of fact, the proposed method is able to increase the performance of neural network by approximately 10% through enhancing its initial weights using genetic algorithm which suggests better weights for neural network.

Dutta et al., (2020) [11] proposes an efficient neural network with convolutional layers to classify significantly class-imbalanced clinical data. The data is curated from the National Health and Nutritional Examination Survey (NHANES) with the goal of predicting the occurrence of coronary heart disease (CHD). While the majority of the existing machine learning models that have been used on this class of data are vulnerable to class imbalance even after the adjustment of class-specific weights, our simple two-layer CNN exhibits resilience to the imbalance with fair harmony in class-specific performance. Given a highly imbalanced dataset, it is often challenging to simultaneously achieve a high class 1 (true CHD prediction rate) accuracy along with a high class 0 accuracy, as the test data size increases. We adopt a two-step approach: first, we employ least absolute shrinkage and selection operator (LASSO) based feature weight assessment followed by majority-voting based identification of important features.

Singhal et al., (2018) [12] paper, Convolutional Neural Networks (CNNs) are used to design an early-stage prediction and medical diagnosis system. 13 clinical features are supplied as input to CNN. Modified backpropagation training method is used to train the CNN. During testing, it is observed that CNN offers more than 95% accurate results by predicting absence and presence of heart disease.

Masethe et al., (2014) [13] The heart disease accounts to be the leading cause of death worldwide. It is difficult for medical practitioners to predict the heart attack as it is a complex task that requires experience and knowledge. The health sector today contains hidden information that can be important in making decisions. Data mining algorithms such as J48, Naïve Bayes, REPTREE, CART, and Bayes Net are applied in this research for predicting heart attacks.

Proposed method

A common LSTM unit is composed of a cell, an input gate, an output gate and a forget gate.The cell remembers values over arbitrary time intervals, and the three gates regulate theflow of information into and out of the cell.Figure 1 shows the structure of a traditional LSTM cell and illustrates the operations ofthe gates. There are three gates (input, forget and output) in the basic cell of LSTM, and eachgate has a sigmoid activation function and a point-wise multiplication operation.

To use LSTM to process sequence data with irregular time intervals, we first adapt thethreshold structure of the LSTM unit to learn the temporal characteristics associated withCVD evolution at different time intervals. After that, we propose to use the target repeatprediction method for the output of hidden layer at each time step, which can simplify themodel training process with different lengths of time series. Finally, for the output layer ofthe model, the Sigmoid function is introduced as the activation function of the multi-tagoutput, so that the patient's multiple diagnostic tags are predicted as output. Heart rate and cholesterol have been identified to be the major factors of atherosclerosisand thus choosing these values as the attributes while using the classificationalgorithm.

Recurrent Neural Networks (RNNs) are connection models that capture the progression of arrangements by means of cycles among the connected nodes in the figure 1. Dissimilar to feedforward neural systems, repetitive systems hold in a state that can speak to data from a subjectively long setting window. Recurrent neural systems have been customized to function using a set of parameters in an arranged structure, by preferring appropriate advanced methods and parallel processing the results can be further improvised. As of late, frameworks in view of long short-term memory (LSTM) and bidirectional (BRNN) models have exhibited weighty execution on errands as changed as image, languages, and recognition.

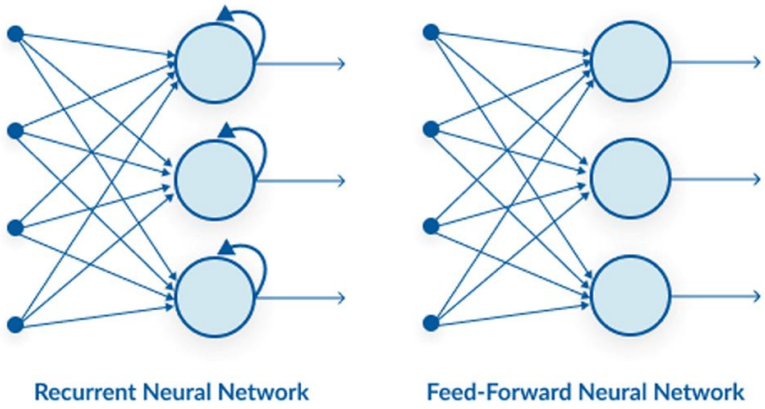

Figure 1:Recurrent Neural Networks (RNNs)

One of the most successful RNN models for sequence learning as of now from 1997 is Long Short-Term Memory introduced by Hochreiter and Schmidhuber. It consists of a memory cell and a unit of calculation that replaces conventional procedures used neurons in the hidden layer of the network. Using these memory cells, network overcomes a few challenges that are faced during the training phase. Next, Bidirectional Recurrent Neural Networks by Schuster and Paliwal present the BRNN architecture in which data from both the future and the past are utilized to decide the output at any time t. Instead, the neurons in the neural network are replaced by memory cells, the figure 3 traditional LSTM . It is used to rectify the gradient vanishing problem across other RNNs. The RNN cannot remember longer sequence and instead have short dependencies and are trained by a separate set of weights for remembering and forgetting outputs.

An LSTM unit reads an input x_t and depends on prior output h_{t-1} and results in an output h_t. It has a memory cell c_t, an input gate i_t, an output gate $o_{t,}$ and a forget gate f_t. Each LSTM cell performs the following functions:

1. Use the current input x_t and the previous hidden state h_{t-1} to decide data to be deleted from the memory vector (c_{t-1}), represented as: $f_t = func(w_f(h_{t-1}x) + b_f)$ where bf is a bias and wf is a set of weights.
2. Using x_t and h_{t-1}, a matrix is constructed that permits a specific information to be updated in c_{t-1}. $i_t = func(w_i.(h_{t-1}x) + b_i)$.

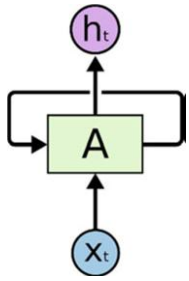

Figure.2. LSTM cell

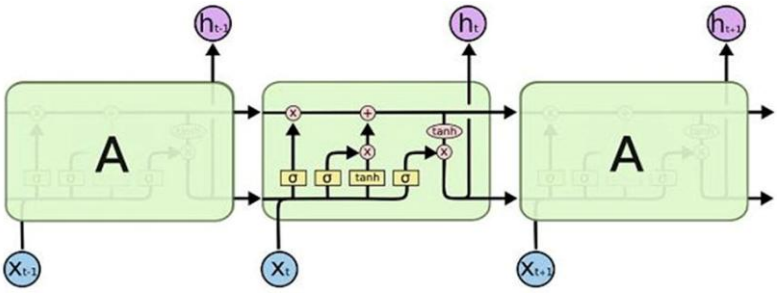

Figure.3. Traditional LSTM

3. Use x_t and h_{t-1} to gather information that should be included. $c_t = func(w_c.(h_{t-1}\,x) + b_c)$.
4. Finally, merge the new information and the old information $c_t = f_t.c_{t-1} + i_t.c_t$.

It can be clearly seen that by using stochastic gradient descent, this model will be used to train so that it can differentiate the information to be forgotten, preserved, or retained.

Improved long short-term memory

In the medical situation, patients with chronic diseases will go to the hospital because of thedevelopment of the disease, such as deterioration or

recurrence. However, different patientsmay have different time intervals between hospitalizations due to their physical condition,condition, etc., and the difference may range from less than 1 month to several years. Thelack of time interval brings certain difficulties and challenges to the study of clinical timeseries data.

To solve the problem of irregular time interval, we propose to smooth the time interval toobtain the time parameter vector and use it as the input of LSTM forget gate. The improvedLSTM cell is shown in Figure 4. introduce the forward propagation process of theLSTMnetwork.

The first step in the forward propagation of the LSTM network is the calculation ofthe forgotten threshold. This threshold determines which of the input informationwill be forgotten and will not affect future time step. In detail, the time intervalbetween the time step t-1 and the time step t is smoothed to obtain a three-dimensionalvector, and the time vector is used as an input parameter of the forget gate, as shownin equation (1).

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

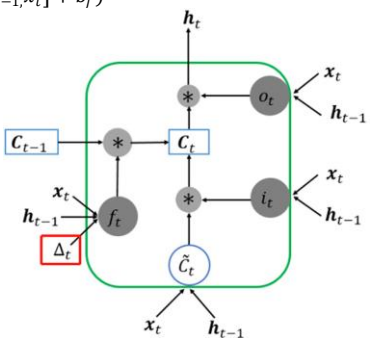

Figure 4.Improved LSTM cell

$$f_t = \sigma\left(W_f [h_{t-1}, x_t] + P_f \mathrm{p}_{\Delta_{t-1:t}} + b_f\right) \quad (2)$$

In equation (2),$P_f \mathrm{p}_{\Delta_{t-1:t}}$ represents a vector after the smoothing of the time interval between time slices, and the smoothing formula is shown in equation (3):

$$\mathrm{p}_{\Delta_{t-1:t}} = \left(\frac{\Delta_{t-1:t}}{60}, \left(\frac{\Delta_{t-1:t}}{180}\right)^2, \left(\frac{\Delta_{t-1:t}}{365}\right)^3\right) \quad (3)$$

In equation (3), $\Delta_{t-1:t}$ represents the time interval, in units of days. Because patients rarely rehospitalizein the same month, so we choose two months as the denominator, then half ayear and one year, making the vector $\mathrm{p}_{\Delta_{t-1:t}}$ within a reasonable range.

P_f is a connection weight parameter corresponding to the time interval vector, whichneeds to be optimized for training to handle the memory effect generated by the irregulartime interval.

The second step of forward propagation determines what information is saved in the cellstate. First, you need to generate a temporary state and then update the old cell state. Theformula is shown in equations (6) and (7).

$$\tilde{C}_t = \tanh(W_C [h_{t-1}, x_t] + b_c) \quad (4)$$

$$C_t = f_t * C_{t-1} i_t * \tilde{C}_t \quad (5)$$

where W_c and b_c are the connection weight and offset of the temporary state. $\tilde{C}_t$ is atemporary state containing new candidate values. $\tilde{C}_{t-1}$ is the status information of theprevious time step. C_t is the state of the time step t after the update.

The third step of forward propagation determines the final network output, as shown inequation (6).

$$h_t = o_t * \tanh(C_t) \quad (8)$$

where h_t is the current hidden state, and h_t and C_t will be used as input for the next time step.

Results

LSTM learn the characteristics of data set from training set and predict the classificationlabels of new samples. The hyper parameters of LSTM model need to be set. The proposedimproved LSTM model is defined as T-LSTM-TR. We train and tune the parameters of ourmodel using 10-fold cross-validation method.

The hyper parameters need to be adjusted and optimized during training process,including the number of hidden layer neurons H, the end time slice loss function weight aand the dropout parameter. The model is trained by setting different parameter setsseparately, and then the test results are compared. Finally, the optimal parameters of theT-LSTM-TR model is set as H = 120,a = 0.5 and Dropout = 0.4.

This paper selects the traditional LSTM model as the benchmark model for performancecomparison. As shown in Table 1, the performance of T-LSTM-TR model proposed in thispaper is similar to that of the LSTM model in terms of precision, while the performance ofT-LSTM-TR is significantly superior compared to that of the traditional LSTM model interms of other indicators. The results show that the classification performance of our modelis effectively improved by adapting the departmental structure of traditional LSTM unit. Asshown in Figure 4, we can more clearly compare the performance of T-LSTM-TR and LSTMthrough the ROC curve.

Table 1.The performances ofT-LSTM-TR andLSTM

Models	Precision	Recall	F1	AUC
T-LSTM-TR	0.4992	0.811	0.608	0.896
LSTM	0.478	0.754	0.584	0.844

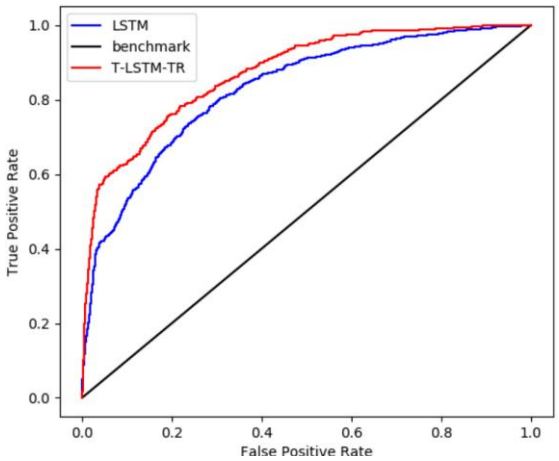

Figure 5.ROC curve ofT-LSTM-TR andLSTM

As shown inFigure 5, it can more clearly compare the performance of T-LSTM - TR and LSTMthrough the ROC curve.

Conclusion

Based on the traditional LSTM, this paper proposed a new model by improving the internalforgetting gate input. First, the irregular time interval is smoothed to obtain the timeparameter vector, and then it is used as the input of the forgetting gate to overcome theprediction obstacle caused by the irregular time interval. The experimental results showthat the dynamic prediction model proposed in this paper has a significant improvement inclassification performance compared with the traditional LSTM model, which verifies theeffectiveness of the proposed model.

REFERENCE

[1]. Hauptmann, Andreas, Simon Arridge, Felix Lucka, Vivek Muthurangu, and Jennifer A. Steeden. "Real-time cardiovascular MR with spatio-temporal artifact suppression using deep learning–proof of concept in congenital heart disease." *Magnetic resonance in medicine* 81, no. 2 (2019): 1143-1156.

[2]. Hu, Xiao-jing, Xiao-jing Ma, Qu-ming Zhao, Wei-li Yan, Xiao-ling Ge, Bing Jia, Fang Liu et al. "Pulse oximetry and auscultation for congenital heart disease detection." *Pediatrics* 140, no. 4 (2017).

[3]. Borkin, Michelle, Krzysztof Gajos, Amanda Peters, Dimitrios Mitsouras, Simone Melchionna, Frank Rybicki, Charles Feldman, and Hanspeter Pfister. "Evaluation of artery visualizations for heart disease diagnosis." *IEEE transactions on visualization and computer graphics* 17, no. 12 (2011): 2479-2488.

[4]. Shouman, Mai, Tim Turner, and Rob Stocker. "Using data mining techniques in heart disease diagnosis and treatment." In *2012 Japan-Egypt Conference on Electronics, Communications and Computers*, pp. 173-177. IEEE, 2012.

[5]. Gandhi, Monika, and Shailendra Narayan Singh. "Predictions in heart disease using techniques of data mining." In *2015 International Conference on Futuristic Trends on Computational Analysis and Knowledge Management (ABLAZE)*, pp. 520-525. IEEE, 2015.

[6]. Hauptmann, Andreas, Simon Arridge, Felix Lucka, Vivek Muthurangu, and Jennifer A. Steeden. "Real-time cardiovascular MR with spatio-temporal artifact suppression using deep learning–proof of concept in congenital heart disease." *Magnetic resonance in medicine* 81, no. 2 (2019): 1143-1156.

[7]. Guo, Chunyan, Jiabing Zhang, Yang Liu, Yaying Xie, Zhiqiang Han, and Jianshe Yu. "Recursion enhanced random forest with an improved linear model (rerf-ilm) for heart disease detection on the internet of medical things platform." *IEEE Access* 8 (2020): 59247-59256.

[8]. Latha, C. Beulah Christalin, and S. Carolin Jeeva. "Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques." *Informatics in Medicine Unlocked* 16 (2019): 100203.

[9]. Tao, Rong, Shulin Zhang, Xiao Huang, Minfang Tao, Jian Ma, Shixin Ma, Chaoxiang Zhang et al. "Magnetocardiography-based ischemic heart disease detection and localization using machine learning methods." *IEEE Transactions on Biomedical Engineering* 66, no. 6 (2018): 1658-1667.

[10]. Arabasadi, Zeinab, Roohallah Alizadehsani, Mohamad Roshanzamir, Hossein Moosaei, and Ali Asghar Yarifard. "Computer aided decision making for heart disease detection using hybrid neural network-Genetic algorithm." *Computer methods and programs in biomedicine* 141 (2017): 19-26.

[11]. Dutta, Aniruddha, Tamal Batabyal, Meheli Basu, and Scott T. Acton. "An efficient convolutional neural network for coronary heart disease prediction." *Expert Systems with Applications* 159 (2020): 113408.

[12]. Singhal, Shubhanshi, Harish Kumar, and Vishal Passricha. "Prediction of heart disease using CNN." *Am. Int. J. Res. Sci. Technol. Eng. Math* 23, no. 1 (2018): 257-261.

[13]. Masethe, Hlaudi Daniel, and Mosima Anna Masethe. "Prediction of heart disease using classification algorithms." In *Proceedings of the world Congress on Engineering and computer Science*, vol. 2, pp. 22-24. 2014.