**AN IMPROVED LSTM BASED FRAME WORK FOR CARDIOVASCULAR DISEASES RISK PREDICTION IN IMBALANCED BIG DATA**

The busy schedule of the modern era leads to an unhealthy life style which causes anxiety and depression. In order to overcome these conditions, there is a tendency to resort to excessive smoking, drinking and taking drugs. All these things are the root cause of many dangerous diseases including cardiovascular diseases, cancer etc. According to the World Health Organization (WHO), cardiovascular diseases (CVDs) have the highest number of death rates, globally. Over a period of time, they have become very common and are now overstretching the healthcare systems of countries. At this stage, fast, accurate and early clinical assessment of the disease severity is vital. To support decision making and logistical planning in healthcare systems, this work proposed an effective data prediction by using Deep learning-based approach. Apply our technique on the publicly available MIMIC-II database and show the effectiveness of the LSTM classifier. Experiments show that our proposed scheme improves the accuracy of prediction.

# INTRODUCTION

A tremendous volume of data being generated in the healthcare sector is growing at a rapid rate. The rise of data comes in response to the digitization of healthcare information that includes biomedical images, clinical text, genomic data, EHRs, sensing data, biomedical signals, and social media which generates the large scale of primary and secondary data within the healthcare industry [1,2]. The overall data generated across the world is expected to dramatically rise in the upcoming years, reaching 175 zetabytes by 2025, leading to a compounded annual growth rate of 61% [3]. As per the 2012 Digital Universe Study by IDC, only 22 % of overall data had the potential for analysis. The percentage of beneficial data would jump to 37% by 2020 [4]. This has generated tremendous interest in exploiting healthcare data access to enhance patient quality and reduce costs. This explosive increase in transient or stored data has created an immediate requirement of the need for automated tools as well as novel techniques that can be helpful in the transformation of vast volumes of data into beneficial information and knowledge in an intelligent way [5]. The healthcare industry today generates large amounts of complex data related to a patient disease & diagnosis. Data resources from the hospital and medical devices are difficult to process by manual methods and it is time consuming and expensive, to load into a traditional relational database for analysis [6].

Statistics and data mining are the leading fields of study that are supporting the empowered individual to discover hidden information for effective decision making. Statistics provide a strong fundamental background for quantification and evaluation of results. However, algorithms based on statistics need to be modified and scaled before they are applied to data mining [7]. Data mining is one of the most useful techniques that can help researchers, entrepreneurs, and individuals for extracting valuable information from large sets of data [8]

Cardiovascular Diseases (CVDs) are the most common and prevalent diseases in India, as well as globally [11]. As per the World Health Organization (WHO), mortalities occurring every year across the world, because of heart problems, is found to be greater than 12 million [12]. CVD mortalities were estimated to be 17.9 million, which would increase to 24.2 million by 2030 [13,15]. The term "heart disease “is often used interchangeably with the term “ cardiovascular disease” that includes a wide range of conditions that affect the heart and the blood vessel [14]. The CVD includes Ischemic Heart Disease, Rheumatic Heart Disease, Congenital Heart Disease, Cardiomyopathy, Valvular Heart Disease, Aortic Valve Sclerosis and Stenosis, Atherosclerosis [15].

This work is exclusively is focused on statistical and data mining tools and techniques for CVDs using echocardiography records. Therefore, the following section will review the importance of echocardiography data for the prediction of CVDs*.*

# Related work

Ema, R. R et al proposed a new hybrid model based on Fuzzy C-means and Artificial Neural Networks (ANNs) with Principle Component Analysis that is capable to predict heart disease. The Principal Component Analysis is used to select the important features from the dataset. Then Fuzzy C-Means Clustering is used to cluster the extracted data from PCA and finally, Artificial Neural Network is used to predict Cardiovascular Disease.

Xu, S et al focus on practical problem of Chinese hospital dealing with cardiovascular patients’ data to make an early detection and risk prediction. To better understand the prescription and advice in Chinese, basic natural language processing method was used to synonym recognition and attribute extraction in Ultrasonic echocardiography.

Joo, G et al assessed the effectiveness of various ML methods in predicting the 2-year and 10-year risk of CVD such as atrial fibrillation, coronary artery disease, heart failure, and strokes. To develop prediction models, we considered the usual medical examination data, questionnaire survey results, comorbidities, and past medication information available in the KNHSC data.

Athanasiou, M et al present study is to develop and evaluate an explainable personalized risk prediction model for the fatal or non-fatal CVD incidence in T2DM individuals. An explainable approach based on the eXtreme Gradient Boosting (XGBoost) and the Tree SHAP (SHapley Additive exPlanations) method is deployed for the calculation of the 5-year CVD risk and the generation of individual explanations on the model’s decisions.Bhatt, A et al focuses on analyzing cardiovascular health of rural and urban residents for early prediction of cardiac ailments through calcium score health indicator. Coronary Angiography is performed and Patients’ Calcium Score results are taken randomly. Calcium score is also termed as Coronary Artery Calcium (CAC). This score is analyzed sex and age-wise in order to predict cardiovascular health issues at early stage.

Nikam, A. et al proposed machine learning techniques to predict cardiovascular disease using features. BMI is one of the highlighting features we used for prediction. BMI is important in predicting cardiovascular disease. The main focus of the article is the effect of BMI on the prediction of cardiovascular disease. The model has proposed with different features as well as regression and classification techniques. Conclude that BMI is a significant factor while predicting cardiovascular disease.

Bhuvaneswari Amma N G et al proposed a medical diagnosis system to predict the risk of cardiovascular diseases with high prediction accuracy. This system is built using an intelligent approach based on Principal Component Analysis (PCA) and Adaptive Neuro Fuzzy Inference System (ANFIS). This system has two stages: In the first stage, dimension of heart disease dataset that has 13 attributes is reduced to 7 attributes using PCA. In the second stage, diagnosis of heart disease is conducted using ANFIS. Rahim, A et al proposed a MaLCaDD (Machine Learning based Cardiovascular Disease Diagnosis) framework for the effective prediction of cardiovascular diseases with high precision. Particularly, the framework first deals with the missing values (via mean replacement technique) and data imbalance (via Synthetic Minority Over-sampling Technique - SMOTE). Subsequently, Feature Importance technique is utilized for feature selection.

Li-Na Pu et al overviewed the eligible genome-wide association studies for CVD outcomes/traits . Clinical trials on CVD prediction using genetic information will be summarized from overall aspects. As yet, most of the single or multiple genetic markers, which have been evaluated in the follow-up clinical studies, did not significantly improve discrimination of CVD. Pham, T. D. et al introduces a computational methodology for predicting such events in the context of robust computerized classification using mass spectrometry data of blood samples collected from patients in emergency departments. Applied the computational theories of statistical and geostatistical linear prediction models to extract effective features of the mass spectra and a simple decision logic to classify disease and control samples for the purpose of early detection.

Park, H. D et al propose a frequency-aware based Attentionbased LSTM (FA-Attn-LSTM) that weighs on important medical features using an attention mechanism that considers the frequency of each medical feature. Our model predicts the risk for cardiovascular disease using the ejection fraction as a prediction target and shows RMSE = 3.65 and MAE = 2.49.

Mostafa, N et al analyzed some common physiological attributes to identify a pattern among the people having a cardiovascular disease which, in further, has been used to distinguish whether a person has a risk of developing cardiovascular disease or not.

Zhu, C.-Y. et al designed a risk assessment model for patients, followed by the design and development of readmission risk assessment system for patients with cardiovascular disease. The risk assessment model includes three parts: risk prediction, clustering analysis and regression analysis of risk factors, which can automatically predicate the risk level and risk factors for the discharged patients in thirty days.

Mendonca, F et al provides a novel method to predict cardiovascular diseases using machine learning. A comparison between various machine learning algorithms is made to analyse the performance on the dataset and the proposed method uses s K Nearest Neighbour which has an accuracy of 92.30%. P, A., & Kalyani David et al portrays a new algorithm, ModifiedBoostARoota, developed similar to BoostARoota, differing in the feature elimination process. Also, by choosing XGBoost and catboost as base models in both BoostARoota and ModifiedBoostARoota, a comparison of both the algorithms’ performances are done. ModifiedBoostARoota algorithm has faster performance compared to BoostARoota, when catboost is chosen as the base model. Also, the XGBoost and CatBoost classifiers modelled on features selected by ModifiedBoostARoota gave better accuracy than that of BoostARoota.

**Proposed System**

Using these structured data and deep learning models to predict CVD which is an important issue in worldwide. In order to solve the problem of low accuracy of Long-Short Term Memory (LSTM) model in CVD prediction, this chapter presented a proposed model of LSTM model based on attention mechanism. The proposed model can learn the importance of each past value to the current value from the long sequence of CVD data at the past moment, which makes it possible to extract more valuable features. Constructed a dataset using the CVD data in the core section of Wuhan for experiments, and the performance of the improved model is compared with the original LSTM model.

CONSTRUCTION OF ATTENTION-LSTM MODEL

LSTM Model

We will briefly introduce the principle of LSTM model. LSTM is a kind of recurrent neural network, as shown in Fig. 1.

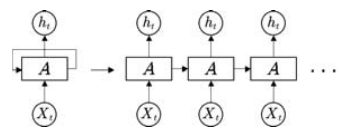


Figure 1 Recurrent neural network (RNN)

However, compared with the conventional RNN, the structure of this repeated module A of LSTM is more complicated, as shown in Fig. 2.

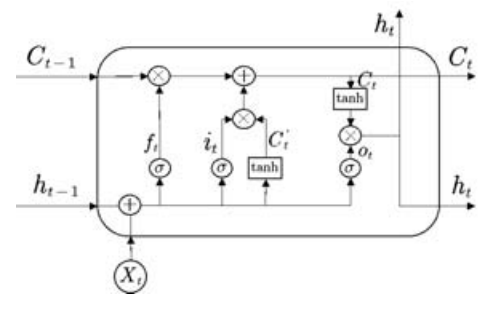


Figure The structure of LSTM cell.

This module consists of three parts, the forgotten gate, the input gate and the output gate. σ is the Sigmoid function, output a value between 0 and 1, describing how much of each part can pass.













Among them, ft determines how much information we want to discard. it determines how much new information we should add. ot determines how much information we want to output. xt is the input at time t. ht−1 is the output of the previous gate, Wf , Wi, Wc and Wo is the weight, bf , bi , bc and bo is the bias, Ct−1 is the cell state at the previous moment, Ct is the cell state at the current moment.

Attention-LSTM Model

Since the model is difficult to learn information at a time far from the current time, and it may be important for the current value. To overcome the weakness, we tried to add an attention layer to the LSTM network. Referring to the attention implementation steps of [9], we can apply it to the LSTM model. As shown in Fig. 3, the attention layer is added to the LSTM model.

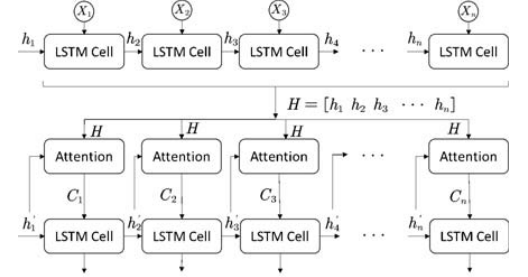


Figure The process of adding an attention mechanism to the LSTM model.

Among them, Xi, i ∈ (1, n) is the input, hi is the intermediate output result of each cell, hi are input into each attention model as H, and the elements of the next layer h i are used as H to calculate the similarity and weight coefficient, and finally get the attention coefficient. The specific attention model is shown in below figure.

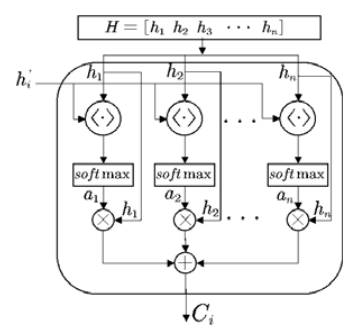


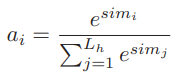
Figure The internal structure of attention model

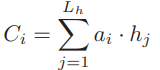
Where ,the similarity between the current element and the intermediate output result in the previous layer, and then normalized by the softmax function to obtain the corresponding weight coefficient ai. Finally, a weighted summation operation is performed to obtain the Attention value Ci. The formula used in the attention layer is as follows:











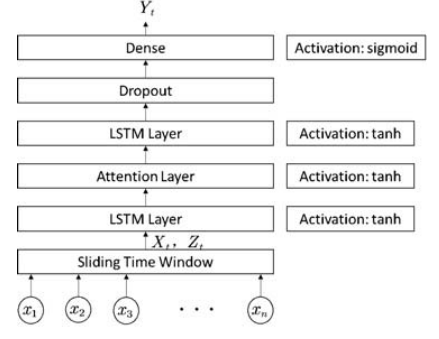
In the above equations, uses vector H and H to calculate similarity to obtain weights, uses the softmax function to normalize the weight, uses the normalized weight ai and hi weighted sum. The result of weighted summation is the attention weight value Ci.The implementation of the Attention layer is to retain the intermediate output results of the input sequence by the LSTM encoder, and then calculate the similarity between the intermediate output results of the previous layer and the current output to obtain the weight factor, and finally obtain the attention coefficient. 

Figure Attention-LSTM network architecture.

As shown in above Fig, we use the original traffic flow data to construct a feature matrix and label vector through a sliding time window method, and obtain the attenton weights based on the correlation between the values in matrix X and the values in vector Z through the attention layer and generate the final prediction Y .

**Proposed algorithm 1**

Input: CVD data

Output: A trained Attention-LSTM model.

1: Construct a dataset with a sliding time window,including Xt and Zt.

2: Normalization Xt and Zt.

3: Input features matrix Xt and current disease vector Zt to A-LSTM network.

4: while training epoch does not reach the set value do

5: Put (Xt, Zt) into the Attention-LSTM network for forward propagation.

6: Calculate the attention weight corresponding to each element

7: Generate Yt

8: Caculate mean square error.

9: Use RMSProp update weights for A-LSTM network.

10: end while

11: return A trained Attention-LSTM model.

The performance of the LSTM model based on the attention mechanism is verified for long time series and large prediction lag time. All prediction models use the same data set and are built in the same way. In the LSTM model, we set 2 hidden layers, the number of hidden layer neurons is 64 and 64, and the learning rate is 0.05. The network optimizer is also RMSprop. The process of Attention-LSTM model training is shown in Algorithm 1.

**SIMULATION RESULT**

**CONCLUSION**

In this project, an attention layer is added to the existing LSTM model to constructed an Attention-LSTM model. And the validity of predicting long-sequence data is verified by experiments. We introduced the process of constructing the Attention-LSTM model and verified its performance using real CVD data sets. Experiments show that our proposed scheme improves the accuracy of prediction. This study only considered the application of the model with the attention layer on the time series. In future work, we can consider the spatial correlation of traffic flow and apply attention mechanisms in it.

**REFERENCES**

1. Zhu, C.-Y., Chi, S.-Q., Li, R.-Z., Tong, D.-Y., Tian, Y., & Li, J.-S. (2016). *Design and Development of a Readmission Risk Assessment System for Patients with Cardiovascular Disease. 2016 8th International Conference on Information Technology in Medicine and Education (ITME).*
2. Park, H. D., Han, Y., & Choi, J. H. (2018). *Frequency-Aware Attention based LSTM Networks for Cardiovascular Disease. 2018 International Conference on Information and Communication Technology Convergence (ICTC).*
3. Mostafa, N., Mostafa, N., Azim, M. A., Azim, M. A., Kabir, M. R., Kabir, M. R., … Ajwad, R. (2020). *Identifying the Risk of Cardiovascular Diseases From the Analysis of Physiological Attributes. 2020 IEEE Region 10 Symposium (TENSYMP).*
4. Pham, T. D., Honghui Wang, Xiaobo Zhou, Dominik Beck, Brandl, M., Hoehn, G., … Wong, S. T. C. (2008). *Computational Prediction Models for Early Detection of Risk of Cardiovascular Events Using Mass Spectrometry Data. IEEE Transactions on Information Technology in Biomedicine, 12(5), 636–643.*
5. Li-Na Pu, Ze Zhao, & Yuan-Ting Zhang. (2012). *Investigation on Cardiovascular Risk Prediction Using Genetic Information. IEEE Transactions on Information Technology in Biomedicine, 16(5), 795–808.*
6. Rahim, A., Rasheed, Y., Azam, F., Anwar, M. W., Rahim, M. A., & Muzaffar, A. W. (2021). *An Integrated Machine Learning Framework for Effective Prediction of Cardiovascular Diseases. IEEE Access, 9, 106575–106588.*
7. Bhuvaneswari Amma N G. (2013). *An intelligent approach based on Principal Component Analysis and Adaptive Neuro Fuzzy Inference System for predicting the risk of cardiovascular diseases. 2013 Fifth International Conference on Advanced Computing (ICoAC).*
8. Nikam, A., Bhandari, S., Mhaske, A., & Mantri, S. (2020). *Cardiovascular Disease Prediction Using Machine Learning Models. 2020 IEEE Pune Section International Conference (PuneCon).*
9. Loizou, C. P., Kyriacou, E., Griffin, M. B., Nicolaides, A. N., & Pattichis, C. S. (2021). *Association of Intima-Media Texture With Prevalence of Clinical Cardiovascular Disease. IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control, 68(9), 3017–3026.*
10. Bhatt, A., Kumar Dubey, S., & Kumar Bhatt, A. (2021). *Systematic Cardiovascular Health Analysis of Rural and Urban Residents for Early prediction of Cardiac Ailments. 2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence).*
11. Athanasiou, M., Sfrintzeri, K., Zarkogianni, K., Thanopoulou, A. C., & Nikita, K. S. (2020). *An explainable XGBoost–based approach towards assessing the risk of cardiovascular disease in patients with Type 2 Diabetes Mellitus. 2020 IEEE 20th International Conference on Bioinformatics and Bioengineering (BIBE).*
12. Joo, G., Song, Y., Im, H., & Park, J. (2020). *Clinical Implication of Machine Learning in Predicting the Occurrence of Cardiovascular Disease Using Big Data (Nationwide Cohort Data in Korea). IEEE Access, 8, 157643–157653.*
13. Xu, S., Shi, H., Duan, X., Zhu, T., Wu, P., & Liu, D. (2016). *Cardiovascular risk prediction method based on test analysis and data mining ensemble system. 2016 IEEE International Conference on Big Data Analysis (ICBDA).*
14. P, A., & Kalyani David, V. (2021). *Feature selection using ModifiedBoostARoota and prediction of heart diseases using Gradient Boosting algorithms. 2021 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS).*
15. Ema, R. R., & Shill, P. C. (2020). *Integration of Fuzzy C-Means and Artificial Neural Network with Principle Component Analysis for Heart Disease Prediction. 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT).*
16. Mendonca, F., Manihar, R., Pal, A., & Prabhu, S. U. (2019). *Intelligent Cardiovascular Disease Risk Estimation Prediction System. 2019 International Conference on Advances in Computing, Communication and Control (ICAC3).*