IJRAR.ORG E-ISSN: 2348-1269, P-ISSN: 2349-5138

INTERNATIONAL JOURNAL OF RESEARCH AND ANALYTICAL REVIEWS (IJRAR) | IJRAR.ORG

An International Open Access, Peer-reviewed, Refereed Journal

Bi-LSTM BASED FRAME WORK FOR CARDIOVASCULAR DISEASES RISK PREDICTION IN IMBALANCED BIG DATA

Submitted by

E.PADMA SUNDARI

MASTER OF TECHNOLOGY ININFORMATIONTECHNOLOGY FRANCIS XAVIER ENGINEERING COLLEGE

Dr.J.Shajilin Loret

Associate Professor

Francis Xavier Engineering college

ABSTRACT

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 a 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 Bi-LSTM classifier. 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.

CHAPTER-1

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]. It refers to the act of searching for huge information stores automatically, to identify patterns and trends which go beyond simple analytical procedures. It makes use of complicated mathematical algorithms for the segments of data and predicts the likelihood of future events. This has further, implemented several techniques from other areas like Machine Learning, Statistics, Visualization, Pattern Recognition, Artificial Intelligence, Database and Data Warehouse systems showed in Figure 1.1.

The scope of these areas makes it hard to understand the enormous change that has been achieved in recent decades [9, 10].

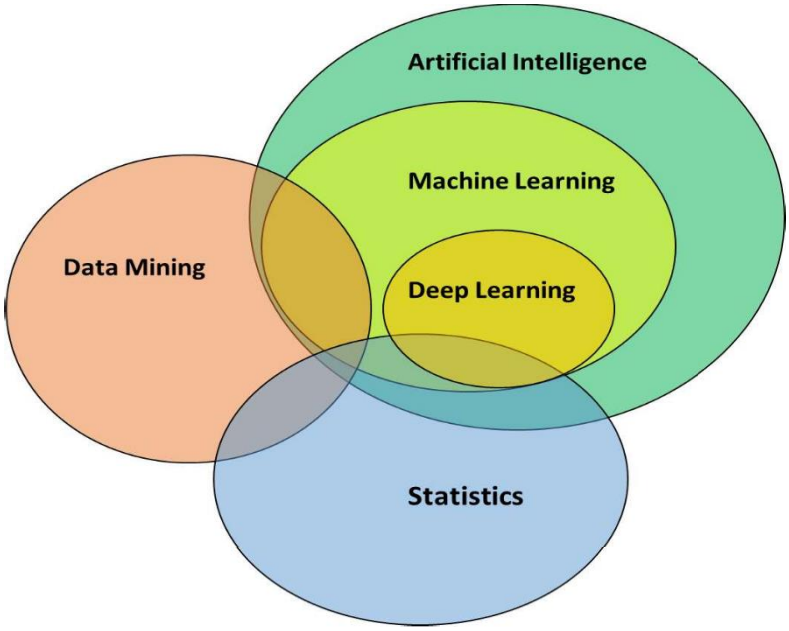


Figure 1.1. General phases of data mining techniques

1.1. Cardiovascular Diseases

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]. These CVDs are diagnosed using several techniques such as Echocardiography (ECHO), Tread Mill Test (TMT), Electrocardiogram (ECG) and Holter Monitoring (HM) tests can help doctors diagnose heart and blood vessel diseases and conditions in adults and children. The timely diagnosis of CVDs patients is the most challenging and complicated task for medical fraternity [16]. The prediction of cardiovascular disease is regarded as one of the most important subjects in healthcare. In this study, ECHO data is processed using statistical and data mining techniques to provide a predictive model for the likelihood of CVDs.

1.1.1. Global Burden of Cardiovascular Diseases

CVDs are the largest cause of mortality, accounting for around half of all deaths resulting from Noncommunicable Diseases (NCDs) and are the leading causes of death in the world, 24.8 % incidences of CVDs have gone up significantly for people between the age 25 and 69. The majority of these deaths are preventable, and despite preconceptions that men are more susceptible, women are equally likely to be affected [17]. There were associated inequalities in disease burden with disability-adjusted life years per 100,000 people due to CVDs over three times as high in middle-income [7160 (IQR 5655 - 8115)] compared with high-income [2235 (IQR 1896 - 3602)] countries. CVDs mortality was also higher in middle-income countries where it is accounted for a greater proportion of potential years of life lost compared with high-income countries in both females (43% vs. 28%) and males (39% vs. 28%) [18].

1.1.2 Cardiovascular Diseases Statistics in India

World Bank epidemiological modelling estimates India to have the second highest CVD mortality worldwide, at 2.5 million new cases occurring annually [12]. As per the WHO survey, the recent data suggests that age-standardized mortality rates of CVD in India, per 100,000, among females and males, are 181-281 and 363-443 respectively [17]. In India, the age-standardized mortality rate of CVD being, 272 per 100,000 population is greater than the world average recorded as 235 for 100,000 population [19]. The rapid urbanization in metropolitan cities in India has led to a range of concerns such as decreased physical activity, changed lifestyle, obesity, alcohol consumption, smoking and hypertension [20]. The National Health Policy 2017 of India aims to reduce 25 % of CVD premature deaths, by screening and treating 80 % of patients with hypertension, by 2025 [16].

1.1.4 Types of Cardiovascular Disease

CVDs are those disorders that affect cardiovascular function adversely, in the cardiovascular system and based on the tissue of target, several causes may result in these diseases. Blood test, Echocardiogram, ECG, Holter monitor, Ambulatory blood pressure monitoring, Transesophageal echocardiography, Chest X-ray, Cardiac MRI & Catheterization, CT scan, Treadmill and Angiogram are commonly used diagnostic methods for CVDs. The patient report

comprising of unstructured, structured and semi-structured data were recognized in the electronic health records.

1.1.4.1 Rheumatic Heart Disease (RHD)

Rheumatic heart disease is a condition in which the heart valves have been permanently damaged by rheumatic fever. It can affect connective tissue throughout the body, especially in the heart, joints, brain and skin. As a co-morbidity, RHD can permanently weaken the heart valves typically affecting children of age, 5-15 years. Streptococcal infections left untreated can raise the risk of heart failure in rheumatism and have been infrequent in the developing countries [21].

1.1.4.2 Ischemic Heart Disease (IHD)

Ischemic heart disease also called coronary artery disease or coronary heart disease is characterized as insufficient blood supply in heart regions due to blockage in the vessels supplying blood to the heart muscle. Anginal pain is a common indication of IHD and involves further laboratory test evidence like coronary angiography. Though narrowing may result from a blood clot or constriction of the blood vessel, it is most frequently caused due to plaque build-up, known as atherosclerosis. The complete blockage of supply of blood to the heart muscles results in the necrosis heart muscle cells, which is considered as myocardial infarction (MI) or heart attack [22].

1.1.4.3 Congenital Heart Disease

Congenital heart disease is a defect of heart existing at the time of birth and this is the most usual birth defect type that may cause changes in the normal flow of blood throughout the heart, prevailing in almost 1% of live births [23].

1.1.4.4 Cardiomyopathy

Cardiomyopathies are considered as "a heterogeneous myocardial disease group related to electrical or mechanical dysfunction that generally demonstrates dilation or improper ventricular hypertrophy, and is often genetic [24].

1.1.4.5 Valvular Heart Disease (VHD)

Valvular heart disease is a congenital defect in mitral, aortic, tricuspid, and pulmonary heart valves that have a common function to promote blood flow into the heart without obstruction. Stenosis and regurgitation are the damaged values that can cause diseases. Clinical procedure is important for evaluating the diagnosis, signs and identification of VHD by auscultation of the patient. Echocardiography plays a very significant role in the disease recognition and assessment of incidence as well as prognosis [22,25].

1.1.4.6 Atherosclerosis

Atherosclerosis is a disorder that causes plaque to build up in the interior of the arteries. It is capable of affecting any artery of the body consisting of brain, heart, legs, arms and pelvis arteries [26].

1.1.4.7 Aortic Valve Sclerosis and Stenosis

Aortic valve sclerosis and aortic valve stenosis, occurring most commonly in the elderly people are characterized by an increased thickness of the leaflet, rigidity, and calcification. With atherosclerosis and aortic stenosis appearing to be similar, several biochemical and clinical factors related to aortic sclerosis that seem to correspond to classical risk factors of atherosclerosis have been identified [27]. Aortic sclerosis has recently been found to be related to a substantial rise in the risk of myocardial infarction, heart failure and cardiovascular death [28].

1.2 Data Mining for Cardiovascular Diseases

Data Mining or Knowledge Discovery in Databases (KDD), refers to the process of investigation of hidden information patterns from different perspectives for categorization into useful information. Currently, data mining and KDD are utilized interchangeably by statisticians, data analysts, and information systems experts. This process includes different types of services like web mining, pictorial data mining, audio and video mining, text mining and social media mining [5,29]. The biggest challenge is to analyze the large data to extract important information that can be used to generating predictive knowledge. Data mining offers a set of techniques and tools for finding patterns and extracting knowledge in the CVD dataset that are difficult to detect with traditional statistical methods [30]. Hence, Data mining provides the methodology and technology

to predict the risk of cardiovascular diseases with high accuracy and less costs.

Data Preprocessing

The real-world databases are highly susceptible to missing, inconsistent and noisy data due to their typically huge size and their likely origin from multiple, heterogeneous sources [31]. Data preprocessing is a data mining technique that is, used to transform the CVDs patient dataset in a useful and efficient format [32]. The commonly used phases in the process of data mining for extracting knowledge is shown in Figure 1.2.

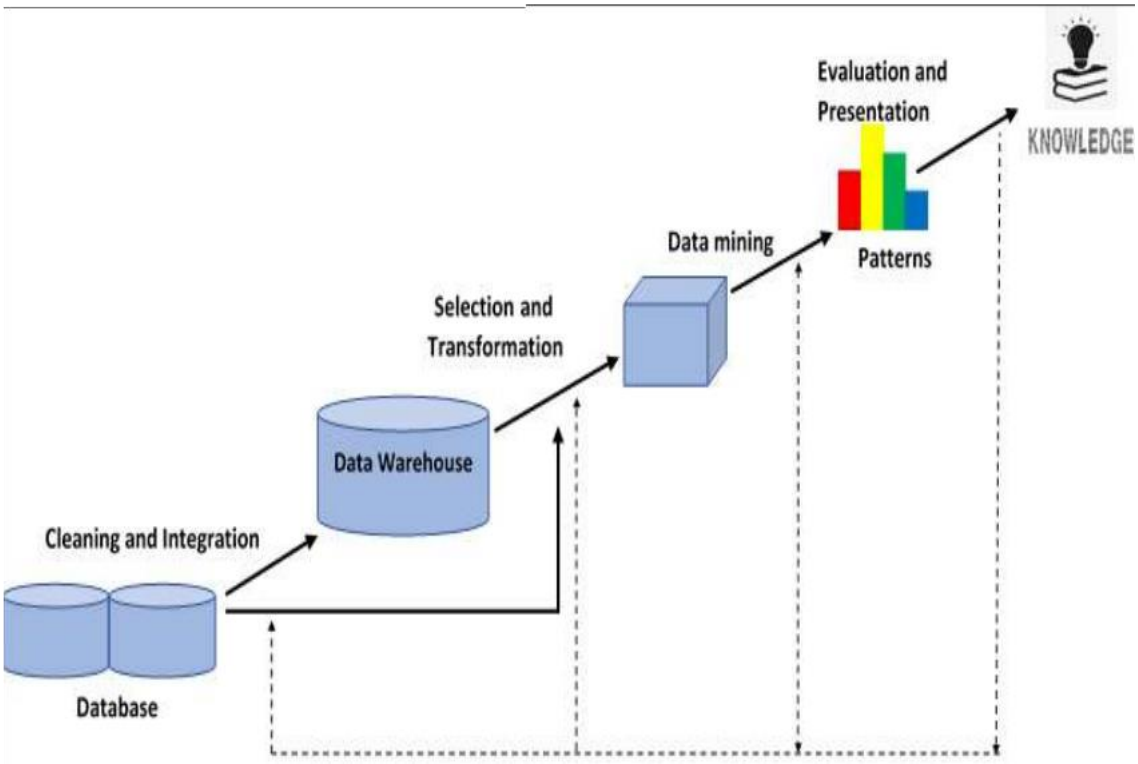


Figure 1.2. Data mining as a step in the process of knowledge discovery [5]

1.2.1 Data Mining Techniques

Data mining techniques give a confidence level about the predicted solutions in terms of consistency of prediction and accuracy. The interdisciplinary nature of data mining research and development contributes significantly to the success of data mining and its extensive applications [32,33]. The following are a few techniques of data mining:

(i) Statistics: Statistics is the branch of mathematics, which deals with the collection and analysis of numerical data by using various methods and techniques. It is used in various stages of the data

mining. For example, we can use statistics in the collection, sampling methods and analysis stage. Data summarization, point estimation, and Testing a hypothesis are the statistical techniques that find its extensive usage in data mining.

(ii) Machine Learning: Machine learning refers to the process of generating a computer system that has the capability of acquiring data independently and integrating that data to generate useful knowledge. The use of machine learning in data mining is that machine learning enables us to discover new and interesting structures and formats about a set of previously unknown data.

(iii) Database Systems and Data Warehouses: The database system mainly emphasis on creating, maintaining and use of organizational and end-user databases. The principles of database systems are high in the data models, query processing, query languages, storing data, methods of optimization, indexing, and accessibility methods. Recently database systems have made use of data mining and data warehousing facilities to build systematized analytical capabilities on the database.

(iv) Information Retrieval: Information retrieval is the process of searching for information in the documents. Documents may be in the multimedia or text form or may reside on the web. Effective search and analysis have raised many challenging issues in data mining. Therefore, multimedia mining and text data mining integrated with information retrieval methods have become increasingly important.

(v) Artificial Intelligence: It is a branch of computer science that create intelligent machines which can behave like a human, think like humans, and able to make decisions. Artificial intelligence machines can be trained to accomplish multiple tasks and improve themselves by processing large scale data.

1.2.3Data Mining Tasks

Data mining uses several algorithms for performing a variety of tasks. These algorithms examine the sample data of a problem and determine a model that fits close to solving the problem. The model that determines to solve a problem is classified as predictive and descriptive (Figure1.3).

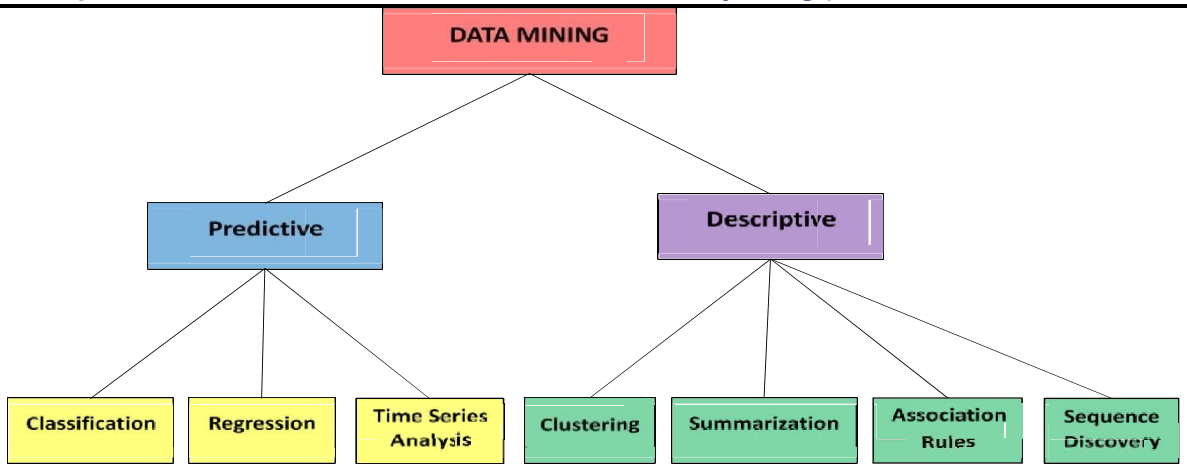


Figure 1.3. Data mining tasks

(a)**Predictive Model:** A predictive model enables the prediction of data values by using the results which are known from the different sample dataset [34-36]. The data mining tasks that form the part of the predictive model are:

(i)**Classification:** Classification refers to the process of determining a model for describing the concepts or data classes. The model is derived based on the analysis of a set of training data. The classification task allows not only the review and analysis of current data but also enable to predict the future behavior of sample data. The derived model can be represented in different forms, like classification rules, that is, IF-THEN rules, mathematical formulae, decision trees, naïve bayes and neural networks.

(ii)**Regression**: Regression Analysis is the most common statistical modeling approach used in data mining which is used to measure the average relationship between two or more variables in terms of the original unit of data. It allows the future data values to be forecasted depending upon the current as well as the past data values. This technique is used by researchers in many areas, especially in health sciences.

(iii)**Time Series Analysis:** Time series analysis is one of the data mining predictive models enables to predict future values for the current set of values and which is evenly distributed as hourly, daily, weekly, monthly and yearly to draw a time series plot. This analysis uses the present and the past data sample to estimate future value.

(b)**Descriptive Model**: A descriptive model allows us to determine the relationships and patterns in a data sample [5,37]. Descriptive model in data mining tasks are:

(i)**Clustering**: Data items that have close resemblance with one another are clubbed together in a single group is known as clusters which are referred to as unsupervised machine learning techniques. It enables to create new classes and groups depending on the study of patterns and relationships between data values in a database.

(ii)**Summarization**: Summarization is one of the descriptive models in data mining also be referred to as generalization or characterization. The use of summarization allows us to summarize a large chunk of data containing in a web page or a document. This task searches for specific data attributes and characteristics in the given large data volumes which can be summarized.

(iii) **Association Rules**: Association rules enable the establishment of association and relationships formed between large unclassified data items depending on certain features and attributes. It defines certain rules of associativity between data sets and then uses those rules to establish relationships.

(iv) **Sequence Discovery**: Sequence discovery is allowing us to determine the sequential patterns that might exist in a large and unorganized database. It discovers the sequence in a data bank by using the time factor i.e., associate the data items by the time at which it was generated.

1.2.4 Data Mining Applications

Data mining finds its applications in various fields in our day-to-day life. It is more useful for healthcare and pharmaceutical organizations, which produce large volumes of data in their diagnostic and clinical activities to access the doctors in making their clinical decision. Almost all the present-day varied organizations use data mining in all the phases of their work and which can apply the various concepts, methods, tasks, and techniques of data mining for decision-making. It can be utilized in several applications like Insurance, Education, Business analysis, Fraud Detection, Computer security, News & entertainment, Production Control, Health and Science Exploration [7,38].

1.3Statistical and Data Mining Tools

There are many computer programs available for analysis [36,39-41]. Some of these listed below software are commonly used for statistical and datamining are shown in Figure 1.4.

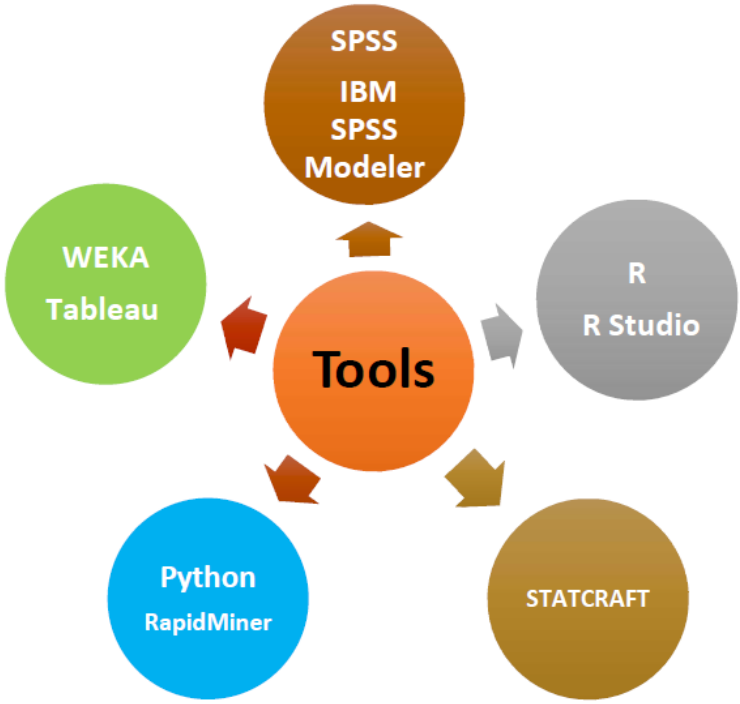


Figure 1.4. Statistical and data mining tools

(i)**R and R Studio Programming Language**: R is a free software environment for widely used statistics programming language among data miners for large scale data analysis. It offers graphical and statistical tools in addition to data mining. R Studio is an Integrated Development Environment (IDE) for R, it allows users to develop and edit programs in R software.

(ii)**STATCRAFT**: STATCRAFT is a web-server based platform that allows users to run data analytics in R from a browser-based GUI that eliminates the need to write complex R codes. STATCRAFT makes it easy for analysts to concentrate on analytics rather than coding.

(iii) **SPSS and IBM SPSS Modeler**: SPSS is the Statistical Package for Social Science and is useful for the organizations and researchers for analyzing complex statistical data. IBM SPSS Modeler is a data mining framework that provides a systematic method for discovering useful associations in large data sets.

(iv) **WEKA**: WEKA is open-source software that offers tools for data mining, machine learning and visualization. This software makes it easy to work with large data and train a machine using algorithms of machine learning which can be directly applied to a dataset.

(v) **Tableau**: Tableau is one of the most effective data mining tools, which helps optimize the task of visualization in the large data set. It is the easiest way to transform the raw data set into an easily comprehensible format with zero technological expertise and coding skills. Tableau may provide a connection to files, Big Data and relational sources.

(vi) **Python**: Python is a general-purpose interpreted, interactive, object-oriented, and high-level programming language. The Python interpreter and the extensive standard library are freely available in source or binary form for all major platforms from the Python web site.

(vii) **RapidMiner**: RapidMiner is one of the popular data mining tools and it is written in Java but no coding is required to operate it. It also offers numerous functionalities of data mining such as pre- processing of data, representation of data, filtering, as well as clustering.

This thesis 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.

CHAPTER-2

LITERATURE SURVEY

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. The simulation results confirm the effectiveness of the proposed method not only in terms of accuracy but also in terms of generalizability of the obtained models.

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. After data preprocessing, over 50 data mining techniques was tested for real patents dataset. Totake full advantage of multi-methods and reduce bias, top 6sub classifiers was selected to form an ensemble system, adjusted voting mechanism was used to make a final result, which consists of risk prediction and confidence.

System has a high precision of 79.3% for 2628 cases of real patents in experiment. Therisk prediction confidence and algorithm accuracy shows great significance in practical use for doctors' diagnosing.

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. We developed various ML-based prediction models using logistic regression, deep neural networks, random forests, and LightGBM, and validated them using various metrics such as receiver operating characteristic curves, precision-recall curves, sensitivity, specificity, and F1 score. Experimental results showed that all ML models outperformed the baseline method derived from the ACC/AHA guidelines for estimating the 10-year CVD risk, demonstrating the usefulness of ML methods. In addition, in our analysis, whether we included the past medication information as a feature or not, the prediction accuracy of all ML models was comparable to each other. Since the use of medications by the physicians provided important information on the occurrence of diseases, when we included it as a feature, all prediction models achieved a slightly higher prediction accuracy.presented as intervals of values for individual groups of surgical diseases and various age intervals. The total index is related to the determination of the risk of occurrence of cardiovascular complications of all levels of severity, and the lethal index to determination of the risk of lethal (severe) cardiovascular incidents.

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. Data from the 5- year follow up of 560 patients with T2DM are used for development and evaluation purposes. The obtained results (AUC=71.13%) indicate the potential of the proposed approach to handle the unbalanced nature of the used dataset, while providing clinically meaningful insights about the model's decision process 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. It is evident from the research study that males are affected more than twice of females by the cardiac health issues. The paper tries to figure out various factors affecting cardiac health among rural and urban residents of different age groups. The research outcomes motivates both rural and urban residents towards following a healthy routine and lifestyle in order to avoid such severity of cardiac health issues in future.

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. In ANFIS, the learning capabilities of neural network and reasoning capabilities of fuzzy logic is combined inorder to give better prediction. The heart disease dataset used is Cleveland Heart Disease dataset provided by the University of California, Irvine (UCI) Machine Learning Repository. The obtained classification accuracy using this approach is 93.2%.

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. Finally, an ensemble of Logistic Regression and K-Nearest Neighbor (KNN) classifiers is proposed for prediction with higher accuracy. The validation of framework is performed through three benchmark datasets (i.e. Framingham, Heart Disease and Cleveland) and the accuracies of 99.1%, 98.0% and 95.5 % are

achieved respectively. Finally, the comparative analysis proves that MaLCaDD predictions are more accurate (with reduced set of features) as compared to the existing state-of-the-art approaches. Therefore, MaLCaDD is highly reliable and can be applied in real environment for the early diagnosis of cardiovascular diseases

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. However, the potential clinical utility of genetic information has been uncovered initially and is expected for further development.

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. While the statistical and geostatistical techniques provide better results than those obtained from some other methods, the geostatistical approach yields superior results in terms of sensitivity and specificity in various designs of the data set for validation, training, and testing. The proposed computational strategies are very promising for predicting major adverse cardiac events within six months.

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. To enhance the performance of the algorithm models, we have generated a secondary dataset based on the output of the classification model, pushing the accuracy of the model to 97.03%. We have also evaluated the correlation of the attributes to the chance of having cardiovascular disease and found some general observation. Producing a secondary dataset, the analysis leading to the observable

patterns among the attributes and, defining general observation for cardiovascular disease using machine learning models make this study unique.

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. The model was accurate 90.62% of the time. Combined the model assessment results with risk control knowledge base, a personalized health management and health guidance given by care workers can be put forward intelligently, which can not only help medical personnel in the rational allocation but also guide patients to carry out self-management better, resulting in the decrease of readmission rate

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.

P. Kaur, et al explores the key ML algorithms for handing missing and incomplete corrupt health care data to understand its impact on real-time decision making for medical diagnostics. In this research we try to identify efficient ML algorithms that can work on real world data to adhere to the requirement for classification / prediction with least delay.Three methods kNN, MICE and Amelia are considered for imputing missing values of a cardiovascular dataset. Classification experiments are carried out using three popular algorithms decision tree, logistic regression and Naive Bayes along with deep autoencoders as a special case. Among non-DAE approaches decision tree was able to achieve better results with kNN for imputing with an improvement of 1.65%. DAE was able to achieve an improvement of 7.2% with retrained weights.

P. Swathi et al presents a thorough overview of both traditional and modern machine learning algorithms. Furthermore, it suggests a cardiovascular disease prediction model that is both efficient and accurate, notably for diabetes patients. This approach suggested uses the support vector machines and random forests. The experiments were performed, and the findings demonstrate how effective the suggested framework.

D. Vora et al proposed a decision support system based on machine learning helps physicians efficiently diagnose patients with heart disease. However, these diseases can be predicted using various machine learning models. Performance is evaluated using logistic regression, K-nearest neighbor method, random forest, and ANN. The accuracy of the random forest algorithm is 83.15%. This was far more accurate than the other algorithms described earlier. The proposed Ensemble learning is used to improve where more classification algorithms can be used simultaneously on a single dataset. The accuracy of the proposed model is 86.41 %. The proposed model helps in predicting the heart disease of various people with various complications.

S. Mishra, et al studied upon the Cleveland HD dataset to classify patients into five classes ranging from 0 to 4. Further, the Logistic Regression (LR), K-Nearest Neighbors (KNN), and Support Vector Machine (SVM) algorithms are implemented in their original forms and the performances are compared by implementing a One-vs-One (OVO) approach in terms of the recall provided. It was observed that the highest recall of 77.2% was provided by the OVO implementation of KNN algorithm.

M. U. Siregar, et al proposed a Random Forest based heart failure prediction from Kaggle. We experimented with two iterations for every nine combinations of the parameters. We compared the results of optimized random forest, stand-alone random forest, decisiontree, and Naïve Bayes algorithms. Our finding is that the optimized method is slightly better than the other algorithms. The best F1-score is obtained atthe second iteration which is 0.90789 compared to0.89404 obtained with the sole random forest, 0.85034 obtained with the decision tree, and 0.86195 obtained with Naive Bayes. The best recall value is 0.91925obtained in the first iteration, and in the second iteration. The best recall is also obtained with the sole random forest algorithm. The best precision valueis 0.89937 which was obtained in the first. By these results, the optimized random forest algorithm could be used to result in reliable predictions about heart failure.

Z. Ali, et al aimed to develop a system that integrates multiple machine learning algorithms, including K-nearest Neighbor, Naïve Byes, Linear Regression, Decision Tree, and Random Forest, which are used to detect cardiovascular disease. Five machine learning algorithm models were developed and their performances were observed based on several other performance indicators like accuracy, Precision, F1-score, Macro Average, and Weighted average among two target classes i.e. Presence and absence of cardiovascular disease. Classification reports generated against each model were utilized to assess the efficacy and strength of the constructed model.

M. Mesinovic et al propose a multi-label framework to predict the occurrence of 5 complications following admission of 1,700 patients after suffering an AMI episode. We evaluate the models using several multi-label prediction metrics as a test of robustness of our method beating numerous other alternatives and comment on the balance of cost-effectiveness of a compact deep learning model versus shallow machine learning in the multi-label context. Our neural network outperformed 13 other algorithms across all metrics, except Hamming loss. We also implement Shapley value analysis to this multi-label problem and observe interesting behaviour such as the duration of arterial hypertension and time elapsed from the beginning of the attack to the hospital being key predictive features of lethal outcome. This framework presents a novel approach in using multi-label learning, and especially compact cost-effective deep learning, simultaneous for prediction of several AMI complications which has not been explored previously.

Bin Ashraf, et al used a dataset that contains the clinical records of patients who have been admitted into a hospital with a heart problem and experimented with different classification algorithms to predict the type of heart problem that the patient got. We have experimented with the dataset from a different perspectives and a thorough discussion reveals that XGB ensemble classification performs best for this multi-class classification problem. This algorithm gives the best evaluation metric of 99% balanced accuracy, 0.99 ROC AUC, and a perfect F1 score.

CHAPTER 3

EXISTING SYSTEM

Neural Network

The field of Neural Networks has arisen from diverse sources. That is ranging from understanding and emulating the human brain to broader issues. That is of copying human abilities such as speech and use in various fields.

Generally, neural networks consist of layers of interconnected nodes. That each node producing a non-linear function of its input. And input to a node may come from other nodes or directly from the input data. Also, some nodes are identified with the output of the network.

On the basis of this, there are different applications for neural networks present. That involve recognizing patterns and making simple decisions about them.

a. Classification Algorithms in Data Mining

It is one of the Data Mining. That is used to analyze a given data set and takes each instance of it. It assigns this instance to a particular class. Such that classification error will be least. It is used to extract models. That define important data classes within the given data set. Classification is a two-step process.

During the first step, the model is created by applying a classification algorithm. That is on training data set.

Then in the second step, the extracted model is tested against a predefined test data set. That is to measure the model trained performance and accuracy. So classification is the process to assign class label from a data set whose class label is unknown.

b. ID3 Algorithm

This Data Mining Algorithms starts with the original set as the root hub. On every cycle, it emphasizes through every unused attribute of the set and figures. That the entropy of attribute. At that point chooses the attribute. That has the smallest entropy value.

The set is S then split by the selected attribute to produce subsets of the information.

This Data Mining algorithms proceed to recurse on each item in a subset. Also, considering only items never selected before. Recursion on a subset may bring to a halt in one of these cases:

- Every element in the subset belongs to the same class (+ or -), then the node is turned into a leaf and
- labeled with the class of the examples
- If there are no more attributes to select but the examples still do not belong to the same class. Then the node is turned into a leaf and labeled with the most common class of the examples in that subset.
- If there are no examples in the subset, then this happens. Whenever parent set found to be matching a specific value of the selected attribute.
- For example, if there was no example matching with marks >=100. Then a leaf is created and is labeled with the most common class of the examples in the parent set.

Working steps of Data Mining Algorithms is as follows,

- Calculate the entropy for each attribute using the data set S.
- Split the set S into subsets using the attribute for which entropy is minimum.
- Construct a decision tree node containing that attribute in a dataset.
- Recurse on each member of subsets using remaining attributes.

c. C4.5 Algorithm

C4.5 is one of the most important Data Mining algorithms, used to produce a decision tree which is an expansion of prior ID3 calculation. It enhances the ID3 algorithm. That is by managing both continuous and discrete properties, missing values. The decision trees created by C4.5. that use for grouping and often referred to as a statistical classifier.

C4.5 creates decision trees from a set of training data same way as an Id3 algorithm. As it is a supervised learning algorithm it requires a set of training examples. That can see as a pair: input object and the desired output value (class).

The algorithm analyzes the training set and builds a classifier. That must have the capaci ty to accurately arrange both training and test cases.

A test example is an input object and the algorithm must predict an output value. Consider the sample training data set S=S1, S2,…Sn which is already classified.

Each sample Si consists of feature vector (x1,i, x2,i, …, xn,i). Where xj represent attributes or features of the sample. The class in which Si falls. At each node of the tree, C4.5 selects one attribute of the data. That most efficiently splits its set of samples into subsets such that it results

in one class or the other.

The splitting condition is the normalized information gain. That is a non-symmetric measure of the difference. The attribute with the highest information gain is chosen to make the decision. General working steps of algorithm is as follows,

Assume all the samples in the list belong to the same class. If it is true, it simply creates a leaf node for the decision tree so that particular class will select.

None of the features provide any information gain. If it is true, C4.5 creates a decision node higher up the tree using the expected value of the class.

An instance of previously-unseen class encountered. Then, C4.5 creates a decision node higher up the tree using the expected value.

d. K Nearest Neighbors Algorithm

The closest neighbor rule distinguishes the classification of an unknown data point. That is on the basis of its closest neighbor whose class is already known.

In which nearest neighbor is computed on the basis of estimation of k. That indicates how many nearest neighbors are to consider to characterize.

It makes use of the more than one closest neighbor to determine the class. In which the given data point belongs to and so it is called as KNN. These data samples are needed to be in the memory at the runtime.

Hence they are referred to as memory-based technique.

The training points are assigned weights. According to their distances from sample data point. But at the same, computational complexity and memory requirements remain the primary concern.

To overcome memory limitation size of data set is reduced. For this, the repeated patterns. That don't include additional data are also eliminated from training data set.

To further enhance the information focuses which don't influence the result. That are additionally eliminated from training data set.

The NN training data set can organize utilizing different systems. That is to enhance over memory limit of KNN. The KNN implementation can do using ball tree, k-d tree, and orthogonal search tree.

The tree-structured training data is further divided into nodes and techniques. Such as NFL and tunable metric divide the training data set according to planes. Using these algorithms we can expand the speed of basic KNN algorithm. Consider that an object is sampled with a set of different attributes.

Assuming its group can determine from its attributes. Also, different algorithms can use to automate the classification process. In pseudo code, k-nearest neighbor algorithm can express,

K ← number of nearest neighbors

For each object Xin the test set do

calculate the distance D(X,Y) between X and every object Y in the training set

neighborhood ← the k neighbors in the training set closest to X

X.class ← SelectClass (neighborhood)

End for

e. Naïve Bayes Algorithm

The Naive Bayes Classifier technique is based on the Bayesian theorem. It is particularly used when the dimensionality of the inputs is high.

The Bayesian Classifier is capable of calculating the possible output. That is based on the input. It is also possible to add new raw data at runtime and have a better probabilistic classifier.

This classifier considers the presence of a particular feature of a class. That is unrelated to the presence of any other feature when the class variable is given.

For example, a fruit may consider to be an apple if it is red, round.

Even if these features depend on each other features of a class.

A naive Bayes classifier considers all these properties to contribute to the probability. That it shows this fruit is an apple. Algorithm works as follows,

Bayes theorem provides a way of calculating the posterior probability, P(c|x), from P(c), P(x), and P(x|c). Naive Bayes classifier considers the effect of the value of a predictor (x) on a given class (c). That is independent of the values of other predictors.

P(c|x) is the posterior probability of class (target) given predictor (attribute) of class.

P(c) is called the prior probability of class.

P(x|c) is the likelihood which is the probability of predictor of given class.

P(x) is the prior probability of predictor of class.

f. SVM Algorithm

SVM has attracted a great deal of attention in the last decade. It also applied to various domains of applications. SVMs are used for learning classification, regression or ranking function.

SVM is based on statistical learning theory and structural risk minimization principle. And have the aim of determining the location of decision boundaries. It is also known as a hyperplane. That produces the optimal separation of classes. Thereby creating the largest possible distance between the separating hyperplane.

Further, the instances on either side of it have been proven. That is to reduce an upper bound on the expected generalization error.

The efficiency of SVM based does not depend on the dimension of classified entities. Though, SVM is the most robust and accurate classification technique. Also, there are several problems.

The data analysis in SVM is based on convex quadratic programming. Also, expensive, as solving quadratic programming methods. That need large matrix operations as well as time-consuming numerical computations.

Training time for SVM scales in the number of examples. So researchers strive all the time for more efficient training algorithm. That resulting in several variant based algorithm.

SVM can also extend to learn non-linear decision functions. That is by first projecting the input data onto a high-dimensional feature space. As by using kernel functions and formulating a linear classification problem. The resulting feature space is much larger than the size of a dataset. That is not possible to store on popular computers.

Investigation of this issues leads to several decomposition based algorithms. The basic idea of decomposition method is to split the variables into two parts: a set of free variables called as a working set. That can update in each iteration and set of fixed variables. That are fix during a particular. Now, this procedure have to repeat until the termination conditions are met

SVM was developed for binary classification. And it is not simple to extend it for multi-class classification problem. The basic idea to apply multi- classification to SVM. That is to decompose the multi-class problems into several two-class problems. That can address using several SVMs.

g. ANN Algorithm

This is the types of computer architecture inspire by biological neural networks. They are used to approximate functions. That can depend on a large number of inputs and are generally unknown. They are presented as systems of interconnected "neurons". That can compute values from inputs. Also, they are capable of machine learning as well as pattern recognition. Due to their adaptive nature.

An artificial neural network operates by creating connections between many different processing elements. That each corresponding to a single neuron in a biological brain. These neurons may actually construct or simulate by a digital computer system.

Each neuron takes many input signals. Then based on an internal weighting. That produces a single output signal that is sent as input to another neuron.

The neurons are interconnected and organized into different layers. The input layer receives the input and the output layer produces the final output.

In general, one or more hidden layers are sandwiched between the two. This structure makes it impossible to forecast or know the exact flow of data.

Artificial neural networks start out with randomized weights for all their neurons. This means that they need to train to solve the particular problem for which they are proposed. A back-propagation ANN is trained by humans to perform specific tasks.

During the training period, we can test whether the ANN's output is correct by observing a pattern. If it's correct the neural weightings produce that output is reinforced. if the output is incorrect, those weightings responsible diminish.

Implemented on a single computer, a network is slower than more traditional solutions. The ANN's parallel nature allows it to built using many processors.That gives a great speed advantage at very little development cost.

The parallel architecture allows ANNs to process amounts of data very in less time. It deals with large continuous streams of information. Such as speech recognition or machine sensor data. ANNs can operate faster as compared to other algorithms.

An artificial neural network is useful in a variety of real-world applications. Such as visual pattern recognition and speech recognition. That deals with complex often incomplete data.

Also, recent programs for text-to-speech have utilized ANNs. Many handwriting analysis programs are currently using ANNs.

A decision tree is a predictive machine-learning model. That decides the target value of a new sample. That based on various attribute values of the available data. The internal nodes of a decision tree denote the different attributes.

Also, the branches between the nodes tell us the possible values. That these attributes can have in the observed samples. While the terminal nodes tell us the final value of the dependent variable.

The attribute is to predict is known as the dependent variable. Since its value depends upon, the values of all the other attributes.
The other attributes, which help in predicting the value of the dependent variable. That are the independent variables in the dataset.
The J48 Decision tree classifier follows the following simple algorithm. To classify a new item, it first needs to create a decision tree. That based on the attribute values of the available training data.

So, whenever it encounters a set of items. Then it identifies the attribute that discriminates the various instances most clearly.
This feature is able to tell us most about the data instances. So that we can classify them the best is said to have the highest information gain.
Now, among the possible values of this feature. If there is any value for which there is no ambiguity. That is, for which the data instances falling within its category. It has the same value for the target variable. Then we stop that branch and assign to it the target value that we have obtained.

1. Support Vector Machines

Support Vector Machines are supervised learning methods. That used for classification, as well as regression. The advantage of this is that they can make use of certain kernels to transform the problem. Such that we can apply linear classification techniques to non-linear data.
Applying the kernel equations. That arranges the data instances in a way within the multi-dimensional space. That there is a hyperplane that separates data instances of one kind from those of another.

The kernel equations may be any function. That transforms the non-separable data in one domain into another domain. In which the instances become separable. Kernel equations may be linear, quadratic, Gaussian, or anything else. That achieves this particular purpose.

Once we manage to divide the data into two distinct categories, our aim is to get the best hyperplane. That is to separate the two types of instances. This hyperplane is important, it decides the target variable value for future predictions. We should decide upon a hyperplane that maximizes the margin. That is between the support vectors on either side of the plane.

Support vectors are those instances that are either on the separating planes. The explanatory diagrams that follow will make these ideas a little more clear.

In Support Vector Machines the data need to be separate to be binary. Even if the data is not binary, these machines handle it as though it is. Further completes the analysis through a series of binary assessments on the data.

CHAPTER-4

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.

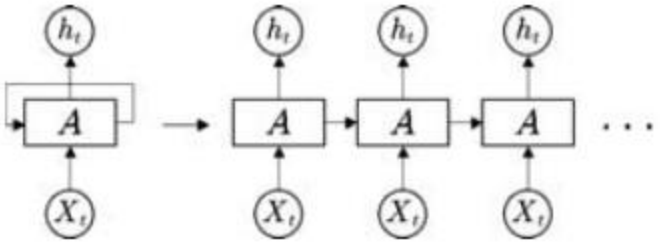


Figure 1 Recurrent neural network (RNN)

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

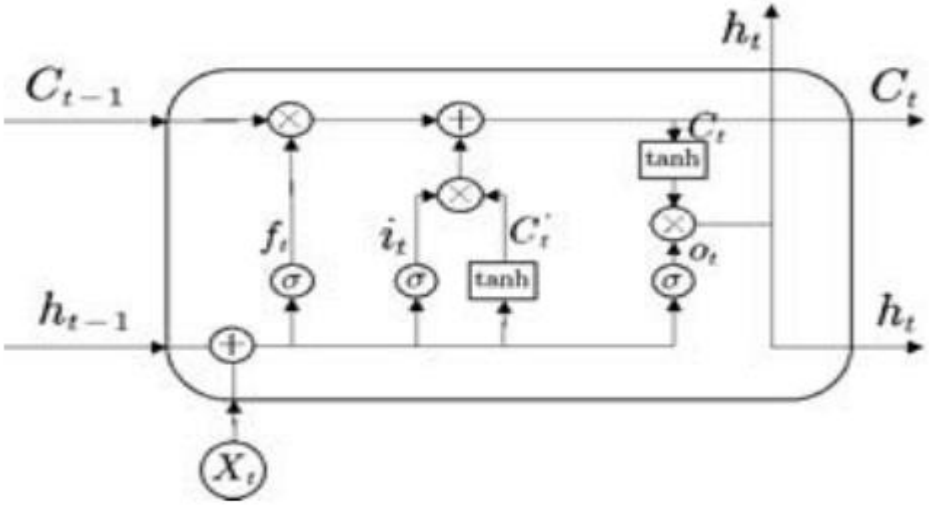


Figure The structure of LSTM cell.

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.

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

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

$$\tilde{C}_t = \tanh\left(W_c \cdot [h_{t-1}, x_t] + b_c\right)$$

$$C_t = f_t \cdot C_{t-1} + i_t \cdot \tilde{C}_{t'}$$

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

$$h_t = o_t \cdot \tan \mathrm{h}\left(C_t\right)$$

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.

Bi-LSTM Model

BiLSTM stands for Bidirectional Long Short-Term Memory, which is a type of recurrent neural network (RNN) architecture used in deep learning for sequence modeling tasks.

In a traditional LSTM (Long Short-Term Memory) network, information flows only in one direction, either from past to future (i.e., forward direction) or from future to past (i.e., backward direction). In contrast, a BiLSTM network processes input sequences in both directions simultaneously, by maintaining two separate hidden states: one for the forward direction and one for the backward direction. This allows the network to capture information from both the past and the future of the current time step, which can be beneficial in tasks that require context from both directions.

The BiLSTM architecture is capable of capturing long-term dependencies in sequential data, and its bidirectional nature allows it to capture context from both the past and future, making it a powerful tool for modeling sequential data with complex patterns. However, it also comes with increased computational complexity and may require larger amounts of training data compared to simpler models like unidirectional LSTMs or other types of recurrent neural networks.

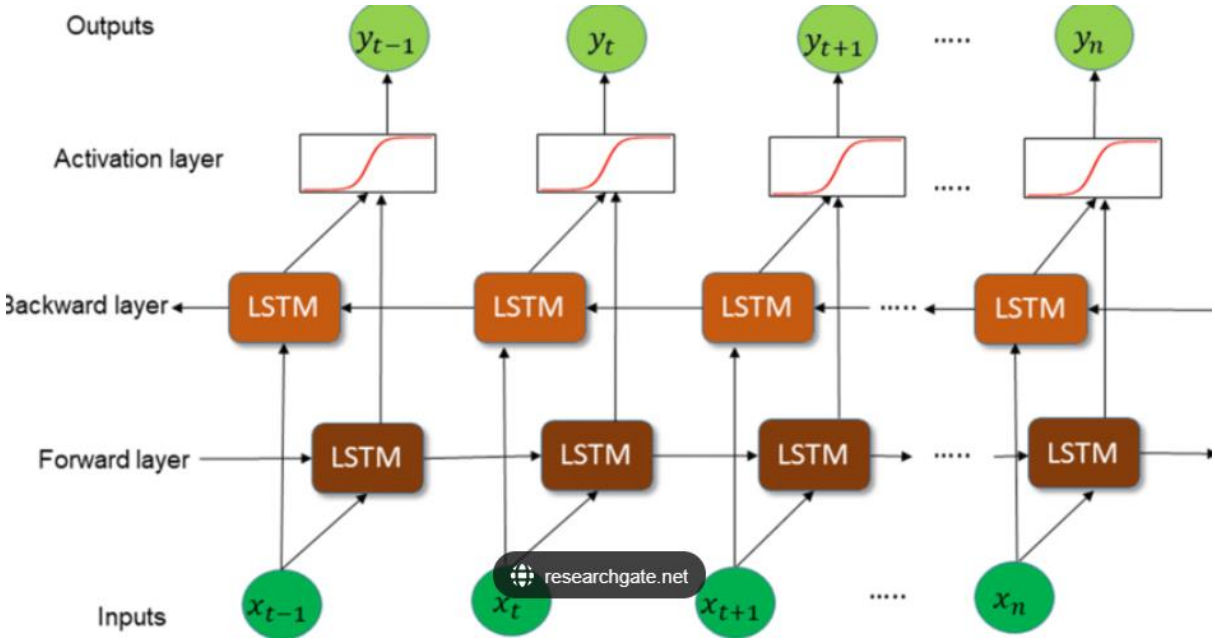


Figure The process of adding an attention mechanism to the LSTM model.

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.

Finally, a weighted summation operation is performed to obtain the Attention value Ci. The formula used in the attention layer is as follows:

$$H = [h_1 \ h_2 \ \cdots \ h_n]$$

$$H_i^{'} = \left[h_i^{'} \ h_i^{'} \ \cdots \ h_i^{'}\right]$$

$$sim_i = H_i^{'} \cdot H^T$$

$$a_i = \frac{e^{sim_i}}{\sum_{j=1}^{L_h} e^{sim_j}}$$

$$C_i = \sum_{j=1}^{L_h} a_i \cdot h_j$$

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

Proposed algorithm 1

Input: Input data

Output: A trained Bi-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.

CHAPTER-5

SOFTWARE SPECIFICATION

SOFTWARE REQUIRED:

- IDLE 1.7
- PYTHON 1.7.6

HARDWARE REQUIRED:

- System : Windows Xp Professional Service Pack 2
- Processor : Up to 1.5 GHz
- Memory : Up to 512 MB RAM

5.1 PYTHON

The Python language had a humble beginning in the late 1980s when a Dutchman Guido Von Rossum started working on a fun project, which would be a successor to ABC language with better exception handling and capability to interface with OS Amoeba at Centrum Wiskunde and Informatica. It first appeared in 1991. Python 2.0 was released in the year 2000 and Python 3.0 was released in the year 2008. The language was named Python after the famous British television comedy show Monty Python's Flying Circus, which was one of Guido's favorite television programmes. Here we will see why Python has suddenly influenced our lives and the various applications that use Python and its implementations.

5.1.1 Why Python?

Now you might be suddenly bogged with the question, why Python? According to Institute of Electrical and Electronics Engineers (IEEE) 2016 ranking Python ranked third after C and Java. As per Indeed.com's data of 2016, the Python job market search ranked fifth. Clearly, all the data points to the ever rising demand in the job market for Python. It’s a cool language if you want to learn just for fun or if you want to build your career around Python, you will adore the language. At school level, many schools have started including Python programming for kids. With new technologies taking the market by surprise Python has been playing a dominant role. Whether it is cloud platform, mobile app development, Big Data, IoT with Raspberry Pi, or the new Block

chain technology, Python is being seen as a niche language platform to develop and deliver a scalable and robust applications.

Some key features of the language are:

- Python programs can run on any platform, you can carry code created in Windows machine and run it on Mac or Linux
- Python has inbuilt large library with prebuilt and portable functionality, also known as the standard library
- Python is an expressive language
- Python is free and open source
- Python code is about one third of the size of equivalent C++ and Java code
- Python can be both dynamically and strongly typed--dynamically typed means it is a type of variable that is interpreted at runtime, which means, in Python, there is no need to define the type (int or float) of the variable

Python applications

One of the most famous platforms where Python is extensively used is YouTube. The other places where you will find Python being extensively used are the special effects in Hollywood movies, drug evolution and discovery, traffic control systems, ERP systems, cloud hosting, e-commerce platform, CRM systems, and whatever field you can think of.

Versions

At the time of writing this book, two main versions of the Python programming language were available in the market, which are Python 2.x and Python 3.x. The stable release as of writing the book were Python 2.7.13 and Python 3.6.0.

Implementations of Python

Major implementations include CPython, Jython, IronPython, MicroPython, and PyPy.

5.1.2 Installation

Here we will look forward to the installation of Python on three different OS platforms, namely, Windows, Linux, and Mac OS. Let's begin with the Windows platform.

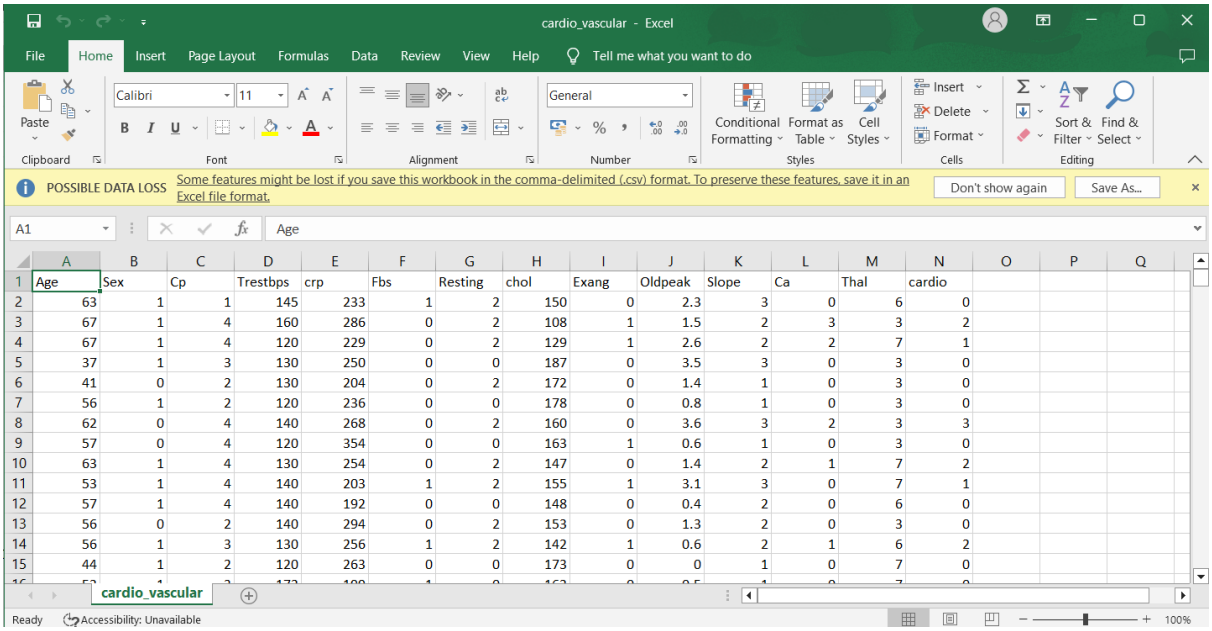
CHAPTER-6

SIMULATION RESULT

Dataset and Per processing

We implement all the approaches based on the data extracted from Xiangya Medical Dataset with Keras 2.2.2 . We split the dataset randomly into the training, validation and testing subset with a ratio of 0.7:0.1:0.2, namely the size of the training, validation and testing subset are 102,407, 14,630 and 29,259 respectively. For each predictive model, we train it in a mini-batch way with 1,024 sequences per epoch and conduct 100 iterations. In order to enhance the models' generalization performance, the data was divided independently and each

model was trained and tested 10 times in our work. Finally we report the mean evaluation metrics on the 10 testing results.



cardio_vascular - Excel

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Age	Sex	Cp	Trestbps	crp	Fbs	Resting	chol	Exang	Oldpeak	Slope	Ca	Thal	cardio
2	63	1	1	145	233	1	2	150	0	2.3	3	0	6	0
3	67	1	4	160	286	0	2	108	1	1.5	2	3	3	2
4	67	1	4	120	229	0	2	129	1	2.6	2	2	7	1
5	37	1	3	130	250	0	0	187	0	3.5	3	0	3	0
6	41	0	2	130	204	0	2	172	0	1.4	1	0	3	0
7	56	1	2	120	236	0	0	178	0	0.8	1	0	3	0
8	62	0	4	140	268	0	2	160	0	3.6	3	2	3	3
9	57	0	4	120	354	0	0	163	1	0.6	1	0	3	0
10	63	1	4	130	254	0	2	147	0	1.4	2	1	7	2
11	53	1	4	140	203	1	2	155	1	3.1	3	0	7	1
12	57	1	4	140	192	0	0	148	0	0.4	2	0	6	0
13	56	0	2	140	294	0	2	153	0	1.3	2	0	3	0
14	56	1	3	130	256	1	2	142	1	0.6	2	1	6	2
15	44	1	2	120	263	0	0	173	0	0	1	0	7	0

Figure : Data set description

```
Epoch 96/100
124/124 [==============================] - 0s 2ms/step - loss: 1.4896e-04 - accuracy: 1.0000
Epoch 97/100
124/124 [==============================] - 0s 2ms/step - loss: 1.2908e-04 - accuracy: 1.0000
Epoch 98/100
124/124 [==============================] - 0s 2ms/step - loss: 1.1651e-04 - accuracy: 1.0000
Epoch 99/100
124/124 [==============================] - 0s 2ms/step - loss: 1.0533e-04 - accuracy: 1.0000
Epoch 100/100
124/124 [==============================] - 0s 2ms/step - loss: 1.0121e-04 - accuracy: 1.0000
```

Figure LSTM Training

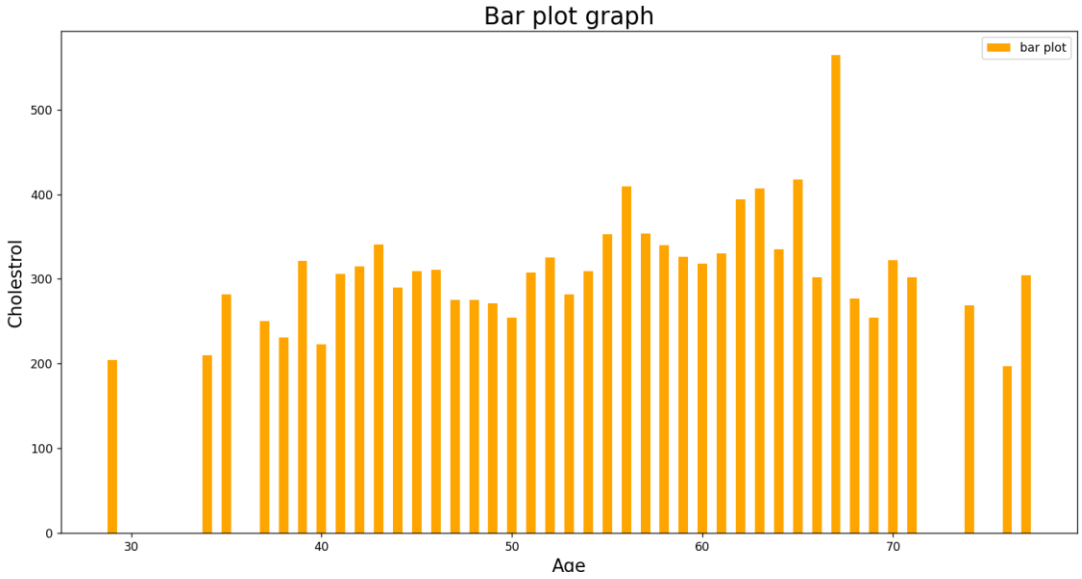


Figure : Cholesterol plot

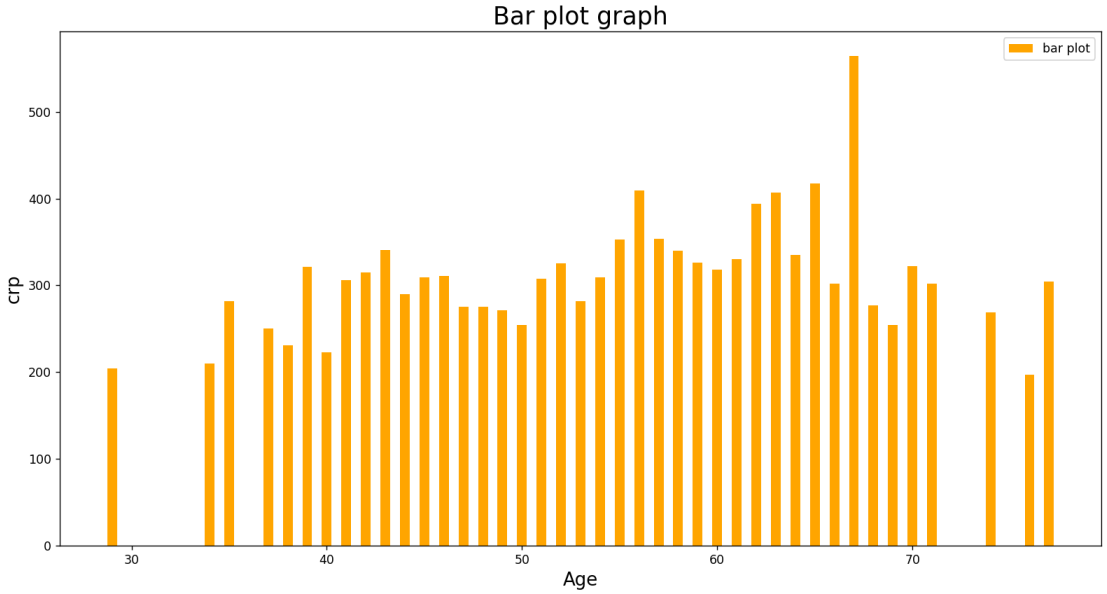


Figure : Crp plot

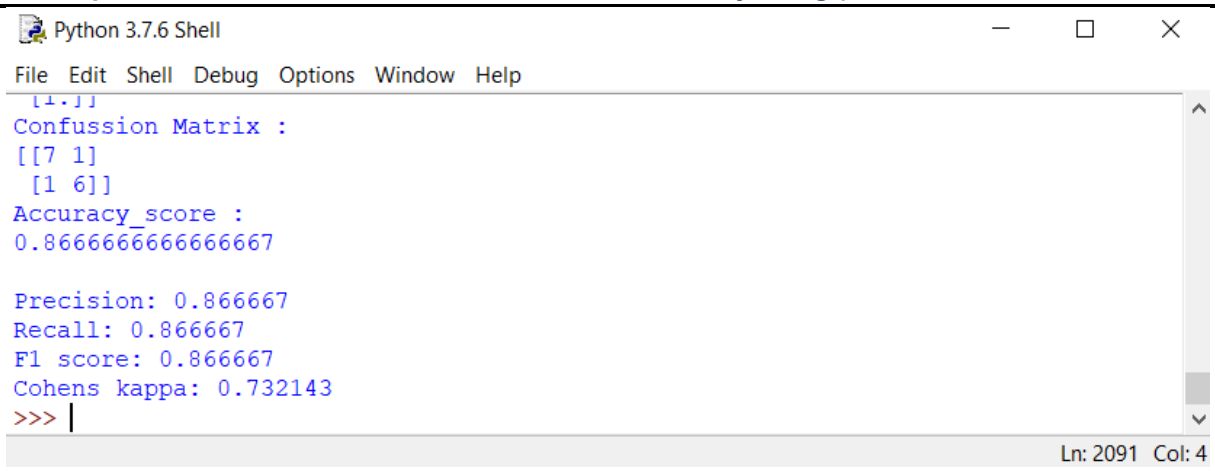


Figure Confusion matrix

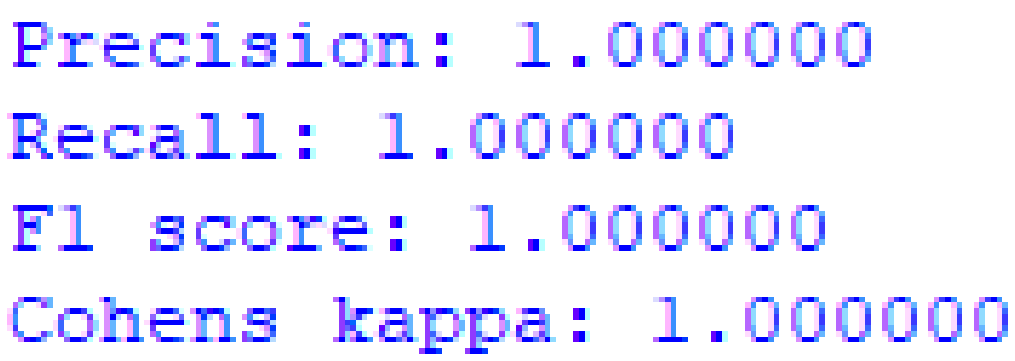
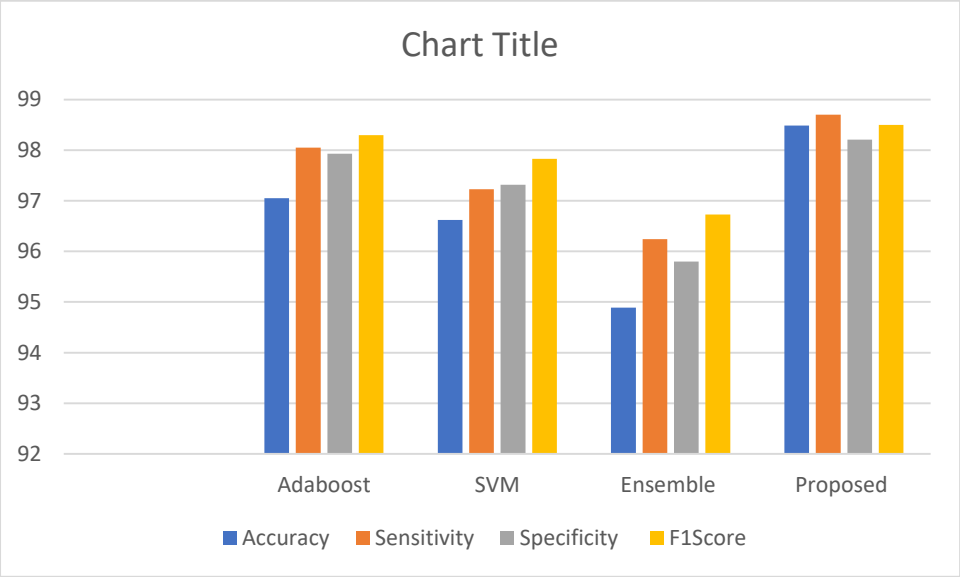


Figure Precision score

```
Real time Testing Started

enter datas separated by space : 63     1       1       145     233     1
2       150     0       2.3     3       0       6


user list is  ['63', '1', '1', '145', '233', '1', '2', '150', '0', '2.3', '3', '
0', '6']
[63.0, 1.0, 1.0, 145.0, 233.0, 1.0, 2.0, 150.0, 0.0, 2.3, 3.0, 0.0, 6.0]
[[0.]]

Cardiovascular disease status : Not Detected
For the given dataset the Predicted Value is Absence of Cardiovascular disease

enter datas separated by space :
```

Figure Testing result

Method	Accuracy	Sensitivity	Specificity	F1Score
Adaboost	97.05	98.05	97.93	98.3
SVM	96.62	97.23	97.32	97.83
Ensemble	94.89	96.24	95.8	96.73
Proposed	98.49	98.7	98.21	98.5



APPENDIX:

```
# Importing the libraries
import numpy as np
import time
import matplotlib.pyplot as plt
import pandas as pd
import tensorflow.keras as tf
from sklearn.preprocessing import StandardScaler
```

```
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import confusion_matrix, accuracy_score
from sklearn import model_selection
from sklearn.naive_bayes import GaussianNB
from sklearn.ensemble import VotingClassifier
from sklearn.metrics import precision_score
from sklearn.metrics import recall_score
from sklearn.metrics import f1_score
from sklearn.metrics import cohen_kappa_score
import sklearn
import scikitplot as skplt
import matplotlib.pyplot as plt
import seaborn as sns
import pickle
# Importing the dataset
dataset = pd.read_csv('heart.csv')
dataset=dataset.dropna(how="any")
print(dataset)
print(dataset.info())
#Data Visualization
#age vs cholestrol
m = dataset['age']
n = dataset['chol']
plt.figure(figsize=(4,4))
plt.title("Bar plot graph",fontsize=20)
plt.xlabel("Age",fontsize=15)
plt.ylabel("Cholestrol",fontsize=15)
plt.bar(m,n,label="bar plot",color=["orange"],width=0.5)
plt.legend(loc='best')
plt.show()
```

```
#age vs chest pain
m = dataset['age']
n = dataset['cp']
plt.figure(figsize=(4,4))
plt.title("Bar plot graph",fontsize=20)
plt.xlabel("Age",fontsize=15)
plt.ylabel("Cp",fontsize=15)
plt.bar(m,n,label="bar plot",color=["orange"],width=0.5)
plt.legend(loc='best')
plt.show()

X = dataset.iloc[:,:-1].values
y = dataset.iloc[:, 13].values
# Splitting the dataset into the Training set and Test set
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.25, random_state = 121)#101
#naive bayes training
print("naive bayes training")
model4 =GaussianNB()
history = model4.fit(X_train, y_train)
time.sleep(7);
y_pred = model4.predict(X_test)
print("ypred of naive bayes ")
print(y_pred)
#confussion Matrix
cm = confusion_matrix(y_test, y_pred)
print("Confussion Matrix for naive bayes")
print(cm)
cm_df = pd.DataFrame(cm,
                     index = ['normal','CVD'],
                     columns = ['normal','CVD'])
```

```
#Plotting the confusion matrix
plt.figure(figsize=(5,4))
sns.heatmap(cm_df, annot=True)
plt.title('Confusion Matrix of Naive Bayes')
plt.show()

nb = accuracy_score(y_test, y_pred
print("naive bayes accuracy is ")
print(nb)
print("")
testy = y_test
yhat_classes = y_pred
precision = precision_score(testy, yhat_classes)
print('Precision: %f' % precision)
recall = recall_score(testy, yhat_classes)
print('Recall: %f' % recall)
f1 = f1_score(testy, yhat_classes)
print('F1 score: %f' % f1)
# kappa
kappa = cohen_kappa_score(testy, yhat_classes)
print('Cohens kappa: %f' % kappa)

print("Training Completed")
```

CHAPTER-7

CONCLUSION AND FUTURE ENHANCEMENT

In this project, an attention layer is added to the existing LSTM model to constructed an Bi-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).

[17] F. Bin Ashraf, T. R. Siam, Z. Nayen and F. U. Zaman, "Identification of Cardiovascular Disorders Using Machine Learning Classification Algorithms," 2022 International Conference on Advancement in Electrical and Electronic Engineering (ICAEEE), Gazipur, Bangladesh, 2022, pp. 1-6, doi: 10.1109/ICAEEE54957.2022.9836433.

[18] M. Mesinovic and K. Yang, "Multi-label Neural Model for Prediction of Myocardial Infarction Complications with Resampling and Explainability," 2022 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI), Ioannina, Greece, 2022, pp. 01-05, doi: 10.1109/BHI56158.2022.9926915.

[19] Z. Ali, N. Naseer and H. Nazeer, "Cardiovascular Disease Detection Using Multiple Machine Learning Algorithms and their Performance Analysis," 2022 International Conference on Emerging Trends in Electrical, Control, and Telecommunication Engineering (ETECTE), Lahore, Pakistan, 2022, pp. 1-7, doi: 10.1109/ETECTE55893.2022.10007319.

[20] M. U. Siregar, I. Setiawan, N. Z. Akmal, D. Wardani, Y. Yunitasari and A. Wijayanto, "Optimized Random Forest Classifier Basedon Genetic Algorithm for Heart Failure Prediction," 2022 Seventh International Conference on Informatics and Computing (ICIC), Denpasar, Bali, Indonesia, 2022, pp. 01-06, doi: 10.1109/ICIC56845.2022.10006987.

[21] S. Mishra, M. Pandey and S. S. Rautaray, "Machine Learning based Cardiovascular Disease Prediction Using One-vs-one Approach," 2022 International Conference on Augmented Intelligence and Sustainable Systems (ICAISS), Trichy, India, 2022, pp. 203-208, doi: 10.1109/ICAISS55157.2022.10011022.

[22] D. Vora, S. Mishra, A. Mukherjee, S. Tiwari, S. Thakur and S. Biswas, "Heart Failure Prediction with Ensembled Learning," 2022 IEEE Pune Section International Conference (PuneCon), Pune, India, 2022, pp. 1-5, doi: 10.1109/PuneCon55413.2022.10014832.

[23] P. Swathi and M. Gunasekaran, "A Methodology For Early Prediction and Classification of Heart Diseases in Diabetic Patients With Machine Learning Techniques," 2022 IEEE 2nd International Conference on Mobile Networks and Wireless Communications (ICMNWC), Tumkur, Karnataka, India, 2022, pp. 1-7,

[24] P. Kaur, Z. Nisa and S. S. Tirumala, "Assessing Machine Learning Approaches for Imputing Missing Values in Cardiovascular Dataset," 2022 IEEE 7th International conference for Convergence in Technology (I2CT), Mumbai, India, 2022, pp. 1-6, doi: 10.1109/I2CT54291.2022.9825194.