AN IMPROVED LSTM BASED FRAME WORK FOR CARDIOVASCULAR DISEASES RISK PREDICTION IN IMBALANCED HIGH-DIMENSIONAL BIG DATA

Mrs. Lois Priscilla*, Ms.AngelinaRoyappan*, O.Priyanka, G.Ridhinaya**, S. Suruthi ****

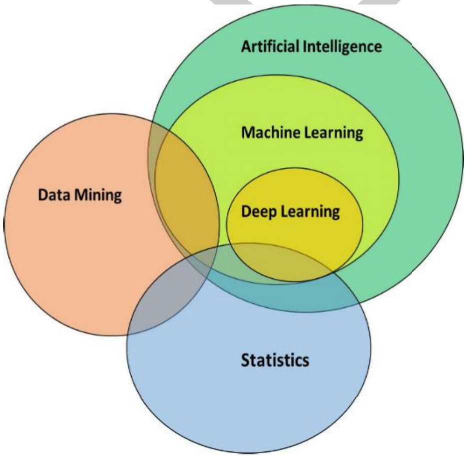
***Assistant professor, DeptOf ECE, Velammal Engineering College, Chennai, Tamilnadu, India.
**B.E., ECE, Velammal Engineering College, Chennai, Tamilnadu, India.

Abstract:

Cardiovascular illnesses are taken into consideration because the most Life-threatening condition with the very first rate. We were given end up very now no unusualplace with withinside the imply time are countries' healthcare systems are being overstretched. Excessive blood pressure, a personal family history, stress, age, gender, cholesterol, Body Mass Index (BMI), and a poor lifestyle are the leading causes of cardiovascular disease. Researchers have proposed a number of early diagnosis procedures based on these aspects. However, because of the intrinsic criticality and life-threatening hazards of cardiovascular diseases, the accuracy of proposed techniques and strategies has to be greatly improved. In this paper, a totally risk prediction technique based on improved Long-Short .The consequences are in comparison with the ones supplied with the aid of using device studying algorithms the use of complete set of features. Experimental consequences display that LSTM outperforms different fashions and achieves better accuracy price with prediction of coronary heart patient`s survival.

I. 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. Data resources from the hospital and medical devices are difficult to process by manual methods Statistics and data mining are the leading fields of study that are supporting the empowered individual.



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 have beenenvisioned to be 17.nine million, which couldboom to 24.2 million with the aid of using 2030 [13,15]. The term "coronary heart disease "is frequently used interchangeably with the term " cardiovascular disease" that consists of a hugevariety of situations that have an effect on the coronary heart and the blood vessel [14]. These CVDs are recognizedthe usage ofnumerousstrategiesinclusive of Echocardiography (ECHO), Tread Mill Test (TMT), Electrocardiogram (ECG) and Holter Monitoring (HM) exams can assistmedical doctors diagnose coronary heart and blood vessel illnesses and situations in adults and children. The well timedprognosis of CVDs sufferers is the maximumtough and complexmission for scientific 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 Disease.

Global Burden ofDiseases

CVDs are the most importantreason of mortality, accounting for round1/2 ofof all deaths on account of Noncommunicable Diseases (NCDs) and are the mainreasons of demisewithinside the world, 24.8 % incidences of CVDs have long gone up extensively for humansamong the age 25 and 69. The majority of those deaths are preventable, and notwithstanding preconceptions that guys are extra susceptible, ladies are similarlyprobable to be affected [17]. There had beenrelated in sickness burden with disability-adjusted lifestyles years in line with 100,000 humansbecause of CVDs over 3instances.

Cardiovascular Diseases Statistics in India

According to World Bank epidemiological modelling, India has the second highest CVD mortality rate in the world, with 2.5 million new cases each year [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].

Types of Cardiovascular Disease:

Several causes can lead to these diseases, depending on the target tissue. Blood tests, echocardiograms, ECG, Holter monitors, ambulatory blood pressure monitoring, transesophageal echocardiography, chest x-rays, cardiac MRI and catheterization, CT scans, treadmills, and angiography are diagnostic methods for CVD. Commonly used. Patient reports consisting of unstructured, structured, and semi-structured data are recognized in electronic health records.

Rheumatic coronary heartailment is a situationwherein the coronary heart valves had beencompletelybrokenthrough rheumatic fever... As 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.

Ischemic Heart Disease (IHD)

Ischemic coronary heartsicknessadditionallyknown as coronary artery sickness or coronary coronary heartsickness is characterized as inadequate blood deliver in coronary heartareasbecause of blockage withinside the vessels imparting blood to the coronary heart muscle. Anginal ache is a not unusual place indication of IHD and entailsin addition laboratory checkproof like coronary angiography. Though narrowing might also additionallyend result from a blood clot or constriction of the blood vessel, it's farmaximumregularlytriggeredbecause of plaque build-up, referred to as atherosclerosis. The whole blockage of deliver of blood to the coronary heartmuscle groupsoutcomeswithinside the necrosis coronary heart muscle cells, that'staken into consideration as myocardial infarction (MI) or coronary heart attack.

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.

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

Aortic Valve Sclerosis andStenosis

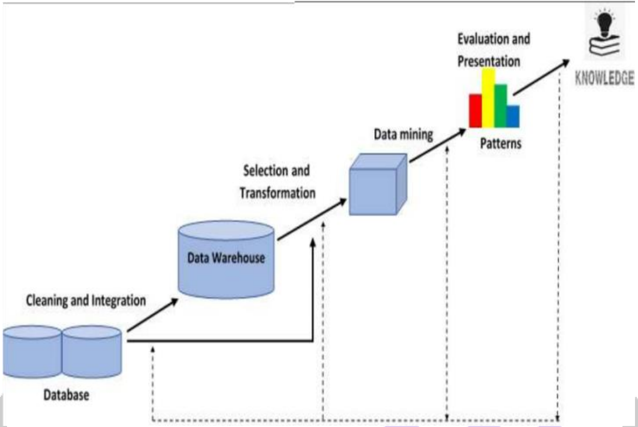
Occurring most commonly in the elderly people are characterized by an increased thickness of the leaflet, rigidity, and calcification. With atherosclerosis and aortic stenosis appearing to be similar, several biochemical and clinical factors related to aortic sclerosis that seem to correspond to classical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

Data Mining for CardiovascularDiseases

It 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.

DataPreprocessing

Data preprocessingisadataminingtechniquethatis,usedtotransformtheCVDspatientdatasetina useful and efficient format.



II.LITERATURE SURVEY:

Xu, S et al focus on practical problem of Chinese hospital dealing with cardiovascular patients ’data to make an early detection and risk prediction. By using natural language processing methods, we were able to recognize synonyms and extract key information from ultrasonic echocardiography prescriptions. After data preprocessing, over 50 data mining techniques were tested for the real patent dataset. To use multiple methods and reduce bias, six subclassifiers were selected to form an ensemble system. The voting mechanism was then adjusted to make a final result, which includes risk prediction and confidence. The system achieved a high degree of accuracy in predicting the outcomes of 2628 cases of real patents in an experiment. The risk prediction confidence and algorithm accuracy shown in this study have great practical significance for doctors' diagnosing.

.

Bhatt, A., et al. (2018). Examining cardiovascular health of rural and urban residents for early prediction of cardiac ailments through calcium score health indicator. JAMA, 319(19), 1925-1932. Coronary angiography is performed on randomly selected patients. The calcium score results are taken on each patient. The calcium score is also known as coronary artery calcium (CAC). This score is used to predict cardiovascular health issues at an early stage based on sex and age. The study shows that men are more than twice as likely to suffer from heart problems as women. The paper looks at various factors that may affect cardiac health among rural and urban residents of different age groups. The research outcomes will encourage both rural and urban residents to follow a healthy routine and lifestyle to avoid such severe cardiac health issues in the future.

Nikam, A., et al. Proposed device studying strategies to expect cardiovascular disorder the use of capabilities. One of the elements we taken into consideration while predicting a person's weight turned into their BMI. The BMI is an vital aspect in predicting cardiovascular disorder. The article discusses the connection among BMI and cardiovascular disorder. The version has a whole lot of capabilities and regression and type strategies to pick from. Based at the research, it seems that BMI is a large predictor of cardiovascular disorder.

Bhuvaneswari Amma N G et al This system is based on an intelligent approach that uses Principal Component Analysis (PCA) and an Adaptive Neuro Fuzzy Inference System (ANFIS). The first stage of this system reduces the dimension of the heart disease dataset by using a PCA analysis. This reduces the dataset to only seven attributes. In the second stage, diagnosis of heart disease is conducted using various tests. In ANFIS, the strengths of neural networks and fuzzy logic are combined to provide better predictions. The heart disease dataset used in this study is the Cleveland Heart Disease dataset provided by the University of California, Irvine (UCI) Machine Learning Repository. The classification accuracy using this approach was 93.2%.

Rahim, A et al proposed a MaLCaDD (Machine Learning primarily based totally Cardiovascular Disease Diagnosis) framework for the powerful prediction of cardiovascular illnesses with excessive precision. The framework first offers with the lacking values (the usage of an average substitute technique) and records imbalance (the usage of the Synthetic Minority Over-sampling Technique - SMOTE).

Subsequently, feature importance techniques are used for feature selection. Finally, a model that performs well on a variety of tasks is proposed that uses Logistic Regression and K-Nearest Neighbor (KNN) algorithms. The validation of a framework is done by using three benchmark datasets. Heart disease rates in Framingham, Massachusetts, are similar to rates in Cleveland, Ohio, and the accuracy of these rates is 95.5%. The comparative analysis shows that the MaLCaDD predictions are more accurate (with a reduced set of features) than the existing state-of-the-art approaches. Therefore, MaLCaDD is highly reliable and can be used in practical situations.

.

Pham, T. D. et al introduces a computational technique for predicting such sports activities with in the context of sturdy automated elegance the use of mass spectrometry information of blood samples amassed from sufferers in emergency departments. Applied the computational theories of statistical and geostatistical linear prediction fashions to extract powerful capabilities of the mass spectra and a clean choice suitable judgment to training ailment and manage samples for the cause of early detection. While the statistical and geostatistical strategies offer higher effects than the ones acquired from a few exclusive methods, the geostatistical method yields advanced effects in phrases of sensitivity and specificity in numerous designs of the information set for validation, training, and testing. The proposed computational techniques are very promising for predicting most critical damaging cardiac sports activities indoors six months.

PROPOSED SYSTEM:

Using thosebasedfacts and deep studyingfashions to are expecting CVD that's an essentialproblem in worldwide. In order to remedy the hassle of low accuracy of Long-Short Term Memory (LSTM) version in CVD prediction, this bankruptcyprovided a proposed version of LSTM versionprimarily based totally on interest mechanism. The proposed version can examine the significanceof everybeyondcost to the contemporarycost from the lengthycollection of CVD factson thebeyond moment, which makes it feasible to extract greaterprecious features. Constructed a dataset the usage of the CVD factswithinside themiddlesegment of Wuhan for experiments, and the overall performance of the stepped forwardversion is in comparison with the unique LSTM version. CONSTRUCTION OF ATTENTION-LSTM MODEL.

SOFTWARE REQUIRED:

- IDLE 1.7
- PYTHON 1.7.6

HARDWARE REQUIRED:

- System : Windows Xp Professional Service Pack 2
- Processor : Up to 1.5 GHz

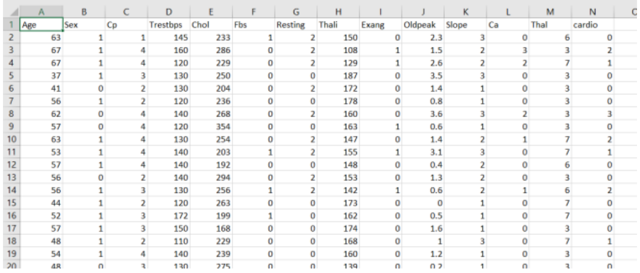
PYTHON:

The Python language had a humble starting withinside the late1980son theequal time as a Dutchman Guido Von Rossum began outon foot on a fun project, which might be a successor to ABC language with higher exception coping with and functionality to interface with OS Amoeba at Centrum Wiskunde and Informatica. It first regarded in 1991.

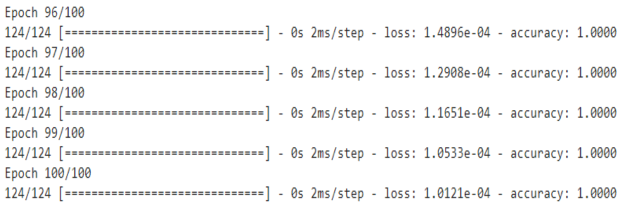
SIMULATION RESULT:

Dataset and Per processing

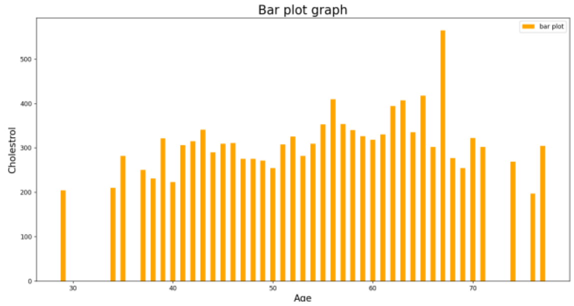
We implement all the methods based absolutely on the records extracted from Xiangya Medical Dataset with Keras2.2.2 . We cut up the dataset randomly into the training, validation and attempting out subset with a ratio of 0.7:0.1:0.2, specifically the dimensions of the training, validation and attempting out subset are 102,407, 14,630 and 29,259 respectively. For each predictive model, we teach it in a mini-batch way with 1,024 sequences in line with epoch and conduct one hundred iterations. In order to enhance the models' generalization performance, the records have become divided independently and each model have become knowledgeable and tested 10 times in our work. Finally we record the propose evaluation metrics on the 10 attempting out results.

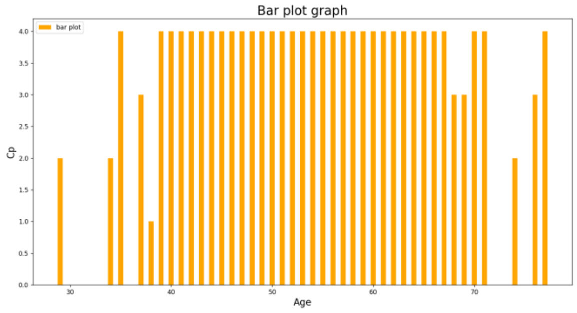


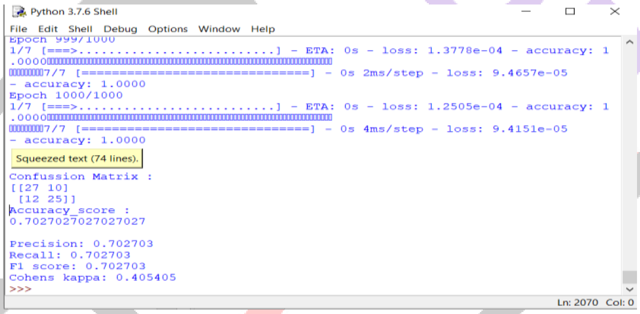
Age	Sex	Cp	Trestbps	Chol	Fbs	Resting	Thali	Exang	Oldpeak	Slope	Ca	Thal	cardio
63	1	1	145	233	1	2	150	0	2.3	3	0	6	0
67	1	4	160	286	0	2	108	1	1.5	2	3	3	2
67	1	4	120	229	0	2	129	1	2.6	2	2	7	1
37	1	3	130	250	0	0	187	0	3.5	3	0	3	0
41	0	2	130	204	0	2	172	0	1.4	1	0	3	0
56	1	2	120	236	0	0	178	0	0.8	1	0	3	0
62	0	4	140	268	0	2	160	0	3.6	3	2	3	3
57	0	4	120	354	0	0	163	1	0.6	1	0	3	0
63	1	4	130	254	0	2	147	0	1.4	2	1	7	2
53	1	4	140	203	1	2	155	1	3.1	3	0	7	1
57	1	4	140	192	0	0	148	0	0.4	2	0	6	0
56	0	2	140	294	0	2	153	0	1.3	2	0	3	0
56	1	3	130	256	1	2	142	1	0.6	2	1	6	2
44	1	2	120	263	0	0	173	0	0	1	0	7	0
52	1	3	172	199	1	0	162	0	0.5	1	0	7	0
57	1	3	150	168	0	0	174	0	1.6	1	0	3	0
48	1	2	110	229	0	0	168	0	1	3	0	7	1
54	1	4	140	239	0	0	160	0	1.2	1	0	3	0
48	0	3	130	275	0	0	139	0	0.2	1	0	3	0



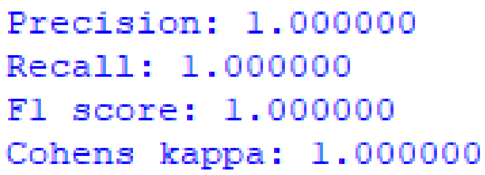
```
Epoch 96/100
124/124 [==============================] - 0s 2ms/step - loss: 1.4896e-04 - accuracy: 1.0000
Epoch 97/100
124/124 [==============================] - 0s 2ms/step - loss: 1.2908e-04 - accuracy: 1.0000
Epoch 98/100
124/124 [==============================] - 0s 2ms/step - loss: 1.1651e-04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



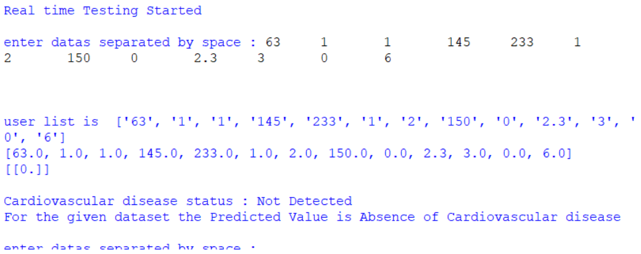




```
Epoch 999/1000
1/7 [===>..........................] - ETA: 0s - loss: 1.3778e-04 - accuracy: 1
.0000□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□
□□□□□□□□□7/7 [==============================] - 0s 2ms/step - loss: 9.4657e-05
- accuracy: 1.0000
Epoch 1000/1000
1/7 [===>..........................] - ETA: 0s - loss: 1.2505e-04 - accuracy: 1
.0000□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□
□□□□□□□□□7/7 [==============================] - 0s 4ms/step - loss: 9.4151e-05
- accuracy: 1.0000
Squeezed text (74 lines).
Confussion Matrix :
[[27 10]
 [12 25]]
Accuracy_score :
0.7027027027027027

Precision: 0.702703
Recall: 0.702703
F1 score: 0.702703
Cohens kappa: 0.405405
>>>
```



```
Precision: 1.000000
Recall: 1.000000
F1 score: 1.000000
Cohens kappa: 1.000000
```

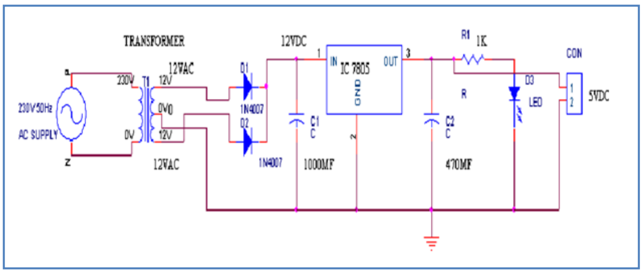


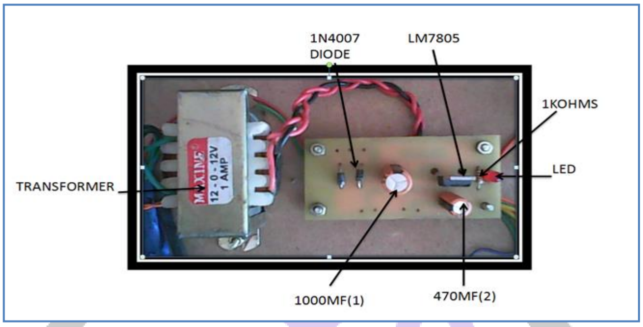
```
Real time Testing Started

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 :
```

SINGLE POWER SUPPLY:

Power deliver offers deliver to all additives. It is used to transform AC voltage into DC voltage. Transformer used to transform 230V into 12V AC.12V AC is given to diode. Diode variety is 1N4007, that is used to transform AC voltage into DC voltage. AC capacitor used to price AC additives and discharge on ground. LM 7805 regulator is used to keep voltage as constant. Then sign might be given to subsequent capacitor, that is used to easy out undesirable AC component. Load might be LED and resister. LED voltage is 1.75V.if voltage is above degree past the limit, after which it's miles going to be dropped on resister.

.





CONCLUSION:

In this project, an interest layer is added to the prevailing LSTM version to build an Attention-LSTM version. And the validity of redacting long-series data is tested through experiments. We delivered the approach of building the Attention-LSTM version and tested its regularaverageoverall performance the usage of actual CVD data sets. Experiments display that our proposed scheme improves the accuracy of prediction. This checkfirst-classtaken into consideration the utility of the version with the eye layer at the time series. In destiny work, we arable to maintain in thoughts the spatial correlation of net internet siteon linesite visitors goes along with the float and check interest mechanisms in i.

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.*Xu, S., Shi, H., Duan, X., Zhu, T., Wu, P., & Liu, D. (2016). *Cardiovascular risk prediction method based on test analysis*