Journal of Computer Science and Technology Studies
ISSN: 2709-104X
DOI: 10.32996/jcsts
Journal Homepage: www.al-kindipublisher.com/index.php/jcsts

JCSTS
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| RESEARCH ARTICLE

Deep Learning Application of LSTM(P) to predict the risk factors of etiology cardiovascular disease

Shake Ibna Abir[1]✉, Shaharina Shoha[1], Sarder Abdulla Al shiam[2], Nazrul Islam Khan[3], Abid Hasan Shimanto[4], Rafi Muhammad Zakaria[4], and S M Shamsul Arefeen[4]

[1]Instructor of Mathematics, Department of Mathematics and Statistics, Arkansas State University, Jonesboro, Arkansas, USA
[2]Department of Management, St Francis College, New York, USA
[3]Department of Mathematics & Statistics, Stephen F. Austin State University, Texas, USA
[4]Department of Management Science and Information Systems, University of Massachusetts Boston, Boston, USA
Corresponding Author: Shake Ibna Abir, **E-mail**: sabir@astate.edu

| ABSTRACT

Cardiovascular vascular disease (CVD) is a leading cause of death in the world. By 2025, it is estimated that 23.6 million people will be attacked by CVD. Thus, the health care industry is established in order to collect a large number of CVD information of cardiovascular disease and to support doctors in finding and recognizing its potential risk factors of CVD through mining and analyzing the information. This structured and unstructured case information may be used to find out potential patterns of diseases and symptoms using deep learning algorithms. The risk factors associated with cardiovascular disease are known from epidemiology, and this is the first prospective investigation on the condition in the community free mobility population. Physicians can anticipate cardiovascular disease and take early action if it can be predicted. Clinical data analysis is therefore considered to be the prediction of cardiovascular disease. The main contents and contributions are as follows: We start off with trying to predict using the classic data mining and machine learning techniques, and the results are not at all ideal. Generally speaking, after analysis, it is mainly caused by imbalance of data sets. To solve the problem of CVD data imbalance, a SMOTE oversampling method is proposed aiming at the imbalance of cardiovascular data collected in the Framingham community. In order to ensure accurate data collection during operation, the relationship between LSTM(P) and unit state is then tried, and the prediction technique of cardiovascular disease utilising LSTM(P) is realised. Lastly, tests are conducted to confirm the 4434 individuals' initial medical information in the data set. The algorithm has an MCC score of 0.96 and an accuracy of about 94%.

| KEYWORDS

Cardiovascular disease; data imbalance, SMOTE; LSTM(P) model; prediction, deep learning

| ARTICLE INFORMATION

ACCEPTED: 12 November 2024 **PUBLISHED:** 14 December 2024 **DOI:** 10.32996/jcsts.2024.6.5.15

1. Introduction

1.1 Research Background and Motivation

The world from the perspective of prediction and forecasting on medical science the future got contributed by deep learning in various scientific research societies. The understanding of strategies that facilitate the way of handling the forthcoming events within the modern research field is offered by the influence of data mining studies. Moreover, cardiovascular disease is the top cause of death worldwide in the healthcare, because the specific number of victims yearly is computed accordingly, approximately 17.9 million [1]. The researchers implemented many prevalence methods in the scientific community because the significance lies in the development of the consciousness techniques for the population who includes the people who are active with their daily life needs and cannot spend much on the expensive medical expenses [2,3,4,5,6]. Deep learning algorithms are used for medical and health informatics to reduce medical cost and provide satisfactory treatment. Over all, deep learning tasks are dedicated to getting the descriptive or predictive results. General properties of the data on the database is what descriptive focuses on. Predictive, takes some predictive evidence on current data in the database to predict on future values. We apply different deep learning techniques to achieve the best information.

Potential value of the hidden knowledge in the database can be exploited to improve the procedures, productivity, or reliability of a number of applications fields. In the healthcare environment, the professionals, pharmaceutical personnel, and medical staff can explore data mining techniques to formulate strategies that help improve the process of medical treatment, use resources optimally, significantly reduce time and cost. The automatic knowledge discovery of healthcare problems is evolutionary, diversified and complicated and should improve medical science. For instance, symptom association analysis for a particular utilization of disease resources can guide the improvement of healthcare services. This extracted knowledge leads healthcare systems to profitable, effective, significantly precise treatment procedures and lower costs. Additionally, healthcare data is diverse because it contains high dimensional data of medical history records from patients who had different symptoms. Furthermore, techniques and procedures are highly needed to solve healthcare data problems, which are also efficient.

1.2 Cardiovascular Disease Status of Patients Prevalence

Cardiovascular disease (CVD) affects the heart, blood vessels, and other essential organs. Deathly conditions include heart attacks, strokes, thromboembolic illnesses, rheumatic heart disease, artery disease, and cardiovascular disease (CVD) at any age. Out of 17.9 million, or 31% of all deaths from CVD, [7] Furthermore, accounting for about 33% of all fatalities worldwide, CVD is the major cause of mortality. Around 300–450 deaths per 100,000 in Pakistan, India, and Bangladesh in 2005 was lower than the 100–200 deaths per 100,000 in emerging countries like China, the United States, and Japan. [8]. Numerous risk factors for CVDs include smoking, obesity, advanced age, leading a poor lifestyle, and more. These days, CVDs affect persons of all ages, males and females alike. The true issue with CVD is identifying its symptoms and underlying cause before we become a victim.

1.3 Research Objectives and Significance

1.3.1ResearchObjectives

The goal of this research is to inform and demonstrate the effective technique for CVD by utilizing deep learning algorithms to get the knowledge to support primary health education, and for supporting to detect the risk factor and decision making in the medical sector. This goal will be examined in three broad aspects by approaching deep learning to the real-life healthcare dataset of CVD patients. The research objective involves the three principle contents as:

(1) An extremely creative approach to bettering CVD therapies and lowering medical expenses is our suggested intelligent healthcare stage, which predicts the risk variables of CVD.

(2 Gathering information from patients is often difficult and complex, and the data is separated into several groups or classes. That is important for clinical consequences and targeting the health sector. Both healthy and ill people can utilise the data.

(3) By using less human and other material resources, the deep learning algorithm LSTM can significantly reduce the cost of identifying the risk factor for CVD.

The Framingham Heart investigation records, which are a lengthy prospective investigation of the aetiology of cardiovascular illness among a population of free-living participants in the society of Framingham, Massachusetts, were collected for our planned research project. In order to assess the model correctness of our suggested deep learning algorithms, we conducted this experiment on a Jupiter notebook using seven performance metrics: accuracy, precision, sensitivity, specificity, F1-score, ROC-AUC, and MCC.

1.3.2ResearchSignificance

Deep learning algorithms are most popular nowadays in medical and digital health informatics. This approach has two big factors as one is practical to develop novel approaches to detect or diagnose medical problems such as to detect breast cancer, brain tumor abnormality, diagnosis of CVD, find out risk factors of different human body disease, and so on. And the second one is to develop the algorithms theoretically.

In developing countries, healthcare is the most significant hitch facing the population and medical doctors. The modeling strategies represented in this research of the deep learning algorithms assist in the initial diagnosis of diabetes and reduce the healthcare burdens in the initial phase. In addition, the most significant research is especially for the affected cardiovascular patients of low socioeconomic status and excessive risk countries who cannot afford the initial expenses of laboratories test.

Furthermore, our proposed research is also significant for the medical specialist or radiologist and decision-making authorities to predict the CVD and to build the medical tools to reduce the healthcare CVD problems. Therefore, the significance of our proposed research work is:

(1) Firstly, Deep Learning algorithms assist to find out the hidden pattern of the CVD and symptoms from the Framingham CVD datasets.

(2) Second, compared to other previously suggested approaches, our suggested deep learning application of LSTM to predict the risk factors of CVD displayed higher performance matrices, had an MCC score of 0.90, and had an accuracy of around 94%.

(3) Thirdly, the findings of the experiment indicate that the LSTM performs better than the other statistical machine learning algorithms for forecasting and identifies important risk factors for CVD, which may encourage patients to modify their behaviour. Ultimately, all of these techniques were combined based on content to effectively search for important medical data to assist the population's primary healthcare system and improve medical care.

2. Literature Review

Numerous intricate datasets that may be employed to forecast and examine the medical sector ratio for the full nation are being encountered and accumulated by the healthcare industry. When necessary, this enormous volume of data must be recovered. Data mining techniques make it simple to extract important and hidden information from datasets related to the healthcare industry. Data mining's primary goal is to classify arrays and prototypes, which facilitates projections and advances medical decision-making for treatment design. Predictive models are used to combine healthcare systems in order to reduce bias and provide decision-making results. Additionally, data mining yields consistent outcomes in the health sector, which leads to the improvement of quality of facilities.

Medical applications including the medical field, therapeutic pragmatics, operation measures, and medical systems can benefit from data mining. Pearce et al. [10] demonstrated the hypothesis of implication of eye illness and other consequences, whereas Saini S et al. [9] mentioned data mining systems based on Ant colony optimisation and multiclass. Declaris, a different researcher, used Neural Networks (NN) for medical data mining in order to create data visualisation pictures. Nikunj et al. [11] used data mining to improve healthcare services while protecting privacy.

Palaniappan and Awang [12] used data mining techniques to analyse cardiac disease prognosis in an intelligent manner. Data mining using evolution algorithms was used by Luukka P et al. [13] to anticipate diagnosis from datasets related to cardiac disease. Cior, a researcher, discovered the essence of healthcare hypermedia classification by using data mining terminology and methodologies [14]. Tsumoto et al. [15] discovered issues with the hospital information systems' utilisation of various medical programs, including missing datasets and other ambiguous information.

Bramerier and Banzhaf investigated neural networks and linear genetic platforms for medical data mining [16]. In terms of data mining, Olukunle et al. [17] looked at the association rule for image processing mining and proposed that experimental findings can identify the kind of things that frequently appear in data sets. Antonio et al. [18] had proposed the classification methods and formally smeared them to predict liver condition and other liver disorders with the aid of data mining. Abir et al. [19] used data mining to examine several algorithms for medical growth. A construction that is positioned by both grid distribution and heterogeneity was developed by Brunei et al. [20]. Abir et al. [21] worked on real-life diabetes patients by the fractional polynomial, in which the study was based on Logistic Regression Modeling.

Podgorelec et al. [22] found another important prediction method to categorise the medical data. Abir et al. [23] used data mining to examine the dominance and insufficiency of data mining technology. The classification system has been used to diagnose cardiovascular disease by Cheng et al. [24]. Wong et al. [25] suggested another innovation in web-based data mining and telemedicine. Bethel et al. [26] demonstrated the advancement of an association rule based on individuals with breast cancer.

A brief discussion of the several methods utilised for CVD prediction, including machine learning and deep learning algorithms. Numerous methods were previously developed for the prediction of heart disease. Two researchers, Chaurasia et al. [27], developed a decision tree and support vector machine, while another researcher, Dhanashree et al. [28], focused on CVD predictions using Naïve Bayes classification methods. Parthiban et al. [29] offered a technique that uses Naïve Bayes and support vector machines to determine the risk of CVDs, whereas Otoom et al. [30] suggested utilising Bayes net and SVM to identify heart illnesses.

Muhammad et al. [31], another researcher, demonstrated the performance of the multiclassification model using three performance matrices: precision, recall, and f1 score. According to Aditi et al. [32], Multi Layer Perception (MLP) in neural networks exhibits great recall and accuracy values. The accuracy level of this model varied from 91%, according to the findings of the authors Aditi et al. Chauhan et al. [33], a Hungarian researcher, employed the Weighted Association Rule (WAR) to predict coronary disease, however their accuracy was extremely poor. This researcher used the medical datasets from the Hungarian Institute of Cardiology in Budapest. The primary issue with this study work is that the model is not suitable for predicting diseases, and the medical dataset contains extremely few characteristics. Purushottam et al. [34] developed a novel technique to determine the patients' risk level by examining their symptoms. The collected dataset used for this research work is from V.A Medical Center, Long Beach, and Cleveland Clinic Foundation and the number of class level used for this research work is five. For their work, they used DT and their accuracy on the prediction model was quite good.

However, in order to predict the CVD, Thomas et al. [35] proposed many classification techniques, such NB K-Nearest Neighbour (KNN), DT, and NN. By choosing the many CVD traits and indicators, the author stated a multi-class approach. Additionally, they used KNN and discovered an accuracy of 80.4%. They employed a number of attributes to obtain the accuracy. Lastly, they unveiled a platform that allows medical professionals to obtain the statement immediately from the prediction's outcomes. where it is also necessary to measure the patients' prior medical histories.

The author Raihan et al. used the design in their study [36] to gather information on an Android application system for Ischaemic Heart Disease (IHD). Based on the methodology, the author deduced that there was a significant correlation between the p-value test results and IHD and the characteristics or symptoms. The data was gathered by the author using a risk score tree and chi square correlation. The ELM algorithm is used to build the heart disease dataset in accordance with the methodology used by Ismaeel et al. [37]. A three-layer neural network with a sigmoid activation function was demonstrated to be ELM.

According to R. Jones, Z. Shen, and Alberti [38], a NN on self-questionnaire data organised a heart disease prediction context that will publish other datasets gathered in self-questionnaire data for the prediction of risk factors in CVD in addition to heart disease's key hazard components. Heart disease is predicted by age, sex, blood pressure, cholesterol, and smoking. However, PS Hong and SY Huang [39] made another CVD forecast and suggested three stages for their process. In addition to using Artificial Neural Networks (ANN) for classification, the author chooses points from thirteen features. We achieved 80% accuracy after developing the framework. Using the same ANN and methodology, Jayshril et al. [40] similarly demonstrated the diagnosis of cardiac illness. This author created a Vector Quantisation Algorithms with three layers and thirteen neurones for the suggested system. When diverse neurones and ages were used, the framework's efficiency and execution were enhanced. On a specific CVD dataset with CVD recognition, feature selection and classification approaches were applied, same like in the works of two other authors, Ghonji and Omid et al. [41,55,56,57,58].

The researcher's methodology consisted of three steps: first, the isolated dataset was divided into two subsets, one for CVD and the other for non-CVD; second, 8192 subsets that contained features of the entire dataset were extracted; and third, Feed Forward Backpropagation was used to find the appropriate pattern. Another study by Manza et al. (42) used an ANN classifier to predict the treatment of heart disease. The author used five steps in this process: collecting patient and physician data, classifying the data into binary labels (either as heart disease-affected or not), determining the function that will be exploited, testing the heart disease dataset to determine classifier performance, and finally recommending medication for the afflicted patients.

The similar model and ANN were reported by Tuly et al. [43], although the neural layer results were inconsistent. Thirteen signs and symptoms are included in the heart disease prediction model. Our average accuracy was about 80.67% without Principal Component Analysis (PCA) and 91.33% with PCA. A novel mixing procedure of the Electrocardiogram (ECG) data with classified clinical causes of heart illness is developed by Ravish et al. [44]. The ECG inputs and clinical data were prepared for training a neural network (NN) to identify anomalies of cardiac illness after the signal was first amplified and then filtered by an arbitrary digital circuit to eliminate noise.

Furthermore, Equardo et al. [45,59,60,61] looked at how a NN might use the developing properties of the P-wave to assess and detect the presence of cardiac illness. To do this, the author suggested using the conventional backpropagation (BPP) method to extract heart disease-related characteristics and symptoms. A brief discussion of two recently published articles [46, 47, 51,52,53,54] has shown the significance and outcomes of CVD risk factors for the benefit of society, stating that when new species emerge, it is imperative that they be able to forecast CVD risk factors in order to lower the enormous number of CVD patients. For this, we show very conveniently the different approaches to predict CVD that we re-collected all the past researcher's work in table 1, to help to give a small review of the works.

Table 1: Different Machine Learning Methods to predict CVDs

Research Paper	Machine Learning Methods	Accuracy
Chaurasia et al.	J48 SVM	84.35% 85.03%
Dhanashree et al.	NB	86.41%
Author Parthiban et al.	NB SVM	74% 94.60%
Otoom et al.	SVM Bayes Net	88.3% 84%
Muhammad et al.	LG SVM RF DT NB	80% 83.8% 68.8% 86.6% 86.7%
Aditi et al	MLP	91%
Chauhan et al.	WAR	53.4%
Purushottam et al.	WAR	53.4%
Thomas et al.	KNN and ID3	80.4%
Ismaeel et al.	ELM	80%
SY Huang et al.	ANN	80%
Jayshril et al.	ANN	85.55%
Omid et al.	ANN	91.94%
Manza et al.	ANN	97%
Tuly et al.	ANN(PCA)	91.33%

This chapter has covered the fundamental ideas and methodological viewpoints on data mining, deep learning, and its applications as well as many approaches or strategies to medical healthcare, particularly in the context of CVD. In this chapter, it was discussed how deep learning and machine learning are used in healthcare, how they efficiently approach the process of diagnosing and predicting cardiovascular disease (CVD) and other illnesses, and how this enhances the effectiveness of medical practitioners. By suggesting data mining and deep learning techniques to detect CVD in its early phases and incorporating epidemic strategies into diverse methodologies, this chapter also showcases the work that has been done by other academics. Even while there isn't a single, all-encompassing algorithm or technique that can accurately forecast every medical condition, deep learning technology still appears to be gaining traction in the medical field and has the potential to make significant and admirable advancements. Nonetheless, there are still a lot of hybrid models that need to be demonstrated in the form of achievements that effectively manage time constraints and the resolve to develop and produce hybrid models for more suitable and precise disease prediction, risk factor identification, clinical inferences, and primary medical applications.

3. Methodology

3.1 Analysis of Imbalanced Cardiovascular Disease Dataset

In our proposed research the Framingham heart study was commenced back in 1948 and around 5209 were enrolled mainly for the research work. All the participants were scrutinized biennially since the initiation of the research as well as all subjects continually investigated with consistent investigation for cardiovascular end points. Experimental examination data in the proposed work have included cardiovascular risk factors and markers of the disease such as blood pressure, blood chemistry, lung function, smoking history, health performances, ECG tracings, Echocardiography (ECHO), etc., as well as medication use. With the consistent surveillance of area hospitals, following contact of participants with patient and alternative secondary death certificate, Framingham heart study adjudicates for analysis of Angina Pectoris, Myocardial Infarction, Heart Failure, and cerebrovascular disease. It is a subset of data collected as part of the Framingham heart study (research laboratory, clinic or health center, questionnaire process, adjudicated event data), approximately 4434 participants. Around 1956 to 1968 we accumulated participant clinic data through three examination epochs, roughly every six years, each six year epoch approximately. For the research experimental work, every participant was followed for a total of 24-year for the outcome of the following events: Cerebrovascular accidents, Angina pectoris (AP), Myocardial infarction (MCI), Atherothrombotic Infarction or Cerebral Hemorrhages (Stroke), or death.

3.2 Analytical Platform and Processing

Each participant has between 1 and 3 clinical observations associated with the number of experimental exams to which the subject appeared leading to 11,627 observations over 4,434 participants in the experiment. Nevertheless, the event data has been added for all participants but without respect to prevalent disease status or time of collection of the pragmatical examination CVD data. In addition, we expect to learn what are the major main risk factors which we cannot address specifically by attributes alone. The Cardiovascular disease dataset attributes are shown in table 2 and according to the attributes it describes the features description, units, ranges or counts, it provides a clear view of the dataset with a pragmatical explanation. We also have studied for our research purpose the following medical event data for every participant. Each of the three medical events "0" denotes that the event did not happen during examination time and "1" denotes that the event did occur during experimental time. The risk factor of our proposed work can be understood from some major medical event data shown in Table 3.

3.3 Defining Incident Events

Epidemiologists may need to study the population at risk for a disease or the occurrence of an event, or individuals early had an In order to calculate new or first events, these have to be excluded from the event being investigated. We compute any of three examinations including start of exam (examination) incidence of first event rates. We will denote the population at risk for the consequence of concentration by the features PREVAP, PREVMI, PREVSTRK, PREVHYP and PREVCHD. Since our variables for PREVAP, PREVMI, PREVCHD, PREVSTRK and PREVHYP are not balanced, we analyze the variables. In our research we look at the columns of PREVAP, PREVMI, PREVCHD and PREVSTRK, and despite the fact that they are quite unbalanced, we do not analyse the number of true events (1) or the number of false events (0). We can go on to look at table 4. The third thing that it had going from table 4 is that the output column PREVHYP is a balanced dataset and another remaining output column is not balanced such as PREVCHD, PREVAP, PREVSTRK, PREVMI. To remedy imbalance we proposed Synthetic Minority Oversampling Technique (SMOTE) in here. Finally, SMOTE is a novel oversampling method (and a famous SMOTE was introduced by [48]). By using SMOTE rather than just repeating the data, the minority is oversampled by creating artificial examples in the feature space based on an example and its K nearest neighbors it can both leverage and also avoid over citing issues.

Table 2: Analysis of the Cardiovascular Dataset many Characters and its Units and Range

Features	Description	Units	Range
Randid	Exceptional number for every participant		2448 - 9999312

Gender	Participant sex [Men (1) or Women (2)]	1 = Men
		2 = Women
			n = 5021
			n = 6605

Period	Experimental examination cycle	1 = First Cycle
		2 = Second cycle
		3 = Third Cycle
			n = 4540
			n = 4000
			n = 3400

Time	No of days since experiment started baseline.		0 – 4855
Age	Years in age		32 – 81
SBP	Systolic Blood Pressure (mmHg)		83.5 – 295
DBP	Diastolic Blood Pressure (mmHg)		30 – 150

BPMEDS	Use of Anti-hypertensive medication at experimental examination	0 = Not in use
		1 = use
			n = 10090
			n = 944

CSMOKE	Existing cigarette smoking at test time	0 = Not smoker
		1 = Smoker
			n = 6600
			n = 5000

CPDAY	Total no. of cigarettes smoked/ day	0 = Not smoker
		1 to 91 cigarettes everyday
			n = 6598
			n = 5029

EDUC	Participated education	1 = 0 to 11-year 2 = High School Diploma 3 = Some College (BSC, or BA)	
TOTCHOL	Total Serum Cholesterin (mg/dL)		107 – 696
HdLC	Lipoprotein Cholesterol of high density (mg/dL)	Accessible for period three only	10 – 189
LdLC	Lipoprotein Cholesterol of low density (mg/dL)	Obtainable for period three only	20 – 565

BMI	Body Mass Index, weight in kilograms/height meters squared		14.43 – 56.8
Glucose	Casual blood serum aldohexose (mg/dL)		39 – 478

Diabetes	Diabetic according to the norms of 1st exam treated or 1st exam with casual glucose of 200 mg/dL or more	0 = Not affected
		1 = affected
			n = 11097
			n = 530

Heart Rate	Heart rate in beats/min		37 – 220

PREVAP	Prevalent Angina Pectoris at exam	0 = Not affected
		1 = affected
			n = 12000
			n = 700

PREVCHD	Coronary Heart Disease demarcated as pre existing Angina pectoris, Myocardial Infarction, and Coronary Insufficiency.	0 = Not affected
		1 = affected
			n = 10900
			n = 900

PREVMI	Prevalent Myocardial Infarction	0 = Not affected
		1 = affected
			n = 11500
			n = 400

PREVSTRK	Prevalent Stroke	0 = Not affected
		1 = affected
			n = 12000
			n = 200

PREVHYP	Subject was denoted as Hypertensive if preserved or if mean Diastolic >=90 mmHg or mean Diastolic 140 mmHg.	0 = Not affected
		1 = affected
			n = 6500
			n = 5500

Table 3: Analysis of the Cardiovascular Disease Dataset with other Significant Attributes during Experiment Time

Variable Name	**Label**
ANGINA	Angina Pectoris Incident
HOSPMI	Hospitalized Myocardial Infarction Incident
MIFCHD	Myocardial Infarction or Fatal Coronary Heart Disease (CHD), Incidence Hospitalized
ANYCHD	Myocardial Infarction; Myocardial Insufficiency (Unstable Angina); Angina Pectoris, Fatal Coronary Heart Disease, Coronary Insufficiency.
STROKE	Stroke Fatal / non-Fatal Incident occuring
CVD	Fatal or Non Myocardial Infarction or Stroke Incident Hospitalized
HUPERTEN	Incident Hypertension [1st exam pretends for high Blood Pressure and 2nd exam in which either Systolic $\geq 140\ mmHg$ or Diastolic $\geq 90\ mmHg$]
DEATH	Death Indicator
TIMEAP	Total number of days from baseline exam to 1st Angina
TIMEMI	Baseline Days - Incident Hospitalized Myocardial Infarction
TIMEMIFC	Baseline Days – Incident Myocardial Infarction Fatal CHD
TIMECHD	Baseline Days – Incident Any CHD
TIMESTRK	Baseline Days – Incident Stroke
TIMECVD	Baseline Days – Incident CVD
TIMEYP	Baseline Days – Incident Hypertension
TIMEDTH	Baseline Days – Death

Table 4: Evaluation of the event output to check the imbalance data

Variable	**Average**	**Count**	**Sum**
PREVCHD	0.072417649	11628	842
PREVAP	0.053926206	11628	627
PREVSTRK	0.013070302	11628	152
PREVMI	0.032166509	11628	374
PREVHYP	0.459619851	11628	5344

3.4 Synthetic Minority Oversampling Technique

Initially look at the dataset PREVAP from the Imbalance datasets of PREVAP, PREVMI, PREVCHD, PREVSTRK. If we pick 2 of 4434 participants, let's say x_1,x_2... samples from the PREVAP dataset. With n number of attributes in each sample. The biggest process we took to get our new instances to balance our dataset was: In the 1st process, we have set the minority class set M, for each given x∈M, and calculate for each x, the K nearest neighbors as calculated Euclidean distance value of M to x. In the second process, the sampling rate A was set according to the imbalanced quantity. For every $x \in M$, number of examples such as $x_1, x_2, x_3, \dots \dots \dots \dots \dots \dots \dots \dots \dots, x_a$ are arbitrarily selected from its K-nearest neighbors, defined this set M_1. In the third steps, $x_k \in M_1$where k=1,2,3,4,5,6,7,8,9, ..., and so on. The following equation was used to create instances:

$$x' = x + rand(0,1) * |x - x_k|$$

(1)

Where, $rand(0,1)$ stands for random numbers between 0 and 1. According to this above process, we can set the execution time for the sampling rate and repeated the above procedure until to find our required solution. For the other attributes such as PREVCHD, PREVAP, PREVSTRK, and PREVMI, we applied the SMOTE to get balanced data set to run our proposed model.

3.4.1 Assessment of PREVAP Dataset

After applying SMOTE in the CVD dataset, we get our required first balanced dataset. In Figure 1, it presents all the attributes in the graphs showed balanced. In PREVAP, we analyzed the number of major and minor classes that were almost closer.

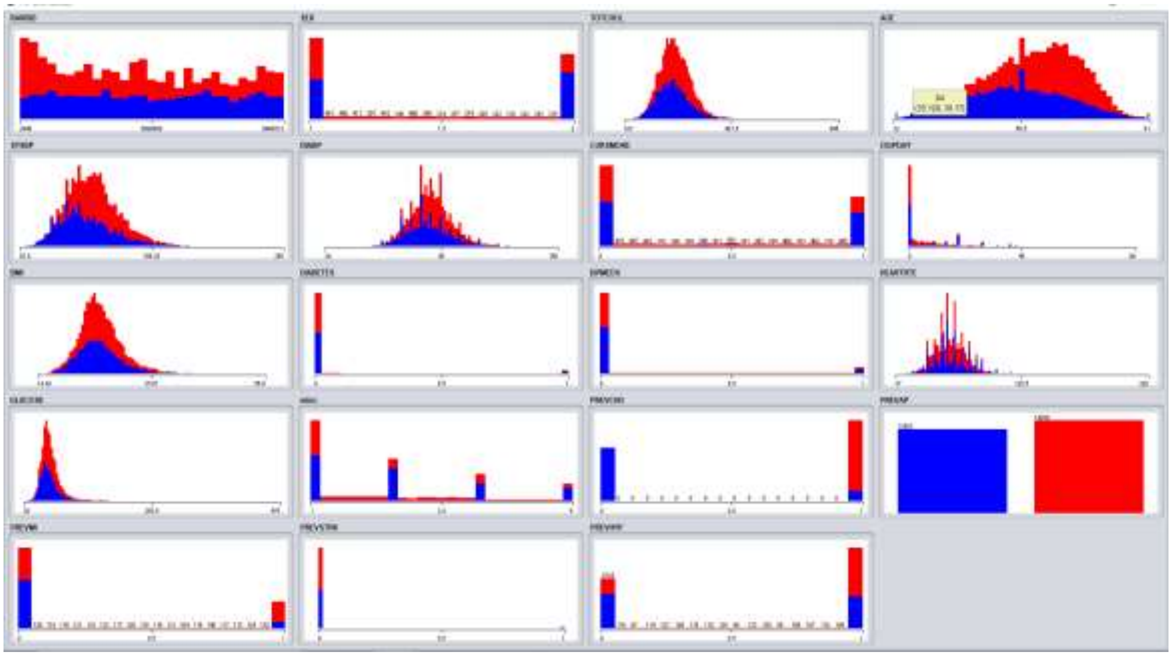


Figure 1: The graphical representation of the feature PREVAP

3.4.2 Assessment of PREVCHD Dataset

In the Figure 2, after implementing SMOTE in our CVD dataset, attribute PREVCHD creates the positive and negative classes, and SMOTE helps to create artificial instances to get the balanced data to prepare for deep learning model to get better accuracy level. In the PREVCHD dataset both numbers of classes now close to each other.

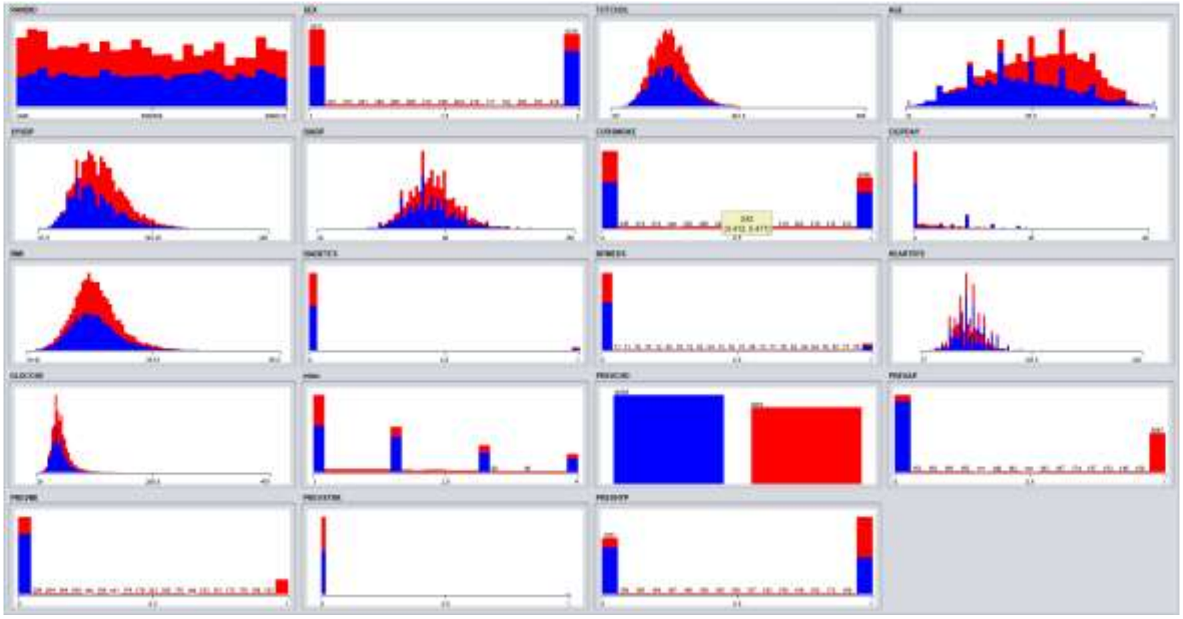


Figure 2: The graphical representation of the feature PREVCHD

3.4.3 Assessment of PREVHYP Dataset

By analyzing the graphical representation of the attribute PREVHYP in Figure 3, each variable of our CVD dataset shows balanced. From the graph, we observed that the number of red and blue colors for every variable has been reflected equal number of instances. SMOTE algorithms helped to create synthesis instances for every variable. The number of minority and majority classes is closer in the PREVHYP class. By SMOTE strategy, which is a tremendously understood oversampling methodology for adjusting an imbalanced CVD dataset to create an adjusted dataset. This method has invented a supportive approach to augment arbitrary oversampling by appropriating the cases for the overriding part class and the minority class correspondingly. Demolished creates manufactured samples or illustrations of the minority class and has a propensity to develop the perceptive exactness over the minority class of the CVD dataset.

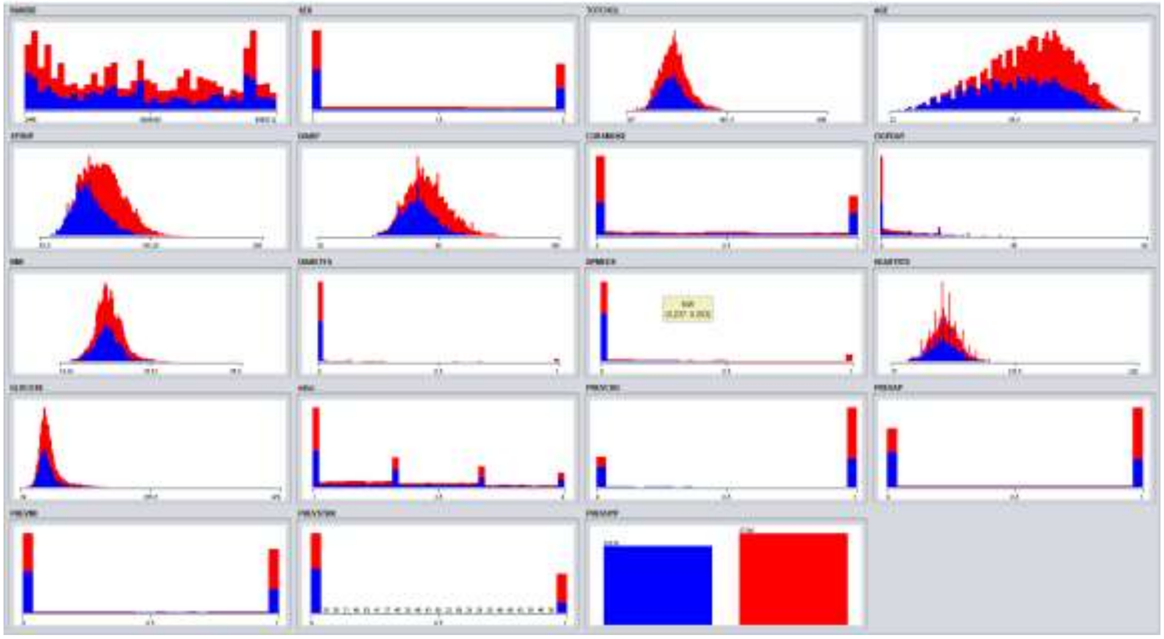


Figure 3: The graphical representation of the feature PREVHYP

3.4.4 Assessment of PREVMI Dataset

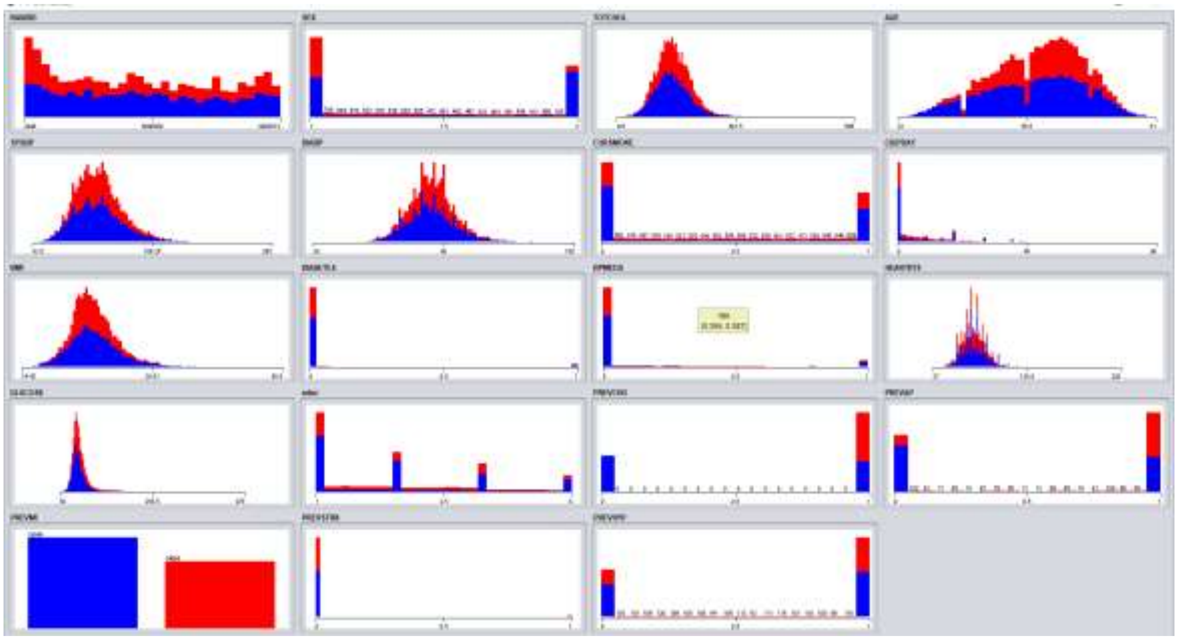


Figure 4: The graphical representation of the feature PREVMI

In the above Figure 4, the attribute of PREVMI, number of minority and major classes are neighbor to each other. All the other variables in our CVD dataset also get balanced after implementing SMOTE algorithms.

3.4.5 Assessment of PREVSTRK Dataset

After balancing the five attributes PREVAP, PREVMI, PREVCHD, PREVHYP, and PREVSTRK with the help of SMOTE algorithms our deep learning model is to prepare to predict the risk factor of Cardiovascular disease. In Figure 5, the attribute PREVSTRK shows the number of minority classes and majority classes is trending closer.

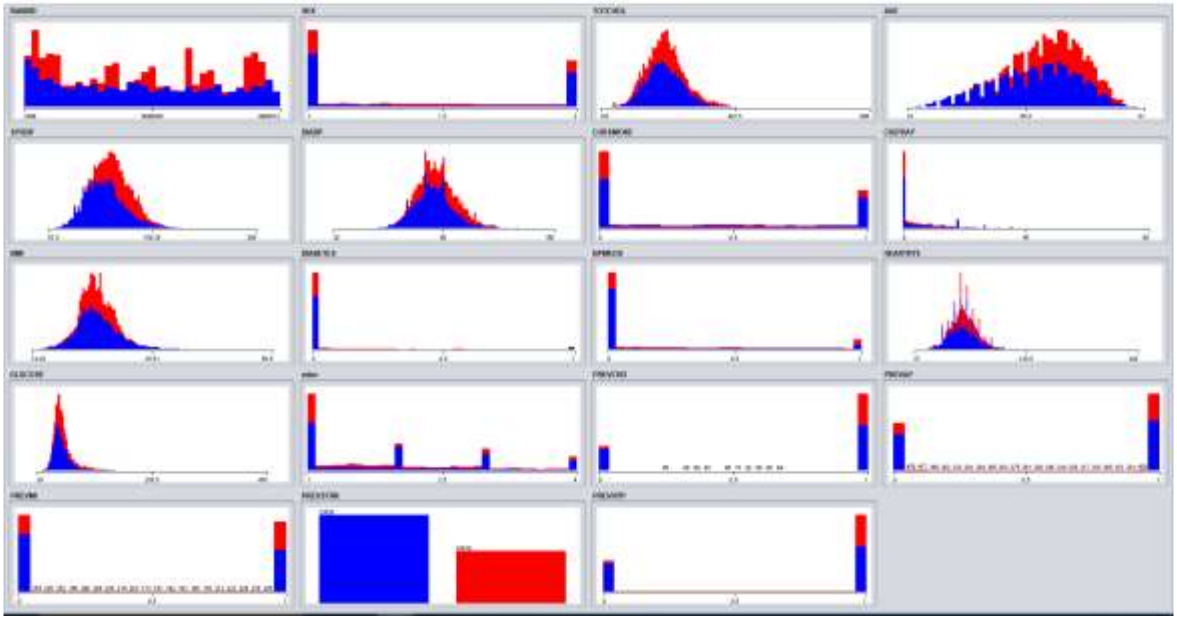


Figure 5: The graphical representation of the feature PREVSTRK

3.4.6 Measurements

To run this experiment, WEKA 3.7.3 software was used. Software WEKA is an associate assembly of machine learning algorithms for data mining applications. This data mining software contains applications for data pre-processing, feature selection, clustering, classification, association, and also for visualization. From table 5, and 6, we observed that the number of correctly classified instances are 25280. Moreover, we applied 10 fold cross validation technique with parameters batch size=100 and confidence interval=0.25% on 10 different information data mining classification algorithms on 27192 cardiovascular patients. This evaluation took 1.19 seconds to run, kappa statistic 0.8584, average true positive rate 0.93, average false positive rate 0.07, average accuracy, precision, recall, and F measure 0.93, MCC is 0.86 and average ROC and PRC area is 0.95. Accordingly, the final accuracy using equation (accuracy equation) was 0.93 and error rate was 0.2. We also analyzed the number of incorrectly classified instances, which was 1912 patients, from this classification assessment. True positive (TP) and false positive (FP) constitute the structure of our proposed study.

Table 5: The Classification Assessment

Correctly Classified Instances	25280		92.9685 %
Incorrectly Classified Instances	1912		7.0315 %
Kappa statistics		0.8584	
Mean absolute error		0.0994	
Root mean squared error		0.2427	
Relative absolute error		19.9306 %	
Root relative squared error		48.6042 %	
Total Number of Instances	27192		

In this section, the aim is to develop SMOTE algorithm for balancing our Cardiovascular datasets from imbalanced. The imbalanced cardiovascular dataset is a significant issue in data mining research and many machine learning or deep learning algorithms can hardly contend with the imbalanced class distribution. To solve this problem, we proposed SMOTE in our research study to solve this problem and make our cardiovascular data accurate to dig out the risk factor and appropriate it for deep learning algorithms. SMOTE enhances the minority class by engendering new synthetic cases based on its number of nearest neighbors.

In our research work, SMOTE, as executed in data mining application software WEKA, was implemented to generate synthetic instances. For our study k-nearest neighbors of a real existing case were used to determine a new artificial one. After using several classification (J48) assessments, we achieved an accuracy of 93%, whereas, the error rate and false-positive ratio are (0.2 and 0.07). Besides, Kappa statistics, PRC, MCC, precision, recall, F-matrix, and true positive ratios are closer to 1.

Table 6: Detailed Accuracy by Class

	TP Rate	FP Rate	Precision	Recall	F-measure	MCC	Area ROC	Area PRC
	0.881	0.026	0.968	0.881	0.922	0.862	0.959	0.964
	0.974	0.119	0.901	0.974	0.936	0.862	0.959	0.944
Average weighted	0.930	0.075	0.933	0.930	0.929	0.862	0.959	0.953

3.5 LSTM Framework in Cardiovascular Disease

3.5.1 Long Short-Term Memory

The LSTM architecture possesses another repeatedly associated elements, which are called memory cells or blocks. These multiple blocks can serve as a differentiable type of memory chips in a contemporary computer. Each block contains one or more self-connected memory cells and substantial multiple factors, these are input 'gets,' the output 'gets' and the forget 'gets' that provide constant analogs for writing, reading and erase modes for the cells. Because the multiplicative gates allow LSTM memory to build and retrieve information for as long as required. It reduces the vanishing gradient problem Essentially it also helps in the training process. For example, for a long time, the input gate of the LSTM cell is shut, therefore the activation of the cell will not be replaced by the new input received in the network system, and can hence be made accessible to the information in later time step, by opening the output gate. LSTM has three gates as input gate, forget gate, and output gate. First and Foremost, $x_t(\text{new input})\ and\ h_{t-1}$ $(\text{output from the previous timestamp})$will pass through from a sigmoid layer and the outputs a number 0 and 1 is in cell state C_{t-1} .Where 0 means don't allow any data to flow (gate closed) and 1 means allow everything for further proceed (gate open). After that, new information will be in the cell state (Figure 6).

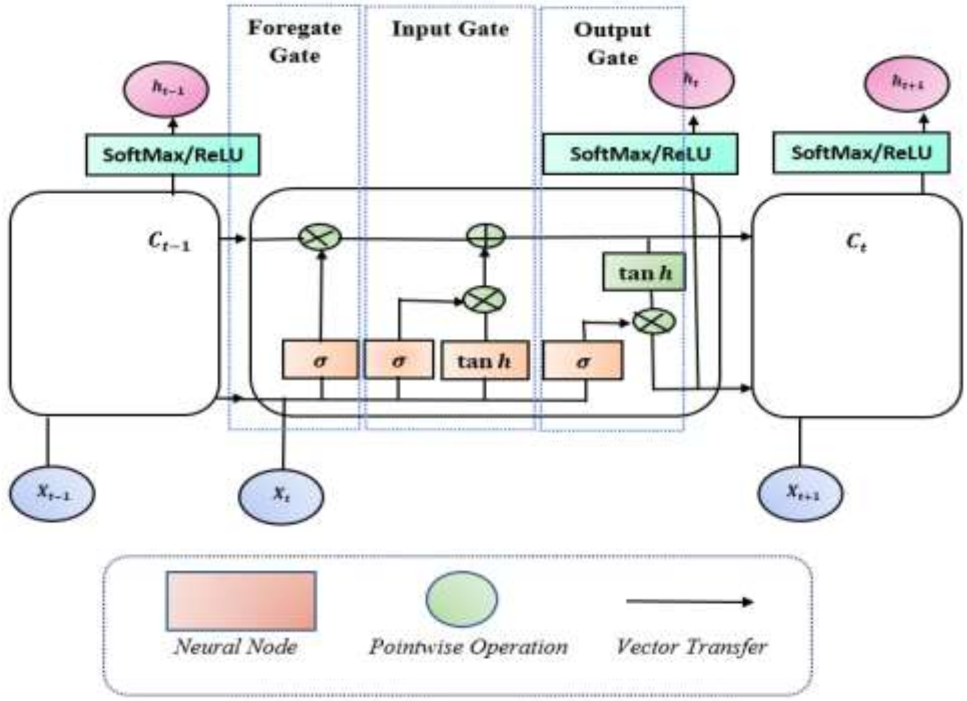


Figure 6: Original LSTM Architecture

It has two parts. First, a sigmoid layer called input gate that means it helps to decide which value will continue to keep update and next, a $\tan h$ layer creates a vector of new candidate values C'_t , It could be added to the state. In the next, we will combine these two to create an update to the state. Now update the old cell state, C_{t-1}, into the new cell state C_t. Then C_{t-1} is multiplied by forgetting gate f_t after that add this value with $i_t * C'_t$. There are new candidate values. In the end, we run the first sigmoid layer which decides what parts of the cell state are going to output. Then, we continue the cell state C_t through $\tan h$ (To find the value between -1 to 1) and multiply it by O_t. The tanh, Softmax, or ReLU are used as activation functions in the network layer, to adjust the weights of each layer, a gradient descent algorithm is proposed. To compute the derivative of weights, backpropagation algorithms are analyzed in the hidden layer. Therefore, the LSTM memory cell is computed by the following equations:

$$f_t = \sigma\left(W_f x_t + W_f h_{t-1} + b_f\right) \tag{2}$$

$$i_t = \sigma(W_i x_t + W_i h_{t-1} + b_i) \tag{3}$$

$$O_t = \sigma(W_o x_t + W_o h_{t-1} + b_o) \tag{4}$$

$$C'_t = \sigma(W_c x_t + W_c h_{t-1} + b_c) \tag{5}$$

$$C_t = f_t * C_{t-1} + i_t * C'_t \tag{6}$$

$$h_t = \tan h\,(C_t) * O_t$$

(7)

The mathematical notation of the equations from 2 to 7 are as follows:

$C_{t-1}\ and\ C_t$	Cell State of the LSTM model
W_i	Weight matrices in the input cell
W_f	Weight matrices in forget cell
W_o	Weight matrices in the output cell
W_c	Weight matrices in candidate cell
b_f	Bias vectors in forget cell

b_c	Bias vectors in candidate cell
b_o	Bias vectors in the output cell
b_i	Bias vectors in the input cell
$" * "$	Pointwise multiplication of two vectors
C_t'	Candidate cell
i_t	Input gate in the LSTM memory
f_t	Forget gate in the LSTM memory
h_{t-1}	Value of the previous memory cell at time t-1
h_t	Present value of the LSTM memory at time t
σ	Sigmoid function
$tanh$	Tanh activation function

3.5.2 LSTM with Peepholes Connection

In the above example, LSTM cell does not contain any direct connection to cell state. Because of this lack of crucial information that impacts the network's performance, there is a missing important piece of the puzzle. To make it simpler, Gers and Schmid Huber et al [49, 50] introduced LSTM(P) cell by adding peephole connection as indicated in Figure 7. The mathematical expressions obtained by figure 7, and the equations can be expressed as follows:

$$f_t = \sigma\left(W_f x_t + W_f h_{t-1} + P_f * C_{t-1} + b_f\right) \tag{8}$$

$$i_t = \sigma(W_i x_t + W_i h_{t-1} + P_i * C_{t-1} + b_i) \tag{9}$$

$$O_t = \sigma(W_o x_t + W_o h_{t-1} + P_o * C_t + b_o) \tag{10}$$

$$C'_t = \sigma(W_c x_t + W_c h_{t-1} + b_c)$$

(11)

$$C_t = f_t * C_{t-1} + i_t * C'_t \tag{12}$$

$$h_t = \tan h\,(C_t) * O_t \tag{13}$$

The LSTM cell was able generate these peephole connections to examine its current internal situation. As a result, with no trainers at hand, the LSTM(P) with peephole connection learns the stability and the timing algorithms. The mathematical notation of the new LSTM(P) model with peephole connections are as follows:

P_i	Peephole weights for the input gate
P_f	Peephole weights for the forget gate
P_0	Peephole weights for the output gate
$P_f * C_{t-1}$	Peephole weights for the forget gate multiplied with the previous cell
$P_i * C_{t-1}$	Peephole weights for the input gate multiplied with the previous cell
$P_o * C_t$	Peephole weights for the output gate multiplied with cell state of LSTM model

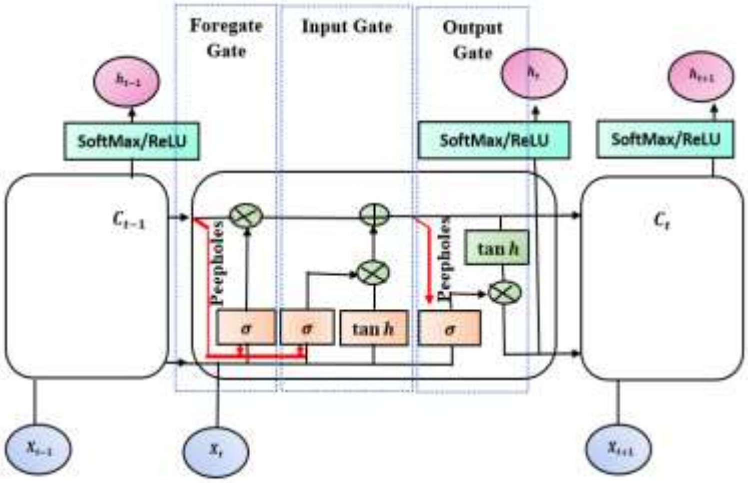
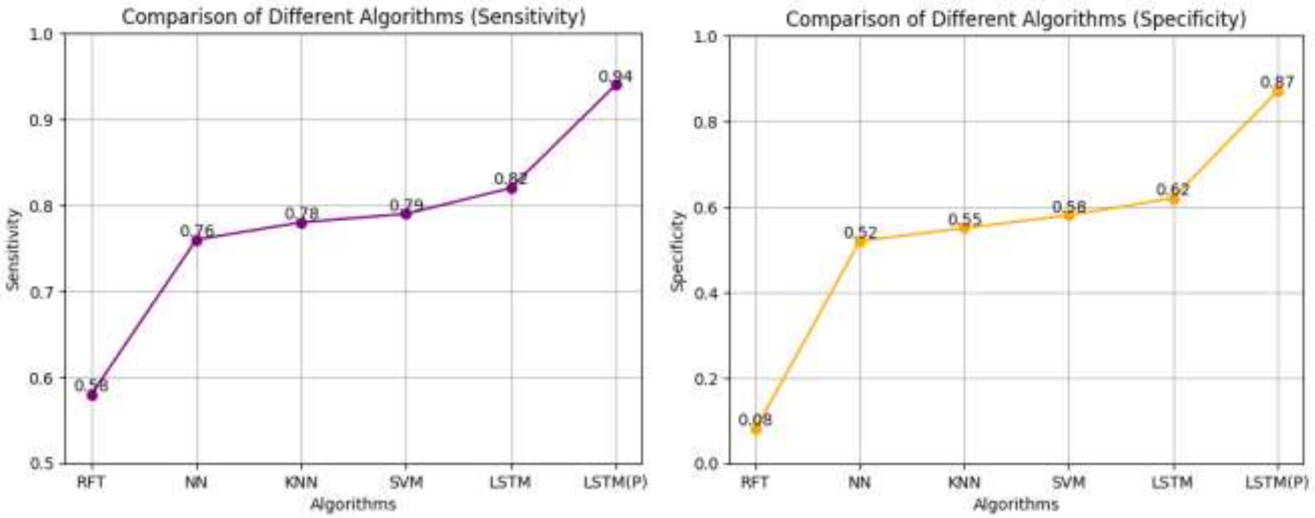


Figure 7: LSTM(P) Memory with peephole connection

4. Results and Discussion

In our proposed experiment, the cross validation strategies are used to train with the eighty percent (80%) data, and use the remaining twenty percent (20%) data to test the validity in the prediction effectiveness of the model. Using evaluation performance metrics such as accuracy, precision, sensitivity, specificity, F1-score, ROC and AUC, and MCC to select the best method of predicting major risk factors of CVD, we test our experiment performance. Our experimental result analysis is given in table 6. Figure 8 shows our proposed deep learning algorithm LSTM(P) model which scored higher performance matrices of efficiency than other machine learning model with lower accuracy level.

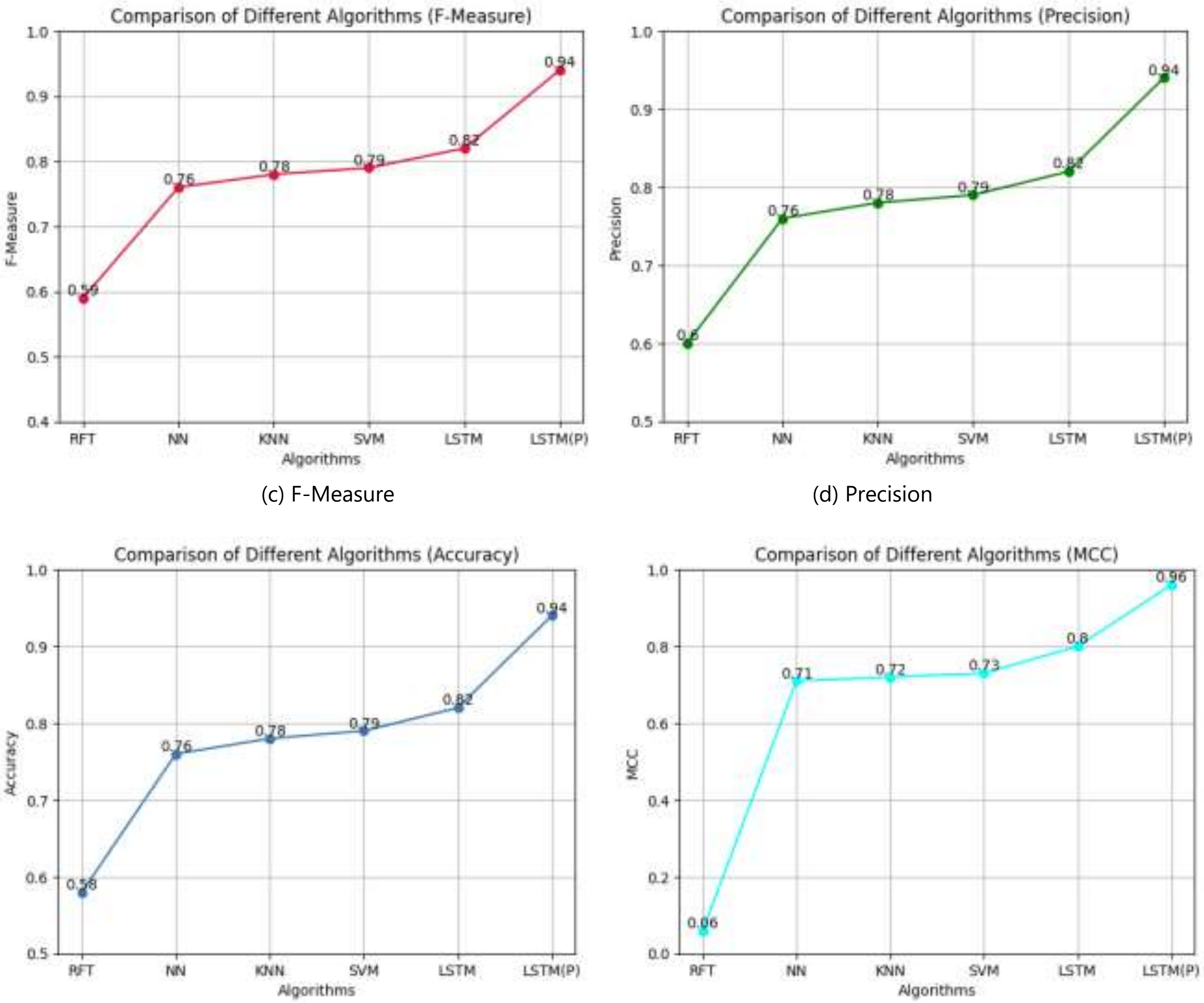
For further examination, we need to study the following two tables and compare the performance scores for every model. In table 7, every model indicated specific values of the performance metrics. After taking a closer look at each table, only the LSTM(P) model gives better performances, and our proposed model would be perfect for the CVD prognosis. As a result, LSTM(P) gives 94% accuracy and also provides a score of more than 0.87 in all other assessment parameters.



(a) Sensitivity

(b) Specificity

		Performance Metrics			
Models	Accuracy	Sensitivity	Specificity	Precision	F1-score
RFT	0.58	0.58	0.08	0.60	0.59
NN	0.76	0.76	0.52	0.76	0.76
KNN	0.78	0.78	0.55	0.78	0.78
SVM	0.79	0.79	0.58	0.79	0.79
LSTM	0.82	0.82	0.62	0.82	0.82
LSTM(P)	**0.94**	**0.94**	**0.87**	**0.94**	**0.94**



(c) F-Measure

(d) Precision

(e) Accuracy.

(f) MCC

Figure 8: Assessment of the performances of the models (a) Sensitivity, (b) Specificity, (c) F-Measure, (d) Precision, (e) Accuracy, and (f) MCC

Table 7: Performance Metrics Evaluation Comparison

On the other hand, all machine learning methods have shown an average presentation with all the assessments of the performance evaluation. On the contrary, table 8 shows a clear explanation, here other machine learning models such as RFT, NN, KNN, SVM, and LSTM gives average performance result whereas LSTM(P) for both performance metrics such as ROC-AUC and MCC provides the best performance score of 0.99 and 0.96 which is closer to 1.

Table 8: Comparison of Performance based on ROC-AUC and MCC

	Performance Metrics	
Models	ROC-AUC Score	MCC
RFT	0.69	0.06
NN	0.76	0.70
KNN	0.79	0.71
SVM	0.86	0.73
LSTM	0.87	0.80
LSTM(P)	0.99	0.96

For prognosis CVD major risk factors, we proposed the LSTM(P) model in our research study. From the experimental result analysis, we found that our proposed LSTM(P) model to dig out the major significant risk factors of CVD. In Figure 9, the major risk factors are shown with major features including the weights given by our proposed deep learning model. From the following graph, we observe that PREVCHD (Prevalent Coronary Heart Disease, chest abnormality or pain) and PREVSTRK (stroke) are in the uppermost. We consider 0.55 as our threshold point that means features or attributes above the red line or greater than 0.55 are the noteworthy risk factors of our CVD dataset.

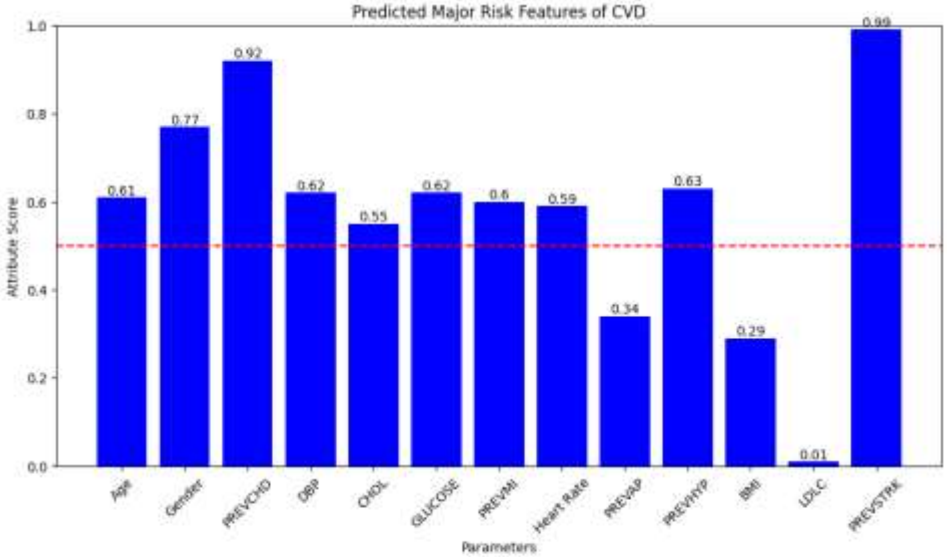


Figure 9: Predicted Major Risk Features of the CVD

The graph also illustrated that Age, Gender, DBP, Heart Rate, and PREVHYP (Prevalent Hypertensive) are also in-between significant factors of the CVD, whereas other features or attributes such as PREVAP (Prevalent Angina Pectoris), BMI (Body Mass Index or weights), LdLC (Lipoprotein Cholesterol of low density) have the slight influence over the CVD. Though cholesterol is also shown at the threshold point 0.55, it is also a critical risk factor among the attributes of CVD. Overall, to develop CVD treatment in healthcare, we need to take into consideration the circumstances, environmental effects, the quantity and quality of the instances also. Our research work assists the CVD patients to change their day-to-day life based on the predictive recommendations formed from the deep learning methods. Several major high-risk factors are shown in table 9 with their feature scores after analyzing our experimental results.

Table 9: Some remarkable high-risk areas for the CVD

Risk Zones	Feature Scores
Men and Women both	0.77
PREVCHD (Prevalent Coronary Heart Disease, chest abnormality)	0.92
BP (Blood Pressure)	0.62
PREVMI (Myocardial Infarction)	0.60
Heart Rate (Abnormalities)	0.59
PREVHYP (Hypertension)	0.63
PREVSTRK (Stroke)	0.99

The other additional features approached in this work, the outcome are first findings on the prevalence of cardiovascular in Framingham, Massachusetts, including diagnosed cases and stratified by age, gender, chest abnormality, PREVAP, PREVCHD, PREVMI, PREVSTRK, and PREVHYP. It has been detected that a higher prevalence of cases was found in PREVCHD (Prevalent Coronary Heart Disease, chest abnormality or pain) and PREVSTRK (stroke). Additionally, other major features were also found during our research work. In conclusion, future research can be conducted by utilizing this approach on other deep learning algorithms. Predominantly, our experimental research can use the proposed approach to classify CVD for appropriate planning and management of CVD risk by authorities, policymakers, and non-governmental organizations.

5. Conclusion

CVD is an exponentially increasing disease and its prevalence has been becoming a wide-reaching challenge. Nowadays, millions of people are suffering from CVD complications. The goal and objective of this research are to achieve a better prediction strategies model to support healthcare and to detect risk factors for medical decision-making. This epidemic research demonstrates the CVD prevalence method by deep learning approaches, which contains a prominent influence on the healthcare sector for the population besides the implication techniques for medical specialists. In addition, the approaches were developed on a deep learning platform to measure the best performance result. Furthermore, it is significant for clinical implications and healthcare research especially in the states with low socioeconomic status and substantial epidemiological risk around the world. The main objectives of this work are to progress healthcare to offer useful scientific information for professional detection of risk factors. The significant innovations of this research include: Firstly, we presented a predicting algorithm for CVD based on the LSTM(P), which describes the relationship variable with the highest accuracy level and the goal that the development can be received for inspection and screening of the asymptomatic people. Secondly, SMOTE algorithms helped to balance the imbalanced CVD datasets. In addition, the classification assessment and detailed accuracy by class were performed, and an accuracy of 93% was obtained. Also, other performances assessment in Kappa statistics, PRC, MCC, precision, recall F-matrix and true positive ratios are very close to 100%. They give better performance result. Third, we detail the development of an accurate algorithm based on using the available healthcare datasets to identify some pertinent high risk factors related to cardiovascular patients which will help in prevalence. By using our proposed model, the patients can detect the presence and detect CVD early, thereby informing the doctors to undertake early actions for a correct treatment. We used 0.55 as our threshold point to know major risk factors of CVD, some features are having much higher value than the threshold point, which might be the remarkable risk factors. Moreover, the sensitivity, MCC were computed and other performance metrics were obtained with significant scores. The performances based on the comparison can lead our LSTM(P) model to have significant result and better detection level of major risk factors of CVD. Early prediction treatment can inspire the entire nation to understand the common and modified healthcare. It also serves as a possible routine for CVD patient even those who already entered in to the CVD danger zone due to their rebellious lifestyle in the young generation. We should now be concentrating on the prognosis of the major risk factors rather than detection of the CVD, given that the latter can be accomplished using our proposed module by means of deep learning algorithms.

Funding: This research received no external funding.
Conflicts of Interest: The authors declare no conflict of interest.
ORCID iD: Shake Ibna Abir[1] (https://orcid.org/my-orcid?orcid=0009-0004-0724-8700), Shaharina Shoha[1] (https://orcid.org/0009-0008-8141-3566)

References

[1] Institute of Medicine (US) Committee on Preventing the Global Epidemic of Cardiovascular Disease: Meeting the Challenges in Developing Countries, Promoting Cardiovascular Health in the Developing World: A Critical Challenge to Achieve Global Health, in: V. Fuster, B.B. Kelly (Eds.), Epidemiology of Cardiovascular Disease, Vol. 2, National Academies Press (US), Washington (DC).

[2] WILLIAMS K V., BECKER D J, ORCHARD T J et al. Persistent C-peptide levels and microvascular complications in childhood onset type 1 diabetes of long duration[J]. Journal of Diabetes and its Complications, Elsevier Inc., 2019, 33(9): 657–661.

[3] SHARMA A, VAS P, COHEN S et al. Clinical features and burden of new onset diabetic foot ulcers post simultaneous pancreas kidney transplantation and kidney only transplantation[J]. Journal of Diabetes and its Complications, Elsevier Inc., 2019, 33(9): 662–667.

[4] REN H, SHAO Y, MA X et al. Expression levels of serum vasohibin-1 and other biomarkers in type 2 diabetes mellitus patients with different urinary albumin to creatinine ratios[J]. Journal of Diabetes and its Complications, Elsevier Inc., 2019, 33(7): 477–484.

[5] SAULNIER P J, BRIET C, GAND E et al. No association between fear of hypoglycemia and blood glucose variability in type 1 diabetes: The cross-sectional VARDIA study[J]. Journal of Diabetes and its Complications, Elsevier Inc., 2019, 33(8): 554–560.

[6] UBA M M, REN J, SOHAIL N M et al. Principal Component Analysis of Categorized Polytomous Variable-Based Classification of Diabetes and Other Chronic Diseases[J]. Environmental Research and Public Health, 2019, 5(3): 24-32.

[7] WHO, https://www.who.int/hrh/work forc e_mdgs / en.

[8] WHO, https://www.who.int/hrh/work forc e_mdgs / en.

[9] SAINI S. Hybrid Model Using Unsupervised Filtering Based On Ant Colony Optimization And Multiclass Svm By Considering Medical Data Set[J]. International Research Journal of Engineering and Technology(IRJET), 2017, 4(6): 2565–2571.

[10] PEARCE I, SIMÓ R, LÖVESTAM-ADRIAN M et al. Association between diabetic eye disease and other complications of diabetes: Implications for care. A systematic review[J]. Diabetes, Obesity and Metabolism, 2019, 21(3): 467–478.

[11] SHALVI D, DECLARIS N. Unsupervised neural network approach to medical data mining techniques[C]//IEEE International Conference on Neural Networks - Conference Proceedings. IEEE, 1998, 1: 171–176.

[12] Domadiya, N., Rao, U.P. Improving healthcare services using source anonymous scheme with privacy preserving distributed healthcare data collection and mining. Computing 103, 155–177 (2021).

[13] Palaniappan S, Awang R (2013) Intelligent heart disease prediction system using data mining techniques. Int J Healthc Biomed Res 1:94.

[14] Luukka P, Lampinen J (2010) A classification method based on principal component analysis and differential evolution algorithm applied for prediction diagnosis from clinical emr heart data sets. In: Computational intelligence in optimization, Springer, pp 263–283.

[15] CIOS K J. From the guest editor: Medical data mining and knowledge discovery[J]. IEEE Engineering in Medicine and Biology Magazine, Institute of Electrical and Electronics Engineers Inc., 2000, 19(4): 15–16.

[16] TSUMOTO S. Problems with mining medical data[C]//Proceedings - IEEE Computer Society's International Computer Software and Applications Conference. IEEE Comput. Soc, 2000: 467–468.

[17] BRAMEIER M, BANZHAF W. A comparison of linear genetic programming and neural networks in medical data mining[J]. IEEE Transactions on Evolutionary Computation, 2001, 5(1): 17–26.

[18] OLUKUNLE A, EHIKIOYA S. A fast algorithm for mining association rules in medical image data[C]//Canadian Conference on Electrical and Computer Engineering. 2002, 2: 1181–1187.

[19] Noman M. Sohail, Ren Jiadong, Irshad. M., Shake Ibna Abir; Data mining techniques for Medical Growth: A Contribution of Researcher reviews; [J] Int. J. Comp. Sci. Network Sec; v. 18(10): 5-10, 2018. (ESCI; chapter 2).

[20] BRUNETTI A, CARNIMEO L, TROTTA G F et al. Computer-assisted frameworks for classification of liver, breast and blood neoplasias via neural networks: A survey based on medical images[J]. Neurocomputing, Elsevier B.V., 2019, 335: 274–298.

[21] Noman M. Sohail, Ren Jiadong, Musa Uba. M., Shake Ibna Abir., Irshad M., Musavir B.; Group covariates assessment on real-life Diabetes patients by Fractional Polynomials: a study based on Logistic Regression Modeling; [J] Biotech Research; v.10: 116-125, 2019. (EI; chapter 3).

[22] BRUNIE L, MIQUEL M, PIERSON J M et al. Information grids: Managing and mining semantic data in a grid infrastructure; Open issues and application to geo-medical data[C]//Proceedings - International Workshop on Database and Expert Systems Applications, DEXA. Institute of Electrical and Electronics Engineers Inc., 2003, 2003-Janua: 509–515.

[23] Noman M. Sohail, Ren Jiadong, Shake Ibna Abir, Wasim I., Usman A., Tahir R., Anthony J. V.; Why only data mining? A pilot study on inadequacy and domination of data mining technology; [J] Int. J. Recent Sci. Research; v. 9(10B): 29066-29073, 2018. (Scopus; chapter2).

[24] CHENG T H, WEI C P, TSENG V S. Feature selection for medical data mining: Comparisons of expert judgment and automatic approaches[C]//Proceedings - IEEE Symposium on Computer-Based Medical Systems. 2006, 2006: 165–170.

[25] WONG KOK SENG, BIN BESAR R, ABAS F S. Collaborative Support for Medical Data Mining in Telemedicine[C]//Institute of Electrical and Electronics Engineers (IEEE), 2006: 1894–1899.

[26] BETHEL C L, HALL L O, GOLDGOF D. Mining for implications in medical data[C]//Proceedings - International Conference on Pattern Recognition. 2006, 1: 1212–1215.

[27] V. Chaurasia and S. Pal, "Data Mining Approach to Detect Heart Dieses," International Journal of Advanced Computer Science and Information Technology (IJACSIT), vol. 2, no. 4, pp. 56-66, 2013.

[28] M. Dhanashree S, B. Mayur P, and D. Shruti D, "Heart Disease Prediction System using Naive Bayes," International journal of Enhanced Research In Science Technology & Engineering, vol. 2, no. 3, pp. 290-294, 2013.

[29] G. Parthiban and S. K. Srivatsa, "Applying Machine Learning Methods in Diagnosing Heart Disease for Diabetic Patients," International Journal of Applied Information Systems, vol. 3, no. 7, pp. 25-30, 2017.

[30] A. F. Otoom, E. E. Abdallah, Y. Kilani, A. Kefaye, and M. Ashour, "Effective diagnosis and monitoring of heart disease," International Journal of Software Engineering and its Applications, vol. 9, no. 1, pp. 143-156, 2015.

[31] M. Saqlain, W. Hussain, N. A. Saqib, and M. A. Khan, "Identification of Heart Failure by Using Unstructured Data of Cardiac Patients," in International Conference on Parallel Processing Workshops, pp. 426- 431, 2016.

[32] A. Gavhane, G. Kokkula, I. Pandya, and P. K. Devadkar, "Prediction of Heart Disease Using Machine Learning," in International Conference on Electronics, Communication and Aerospace Technology (ICECA), pp. 1275-1278, 2018.

[33] A. Chauhan, A. Jain, P. Sharma, and V. Deep, "Heart Disease Prediction using Evolutionary Rule Learning," in International Conference on Computational Intelligence & Communication Technology (CICT), pp. 1-4,2018.

[34] Purushottam, K. Saxena, and R. Sharma, "Efficient heart disease prediction system using decision tree," in International Conference on Computing, Communication & Automation, pp. 72-77, 2015.

[35] J. Thomas and R. T. Princy, "Human heart disease prediction system using data mining techniques," in International Conference on Circuit, Power and Computing Technologies (ICCPCT), pp. 1-5, 2016.

[36] M. Raihan, S. Mondal, A. More, and M. Sagor et al., "Smartphonebased ischemic heart disease (heart attack) risk prediction using clinical data and data mining approaches, a prototype design," in International Conference on Computer and Information Technology (ICCIT), pp. 299-303, 2016.

[37] S. Ismaeel, A. Miri, and D. Chourishi, "Using the Extreme Learning Machine (ELM) technique for heart disease diagnosis," in IEEE Canada International Humanitarian Technology Conference, pp. 1-3, 2015.

[38] Z. Shen, M. Clarke, R. W. Jones, and T. Alberti, "Detecting the risk factors of coronary heart disease by use of neural networks," in International Conference of the IEEE Engineering in Medicine and Biology Societ, pp. 277-278, 1993.

[39] Z. Shen, M. Clarke, R. W. Jones, and T. Alberti, "Detecting the risk factors of coronary heart disease by use of neural networks," in International Conference of the IEEE Engineering in Medicine and Biology Societ, pp. 277-278, 1993.

[40] J. S. Sonawane and D. R. Patil, "Prediction of heart disease using learning vector quantization algorithm," in Conference on IT in Business, Industry and Government: An International Conference by CSI on Big Data, CSIBIG, pp. 1-5, 2014.

[41] M. G. Feshki and O. S. Shijani, "Improving the heart disease diagnosis by evolutionary algorithm of PSO and Feed Forward Neural Network," in Artificial Intelligence and Robotics (IRANOPEN), pp.48-53, 2016.

[42] S. A. Hannan, A. V. Mane, R. R. Manza, and R. J. Ramteke, "Prediction of heart disease medical prescription using radial basis function," in IEEE International Conference on Computational Intelligence and Computing Research, ICCIC, pp. 735-740, 2010.

[43] T. Karayilan and O. Kilic, "Prediction of Heart disease using neural network," in International Conference on Computer Science and Engineering, pp. 719-723, 2017.

[44] D. K. Ravish, K. J. Shanthi, N. R. Shenoy, and S. Nisargh, "Heart function monitoring, prediction and prevention of Heart Attacks: Using Artificial Neural Networks," in International Conference on Contemporary Computing and Informatics, pp. 1-6, 2014.

[45] E. d. A. Botter, C. L. Nascimento, and T. Yoneyama, "A neural network with asymmetric basis functions for feature extraction of ECG P waves," IEEE Transactions on Neural Networks, vol.12, no. 5, pp. 1252-1255,2001.

[46] G. Shanmugasundaram, V. M. Selvam, R. Saravanan, and S. Balaji, "An Investigation of Heart Disease Prediction Techniques," in IEEE International Conference on System, Computation, Automation and Networking (ICSCA), pp. 1-6, 2018.

[47] H. Kahtan, K. Z. Zamli, W. N. A. W. A. Fatthi, and A. Abdullah et al., "Heart Disease Diagnosis System Using Fuzzy Logic," in International Conference on Software and Computer Applications, pp. 297-301, 2018.

[48] K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," Journal of Artificial Intelligence Research, vol. 16, no. 1, pp. 321–357, 2002.

[49] Gers, F. (2001). Long short-term memory in recurrent neural networks. PhD diss., Beuth Hochschule fur Technik Berlin.

[50] Gers, F.A.,&Schmidhuber, J. (2001). LSTM recurrent networks learn simple contextfree and context-sensitive languages. IEEE Trans. Neural. Netw., 12(6), 1333–1340.

[51] Abir, S. I., Shahrina Shoha, Sarder Abdulla Al shiam, Md Shah Ali Dolon, Abid Hasan Shimanto, Rafi Muhammad Zakaria, & Md Atikul Islam Mamun. (2024). Deep Neural Networks in Medical Imaging: Advances, Challenges, and Future Directions for Precision Healthcare . Journal of Computer Science and Technology Studies, 6(5), 94–112. https://doi.org/10.32996/jcsts.2024.6.5.9

[52] Shaharina Shoha, Abir, S. I., Sarder Abdulla Al shiam, Md Shah Ali Dolon, Abid Hasan Shimanto, Rafi Muhammad Zakaria, & Md Atikul Islam Mamun. (2024). Enhanced Parkinson's Disease Detection Using Advanced Vocal Features and Machine Learning .Journal of Computer Science and Technology Studies,6(5), 113–128. https://doi.org/10.32996/jcsts.2024.6.5.10

[53] Abir, Shake Ibna and Shoha, Shaharina and Dolon, Md Shah Ali and Al Shiam, Sarder Abdulla and Shimanto, Abid Hasan and Zakaria, Rafi Muhammad and Ridwan, Mohammad, LungCancer Predictive Analysis Using Optimized Ensemble and Hybrid Machine Learning Techniques. Available at SSRN: https://ssrn.com/abstract=4998936or http://dx.doi.org/10.2139/ssrn.4998936

[54] S. I. Abir, S. Shoha, S. A. Al Shiam, M. M. Uddin, M. A. Islam Mamun and S. M. Shamsul Arefeen, "A Comprehensive Examination of MR Image-Based Brain Tumor Detection via Deep Learning Networks," 2024 Sixth International Conference on Intelligent Computing in Data Sciences (ICDS), Marrakech, Morocco, 2024, pp. 1-8, https://doi.10.1109/ICDS62089.2024.10756457

[55] S. I. Abir, S. Shoha, S. A. Al Shiam,M. M. Uddin, M. A. Islam Mamun and S. M. Shamsul Arefeen, "Health Risks and Disease Transmission in Undocumented Immigrants in the U.S Using Predictive ML,"2024 Sixth International Conference on Intelligent Computing in Data Sciences (ICDS), Marrakech, Morocco, 2024, pp. 1-6, https://doi.10.1109/ICDS62089.2024.10756308

[56] Abir, Shake Ibna, Richard Schugart, (2024). Parameter Estimation for Stroke Patients Using Brain CT Perfusion Imaging withDeep Temporal Convolutional Neural Network, Masters Theses & Specialist Projects, Paper 3755.

[57] Sohail, M. N., Ren, J., Muhammad, M. U., Rizwan, T., Iqbal, W., Abir, S. I., and Bilal, M, (2019). Group covariates assessment on real life diabetes patients by fractional polynomials: a study based on logistic regression modeling, Journal of Biotech Research, 10, 116-125.

[58] Sohail, M. N., Jiadong, R., Irshad, M., Uba, M. M., and Abir, S. I, (2018). Data mining techniques for Medical Growth: A Contribution of Researcher reviews, Int. J. Comput. Sci. Netw. Secur, 18, 5-10.

[59] Sohail, M. N., Ren, J. D., Uba, M. M., Irshad, M. I., Musavir, B., Abir, S. I., and Anthony, J. V, (2018). Why only data mining? a pilot study on inadequacy and dominationof data mining technology, Int. J. Recent Sci. Res, 9(10), 29066-29073.

[60] Abir, S. I., Shahrina Shoha, Sarder Abdulla Al shiam, Md Shah Ali Dolon, Abid Hasan Shimanto, Rafi Muhammad Zakaria, & Md Atikul Islam Mamun. (2024). Deep Neural Networks in Medical Imaging: Advances, Challenges, and Future Directions for Precision Healthcare . Journal of Computer Science and Technology Studies, 6(5), 94-112. https://doi.org/10.32996/jcsts.2024.6.5.9

[61] Abir, S. I., Shaharina Shoha, Sarder Abdulla Al Shiam, Shariar Islam Saimon, Intiser Islam, Md Atikul Islam Mamun, Md Miraj Hossain, Syed Moshiur Rahman, & Nazrul Islam Khan. (2024). Precision Lesion Analysis and Classification in Dermatological Imaging through Advanced Convolutional Architectures. *Journal of Computer Science and Technology Studies*, *6*(5), 168-180.