Section A-Research paper

Adjusted Crossover Model to Expand Clinical Decisions Accuracy for Diagnosisof Chronic Diseases

Shalini[1*], Dr.Kavita[2]

Abstract:
People's way of life and diet have slowly changed as a result of the rapid growth of the economy in modern society. They have a bad habit and don't exercise enough, which has a negative impact on their health. In recent years, chronic diseases such as diabetes, cardiovascular sickness, and eternal kidney disease have become major public health issues worldwide. Huge published efforts in the medical field have indicated that diagnosis of these diseases at an early stage can greatly enhance patient outcomes and reduce healthcare costs. Furthermore, investigated efforts have also denoted that by consuming clinical data, data mining algorithms competently offer great potential in the prediction and diagnosis of a vast number of diseases. However, in a real-time frame, several approaches are available in an easy mode that are rapidly utilized by experts to predict and diagnose such diseases, but the associated flaws of former techniques such as some approaches taking a significant amount of time to compute while other methods are quick but inaccurate, have generate the need for additional research. This paper proposes a more precise clinical decision support model for the prediction of heart and diabetes disease with addressing the issue of existing methods. For this purpose, disease affected patient data has accessed from UCI's open repositories and taking into account the disease prediction capabilities of five classifiers, including NB, DT, SVM, IBK, and ANN. Additionally, the standard K-fold cross-validation method has used to train and validate the proposed processes. Attained upsurges amount of disease prediction accuracy have symbolized the proficiency of the proposed approach with attaining accuracy 94% for recognize diabetic patients and 93% accuracy for identify heart illness.

Index Terms: Cardiac Arrest,Diseases Prediction,ML,Data-Mining, Classification Techniques, Type 1 & 2 Diabetes.

[1*]PhD Scholar, Department of Science and Technology, Jayoti Vidyapeeth Women's University, Jaipur, Rajasthan, India
[2]Professor, Department of Science and Technology, Jayoti Vidyapeeth Women's University, Jaipur, Rajasthan, India
***Corresponding Author:**Shalini
*PhD Scholar, Department of Science and Technology, Jayoti Vidyapeeth Women's University, Jaipur, Rajasthan, India

I. Introduction

Around the globe, every healthcare provider faces the challenge of chronic diseases. A clinical proclamation points out that people are passing at a faster pace due to constant infections. During 2019, cardiovascular diseases (CVDs) triggered to 17.9 million expiries, according to the World Health Organization. Conferring by IDF (International Diabetes Federation), in 2019, there were around 463 million grown person ageing 20 to 79 with diabetes, and that quantity is likely to rise to 700 million till 2045 [1-3]. An additional investigation revealed that more than 84 million Americans age 18 or older have prediabetes [4]. In both cases, human health is significantly affected.

As a result of digestion, glucose is released by the human body. Cells consume blood glucose as it travels from the blood to them and renovate it into liveliness through insulin, a hormone found in the blood. Unsatisfactory production of insulin via pancreas stops the cells from absorbing glucose, which leaves glucose in the blood. Due to this, glucose and sugar levels in the blood rise to an extremely unacceptable level. Having an extremely high blood sugar level in the body can cause diabetes [5]. Diabetes can be divided into two categories: insulin- dependent diabetes mellitus, which is too acknowledged as type1 form of diabetes and requires insulin injections because the body does not produce enough insulin; and non-insulin- dependent diabetes mellitus, too acknowledged as type2 form of diabetes, in which the body's cells are incapable to effectually usage the insulin. Gestational diabetes, which develops during pregnancy, monogenic diabetes, which is brought on by a genetic mutation, and secondary diabetes are three other less common types of diabetes [6]. Diabetes is a significant risk factor for renal failure, cardiac arrest, vision issues, and lumbar spine amputation.

Contrarily, cardiovascular disease (CVD) is the term used to describe abnormalities in the heart's blood vessels brought on by plaque formation, a substance that constricts the arteries and veins that deliver blood to and from the heart. This makes it harder for blood to flow and might even cause blockages, which could cause a cardiac arrest. Hypertension, a faulty diet, idleness, elevated cholesterol levels, excessive drinking, nicotine consumption, being overweight, and abnormal habits of living are all risk factors for cardiovascular disease. Several tests are necessary for the proper diagnosis of CVD. It is difficult to spot someone in time, particularly in underdeveloped countries It can be challenging to diagnose someone in time, especially in nations with poor infrastructure where there's aren't enough skilled healthcare providers, diagnostic equipment, or other resources to identify and treat cardiac abnormalities [7]. Change has accelerated in the healthcare sector since the introduction of information technology (IT). Integrating IT into healthcare is intended to simplify and reduce costs for individuals' lives. Numerous healthcare organizations are currently utilizing computer-based predictive modelling. In addition, complex algorithms are identifying patterns and processes that are not visible. This helps scientists in creating novel medicines and drugs. Information mining, AI, and measurements are utilized in prescient displaying to track down designs in information and ascertain the probability of specific results. The formation of a diabetes and CVD predictive model is the main goal of this paper. Using the disease prediction capabilities of five classifiers—NB, DT, SVM, IBK, and ANN—and accessing disease-affected patient data from UCI's open repositories, this goal was accomplished. Moreover, the proposed processes were trained and validated using the standard K-fold cross-validation approach.

The remaining sections of this paper are arranged as follows: The second section covers the earlier work. The model design is covered in section 3, while Section 4 presents the experimental findings. The obstacles that still need to be overcome are listed in Section 5, and the experiment as a whole is wrapped up in Section 6.

II. Literature Review

The study that informed the development of the suggested model for chronic disease prediction is covered in this section. The discussion that follows builds on an analysis of the recently published work and strengthens and improves the creation of the suggested framework.

Numerous researchers in the medical research field used Pima Indian diabetes dataset (PIDD) to predict diabetes using a comprehensive version of data mining algorithms [8–12]. Using Logistic Regression and SVM, the authors of [13] created a diabetes estimation model. Data was pre-processed with the intention of an attractive fallouts. They initiate that applied scheme SVM outperformed, with a 79% accurateness. Researchers in [14] used the Naive Bayes, SVM, and Decision Tree methods. The practice of cross authentication (ten-fold) was made in order to expand recital. The Naive Bayes procedure engendered peak accuracy, 76.30%. For the purposes of the investigation, the Pima Indian Diabetes dataset was examined. For diabetic prediction, the authors of [15] used logistic regression on PIDD. They discovered that among all PIDD features, the most important variables for diabetes prediction are the number of pregnancies, BMI, and glucose level. RStudio is utilized to process and display the analysis results of the Pima Indian Diabetes dataset. With an accuracy of 75.32 percent, their model is making pretty accurate predictions. All of the patient data in study [16] are trained and tested through ten cross-validations

made with NB and a decision tree. The performance was then examined, evaluated, and contrasted with other WEKA classification algorithms. With an accuracy rate of 76.3021%, the study's findings showed that Naive Bayes was the most efficient method. The authors of [17] used Random Forest, Decision Tree, and ANN to classify PIDD after decreasing features using PCA & mRMR approaches. They found that the RF technique in the company of mRMR feature decreasing strategy created the best precision for Pima Indians, 77.21%.

Machine learning was used in a separate study to measure CVD in dialysis patients [18]. Italian and American datasets were used to test an act of various algorithms. With an accuracy of 95.25 percent in the Italian Dataset and 92.15 percent in the American Dataset, the Support Vector Machine (SVM) with the RBF Kernel algorithm produced the most accurate results. However, the bias in the Italian dataset could affect predictions' accuracy. Using the concept of Heart Rate Variability (HRV), a separate group in the medical research field has proposed a method for predicting cardiac arrest in smokers [19]. The non-invasive HRV method is used to examine how the heartbeat is controlled. To get the right data points, you need the best conditions and time. Comparisons were made between the Decision Tree, Logistic Regression, and Random Forest outcomes. The 10-overlay approval strategy is utilized to gauge the exhibition of the whole grouping methods. The results showed that Logistic Regression, Decision Tree, and Random Forest were all accurate to 93.61%, 89.7%, and 92.59%, respectively. It was determined that Random Forest was the most effective of these methods.

A learning approach that is mentioned in study [20] proposed for accurately predicting CVD and to overcome missing values in the medical dataset. With the use of Naive Bayes, SVM, Decision Tree, Logistic Regression, RBF, and Random Forest algorithms, the CVD was predicted. Despite missing values, the results show that RF was superior to other approaches with a sensitivity, specificity, and accuracy of 88%, 87.6%, and 88% respectively. Several published efforts [21-27] denoted that through proper utilization of machine learning techniques in this area may aid to experts in the early detection and prevention of these disease. Therefore, a naïve endeavor has been made in this paper using the prognostic act of diverse ML technique under a unique format to augment the diseases recognition accuracy.

III. Methodology

Gigantic investigations gave a relative examination among different AI calculations. Cross-validation and data pre-processing were used by some to improve accuracy, but they all focused more on comparing the performance of different models than on improving a single model. In this paper, we focused on a single model and looked at ways to increase forecasting performance of heart and diabetes diseases at increasing execution speed and accuracy simultaneously. This paper shows that notwithstanding calculation determination, pre-and post-handling of information assume a significant part in the general improvement of the model. Two fundamental datasets are utilized in this paper. The first is the PIMA Indian Diabetes Dataset [28], which includes data of 768 patients, 268 of whom have diabetes, and 500 of whom do not. There are nine characteristics or variables in total, eight of which are predictive variables and one of which is the objective variable. Such attributes are listed below:

Pregnancies: Number of times pregnant
Glucose: Plasma glucose concentration a 2 hours in an oral glucose tolerance test
BloodPressure: Diastolic blood pressure (mm Hg)
SkinThickness: Triceps skin fold thickness (mm)
Insulin: 2-Hour serum insulin (mu U/ml)
BMI: Body mass index (weight in kg/(height in m)^2)
DiabetesPedigreeFunction: Diabetes pedigree function
Age: Age (years)
Outcome: Class variable (0 or 1)

Figure 1. Attributes of PIMA Dataset

Figure 2, heatmap illustrates the connection between the characteristics of the PIMA dataset. More correlation is represented by lighter colors, whilst less correlation is represented by darker colors. Furthermore, one most notable exploration dataset of coronary illness offered by Cleveland Facility Establishment has taken in account [29]. This paper investigates a variety of approaches to increasing the percentage of accurate predictions for these diseases. The process by which the proposed model's implementation will be carried out is depicted in Figure 3.

As illustrated the methodology in figure 4, to increase the accuracy of heart and diabetes illness prediction, various data classification approaches are evaluated in a range of feature selection practices. Initially whole considered classical schemes evaluated with the suggested feature set that offered by each adopted practice. This evaluation has done to find out and elect most suited practice for further optimization of prediction performance. This procedure has a direct impact on the results, eliminates the possibility of selecting the incorrect method for disease prediction and saves investigators time and effort. On the base of such evaluation process a set of approaches has elected which offered top 2 prediction accuracy of such diseases. Furthermore, a hybrid approach is build-up

by utilizing the fetched algorithms into a layered form. In build approach each approach executed with the separate feature set. After building the proposed hybrid diseases prediction process it has assessed in front of classical as well as some of current offered scheme and the significant fallout of build approach reveal its supremacy.

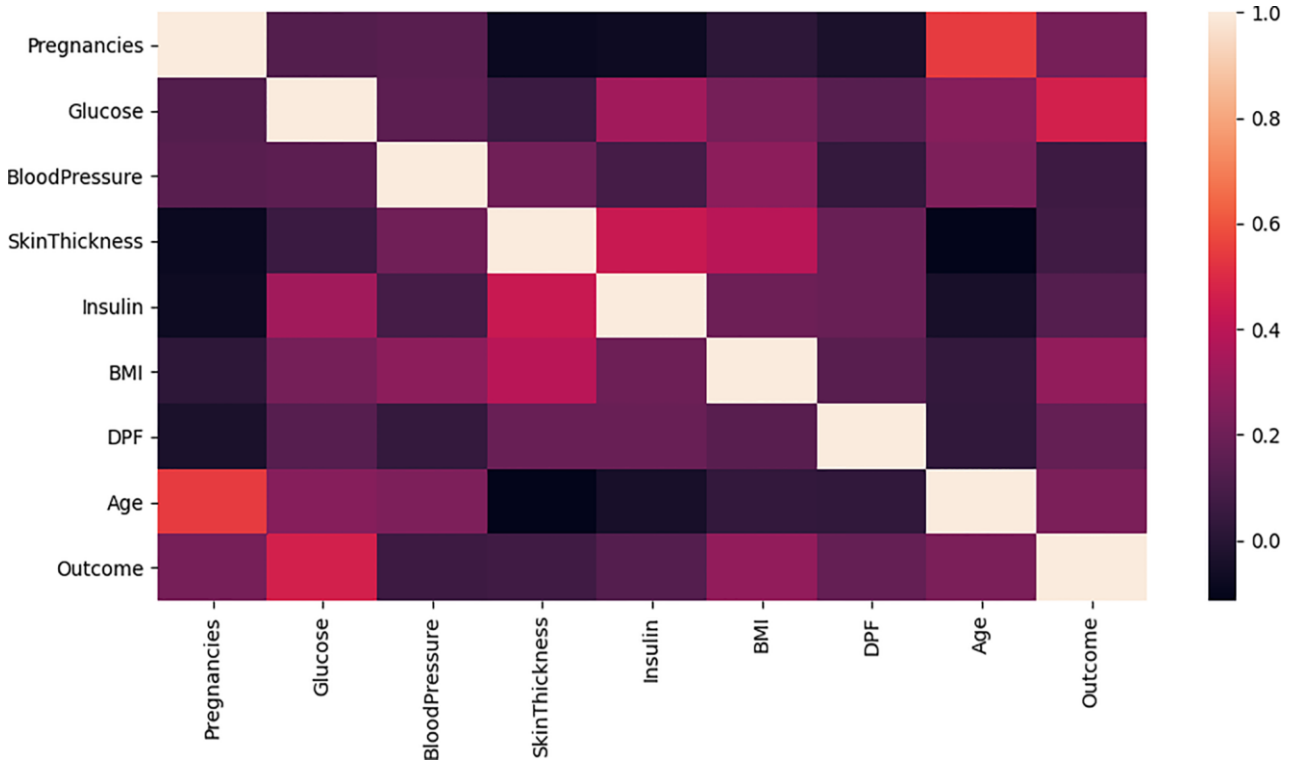

Figure2Relationship between the characteristics of the PIMA dataset [30]

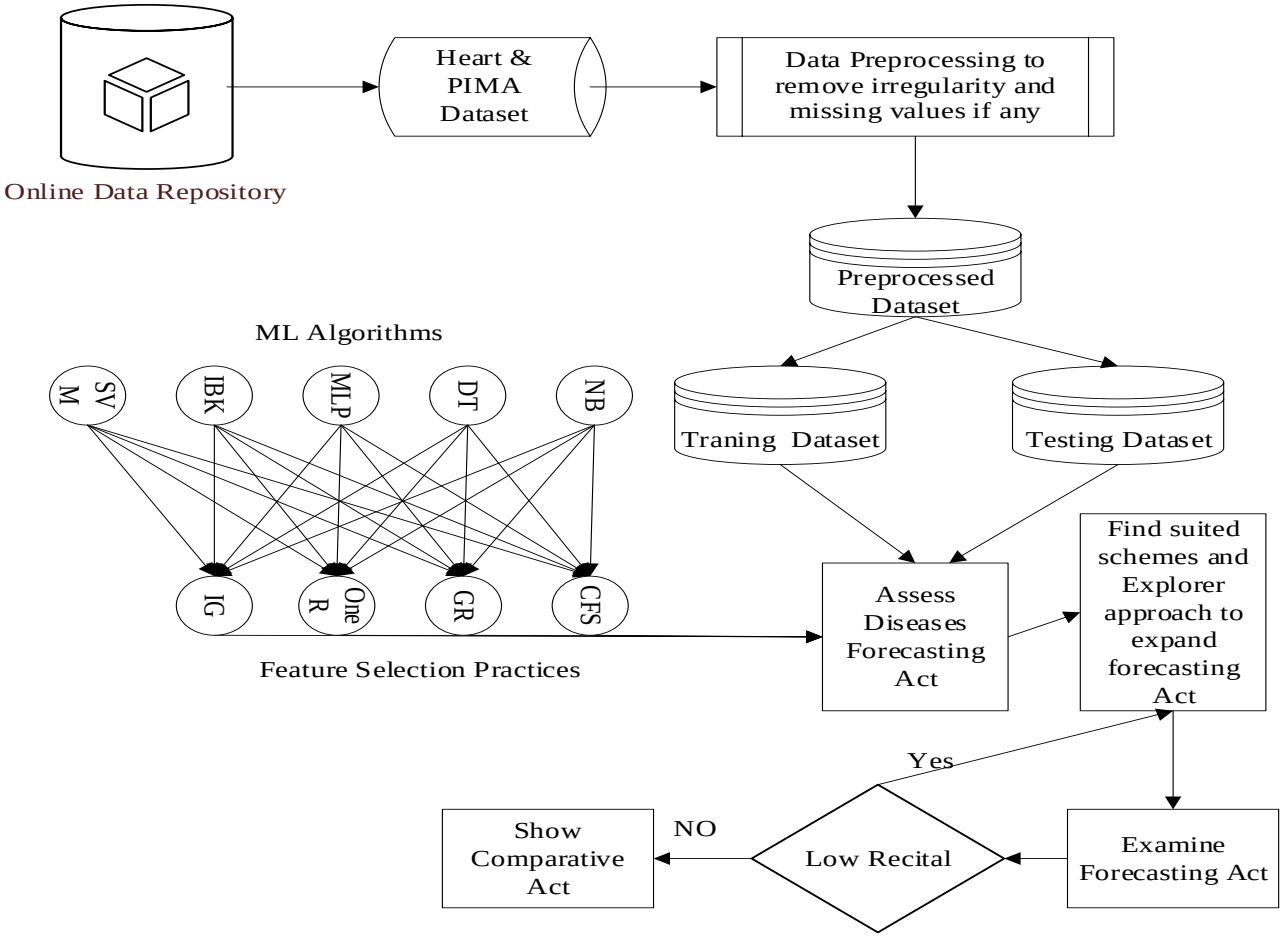

Figure 3 Methodology for Augmenting Diseases forecasting act

IV. Performance Evaluation and Discussion

During the assessment procedure, confusion matrix, accuracy, and precision score are utilized. A confusion matrix is a tabular representation that comprises actual and predicted values, namely true positive and true negative, depicted in figure 4.

		Predicted value	
		P	N
True value	P	TP	FN
	N	FP	TN

Fig. 4. Confusion Matrix

true positives (TP), which are occurrences that have been proven to be genuine, are the first component of the confusion matrix. The second factor false positives, or FPs, are occasions where something was thought to be true but was actually false. A false negative (FN), which is the third element, is a circumstance that was genuine but was seen negatively. The fourth component, known as a true negative (TN), refers to incidents that were clearly characterised as negative.

Method for verdict an accuracy is

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN) \quad (1)$$

The ratio of correct predictions to the total correct values, including true and false predictions, is referred to as the precision or positive predictive value (PPV) and can be represented mathematically as follows:

$$\text{Precision} = TP/ (TP + FP) \quad (2)$$

The ratio of the right predicted values to the sum of the proper positive forecasts and the erroneous negative predictions, also known as the recall, sensitivity, or true positive rate (TPR), is represented mathematically as follows:

$$\text{Recall} = TP/ (TP + FN) \quad (3)$$

The F score's precision and recall parts are combined and weighted by a factor β. The F1 score is more important than accuracy when there is an imbalance in the distribution of classes. This is especially true when the number of false positives and false negatives is different. The F1 score is characterized numerically as follows:

$$F1 \text{ score} = 2((\text{Precision}^* \text{ Recall}) / (\text{Precision} + \text{Recall})) \quad (4)$$

As was mentioned earlier, two distinct disease datasets—the Cleveland heart disease dataset and the PIMA diabetes dataset—were taken into consideration during the experimental procedure. In addition, the pre-processing process has taken into consideration for removing any associated flaws and only obtaining the best features. The attributes of the dataset, as well as the complete and suggested features set, as determined by the CFS, GR, OneR, and IG schemes for the PIMA Indian Diabetes datasets, are shown in Table 1. Figure 5 portrayed the evaluate aftermaths of the proposed strategy and the other considered methods for example NB,DT,MLP,IBK and the SVM.

TABLE I. Full and Suggested Features of PIMA datasets

	Suggested Attributes	Full Attributes
CFS suggested Attributes	Plas Mass Pedi Age	Preg Plas Pres Skin Insu Mass Pedi Age Class
GR suggested Attributes	Plas Mass Pedi	
OneR suggested Attributes	Preg Plas Age Insu Mass	
IG suggested Attributes	Plas Mass Age Insu	

The outcomes of the NB, DT, MLP, IBK, and the SVMare shown in Table 2 &Figure 5.

TABLE II. Diabetic Patients Prediction Act of Assorted Techniques

Evaluating Approaches	Accuracy (%) With Suggested Features set			
	CFS	GR	IG	OneR
NB	**75.52**	74.22	73.05	73.05
DT	**74.87**	74.61	73.04	73.05
MLP	75.52	**76.43**	74.1	74.08
IBK	68.36	**70.18**	70.05	70.1
SVM	62.37	63.28	**71.48**	71.4

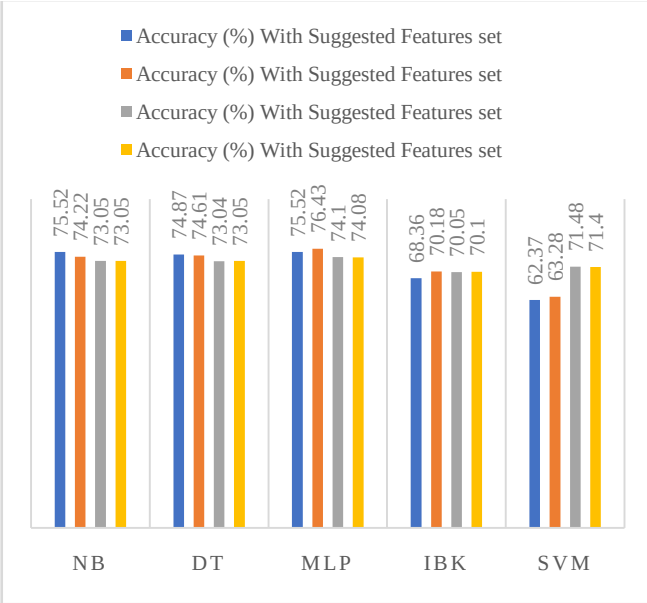

Figure 5. Diabetes Prediction act of Assorted Techniques

With the suggested feature set of the CFS and GR schemes, NB and MLP provided significantly improved results for diabetes prediction (denoted in table 2 and figure 5). As a result, the proposed model takes advantage of these strategies to predict whether or not a person will develop diabetes. Following table 3 and figure 6 meant the significancy among the current and the proposed model.

TABLE III. Diabetic Patients Prediction Act of Proposed and the Assorted Techniques

Evaluating Approaches	Attained Accuracy (%)
NB	75.52
DT	74.87
MLP	76.43
IBK	70.18
SVM	71.48
Proposed Hybrid Approach	**94**

The various best prediction accuracy acts of assorted techniques are compared to the proposed scheme in table 3. When compared to other schemes, the obtained results indicate that the proposed method is superior. Figure 6 present graphical impression of the comparative fallouts.

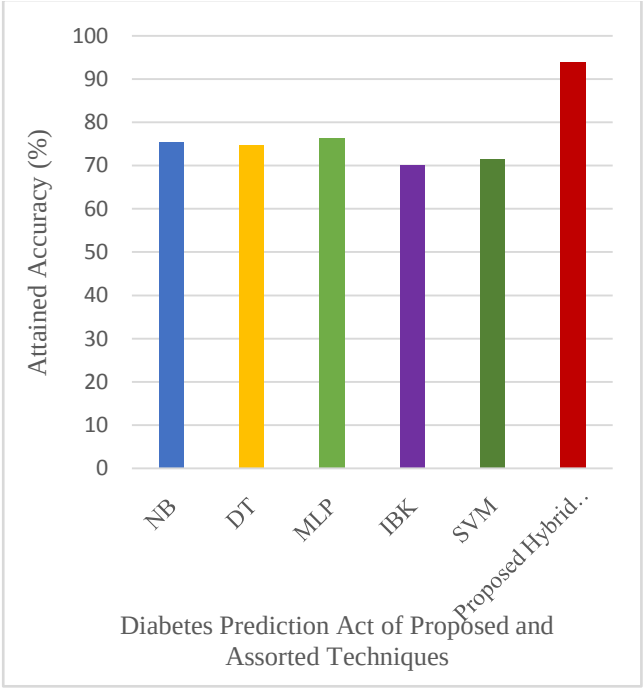

Figure 6. Diabetes Prediction act of Assorted & Proposed Techniques

The comparative analysis of the various approaches shown above showed that the build method is more effective than these strategies. Additionally, the build approach has been evaluated for heart disease prediction. Figures showing how each scheme performed and a comparison between the best-attained fallouts of each approach and the proposed model are presented in the following tables.

TABLE IV. Heart Diseases Prediction Act of Assorted Techniques

Evaluating Approaches	Accuracy (%) With Suggested Features set			
	CFS	GR	IG	OneR
NB	83.17	83.1	**83.83**	83.8
DT	**77.23**	76.57	76.5	76.57
MLP	**82.50**	79.54	82	82.05
IBK	79.21	**79.87**	77.2	77.3
SVM	75.25	**76.57**	74.26	74.2

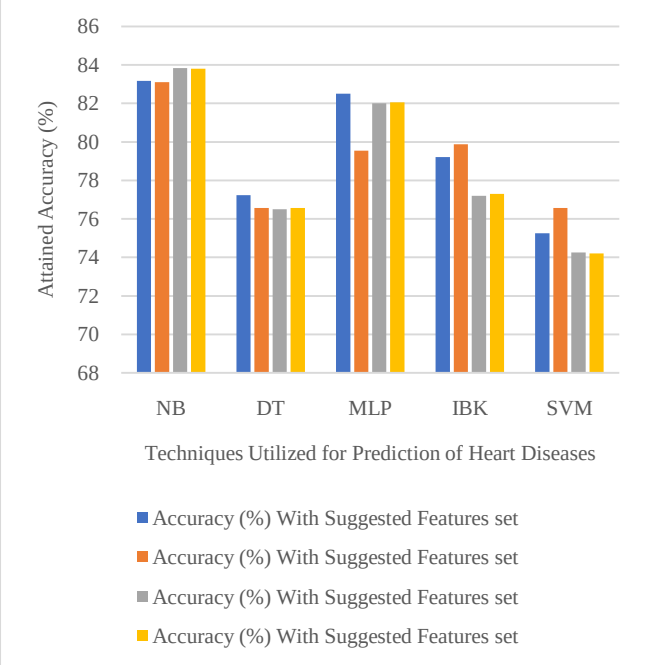

Figure 7. Diabetes Prediction act of Assorted Techniques

With the suggested feature set of the IG and CFS schemes, NB and MLP approached again provided significantly improved results for prediction of heart diseases (denoted in table 4 and figure 7). As a result, the proposed model takes advantage of these strategies to predict heart diseases whether or not a person has a heart illness or not. Following table 5 and figure 8 meant the significancy among the current and the proposed model.

TABLE V. Prediction Act of Heart Illness: Proposed Vs the Assorted Techniques

Evaluating Approaches	Attained Accuracy (%)
NB	83.83
DT	77.23
MLP	82.50
IBK	79.87
SVM	76.57
Proposed Hybrid Approach	**93**

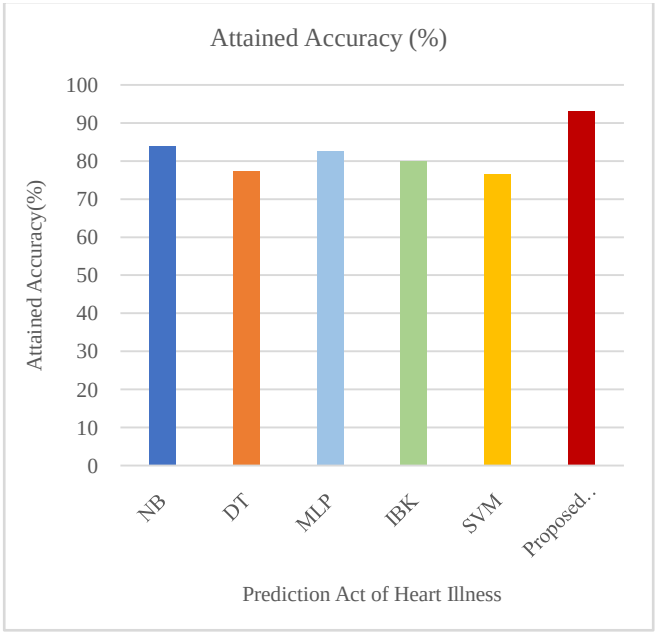

Figure 8. Prediction act of Assorted & Proposed Techniques for Heart Illness

When compared to other methods for predicting diabetes and heart disease, the overall comparison results demonstrated the significance of the proposed method. The outcomes achieved clearly indicate that the proposed strategy is a superior option.

V. Conclusion

The strategy for identifying and predicting the existence of chronic disease was suggested in this research. The suggested system has an advantage over many of the existing methods in that it uses real-world, structured and unstructured data to prepare the data set. The performance of the suggested model is evaluated in this study in comparison to other algorithms, including Naive Bayes, decision trees, IBK, MLP, and SVM algorithms. The outcomes demonstrate that the suggested approach offers a higher level of disease prediction accuracy than the other varying algorithms. It is firmly believed that the suggested system can lower the risk of chronic diseases by diagnosing them sooner and that it may also decrease the cost of healthcare assistance, assessment, and care.

References

Cardiovascular Diseases (Cvds), "World health organization," https://www.who.int/news-oom/fact%20sheets/detail/cardiovascular-diseases-(cvds).

K. Dwivedi, S. A. Imtiaz, and E. R. Villegas, "Algorithms for automatic analysis and classification of heart sounds – a systematic review," IEEE Access, vol. 7, 2019.

A Coronary, "Heart disease," Available from: https://www.aihw.gov.au/reports/australias-health/coronaryheart-disease, 2020.

Priyanka Rajendra, Shahram Latifi "Prediction of diabetes using logistic regression and ensemble techniques" Computer Methods and Programs in Biomedicine Update, Volume 1, 2021.

Jobeda Jamal Khanam, Simon Y. Foo "A comparison of machine learning algorithms for diabetes prediction" ICT Express, Volume 7, Issue 4, 2021, Pages 432-439.

A. Mujumdar, V. Vaidehi, Diabetes prediction using machine learning algorithms, in: International Conference on Recent Trends in Advanced Computing", 2019, ICRTAC, 2019.

Stephen F. Weng, Jenna Reps, Joe Kai, Jonathan M. Garibaldi, Nadeem Qureshi, "Can machine-learning improve cardiovascular risk predict ion using rout ine clinical dat a?", PLoS One 12(4): e0174944, April, 2017.

T.M. Alam, et al., Informatics in medicine unlocked a model for early prediction of diabetes, Inform. Med. Unlocked 16 (2019) 100204.

D. Sisodia, D.S. Sisodia, Prediction of diabetes using classification algorithms, Procedia Comput. Sci. 132 (2018) 1578–1585.

N.P. Tigga, S. Garg, Predicting type 2 Diabetes using Logistic Regression accepted to publish in: Lecture Notes of Electrical Engineering, Springer.

Salim Amour Diwani, Anael Sam, Diabetes forecasting using supervised learning techniques, Adv. Comput. Sci.: Int. J. [S.l.] (ISSN: 2322-5157) (2014) pp. -10–18.

S. Perveen, M. Shahbaz, A. Guergachi, K. Keshavjee, Performance analysis of data mining classification techniques to predict diabetes, Procedia Comput. Sci. 82 (2016) 115–121.

T.N. Joshi, P.M. Chawan, Logistic regression and SVM based diabetes prediction system, Int. J. Technol. Res. Eng. 11 (5) (2018).

D. Sisodia, D.S. Sisodia, Prediction of diabetes using classification algorithms, Procedia Comput. Sci. 132 (2018) pp. 1578–1585.

N.P. Tigga, S. Garg, Predicting type 2 Diabetes using Logistic Regression accepted to publish in: Lecture Notes of Electrical Engineering, Springer.

Salim Amour Diwani, Anael Sam, Diabetes forecasting using supervised learning techniques, Adv. Comput. Sci.: Int. J. [S.l.] (ISSN: 2322-5157) (2014) pp. 10–18.

Q. Zou, K. Qu, Y. Luo, D. Yin, Y. Ju, H. Tang, Predicting Diabetes Mellitus with Machine Learning Techniques, Vol. 9, Frontiers in genetics, 2018, p. 515

Sabrina Mezzatesta , Claudia Torino, Pasquale De Meo, Giacomo Fiumara , Ant onio Vilasi, " A machine learning-based approach for predict ing the outbreak of cardiovascular diseases in pat ients on dialysis" Comput er Met hods and P rograms in Biomedicine, Elsevier, vol. 177, pp. 9-15, August 2019

Shashikant R,Chetankumar P, "Predict ive model of cardiac arrest in smokers using machine learning technique based on Heart Rate Variability parameter", Applied Computing and Informatics, June 2019

Runchuan Li, Shengya Shen, Xingjin Zhang, Runzhi Li, Shuhong Wang, Bing Zhou and Zongmin Wang, " Cardiovascular Disease Risk P redict ion Based on Random Forest ", P roceedings of t he 2nd International Conference on Healthcare Science and Engineering, vol. 536, pp. 31-43, May 2019.

Chen, T.; Guestrin, C. XGBoost: A scalable tree boosting system. In Proceedings of the KDD '16: 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; Association for Computing Machinery: New York, NY, USA, 2016; pp. 785–794.

Gietzelt, M.;Wolf, K.-H.; Marschollek, M.; Haux, R. Performance comparison of accelerometer calibration algorithms based on 3D-ellipsoid fitting methods. Comput. Methods Programs Biomed. 2013, 111, 62–71.

K, V.; Singaraju, J. Decision Support System for Congenital Heart Disease Diagnosis based on Signs and Symptoms using Neural Networks. Int. J. Comput. Appl. 2011, 19, 6–12.

Waigi, R.; Choudhary, S.; Fulzele, P.; Mishra, G. Predicting the risk of heart disease using advanced machine learning approach. Eur. J. Mol. Clin. Med. 2020, 7, 1638–1645.

Shalini, Saini, P.K., Sharma, Y.M. (2021). An Intelligent Hybrid Model for Forecasting of Heart and Diabetes Diseases with SMO and ANN. In: Shorif Uddin, M., Sharma, A., Agarwal, K.L., Saraswat, M. (eds) Intelligent Energy Management Technologies. Algorithms for Intelligent Systems. Springer, Singapore.

Sharma, Y.M., Saini, P.K., Shalini, Sharma, N. (2021). Effective Decision Support Scheme Using Hybrid Supervised Machine Learning Procedure. In: Goyal, D., Gupta, A.K., Piuri, V., Ganzha, M., Paprzycki, M. (eds) Proceedings of the Second International Conference on Information Management and Machine Intelligence. Lecture Notes in

Networks and Systems, vol 166. Springer, Singapore.

Jaiswal, O., Saini, P.K., Shalini, Sharma, Y.M. (2021). Analyze Classification Act of Data Mining Schemes. In: Goyal, D., Gupta, A.K., Piuri, V., Ganzha, M., Paprzycki, M. (eds) Proceedings of the Second International Conference on Information Management and Machine Intelligence. Lecture Notes in Networks and Systems, vol 166. Springer, Singapore.

https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database

https://archive.ics.uci.edu/ml/datasets/heart+disease

Priyanka Rajendra, Shahram Latifi," Prediction of diabetes using logistic regression and ensemble techniques, Computer Methods and Programs in Biomedicine Update, Volume 1, 2021.