Journal of Statistics & Management Systems
ISSN 0972-0510 (Print), ISSN 2169-0014 (Online)
Vol. 27 (2024), No. 2, pp. 455–464
DOI : 10.47974/JSMS-1280

Diabetes prediction using machine learning classifiers with oversampling and feature augmentation

Mehroush Banday *
Sherin Zafar †
Parul Agarwal §
M. Afshar Alam ‡
Department of Computer Science
School of Engineering Sciences and Technology
Jamia Hamdard 110062
New Delhi
India

Abstract

Two of the leading causes of mortality in the US are diabetes and cardiovascular disease. The first step in halting the course of these disorders in patients is to recognize and anticipate them. Using survey data (as well as test findings), we assess the efficacy of machine learning algorithms for identifying patients who are at risk and pinpoint important characteristics in the data that are causing these diseases in the patients. Hunger, thirst, and frequent urination are signs of high blood sugar. Neglecting diabetes can lead to several issues. It's critical to be able to identify diabetes early. In this study, we use supervised machine-learning techniques such as Decision Tree, KNN, (RF)Random Forest, Logistic Regression, Ada Boost, and Gradient Boosting to train on the actual data of 520 diabetics. The proposed work has achieved a more thorough comparative analysis between different datasets and their features that may be carried out to pinpoint all the essential characteristics for predicting diabetes.

Subject Classification: *68T99.*

Keywords: *Gradient boosting, KNN, Ada boost, Cardiovascular disease, Diabetes mellitus.*

* *E-mail:* **mehroushbanday_sch@jamiahamdard.ac.in (Corresponding Author)**
† *E-mail:* **sherin.zafar@jamiahamdard.ac.in**
§ *E-mail:* **pagarwal@jamiahamdard.ac.in**
‡ *E-mail:* **aalam@jamiahamdard.ac.in**

1. Introduction

To Diabetes mellitus is still rising and is ranking among the most serious and costly chronic illnesses. Cardiovascular diseases (CVD)are the main cause of diabetes-related mortality in individuals with diabetes, and there is a close association between both conditions. Most people in the world suffer from heart disease, which has a high rate of morbidity and death. The volume of medical data is enormous and continues to grow, making it exceedingly challenging to manage and evaluate. Machine learning methods are thus a logical choice for handling this kind of data [1]. Diabetes can increase the risk of heart failure by increasing the incidence of cardiovascular disease (CVD), dyslipidemia, obesity, high blood pressure, and other risk factors. Globally, the prevalence of diabetes mellitus (DM) is significantly rising. By 2030, 596 million people or 2 in 10 worldwide will have diabetes, according to the most recent projections from the IDF Project. Type 1 and type 2 diabetes are both on the rise, however, Type 2 has contributed proportionally much more to globally increase in DM than Type1.The most common cause of morbidity and death among people with diabetes is CVD. Adults (over the age of 18) with diabetes mellitus (DM) had 1.7 times the mortality rate from CVD in the United States compared to those which are without the diagnosis. When compared to people without diabetes, one with diabetes had a relative risk of 2 to 4 for men and 3 to 6 for women for Cardiovascular Disease death and morbidity. The introduction portion of the paper is focused on ideas of diabetes and its classification based on their severity on health; additionally, relevant analysis was done on the various forms of research carried out for machine learning in healthcare in Section 2. Section 3 contains the suggested study objectives, followed by Section 4 which is about the analysis of the study and will give a full examination of the planned research work's numerous aims. Section 5 gives us the conclusion of the research. In the final section of the paper, a list of references is provided.

2. Literature Review

The related work section summarizes previous research on predictive analytics-based diabetes prediction and cardiovascular systems. Several studies have tried to employ MLA to address the problems of diabetes and cardiovascular prediction. ANN, NB, SVM, Decision Tree, and SVM studies are a few of them. Early diagnosis is crucial for diabetes and cardiovascular disease since it may be identified early. Some of the researchers have conducted a thorough examination of machine learning techniques for diabetes diagnosis and cardiovascular diagnosis, which is shown below:

S. No	Author	
1	Sharma et al., 2020	This review discusses the connections between diabetes and cardiovascular disease, looks at possible therapeutic approaches, reviews current care and therapy, and outlines future research possibilities [2].
2	Leon, 2015	This paper outlines the relationship between diabetes and heart attack or stroke, as well as possible causes for disease progression, current treatment recommendations, and future research objectives. [3].
3	Khan et al., 2022	This review fostered a troupe approach which is called as "Stacking Classifier" to work on the exhibition of coordinated adaptable individual classifiers and lessening the probability of misclassifying a solitary case. Bayes, Direct Discriminant Examination LDA and DT are only a couple of the classifiers utilized in this review. As a meta-classifier, SVM and Irregular Woodland are utilized [4].
4	Dinh et al., 2019	Our study looks at data-driven methods for identifying individuals with these illnesses that make use of supervised machine-learning algorithms. To create models for the diagnosis of cardiovascular disease, prediabetes, and diabetes, we do a thorough search of all feature variables present in the (NHANES) dataset[5].
5	Sekar & Selvarajan, 2023	The suggested study predicts and diagnoses cardiac and diabetes problems using Kaggle data. The two main causes of increased mortality rates among adults are diabetes and heart disease. Target characteristics for heart disease diagnosis are included in the dataset[6].

3. Proposed Approach

In Figure 1 the proposed system has been discussed by starting with the collection of data and finally by comparing the disease. The following steps have been taken for the proposed system.

- Data Collection: The real-world data consists of both unstructured and structured information, such as the patient's symptoms and medical consultations.To protect patient privacy, the data set does not include the patient's name, ID, or location.

- Construction of Machine Learning Model: Since the majority of the structured data are missing values, the acquired data are preprocessed. Therefore, to improve the quality of the data collection, it is imperative to complete any missing information or to eliminate or alter it[7].
- Constructed Machine Learning Model: The data set includes both unstructured data and structured data, as was previously mentioned. The structured data includes the following: the patient's environment, test results, the disease that they are affected by in tabular format, and patient demographics and information about the aetiology of the condition, such as height, age gender, weight, and so forth. The symptoms of the patient's illness and the textual records of the doctor interviews compose the unstructured data.
- Prediction and Comparison of Diseases Using Machine Learning Models: The first step is to gather data. Our proposed technique gathers data from several sources, both structured and unstructured. They go through preprocessing and are separated into trial and cleaning data sets after data gathering. The training data set is then taught across many epochs utilizing ML methods such as CNN, Random Forest, LR and KNN to enhance the accuracy of the prediction findings.[8].
- Data Testing

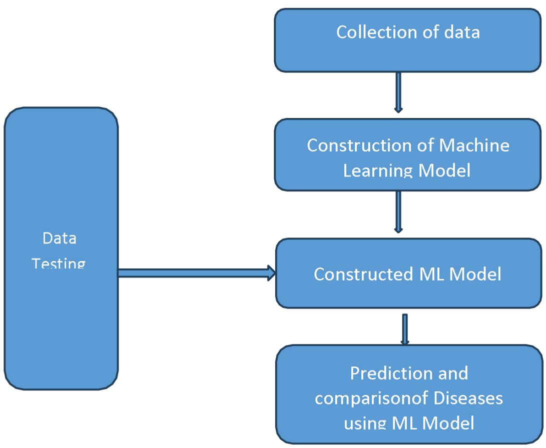

Figure 1

Proposed System

4. Performance Evaluation

Three performance assessment measures are applied to assess the suggested illness prediction model [9]. The true positives (TP), represent the target as a patient with a chronic illness, the true negatives (TN), represent the target as a person without a disease, the FP which is False Positive, represents the wrong assumption of the person who is healthy as a diseased person, and the false negatives (FN), which represent the wrong forecasting of the target as a person in good health, make up the confusion matrix [10]. The three assessment parameters are described as follows.

- F1 Score
- Recall
- Accuracy

As seen in Table 1 and Table 2, parameters recall, F1-score and accuracy are used to assess the performance of metrics in diabetes and cardiovascular disease.

Table 1

Performance Evaluation for Diabetes

Algorithms	Accuracy	F1 Score	Recall
KNN	0.678260869%	0.554216867%	0.554216867%
DecisionTree	0.7084347%	0.595238095%	0.60240963855%
RandomForest	0.7434782%	0.61437980%	0.566265060%
Logistic Regression	0.782608%	0.66666667%	0.6024096%
Ada Boost	0.6913043%	0.5942871%	0.62650200%

Table 2

Performance Evaluation for Cardiovascular Disease

Algorithms	Accuracy	F1 Score	Recall
KNN	0.9050001641%	0.205751633%	0.144509731%
DecisionTree	0.86442877%	0.244093801%	0.257069408%
RandomForest	0.906283713%	0.1752887866%	0.1169665809%
Logistic Regression	0.9161181381%	0.180540705%	0.18054070566%
Ada Boost	0.87033881080%	0.23049086016%	0.228057289753%

5. Analysis

In this section, an analysis has been done based on diabetes prediction concerning heart disease. After being downloaded in raw format, the dataset was transformed into a format that could be read by machines. The kinds and quantities of data that an algorithm may access determine how accurate it is. Before the original data was given to the classifier for prediction, it was cleaned up since it was noisy and unreliable. Pre-processing techniques are therefore crucial for improving data quality, which improves classification accuracy [11]. Many factors in the diabetes dataset are missing, which leads to ambiguity and erroneous findings. Because of this uncertainty, the suggested study data cleansing procedure is essential [12].Values that are missing are addressed, noise is being minimized, and dataset discrepancies are being corrected throughout the data-cleaning process. The diabetes dataset is shown in nine columns and 3000 rows in the diabetes dataset. The last column, which represents the "outcome" and aim characteristic, indicates whether the patient has diabetes or not. In addition, the result column only has two numbers: 0 for those without diabetes and 1 for those with diabetes [13]. The distribution of data seen in Figures 2 and 3 is shown and recognized using the Panda "Matplotlib" created library.

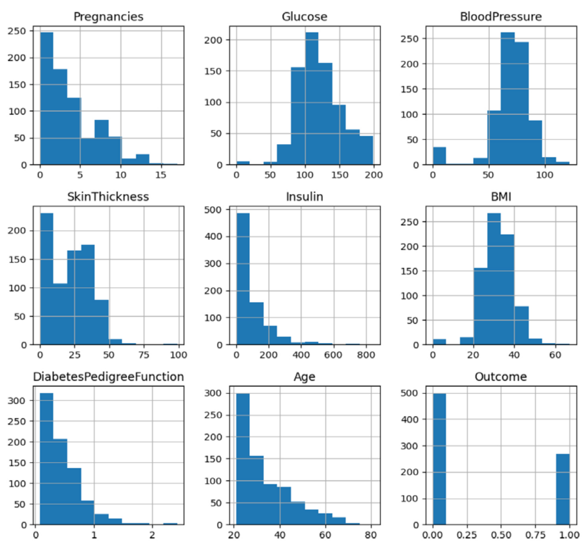

Figure 2

Distribution of Data for Diabetes

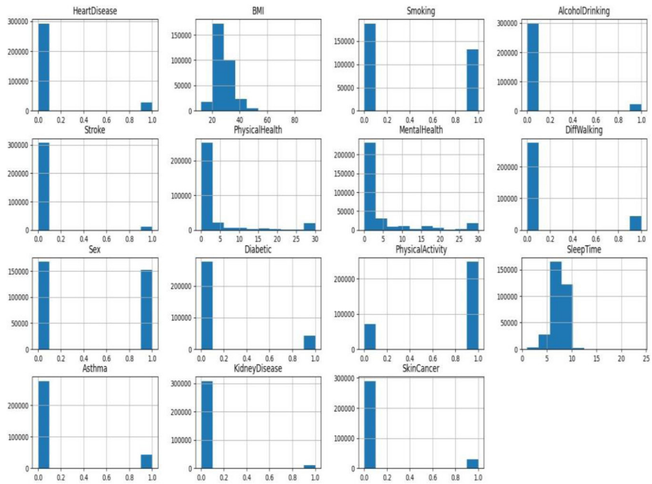

Figure 3

Distribution of Data for Cardiovascular Disease

5.1 *Correlation analysis of cardiovascular disease and diabetes.*

In this research study authors have specified a correlation analysis of diabetes and cardiovascular to analyze how one attribute is related to another and vice versa. Figure 4 and Figure 5 specifies about the attributes of diabetes and cardiovascular disease which are interrelated to each other.

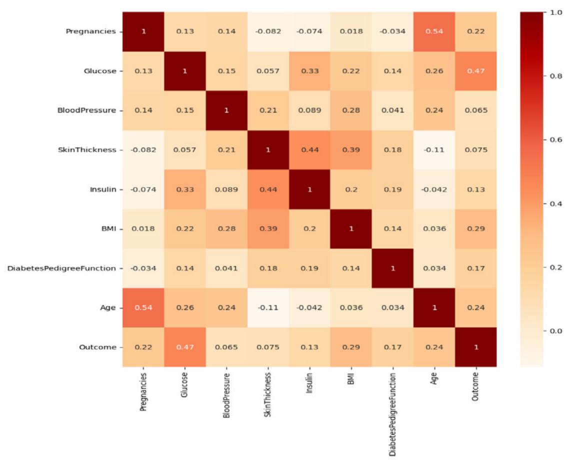

Figure 4

Correlation of attributes in Diabetes

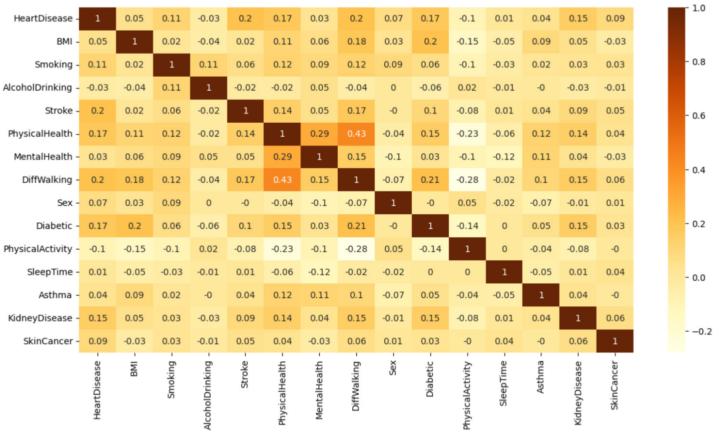

Figure 5

Correlation of attributes in cardiovascular disease

6. Conclusion and Future Work

To compare machine learning models' abilities to identify individuals with diabetes and cardiovascular disease, our study performs a thorough search through secondary data sourced online. In contrast to the Support Vector Machine-based diabetes diagnosis method, the models created in this work (based on non-laboratory data) exhibit a marginal improvement in accuracy [6]. Numerous research have been done in the past to identify cardiovascular disease and diabetes using just one classifier to predict persons with these conditions. However, a stacking classifier strategy was used in this work to classify cardiovascular disease and diabetes. Random Forest and SVM are examples of meta-classifiers, while KNN, Linear Discriminatory Analysis, Naive Bayes Approach and (DT)Decision tree methods are examples of basic classifiers. The recall, accuracy and f1-measure of the suggested model are compared with those of the other cardiovascular disease classifiers, such as NB, LDA, SVM, and DT[14]. Regarding the diagnosis of diabetic disease, the suggested stacking with RF acting as a meta-classifier is contrasted with the other classifiers like KNN, Naive Bayes and DT in the same manner in terms of the classification of metrics. Finding the characteristics that lead to the illnesses is one of the study's main contributions [15]. Our models can distinguish which categories—demographics (such as income and ethnicity), food consumption (such as salt, fibre, and caffeine intake), and physical factors (such as age, size etc)—affect the illness categorization in diabetes patients. The models identify patients with cardiovascular illnesses primarily based on their physical qualities (age, BP, weight, smoking, diabetes etc.), health problems (chest discomfort and hospitalization episodes), and dietary factors (calories, carbs, consumption of fibre etc.). Numerous similarities between the two conditions imply that those with diabetes may also be at risk for cardiovascular problems, and vice versa.

References

[1] A. Sharma, S. Mittal, R. Aggarwal, and M. K. Chauhan, "Diabetes and cardiovascular disease: inter-relation of risk factors and treatment," *Future Journal of Pharmaceutical Sciences*, vol. 6, no. 1, Dec. (2020), doi: 10.1186/s43094-020-00151-w.

[2] A. Sharma, S. Mittal, R. Aggarwal, and M. K. Chauhan, "Diabetes and cardiovascular disease: inter-relation of risk factors and treatment," *Future Journal of Pharmaceutical Sciences*, vol. 6, no. 1, Dec. (2020), doi: 10.1186/s43094-020-00151-w

[3] B. M. Leon, "Diabetes and cardiovascular disease: Epidemiology, biological mechanisms, treatment recommendations and future research," *World Journal of Diabetes*, vol. 6, no. 13, p. 1246 (2015), doi: 10.4239/wjd.v6.i13.1246.

[4] A. Khan, A. Khan, M. M. Khan, K. Farid, M. M. Alam, and M. B. M. Su'ud, "Cardiovascular and Diabetes Diseases Classification Using Ensemble Stacking Classifiers with SVM as a Meta Classifier," *Diagnostics*, vol. 12, no. 11, p. 2595, Oct. (2022), doi: 10.3390/diagnostics12112595.

[5] A. Dinh, S. Miertschin, A. Young, and S. D. Mohanty, "A data-driven approach to predicting diabetes and cardiovascular disease with machine learning," *BMC Medical Informatics and Decision Making*, vol. 19, no. 1, Nov. (2019), doi: 10.1186/s12911-019-0918-5.

[6] U. Sekar and K. Selvarajan, "Machine learning approach for predicting heart and diabetes diseases using data-driven analysis," *IAES International Journal of Artificial Intelligence (IJ- AI)*, vol. 12, no. 4, p. 1687, Dec. (2023), doi: 10.11591/ijai.v12.i4.pp1687-1694.

[7] W. Yu, T. Liu, R. Valdez, M. Gwinn, and M. J. Khoury, "Application of support vector machine modelling for prediction of common diseases: the case of diabetes and pre- diabetes," *BMC Medical Informat-*

ics and Decision Making, vol. 10, no. 1, Mar. (2010), doi: 10.1186/1472-6947-10-16.

[8] M. Bajaj, P. Rawat, S. Vats, V. Sharma, S. Mehta, and B. B. Sagar, "Enhancing patient outcomes through machine learning: A study of lung cancer prediction," *Journal of Information and Optimization Sciences*, vol. 44, no. 6, pp. 1075–1086 (2023), doi: 10.47974/jios-1438.

[9] M. TIWARI, R. Bharuka, P. Shah, and R. Lokare, "Breast Cancer Prediction Using Deep Learning and Machine Learning Techniques," *SSRN Electronic Journal* (2020), Published, doi: 10.2139/ssrn.3558786.

[10] H. Nasir, "Stability analysis and optimal control of a five-state diabetic population model," *Journal of Statistics and Management Systems*, vol. 25, no. 1, pp. 245–267, Jun. (2021), doi: 10.1080/09720510.2021.1881218.

[11] H. Nasir, "A time-delay model of diabetic population: Dynamics analysis, sensitivity, and optimal control," *Physica Scripta*, vol. 96, no. 11, p. 115002, Jul. (2021), doi: 10.1088/1402- 4896/ac1473.

[12] S. Bodiwala and N. Nanavati, "CDIPEM: Cardiovascular disease incidence probability estimation CNN model," *Journal of Information and Optimization Sciences*, vol. 44, no. 5, pp. 865–894 (2023), doi: 10.47974/jios-1175.

[13] L. Ali, A. Rahman, A. Khan, M. Zhou, A. Javeed, and J. A. Khan, "An Automated Diagnostic System for Heart Disease Prediction Based on ${\chi^{2}}$ Statistical Model and Optimally Configured Deep Neural Network," *IEEE Access*, vol. 7, pp. 34938–34945 (2019), doi: 10.1109/access.2019.2904800.

[14] L. Dehyadegari, S. Khajehasani, Z. Bazrafshan, M. A. Sharifi, and M. Yazdanpanah, "Epilepsy Diagnosis from EEG Signals Using Various Machine Learning Methods," *SSRN Electronic Journal* (2022), Published, doi: 10.2139/ssrn.4218393.

[15] B. Shivalkar *et al.*, "Flow Mediated Dilatation and Cardiac Function in Type 1 Diabetes Mellitus," *The American Journal of Cardiology*, vol. 97, no. 1, pp. 77–82, Jan. (2006), doi: 10.1016/j.amjcard.2005.07.111.