International Journal of Computer (IJC)
ISSN 2307-4523 (Print & Online)
https://ijcjournal.org/index.php/InternationalJournalOfComputer/index

Comparative Analysis of Machine Learning Algorithms for Diabetes Prediction: Finding the Optimal Approach

Aftab UL Nabi[a], Neetesh Kumar[b],Waqas Chander[c], Sunil Kumar[d], Muhammad Waqas Pasha[e], Rajesh Kumar[f]*

[a]*Department of Computer science, South China University of Technology, China*
[b]*Department of Computer Science & Information Technology, TIEST, NED University, Pakistan*
[c]*Department of Electrical Engineering, Mehran University of Engineering and Technology, Pakistan*
[d]*Department of Electronics Engineering, Quaid Awam University of Engineering and Technology, Pakistan*
[e]*Department of Computing, Hamdard University, Pakistan*
[f]*Department of Computer Science, University of Palermo, Italy*
[a]*Email: aftab.shahani@mail.scut.edu.cn,* [b]*Email: neeteshkumar@neduet.edu.pk*
[c]*Email: waqaschander445@gmail.com,* [d]*Email: sunilrathi474@gmail.com*
[e]*Email: waqaspasha90@gmail.com,* [f]*Email: rajesh.kumar@unipa.it*

Abstract

Diabetes, as a chronic disease, poses a rapidly escalating risk to human health, stemming from a complex interplay of factors such as obesity, elevated blood glucose levels, and various other triggers. Central to its onset is the disruption of insulin hormone function, resulting in abnormal metabolism and increased blood sugar levels. In this paper, we propose a solution to this pressing issue using machine learning techniques. By applying various machine learning algorithms on the Pima Indian diabetes (PID) dataset, we aim to identify the most effective algorithm for this task. Leveraging powerful machine learning algorithms such as (SVM) Support Vector Machine, (RF) Random Forest and others, we endeavor to forecast the onset of diabetes. Through the amalgamation of these techniques, our objective is to proactively identify individuals at risk, enabling timely intervention and preventive measures to safeguard health. The primary goal of this initiative is to mitigate the risk of diabetes onset by forecasting individuals' susceptibility and advocating for lifestyle and dietary adjustments. This study has dual objectives: firstly, to develop and implement a predictive model for diabetes using machine learning techniques, and secondly, to explore effective strategies for achieving success in this endeavor.

Keywords: Machine learning; Classification; Prediction; Support vector machine.

Received: 4/25/2024
Accepted: 6/10/2024
Published: 6/21/2024

---

* Corresponding author.

1. Introduction

The World Health Organization (WHO) stipulated that approximately 1.6 million individuals succumb to diabetes annually [1]. Diabetes manifests when the blood sugar, or glucose, levels in the body become excessively high. Health experts attribute diabetes to two main causes: inadequate insulin production by the pancreas (Type 1 diabetes) and ineffective utilization of the generated insulin by the body's cells (Type 2 diabetes) [2]. According to data from the Centers for Prevention and Control of Diseases (CDCP), type 2 diabetes saw a twenty-three percent surge in the United States from 2001 to 2009. Organizations, government agencies, and medical organizations worldwide are intensifying efforts toward the control and prevention of chronic diseases to avert premature fatalities. Diabetes is predominantly classified into two types: type I and type II. Type I diabetes, also referred to as Insulin-Dependent Diabetes Mellitus, arises when the body fails to produce sufficient insulin, accounting for 10% of all diabetes cases [3]. Type II diabetes, on the other hand, is distinguished by relative insulin deficiency due to pancreatic β-cell dysfunction and elevated levels of insulin in target organs [4].

According to the statistics released by the Canadian Insulin resistance Association, the prevalence of diabetes in Canada is projected to escalate from two million and fifty thousand to a staggering 3.7 million individuals between 2010 and 2020 [5]. This underscores the imperative for early detection and prevention measures to mitigate the risk of premature mortality associated with diabetes. Consequently, the adoption of machine learning techniques has emerged as a necessity. Machine learning algorithms offer diverse capabilities in classification and prediction [2]. This study undertakes a comparative analysis of seven distinct machine learning algorithms across various categories: probabilistic algorithms, such as Naïve Bayes; vector-based algorithms, including both non-linear and linear kernel Support Vector Machines; decision algorithms, represented by Decision Trees and Random Forests; mathematical algorithms, exemplified by adaptive boosting; predictive algorithms like K-nearest neighbor; and regression algorithms, with Logistic Regression. The primary objective is to discern the most effective algorithmic approach for addressing this specific problem. In the past few years, data mining and machine learning have become indispensable tools in the medical domain. Data mining techniques are employed for preprocessing and feature selection from healthcare datasets, while machine learning algorithms automate diabetes prediction [6]. These methodologies facilitate the extraction of hidden patterns from extensive datasets, enabling accurate decision-making. Data mining encompasses a range of techniques, including machine learning, statistics, and database systems, to unveil patterns within large datasets [7]. Nvidia describes machine learning as a process that employs various algorithms to learn from parsed data and generate predictions [8].
The paper has been organized as follows: Section two provides a comprehensive review of the existing literature. Section three elucidates the methodology employed in this study. The results and ensuing discussion are presented in section four. Finally, the paper concludes with a summary in section five.

2. Related Work

Numerous researchers have applied machine learning (ML) techniques to forecast diabetes using the Pima Indian Diabetes Dataset (PIDD), which consists of nine characteristics and 768 records describing female

patients. Given the possibility of missing values within the dataset, pre-processing methods such as imputation, where missing values are replaced with the mean of existing values, are commonly employed. Normalization and other techniques for pre-processing are also utilized to enhance model efficiency. Additionally, Principal Component Analysis (PCA) is employed to reduce feature dimensionalities [9]. Various other pre-processing techniques, including normalization, scaling, and their combinations, are extensively investigated. Data pre-processing plays a pivotal role in data mining and analysis by facilitating the elimination of missing values and the transformation of continuous data into finite values, thereby improving efficiency in diabetes dataset analysis [10]. The dataset utilized in [11] is sourced from CPCCSSN. A variety of classification techniques are used, with the goal of maximizing the area under the Receiver Operating Characteristics (ROC) curve by the use of appropriate hyperparameters. These algorithms include the Gradient Boost Method (GBM), Logistic Regression (LR), Random Forest (RF), and Rpart. To analyze patterns in forecasting unknown datasets, a 10-fold cross-validation is performed, with Random Forest showing to be particularly significant. In addition, a Support Vector Machine (SVM) classifier is used as a multi-parameter detector to evaluate healthcare variables including blood pressure and heart rate [12]. Classifier has been modified for use with Software Defined Radio and assists in health monitoring [13]. By identifying the most significant traits, SVM is also used to forecast the severity of leukemia cancer [14]. Additionally, a Support Vector Machine (SVM)-based classifier is used to evaluate healthcare. To increase model performance [15], this study elaborates on the Random Forest Classifier algorithm and feature selection strategies, leading to enhanced detection capabilities.

3. Methodology

This section of the paper provides a comprehensive discussion of the methodology, covering all steps in detail, including data collection and the selection of machine learning algorithms for our comparative analysis.

3.1. Data Collection

The collection of data used in this study, the Pima Indian diabetes (PID) dataset, was sourced from Kaggle. It consists entirely of female patients from the community near Phoenix, Arizona, in the USA. The main outcome under investigation was the presence of diabetes, with a total of 2 individuals testing positive and 500 people testing negative. Consequently, there is one target (dependent) variable and eight attributes, as outlined by Tynecki (2018): pregnancy, Oral Glucose The tolerance Test (OGTT), blood pressure, skin thickness, insulin levels, Body Mass Index (BMI), age, and pedigree diabetes function.

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1

Figure 1: Figure1 representation of the dataset

3.2. Data Preprocessing

P In order to optimize data for the creation of reliable machine learning models, which eventually improve accuracy, preprocessing is essential. To improve the quality of the data, this preprocessing entails a number of crucial tasks, such as feature selection, data normalization, outlier rejection, and filling in missing values. There are 268 samples identified as diabetic and 500 samples classified as non-diabetic in the dataset under review.

3.3. Detection of Missing Values

Utilizing both Excel and the Weka tool, we identified missing values within the datasets, as outlined in Table 1. To address this issue, we replaced the missing values with their respective mean values.

Table 1: The number of missing values in PIMA dataset

Attributes	No. of missing values
Preg	0
Glucose	5
BP	35
SkinThickness	227
Insulin	374
BMI	11
DPF	0
Age	0

3.4. Feature Selection

Selecting features is a crucial data preliminary processing step employed in the diabetic complications' dataset. It involves selecting a specific group of features from the entire dataset based on certain numerical scores and eliminating duplication that do not contribute significantly to model performance. The primary objective is to identify the most important features within the dataset. Features with values that are absent are eliminated, and the remaining features are evaluated to determine the number of missing data points. A basic and straightforward technique is employed to measure the distinction between attributes with real numbers, resulting in scores. Features with comparatively high scores are important features for further analysis.

. **Table 2:** Feature Selection Scores

Features	**Score**
Insulin	6948.264513
Glucose concentration	4612.377478
Age	660.4345901
BMI	425.907622
Number of pregnancies	388.169915
Skin thickness	166.858687
Diastolic blood pressure	75.564056
Diabetic pedigree function	16.337401

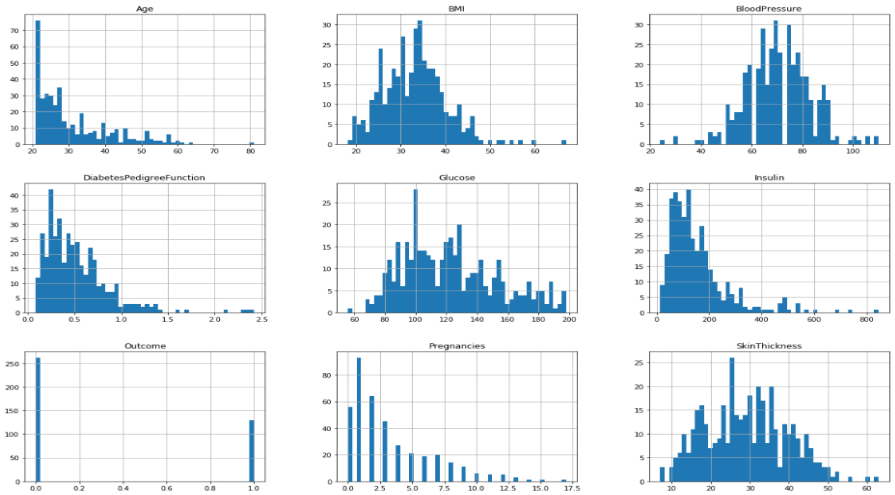

Figure 2: representation of Attribute distribution

3.5 *Metrices*

In In this study, various metrics, including F1-score, recall, precision, and precision, will be used to evaluate machine learning models. Accuracy (A) evaluates the proportion of correct classifications made by a classifier across the entire test set. It is determined as the ratio of true positives (TP) plus true negatives (TN) to the total number of samples (TP + TN + FP + FN), in which TP reflects true positives, TN represents true negatives, FP represents false positives, and FN represents fake negatives. Accuracy is a metric used to measure the overall correctness of a classifier's predictions. It represents the proportion of the number of right predictions (both true positives and true negatives) to the total number of predictions gave by the classifier. The accuracy formula is:

:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \qquad (1)$$

Recall, also known as Sensitivity or True Positive Rate, measures the ability of a classifier to correctly identify positive instances from all actual positive instances. It represents the ratio of true positive predictions to the total number of actual positive instances. The formula for recall is:

$$Recall = \frac{TP}{TP + FN} \qquad (2)$$

Precision measures the proportion of correctly identified positive instances among all instances predicted as positive by the classifier. It represents the ratio of true positive predictions to the total number of positive predictions made by the classifier. The formula for precision is:

$$Precison = \frac{TP}{TP + FP} \qquad (3)$$

The goal of the current study is to analyze four different types of machine learning algorithms: statistical, vector-based, probabilistic, and decision. In order to do this, comparisons have been made using Naive Bayes, Support Vector Machines with both linear and non-linear kernels, Decision Trees, Random Forests, k-near neighbor, Logistic Regression, and adaptive boost classifier.

a) Support Vector Machine (SVM):

SVM, a supervised learning algorithm, relies on a training set with corresponding labels. Once trained, SVM can classify test data into one of two categories. It excels in linear classification and can also handle nonlinear classification by employing kernel techniques to transform inputs into a higher-dimensional feature space. This enables SVM to perform nonlinear classification effectively. The algorithm generates a categorization hyperplane, selected to minimize the distance between adjacent data points on each side [].

b) K-Nearest Neighbors (KNN) Classifier:

The KNN method operates under the assumption that new data points are similar to those in the existing dataset and assigns them to the category most similar to the ones already present. This allows KNN to classify new data quickly and efficiently into appropriate categories. While KNN can be used for both classification and regression tasks, its application in classification is more widespread and popular.

c) Random Forest:

Ensemble learning, employed in both classification and regression tasks, entails training numerous decision trees and then outputting either the mode of the classes or the regression of individual trees. As its name suggests, Random Forest is a classification method that comprises multiple decision trees built on different subsets of the provided dataset. It aggregates the results to enhance the overall accuracy of the dataset.

d) Decision Tree:

A versatile supervised learning technique, the Decision Tree is applicable to both classification and regression tasks, though it's primarily utilized for classification. It operates as a tree-structured classifier, comprising leaf nodes that indicate outcomes, core nodes storing dataset attributes, and branches representing decision rules. Essentially, it provides a graphical representation of potential answers or choices based on specific criteria. Named for its tree-like structure, the decision tree begins with a root node and branches out, mirroring the growth of a tree.

e) Naive Bayes:

Naïve Bayes (NB) is a machine learning algorithm used for classification tasks, including diabetes prediction. It operates based on Bayes' theorem, assuming that features are conditionally independent given the class label. Despite its simplicity, Naïve Bayes can be effective for diabetes prediction, especially with large datasets. It is computationally efficient and requires minimal training data, making it a suitable choice for healthcare applications like diabetes prediction.

4. Results And Discussion

We will showcase all of the implemented classifiers' findings in this section. The classifiers' hyperparameters were adjusted with the use of cross-validation techniques. The previously stated algorithms were trained with a training-to-testing ratio of 80% to 20% on the dataset. The outcomes for every method in Table 3 are shown below.

Table 3: Performance evaluation of various classifier models

Classifier	Precision	Recall	F1 Score	Accuracy (10-fold)
KNN	0.76	0.73	0.75	0.76
SVM	0.73	0.74	0.73	0.75
NB	0.74	0.74	0.74	0.74
DT	0.72	0.71	0.71	0.71
RF	0.70	0.71	0.71	0.71

As depicted in Figure 4 below, KNN classifiers achieved the highest accuracy of 76% after conducting 10-fold cross-validation, outperforming other classifiers such as Decision Trees, Naive Bayes, Support Vector Machines, and Random Forests, which also demonstrated accuracies above 70%.

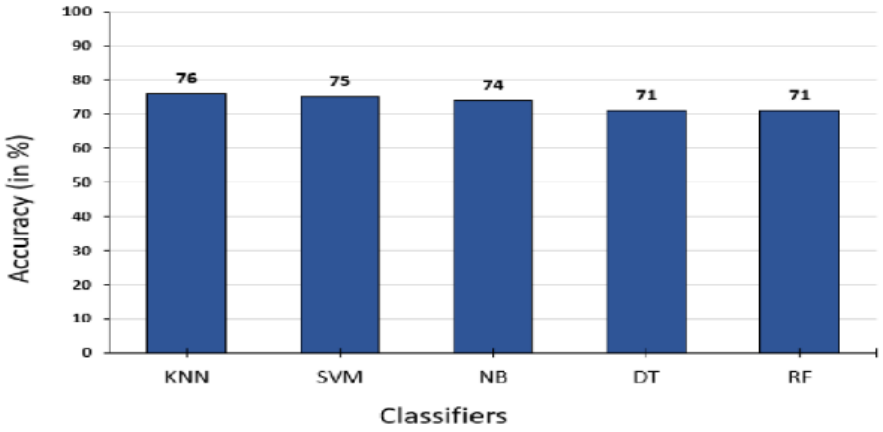

Figure 4: representation of Attribute distribution

5. Conclusion

Early detection of diabetes is one of the most difficult tasks. To create a model in this system that can forecast diseases like diabetes, several experimental techniques are used. Compared to the other algorithms, we obtained the highest accuracy in KNN, with a 76 percent rate. Before the model is trained, the dataset passes through a few crucial pre-processing stages. To identify the optimal attribute, feature selection was done prior to training. An attribute is used in the training technique based on the score obtained. The algorithm that yielded the highest accuracy after training with the three different algorithms will be the most appropriate for the diabetes predictions. It can be used to forecast additional diseases in future research using other categorization schemes or regression algorithms. We may enhance it by predicting additional diseases using essential features as an input, as it is limited to diabetes.

Acknowledgements

I am grateful to all those who have contributed to this research project, providing invaluable support and guidance.

References

[1] World Health Organization. (n.d.). Diabetes. Retrieved from https://www.who.int/health-topics/diabetes.

[2] Medical News Today. (n.d.). How is the pancreas linked with diabetes? Retrieved from https://www.medicalnewstoday.com/articles/325018#how-is-the-pancreas-linked-with-diabetes.

[3] Dunkler, D. (2015). Risk Prediction for Early CKD in Type 2 Diabetes. Clinical Journal of the American Society of Nephrology, 10(8).

[4] Chatterjee, S., Khunti, K., & Davies, M. J. (2017). Type 2 diabetes. The Lancet, 389(10085), 2239–2251.

[5] Alehegn, M., & Joshi, R. (2017). Analysis and prediction of diabetes diseases using machine learning algorithm: Ensemble approach. International Research Journal of Engineering and Technology (IRJET), 04(10).

[6] Swapna, G., Vinayakumar, R., & Soman, K. P. (2018). Diabetes detection using deep learning algorithms. ICT Express, 4(4), 243–246.

[7] Craven, M. W., & Shavlik, J. W. (1997). Using neural networks for data mining. Future Generation Computer Systems, 13(2–3), 211–229.

[8] NVIDIA. (n.d.). What's the Difference Between Artificial Intelligence, Machine Learning, and Deep Learning? Retrieved from https://blogs.nvidia.com/blog/2016/07/29/whats-difference-artificial-intelligence-machine-learning-deep-learning-ai/.

[9] Wei, S., Zhao, X., & Miao, C. (2018). A comprehensive exploration to the machine learning techniques for diabetes identification. In IEEE 4th World Forum on Internet of Things (WF-IoT).

[10] Giri, B., Ghosh, N. S., Majumdar, R., & Ghosh, A. (2020). Predicting Diabetes Implementing Hybrid Approach. In 8th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO).

[11] Lai, H., Huang, H., Keshavjee, K., et al. (2019). Predictive models for diabetes mellitus using machine learning techniques. BMC Endocrine Disorders, 19(101).

[12] Bose, S. S., & Kumar, C. S. (2019). Combining Multiple Features for Improving the Performance of

Multiparameter Patient Monitor. In 5th International Conference on Advanced Computing & Communication Systems (ICACCS).

[13] Abinav, Anil kumar, Naveena Karthika, Pratibha, Ronsen, Gandhiraj R., & Soman K.P. (2010). SVM based Classification of Digitally Modulated Signals for Software Defined Radio. In International Conference on Embedded Systems 2010.

[14] Kavitha K. R., Gopinath, A., & Gopi, M. (2017). Applying Improved SVM Classifier for Leukemia Cancer Classification Using FCBF. In 2017 International Conference on Advances in Computing, Communications, and Informatics (ICACCI).

[15] Thambi, S. V., Sreekumar, K. T., Kumar, C. S., & Raj, P. C. R. (2014). Random forest algorithm for improving the performance of speech/non-speech detection. In First International Conference on Computational Systems and Communications (ICCSC).

[16] Rudra, S., Uddin, M., & Alam, M. M. (2019). Forecasting of Breast Cancer and Diabetes Using Ensemble Learning. International Journal Of Computer Communication And Informatics, 1(1), 1-5.