International Journal of Engineering Sciences & Emerging Technologies, Dec. 2024
ISSN: 22316604 Volume 11, Issue 4, pp: 1078-1084 ©IJESET

EVALUATING MACHINE LEARNING MODELS FOR EFFECTIVE DIABETES DETECTION

Raj Kashyap, Saurabh Srivastava

Department of Computer Science & Engineering
Moradabad Institute of Technology, Moradabad, India

[1]raj.kashyap.tech@outlook.com
[2]srbh.spn@gmail.com

ABSTRACT

Diabetes is immedicable disease where the body either doesn't make enough insulin or can't use the insulin it produces properly. Diabetes is a big health issue worldwide, affecting many people because of things like getting older, genetic, being overweight, and not living a good and healthy lifestyle. According to WHO in the year between 2000-2019 rate of the diabetes is increased by 3%. Diabetes is major cause of heart disease, blindness, and kidney problems. Hospital use historic data to figure out and treat diabetes, but sometimes they need better ways to do it.

Machine learning is a rapidly developing area within data science, focusing on how machines can learn from past experiences. The objective of this study is to be creating a machine learning model that can predict diabetes earlier for patients by comparing various machine learning techniques. Some algorithms include Support Vector Machine, K Nearest Neighbor, and Random Forest are employed. We check how good each algorithm is at predicting diabetes, and then we pick the one that's the best at it to use for making predictions.

KEYWORDS: *Diabetes Detection, Machine Learning, Health Care*

1. INTRODUCTION

One of the most urgent global public health issues is diabetes mellitus, a chronic metabolic disease marked by hyperglycemia. There are serious physical, financial, and social ramifications associated with the startling rise in diabetes prevalence. An outline of the current state of diabetes, including its epidemiology, risk factors, complications, and management techniques, is intended to be provided by this study. Hyperglycemia is a chronic illness that is indicative of diabetes mellitus. Prompt diagnosis and treatment are essential to avoid problems. Compared to conventional methods, machine learning (ML) techniques offer improved accuracy and efficiency in diabetes prediction, making them formidable instruments. This paper highlights different algorithms, datasets, and performance indicators to provide an overview of recent developments in machine learning for diabetes prediction.

There has been a big rise in the number of people with diabetes over the last few decades. Urbanization, changes in lifestyle, and aging populations are the reasons behind the increasing prevalence [1]. The prevalence of diabetes varies greatly by location. The Western Pacific and Middle East and North Africa (MENA) regions have the greatest prevalence rates. For example, the highest rates are found in Kuwait and Saudi Arabia, where approximately 20% of the adult population is affected [2]. Sub-Saharan Africa, on the other hand, has the lowest prevalence, despite a sharp rise in rates [3].
The development of diabetes is significantly influenced by genetic predisposition. Diabetes in the family is a major risk factor, suggesting a substantial hereditary component. Studies have linked a higher chance of getting type 1 and type 2 diabetes to a number of genetic sites [4]. Obesity, physical inactivity, and poor food are examples of lifestyle variables that significantly contribute to the rise in type 2

diabetes. The risk is greatly increased by the modern diet, which is marked by a high calorie intake, an excessive intake of processed foods, and sugary beverages. Sedentary habits, which are common in urban areas, also make the issue worse [5]. Diabetes incidence has been related to environmental factors, such as exposure to air pollution and endocrine-disrupting chemicals (EDCs). Air pollution has been linked to systemic inflammation and insulin resistance, and EDCs can disrupt metabolic processes [6]. Diabetes dramatically raises the risk of peripheral arterial disease, coronary artery disease, and stroke, among other cardiovascular illnesses (CVD). Three of the main risk factors for CVD—hypertension, dyslipidemia, and atherosclerosis—are exacerbated by hyperglycemia [7]. One of the main causes of end-stage renal disease is diabetic nephropathy (ESRD). Proteinuria, a decreasing glomerular filtration rate (GFR), and hypertension are its defining characteristics. The prevention and slow progression of diabetic nephropathy are largely dependent on effective blood pressure and glucose control [8]. A major microvascular consequence of diabetes that can result in blindness and visual impairment is diabetic retinopathy. It happens when long-term hyperglycemia damages the blood vessels in the retina. Preventing eyesight loss requires quick action and routine screening [9].

The foundation of diabetic care is lifestyle adjustment. This include sticking to a healthy weight, exercising frequently, eating a balanced diet, and abstaining from tobacco usage. These actions lower the risk of problems and effectively control blood glucose levels [10]. Pharmacotherapy is the process of controlling blood glucose levels with medicines. Insulin therapy is necessary for the treatment of type 1 diabetes; non-insulin injectables, oral hypoglycemic medications, or insulin therapy can be used to treat type 2 diabetes. The state of the patient and how they respond to treatment determine which medication is best for them [11]. Technological advancements, such as continuous glucose monitoring (CGM) systems and insulin pumps, have revolutionized diabetes management. These innovations improve glycemic control and improve the quality of life for people with diabetes by automating insulin delivery and providing real-time glucose data [12].

Machine learning is a fast-growing field that focuses on how machines may learn from their prior experiences [13][14]. By comparing different machine learning techniques, the goal of this study is to create a machine learning model that can detect diabetes sooner for patients. A few of the algorithms used are Random Forest, K Nearest Neighbor, and Support Vector Machine. We compare the accuracy of each algorithm in predicting diabetes and select the most accurate algorithm to be used in future study.

2. METHODOLOGY

In this section, we'll explore the methods used in machine learning to predict diabetes and discuss our approach to improving accuracy. We employed three primary methods: Support Vector Machine (SVM), K Nearest Neighbor (KNN), and Random Forest. We'll present the results, with a focus on accuracy, to demonstrate the effectiveness of these methods in predicting diabetes. The work flow for predicting diabetes is shown in Figure 1.

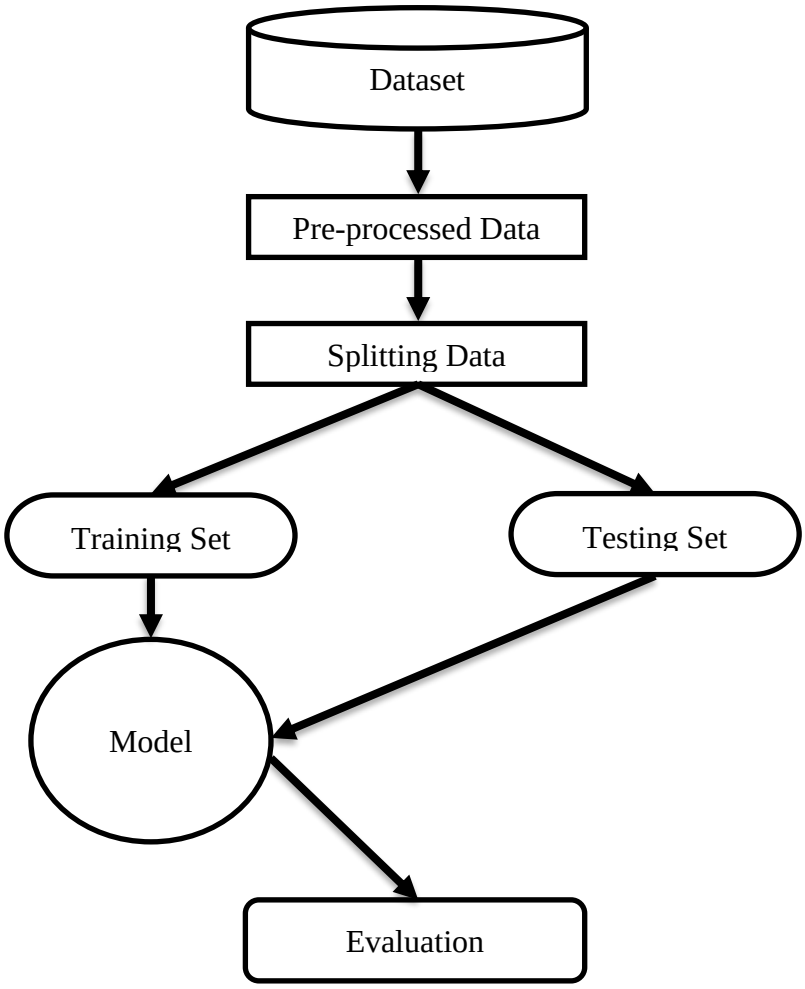

Figure 1. Proposed Model for Diabetes Detection

2.1 Dataset Description

I have acquired a dataset from Kaggle containing essential details such as pregnancies, glucose levels, blood pressure, skin thickness, insulin levels, BMI, diabetes pedigree functions, ages, and outcomes. This study aims to predict whether individuals have diabetes using this dataset. It is a binary dataset, indicating the presence or absence of diabetes with values of 0 and 1 in the outcome (target) attribute, where 0 signifies non-diabetic individuals and 1 represents diabetic individuals. The sample dataset shown in Figure 2.

	A	B	C	D	E	F	G	H	I
1	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
2	6	148	72	35	0	33.6	0.627	50	1
3	1	85	66	29	0	26.6	0.351	31	0
4	8	183	64	0	0	23.3	0.672	32	1
5	1	89	66	23	94	28.1	0.167	21	0
6	0	137	40	35	168	43.1	2.288	33	1
7	5	116	74	0	0	25.6	0.201	30	0
8	3	78	50	32	88	31	0.248	26	1
9	10	115	0	0	0	35.3	0.134	29	0
10	2	197	70	45	543	30.5	0.158	53	1
11	8	125	96	0	0	0	0.232	54	1

Figure 2. Dataset Features

The data set consists of 768 instances, with 9 features each.

- The "Outcome" feature denotes the prediction target, with 0 indicating no diabetes and 1 indicating diabetes.

We plotted the input fields against the outcome to see the relationship between them. The data points were spread out with no clear pattern, as shown in Figure 3.1.2. This dataset is crucial for this study as it serves as the foundation for constructing predictive models to identify diabetes among individuals. We are fully prepared to conduct a thorough analysis of the dataset and utilize it to develop predictive models.

Figure 3. Relationship between input and outcome

2.2 Dataset Pre-Processing

This step is crucial because it helps ensure that our predictive model works accurately. We're carefully preparing our dataset by fixing any missing data, strange numbers, or inconsistencies. We're paying special attention to key factors like blood sugar levels, insulin levels, BMI, and family medical history. By doing this, we're aiming to make our predictive model more reliable, so it can give us better results for diagnosing diabetes.

2.3 Data Splitting

After cleaning up our data, it's time to divide it into two parts: training and testing data at a ratio of 0.80 and 0.20 respectively. We'll use the training data to teach the model to make predictions based on the patterns it finds. Then, we'll use the testing data to see how well the model does on new data it hasn't seen before. This helps us make sure the model works well in real-life situations. Splitting the data like this helps us check if our model can predict outcomes accurately.

2.4 Classification

2.4.1. Support Vector Machine (SVM)

For the classification phase, we implement the Support Vector Machine (SVM) technique to predict diabetes within our dataset. SVM, also known as Support Vector Machine, is highly suitable for binary classification tasks like ours. It operates by identifying a hyperplane line or boundary to separate different groups in the data, aiming to maximize the margin of separation between these groups. The obtained result from SVM classifier is mentioned in Table 1.

Table 1. Accuracy of SVM Classifier

Training Accuracy	0.77
Testing Accuracy	0.76
F1 Score	0.66

2.4.2. K-Nearest Neighbors (KNN)

Following SVM, we utilize the K-Nearest Neighbors (KNN) algorithm for diabetes prediction. KNN is a straightforward yet powerful algorithm that examines the 'k' nearest data points to the one being classified and assigns its group based on the majority of those neighbors. The obtained result from KNN classifier is mentioned in Table 2.

Table 2. Accuracy of KNN Classifier

Training Accuracy	0.79
Testing Accuracy	0.77
F1 Score	0.61

2.4.3. Random Forest

In the final stage of classification, we employ the Random Forest algorithm. Random Forest constructs multiple decision trees and combines their outcomes to yield a more accurate prediction. Each tree independently predicts the outcome, and the final prediction is determined based on the consensus of all the trees. The obtained result from Random Forest classifier is mentioned in Table 3.

Table 3. Accuracy of Random Forest Classifier

Training Accuracy	1.00
Testing Accuracy	0.83
F1 Score	0.72

By leveraging these diverse techniques, we aim to determine the most effective approach for predicting diabetes in our dataset, facilitating informed decisions for future prediction tasks.

2.5 Evaluation

Table 4. Accuracy Comparison

Algorithm	Training Accuracy	Testing Accuracy
SVM	77%	76%
K- Nearest Neighbors	79%	77%
Random Forest	99%	83%

All three methods were reviewed and thereafter, their accuracy rates and F1 scores were compared. Support Vector Machine (SVM) achieved 0.7597 in terms of accuracy which means it is right about 75.97% (approx. 76%) times with an F1 score of 0.66. K-Nearest Neighbor(KNN), on the other hand, had a slightly higher accuracy of 0.7727 meaning that it was correct about 77.27 % time but it gave a lower F1 score of 0.60. Random Forest model, on the other hand, had the highest accuracy score of 0.8311 which shows correctness approximately for about 83.11% times, as well as an F1 score, obtained at level 0f .72.Based on both accuracy and F1 scores, therefore, The best-performing method according to both scores would be Random Forest approach in our dataset suggesting an indication into the future predictions for preferred choices.

3. CONCLUSIONS

In summary, diabetes remains a significant global health concern, with its prevalence increasing steadily over time and posing severe risks such as heart disease, blindness, and kidney complications. While hospitals utilize historical data to address diabetes, there is a recognized need for more effective approaches. This study underscores the potential of machine learning in enhancing early diabetes detection and prediction. By employing diverse machine learning techniques such as Support Vector Machine, K Nearest Neighbor, and Random Forest, a predictive model was developed to assist in identifying diabetes at an earlier stage in patients.

Throughout the study, three different machine learning models were examined and assessed. The results showed that our system work well, especially with Random Forest method, which give an impressive rate of accuracy of 83.11%.

REFERENCES

[1] International Diabetes Federation, "IDF Diabetes Atlas, 10th edn.," Brussels, Belgium: 2021. Available: https://diabetesatlas.org.

[2] S. Al-Hussaini et al., "The high prevalence of diabetes mellitus in the Arabian Gulf countries: A systematic review and meta-analysis," Diabetes Research and Clinical Practice, vol. 165, pp. 107937, Dec. 2020.

[3] A. Ogurtsova et al., "IDF Diabetes Atlas: Global estimates for the prevalence of diabetes for 2015 and 2040," Diabetes Research and Clinical Practice, vol. 128, pp. 40-50, Apr. 2017.

[4] S. M. Sladek et al., "A genome-wide association study identifies novel risk loci for type 2 diabetes," Nature, vol. 445, no. 7130, pp. 881-885, Feb. 2007.

[5] F. B. Hu, "Globalization of diabetes: The role of diet, lifestyle, and genes," Diabetes Care, vol. 34, no. 6, pp. 1249-1257, Jun. 2011.

[6] J. A. Lang et al., "Environmental pollution and diabetes: A review of current evidence," Diabetes & Metabolism, vol. 45, no. 1, pp. 15-25, Feb. 2019.

[7] P. M. Nilsson et al., "Diabetes and cardiovascular disease: Pathophysiology and treatment," Current Opinion in Cardiology, vol. 28, no. 4, pp. 430-435, Jul. 2013.

[8] M. E. Cooper et al., "Diabetic nephropathy: Translating mechanisms to therapy," Nature Reviews Drug Discovery, vol. 20, no. 9, pp. 653-666, Sep. 2021.

[9] A. Aiello et al., "Diabetic retinopathy: A complication of diabetes mellitus," Diabetes Care, vol. 24, no. 8, pp. 1406-1407, Aug. 2001.

[10] American Diabetes Association, "Standards of medical care in diabetes—2021," Diabetes Care, vol. 44, no. 1, pp. S1-S232, Jan. 2021.

[11] R. R. Holman et al., "10-year follow-up of intensive glucose control in type 2 diabetes," New England Journal of Medicine, vol. 359, no. 15, pp. 1577-1589, Oct. 2008.

[12] E. B. Levitt, "Technological advancements in diabetes management," Journal of Diabetes Science and Technology, vol. 13, no. 1, pp. 3-9, Jan. 2019.

[13] S. Srivastava and T. Ahmed, “Similarity-Based Neural Network Model for FBIR System ofOptical Satellite Image Quality Assessment,” 2021.

[14] S. Srivastava and T. Ahmed, “Feature-Based Image Retrieval (FBIR) System for Satellite Image Quality Assessment Using Big Data Analytical Technique.” [Online]. Available: www.psychologyandeducation.net