Available online www.jsaer.com

Journal of Scientific and Engineering Research, 2023, 10(10):8-15

Research Article

ISSN: 2394-2630
CODEN(USA): JSERBR

How can machine learning be used to predict diabetes?

Krish Kapoor

GEMS Modern Academy
Email: krishkapoor1329@gmail.com

Abstract Diabetes is a chronic condition that affects millions of people worldwide and is associated with numerous complications such as cardiovascular disease, blindness, and kidney failure. According to the World Health Organization (WHO), diabetes has already affected 422 million people worldwide. Early detection is key in diabetes because early treatment can prevent serious complications. This paper discusses the use of machine learning in predicting diabetes diagnosis in an individual. We use public dataset from the UCI machine learning repository which uses 520 instances collected from the patients of Sylhet Diabetes Hospital in Sylhet, Bangladesh. We analyze the dataset using several machine learning models: Naive Bayes Algorithm, Decision Trees, Logistic Regression, Neural Networks, Random Forest, Stochastic Gradient, and Support Vector Machines. The results are then evaluated using 10-fold cross validation. Finally, we propose the best machine learning algorithm to use for diabetes diagnosis given specified input parameters, and we discuss the possibility of the deployment of a diabetes diagnosis tool.

Keywords Diabetes Detection, Supervised and Unsupervised Machine Learning

1. Introduction

Diabetes mellitus, a chronic metabolic disorder, is one of the fastest growing diseases and the biggest health crises of our time. Generally, there are two types of diabetes: type 1 and type 2. Type 1 diabetes develops when the immune system mistakenly attacks the pancreatic beta cells, causing the body to produce little or no insulin. It is primarily a genetic condition that manifests itself early in life. Type 2 diabetes, on the other hand, occurs when our bodies either do not produce enough insulin or become insulin resistant. This is primarily a result of one's lifestyle and evolves over time. [8]

According to the World Health Organization (WHO), an estimated 422 million adults worldwide were living with diabetes in 2014, compared to just 108 million in 1980. [9] This represents a nearly four-fold increase in the number of people with diabetes over the past three decades. Diabetes may even exist up to 7 years before clinical diagnosis. [10] Within this period, people can gradually suffer fatal complications such as heart attack, stroke, and eye injuries. However, with early detection and treatment, individuals can manage their blood sugar levels effectively, reducing the risk of these complications. In addition, early detection can also help prevent the development of diabetes in individuals who are at high risk.

According to a report by the International Diabetes Federation (IDF), the global healthcare expenditure on diabetes was estimated to be $760 billion in 2019. This represents 10% of the world's total healthcare expenditure. [11] The report also projected that the global healthcare expenditure on diabetes will increase to $845 billion by 2045 if the current trends continue. [11] Developing countries may also face economic challenges, as the cost of diabetes management can be a significant burden on individuals and healthcare systems. Furthermore, people with diabetes in these countries may face additional challenges such as limited access to healthcare facilities, lack of affordable medications and medical supplies, and inadequate diabetes education and awareness. These challenges can lead to poor diabetes management and increased rates of complications and mortality.

Therefore, early diagnosis plays a pivotal role in the patient outcome. Now, there are existing diabetes tests like continuous glucose monitoring (CGM) system. CGM is a method of continuously monitoring blood sugar levels throughout the day and night by inserting a small sensor under the skin and measuring glucose levels in the interstitial fluid. The sensor transmits data to a receiver or smartphone app, allowing people with diabetes to track their blood sugar levels in real time and make informed diabetes management decisions. [44] While CGM can be a useful tool for managing diabetes, it can be costly. The cost of a CGM system varies depending on the brand and location, but it can range from a few hundred to several thousand dollars per year. In addition to this, some developing countries don’t even provide these tests.

Artificial Intelligence has been rampant in the medical field over the past few years. Companies like IBM, Google, Microsoft, and Amazon have been developing machine learning algorithms for medical image analysis, genomics, and clinical decision support. Their work has already been able to predict and diagnose diseases such as diabetes, and heart disease. AI in medicine is a broad topic and is divided into namely 6 categories: Medical imaging analysis, clinical decision support, drug discovery, health record management, and natural language processing. This paper will mainly be focused on clinical decision support: This involves the use of machine learning algorithms to help clinics diagnose patients based on their medical history, and lab results.

In this modern era of information technology, computers can help us to detect diabetes accurately which has the potential to further save our time and cost. One such example, and is the focus of this paper, is to use machine learning to detect diabetes in an individual. Through training a machine learning model on previously acquired clinical data, we are able to predict, with high accuracy, whether or not an individual has diabetes. In this paper, we analyse a public dataset using different machine learning algorithms (Naïve Bayes, Decision Trees, Neural Networks, Logistic Regression, Stochastic Regression, Support Vector Machines, Random Forest) to find the algorithm that provides the best accuracy. Finally, we propose a tool for end users that uses patients' symptoms and the best algorithm to predict the likelihood of diabetes risk at an early stage.

2. Literature Review

In this section different research papers that were published related to this topic were analysed and provided with their contributions.

[1] **"Machine learning prediction in cardiovascular diseases: a meta-analysis."** This paper reviews studies that used machine learning algorithms to predict cardiovascular disease. The authors found that machine learning algorithms showed promising results in predicting cardiovascular diseases.

[2] **"Predicting the onset of diabetes with Machine Learning Methods"** This study looked at the use of machine learning algorithms to predict the onset of diabetes in prediabetic patients. The authors discovered that advances in machine intelligence can be used to improve understanding of the factors that contribute to the onset of diabetes. The results showed that all models performed well.

[3] **"Prediction of Type 2 Diabetes Based on Machine Learning Algorithm"** This paper examines studies in which machine learning algorithms were used to predict type 2 diabetes in patients. The researchers discovered that machine learning algorithms were highly accurate in predicting type 2 diabetes and could be used to identify patients at high risk for the disease. The model used can provide valuable information on the incidence of T2D to both clinicians and patients ahead of time.

[4] **“Predicting Diabetes Mellitus With Machine Learning Techniques”** The goal of this paper is to predict diabetes using machine learning techniques such as decision trees, random forests, and neural networks. The models were examined using 5 cross validation in the study. When all attributes were used, the results showed that random forest (0.8084) was the most accurate model.

[5] **“Machine Learning Methods to Predict Diabetes Complications”** This paper describes a study conducted as part of the EU-funded MOSAIC project, which used electronic health record data from nearly 1,000 patients to create predictive models of type 2 diabetes mellitus complications. For each complication and time scenario, specialized models were developed, providing an accuracy of up to 0.838 and easy translation to clinical practice.

[6] “A comprehensive review of machine learning techniques on diabetes detection” The purpose of this paper is to review other recent research on the application of machine learning to diabetes prediction and detection. Data inadequacy and model deployment have also been discussed as issues. The review concludes that machine learning has the potential to significantly improve diabetes prediction and detection, and that future research should focus on improving existing models' performance and developing new methods to address the challenges associated with diabetes diagnosis and management.

[7] “A Machine Learning Approach to Predicting Blood Glucose Levels for Diabetes Management” This paper describes an automatic prediction model that uses a physiological model of blood glucose dynamics to generate informative features to warn patients of impending changes in their blood glucose levels. The model outperforms diabetes experts in predicting blood glucose levels and can predict nearly a quarter of hypoglycemic events up to 30 minutes ahead of time.

The above research mainly indicates the accuracy of machine learning models in predicting certain medical conditions like diabetes. However, this paper not only does that but also compares each machine learning model to each other and find out the most accurate model at predicting medical conditions. This paper adds to the growing body of knowledge on the use of machine learning in healthcare and will help improve the industry. Observations and analyses from this research could very well be used in future research in medicine to see faster diagnoses of diseases.

3. Methodology

Proposed System Architecture

The underlying figure depicts the proposed system architecture. The dataset containing patient symptoms will be fed into prediction algorithms such as Naive Bayes, Decision Trees, Logistic Regression, Support Vector Machines, Neural Networks, Stochastic Gradient, and Random Forest. Then the performance of the algorithms will be tested with appropriate evaluation model, in particular, 10-fold Cross-validation. WE will then choose the best algorithm to build the system for the end users using the dataset as Database. The tool will take the symptoms from the user as input and will display classify whether the user has diabetes or not.

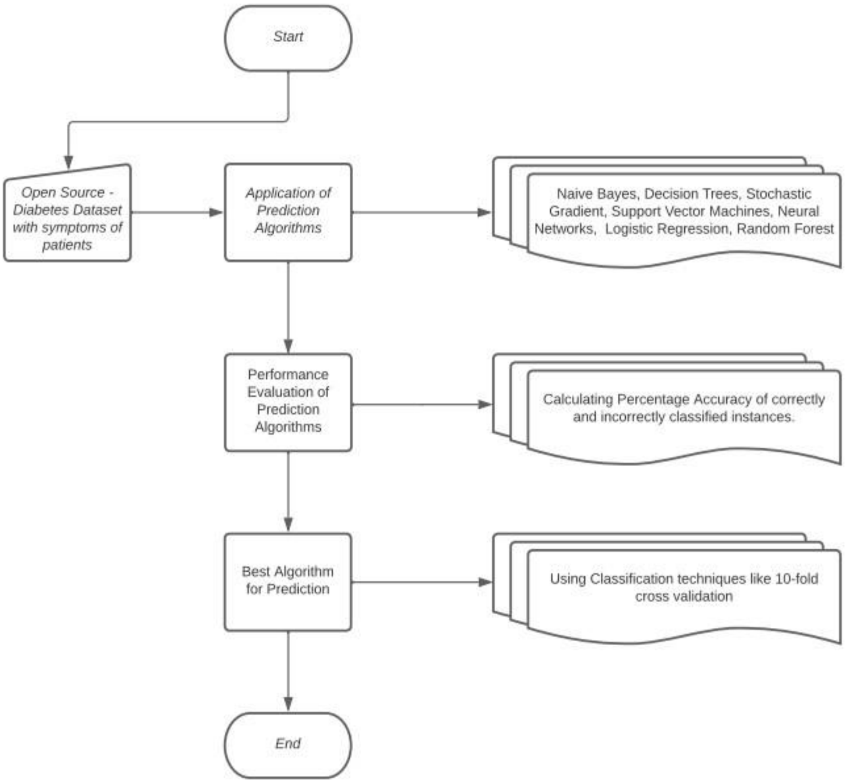
Start
Open Source - Diabetes Dataset with symptoms of patients
Application of Prediction Algorithms
Naive Bayes, Decision Trees, Stochastic Gradient, Support Vector Machines, Neural Networks, Logistic Regression, Random Forest
Performance Evaluation of Prediction Algorithms
Calculating Percentage Accuracy of correctly and incorrectly classified instances.
Best Algorithm for Prediction
Using Classification techniques like 10-fold cross validation
End

Figure 1: Proposed System Architecture

Machine Learning Algorithms

Naïve Bayes - Probabilistic machine learning algorithm that is used to classify data. Using Bayes' theorem, it calculates the probability of an instance belonging to each class. The algorithm makes the "naive" assumption that the features used for classification are independent of one another, which simplifies the calculation of the probability that an instance belongs to a specific class. To apply the algorithm for classification, we train it with a labeled dataset, in which it learns the probabilities of each feature given each class. When a new instance is presented, the algorithm calculates the probability of the instance belonging to each class based on the learned probabilities.

Logistic Regression - Machine learning algorithm that predicts the likelihood of an instance belonging to a specific class based on the input feature values. When a logistic function is fitted to the training data, the input features are converted to a probability score between 0 and 1. To use logistic regression for classification, we train the model on labeled data and use maximum likelihood estimation to learn the logistic function parameters. When presented with a new instance, the algorithm calculates the probability of the instance belonging to the positive class using the learned logistic function and selects the class with the highest probability.

Decision Trees - Machine learning algorithm that recursively partitions data into smaller subsets based on the values of the input features. They are used for classification and regression tasks. During training, the algorithm learns the tree structure and partitioning criteria and can then be used to predict the target variable for new instances by traversing the tree using their input features.

Neural Networks - Type of machine learning algorithm that consists of layers of interconnected neurons that process input data and predict output. During training, the network learns to adjust its weights and biases in order to minimize a loss function that measures the difference between its predicted and true labels. Once trained, the network can be used to predict the target variable for new instances.

Random Forest - Machine learning algorithm that combines multiple decision trees to improve prediction accuracy and robustness. Each tree is built with a random subset of features and training instances, and the final prediction is made by aggregating all of the trees' predictions.

Stochastic Gradient - Stochastic gradient descent is a popular machine learning optimization algorithm for determining the parameters that minimize a model's cost function. It is an iterative algorithm that updates the parameters by taking small steps in the direction of the cost function's negative gradient with respect to the parameters at each iteration, using a randomly selected subset of the training data (i.e., a mini batch).

Support Vector Machines - SVMs are a widely used supervised learning algorithm for classification and regression tasks. SVMs find the best hyperplane in the feature space by maximizing the margin between the closest points from each class. This is accomplished by solving a quadratic optimization problem involving the minimization of a cost function while keeping constraints in mind.

Dataset Details

This dataset contains reports of diabetes-related symptoms of 520 people. It includes data about people including symptoms that may cause diabetes. We have taken this dataset from the UCI machine learning repository.

Table 1: Description of Dataset

	Number of Attributes	**Number of Instances**
Diabetes Symptom Dataset	16	520

Table 2: Description of Attribute

Attributes	**Values**
Sex	1. Male, 0. Female
Polyuria	1. Yes, 0. No
Polydipsia	1. Yes, 0. No
Sudden Weight Loss	1. Yes, 0. No
Weakness	1. Yes, 0. No
Polyphagia	1. Yes, 0. No

Genital Thrush	1. Yes, 0. No
Visual Blurring	1. Yes, 0. No
Itching	1. Yes, 0. No
Irritability	1. Yes, 0. No
Delayed Healing	1. Yes, 0. No
Partial Paresis	1. Yes, 0. No
Muscle Stiffness	1. Yes, 0. No
Alopecia	1. Yes, 0. No
Obesity	1. Yes, 0. No
Class	1. Positive, 0. Negative

4. Results

Table 3: Comparison of Evaluation Metrics using 10-Fold Cross Validation

Evaluation Metrics	**Cross Validation**						
Model	**NB**	**LR**	**DT**	**NN**	**RF**	**SG**	**SVM**
Total Number of Instances	520	520	520	520	520	520	520
Correctly Classified Instances	457 87.9%	480 92.3%	500 96.2%	502 96.5%	504 96.9%	463 89.0%	502 96.5%
Incorrectly Classified Instances	63 12.1%	40 7.7%	20 3.8%	18 3.5%	16 3.1%	57 11.0%	18 3.5%

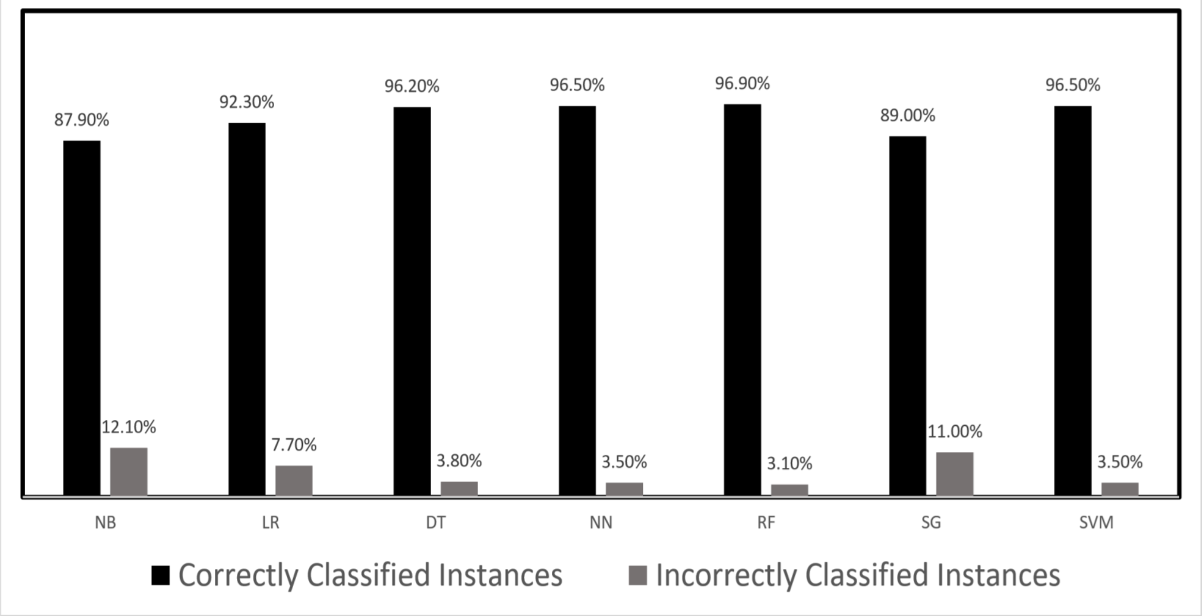

Figure 2: Performance of Classification Algorithms Using Cross-Validation Technique

Table 4: Comparison of Performance Parameters using 10-Fold Cross Validation

Performance Parameters	**Class**	**Weighted Average**						
		NB	**LR**	**DT**	**NN**	**RF**	**SG**	**SVM**
Precision	Positive	0.88	0.91	0.98	0.97	0.99	0.84	0.99
	Negative	0.89	0.95	0.94	0.94	0.91	1.00	0.92
	Weighted Average	0.89	0.92	0.96	0.96	0.96	0.90	0.96
Recall	Positive	0.94	0.97	0.96	0.97	0.96	1.00	0.95
	Negative	0.81	0.86	0.97	0.94	0.97	0.71	0.98
	Weighted Average	0.88	0.92	0.96	0.96	0.96	0.88	0.96
F-measure	Positive	0.91	0.94	0.97	0.97	0.97	0.91	0.97
	Negative	0.85	0.90	0.95	0.94	0.94	0.83	0.95
	Weighted Average	0.88	0.92	0.96	0.96	0.96	0.88	0.96

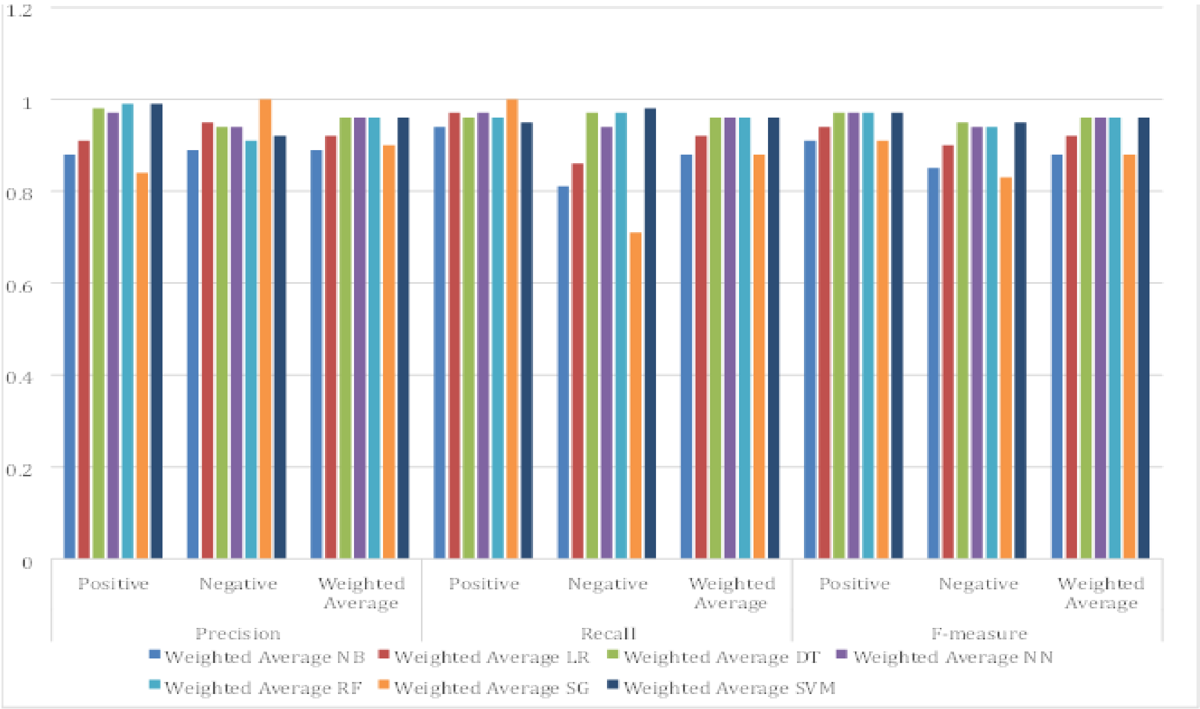

Figure 3: Performance of Parameters using 10-Fold Cross Validation

Table 3 shows us the pure accuracy of each model using 10-fold cross validation. We can clearly see that the Random Forest model classified the greatest number of instances correctly with 504 correct instances out of 520 (96.9% accuracy). This is followed closely by Neural Networks and Support Vector Machines, each classifying 502 instances correctly with a 96.5% accuracy. The least accurate models were Naïve Bayes and Stochastic Gradient, classifying 457 and 463 instances correctly respectively.

Table 4 shows us the precision, recall and f-scores of each model. The models with the highest average precision scores (proportion of positively predicted labels that are actually correct) are Decision Trees, Random Forest, and Support Vector Machines. These three models also have the highest average recall scores (ability to correctly predict the positives out of actual positives). Moving on to f-scores (mean of a system's precision and recall values), the same 3 models appear again but Neural Networks are also included this time.

In statistics, precision, recall, and F-measure are common metrics used to evaluate the performance of a classification model. Precision measures the proportion of true positives (TP) among the instances that are predicted as positive (TP + false positives, FP), and thus reflects the accuracy of the positive predictions. Recall, on the other hand, measures the proportion of true positives among the instances that are actually positive (TP + false negatives, FN), and thus reflects the completeness of the positive predictions.

False positives and false negatives can have different consequences when predicting diabetes with machine learning. A false positive occurs when the model incorrectly predicts that a person has diabetes when they do not. This can result in unnecessary medical treatment as well as increased anxiety and stress for the individual. In contrast, a false negative occurs when the model predicts that a person does not have diabetes when they actually do. This can result in a delayed diagnosis and treatment, increasing the risk of complications and long-term health issues. Therefore, it is important to balance the trade-off between precision and recall depending on the specific context and the costs associated with false positives and false negatives. For example, if the cost of a false negative is much higher than that of a false positive, the model should prioritize recall over precision to minimize the risk of missing true positives.

5. Discussion

Result Analysis

The best result was obtained using the Random Forest Algorithm, with 96.9% of instances correctly classified using 10-fold cross validation. It also had the highest average percentages of precision, recall, and f-measure. The performance of the algorithms using cross-validation evaluation is depicted in the figure above.

Limitations

The main limitation of this machine learning project is the use of an open-source dataset collected from a specific area. Machine learning models trained on such datasets may not perform well when applied to different contexts, potentially leading to inaccurate predictions. The project's outcomes are highly dependent on the quality and representativeness of the dataset, which might not align with the target population of interest. Additionally, the data collection and sanitization processes were performed without our control, introducing potential biases or inconsistencies that can impact the model's performance and reliability. Furthermore, the dataset's small size may limit the model's ability to capture the full complexity and variability of diabetes-related factors, potentially leading to suboptimal predictive accuracy. Finally, relying on open-source algorithms may restrict the model's sophistication and limit its potential performance compared to more advanced, proprietary algorithms.

Proposed Tool

Based on our study's findings, we propose a tool that employs machine learning algorithms, particularly the Random Forest, to predict diabetes. This tool would be user-friendly, allowing individuals to predict their likelihood of developing diabetes by inputting relevant health information. We envision this tool as a quick and easy-to-use solution that leverages the power of machine learning to accurately predict the risk of diabetes. With the rising prevalence of diabetes worldwide, this tool would be an asset to people who want to take control of their health. The tool's intuitive design and easy navigation would ensure that users can easily comprehend the results and take appropriate action.

By utilizing this technology, individuals can diagnose themselves at home and then seek medical advice from a doctor. This approach saves time and resources, allowing doctors to focus on cases that require more urgent attention. Furthermore, in regions where diabetes is a significant health concern, this tool can significantly reduce the load on healthcare systems by enabling people to self-diagnose and manage their condition proactively. This approach not only benefits the individual but also helps to alleviate the burden on healthcare systems, ensuring that people receive timely and appropriate medical attention.

Challenges

Ensuring data privacy is crucial, as sensitive health information is involved. Robust privacy measures, such as anonymization and secure data storage, should be implemented to protect patient privacy and comply with ethical standards. Secondly, determining where the computations run raises considerations about the infrastructure, whether it is a central server, the clinic, or the user's device, and who stores the training dataset. Careful planning and coordination are required to ensure secure data handling and efficient computation. Finally, as self-diagnosing diabetes is not recommended, the design should involve doctor's approval, promoting a collaborative approach that includes healthcare professionals in the decision-making process to maintain ethical standards and ensure accurate diagnosis and appropriate medical guidance.

6. Conclusion

In this paper, we implemented open-source machine learning algorithms on a public dataset of 520 individuals in order to predict whether an individual has diabetes. We evaluated different machine learning algorithms based on their accuracy, precision, recall and f-scores, etc. Our research has shown that machine learning models, particularly Random Forest, can be used to predict diabetes with a high degree of accuracy. This has the potential to improve early diagnosis and prevention of the disease. We have proposed a tool for patients that utilizes this technology to provide personalized risk assessments and recommendations for lifestyle modifications. While there are still limitations to the use of machine learning in healthcare, this study highlights the potential for these models to aid in the management and prevention of chronic diseases. Further research is needed to validate and refine these models, but the results of this study suggest a promising future for the use of machine learning in diabetes prediction and prevention."

Code: **https://github.com/KrishKapoor1329/Machine-Learning**
Machine Learning Algorithms: **https://scikit-learn.org/**
Link: **https://archive.ics.uci.edu/ml/datasets/Early+stage+diabetes+risk+prediction+dataset.**

References

[1]. Chayakrit Krittanawong, Hafeez Ul Hassan Virk, Sripal Bangalore, Zhen Wang, Kipp W. Johnson, Rachel Pinotti, HongJu Zhang, Scott Kaplin, Bharat Narasimhan, Takeshi Kitai, Usman Baber, Jonathan L. Halperin, W. H. Wilson Tang, Chun-Yang Chou, Ding-Yang Hsu, and Chun-Hung Chou. "Machine Learning Prediction in Cardiovascular Diseases: A Meta-Analysis." no. 1, Sept. 2020, https://doi.org/10.1038/s41598-020-72685-1.

[2]. Chun-Yang Chou, Ding-Yang Hsu, Chun-Hung Chou. "Predicting the Onset of Diabetes with Machine Learning Methods." no. 3, Feb. 2023, pp. 406–6, https://doi.org/10.3390/jpm13030406.

[3]. Henock M. Deberneh and Intaek Kim. "Prediction of Type 2 Diabetes Based on Machine Learning Algorithm." no. 6, Mar. 2021, pp. 3317–17, https://doi.org/10.3390/ijerph18063317.

[4]. Quan Zou, Kaiyang Qu, Yamei Luo, Dehui Yin, Ying Ju, Hua Tang. "Predicting Diabetes Mellitus with Machine Learning Techniques." Nov. 2018, https://doi.org/10.3389/fgene.2018.00515.

[5]. Arianna Dagliati, PhD, Simone Marini, PhD, Lucia Sacchi, PhD, Giulia Cogni, MD, Marsida Teliti, MD, Valentina Tibollo, MS, Pasquale De Cata, MD, Luca Chiovato, PhD, and Riccardo Bellazzi, PhD. "Machine Learning Methods to Predict Diabetes Complications." Journal of Diabetes Science and Technology, vol. 12, no. 2, May 2017, pp. 295–302, https://doi.org/10.1177/1932296817706375.

[6]. Toshita Sharma, Manan Shah, Kevin Plis, Razvan Bunescu, Cindy Marling, Jay Shubrook, and Frank Schwartz. "A Comprehensive Review of Machine Learning Techniques on Diabetes Detection." no. 1, Dec. 2021, https://doi.org/10.1186/s42492-021-00097-7.

[7]. Kevin Plis, Razvan Bunescu, Cindy Marling, Jay Shubrook, and Frank Schwartz. "A Machine Learning Approach to Predicting Blood Glucose Levels for Diabetes Management." citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=7e0ad22e8314dbb89695021e505bf3d043ad255a.

[8]. Jill Seladi-Schulman. "What Are the 3 P's of Diabetes?" Healthline, Healthline Media, 19 May 2023, www.healthline.com/health/diabetes/3-ps-of-diabetes.

[9]. World Health Organization (WHO). "Diabetes." 30 October 2018. [Online] Available: https://www.who.int/news-room/fact-sheets/detail/diabetes.

[10]. Hiroyuki Sagesaka, Yuka Sato, Yuki Someya, Yoshifumi Tamura, Masanori Shimodaira, Takahiro Miyakoshi, Kazuko Hirabayashi, Hideo Koike, Koh Yamashita, Hirotaka Watada, Toru Aizawa. "Type 2 Diabetes: When Does It Start?" no. 5, May 2018, pp. 476–84, https://doi.org/10.1210/js.2018-00071.

[11]. "Facts & Figures." idf.org, 2022, idf.org/aboutdiabetes/what-is-diabetes/facts-figures.html.

[12]. Cleveland Clinic. "Continuous Glucose Monitoring (CGM): What Is It & How Does It Work - Cleveland Clinic." 2021, my.clevelandclinic.org/health/drugs/11444-glucose-continuous-glucose-monitoring.