iJRASET
International Journal For Research in
Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 Issue: IV Month of publication: April 2024

DOI: https://doi.org/10.22214/ijraset.2024.61054

Smart Diabetes Detection: A Machine Learning Initiative

Vijay Bochare[1], Saurabh Bhamare[2], Netaji Bhosale[3], Chetan Kharwar[4], Prof. Ishwar Bharambe[5]
Department of Computer Science, Ajeenkya Dy Patil School of Engineering

Abstract: Diabetes mellitus (DM) is a metabolic disease characterized by high blood sugar. The main clinical types are type 1 diabetes and type 2 diabetes. Now, the proportion of young people suffering from type 1 diabetes has increased significantly. Type 1 diabetes is chronic when it occurs in childhood and adolescence, and has a long incubation period. The early symptoms of the onset are not obvious, which may lead to failure to detect in time and delay treatment. Long term high blood sugar can cause chronic damage and dysfunction of various tissues, especially eyes, kidneys, heart, blood vessels and nerves. Therefore, the early prediction of diabetes is particularly important. In this paper, we use supervised machine-learning algorithms like Support Vector Machine (SVM) to train on the actual data of 768 diabetic patients and potential diabetic patients aged 16 to 90. Through comparative analysis of classification and recognition accuracy, the performance of support vector machine is the best.
Keywords: Early Detection, Machine Learning, SVM(Support Vector Machines), Accuracy.

I. INTRODUCTION

Diabetes, also known as diabetes mellitus (DM), is a chronic disorder characterized by high blood glucose levels, due to the inability of the pancreas to generate a sufficient quantity of insulin (Diabetes Mellitus Type-1 (T1DM)) or the failure of cells and tissues to utilize it (Diabetes Mellitus Type-2 (T2DM)) . Apart from T1DM and T2DM, another type is Gestational diabetes, which affects women and develops during pregnancy. Since the prevalence of T2DM in ageing population (i.e., elderly people) is rising ,the analysis in the following sections focuses on such age group which constitutes the participants in Smart-Work .Being one of the main causes of mortality is diabetes mellitus. The current need is for early diabetes detection and diagnosis. An major categorization issue is the diagnosis of the diabetes disease and the interpretation of the diabetes data It is necessary to create a classifier that is accurate, practical, and costeffective. A lot of human ideologies are provided by artificial intelligence and soft computing techniques, which are also used in human-related domains of application. These systems are useful for making diagnoses in medicine. The main topic of this research report is "Diabetes Detection Using Support Vector Machines (SVM)," a sophisticated machine learning technique that has become well-known for its efficiency in resolving challenging classification issues. SVM's capacity to manage high-dimensional data and spot subtle trends within datasets makes it especially well-suited for medical diagnosis applications, such as diabetes detection. However, with the remarkable growth of Machine Learning and advanced information processing techniques, we now have the tools to address this health crisis effectively.

These cutting-edge techniques grant us the ability to predict the onset of polygenic illnesses such as diabetes with a level of precision and efficiency previously deemed unattainable. Additionally, the proactive forecasting of illnesses is the first crucial step toward providing timely intervention, potentiating the potential for a cure. The accuracy of various different approaches employed for the diabetes categorization dataset ranged from 59SVMs have demonstrated impressive performance when using Computer Aided Diagnostic (CAD) systems to enhance diagnostic choices . Vapnik was the first to introduce the Support Vector Machine (SVM), a unique learning device that has lately been used in a number of financial applications, primarily in the field of time series prediction and classification. A medical diagnosis is a classification process. A physician has to analyze lot of factors before diagnosing the diabetes which makes physicians job difficult. In recent times, machine learning and data mining techniques have been considered to design automatic diagnosis system for diabetes. Recently, there are many methods and algorithms used to mine biomedical datasets for hidden information including Neural networks (NNs), Decision Trees (DT), Fuzzy Logic Systems, Naive Bayes, SVM, cauterization, logistic regression and so on. These algorithms decrease the time spent for processing symptoms and producing diagnoses, making them more precise at the same time. There is a great variety of methods related to diagnosis and classification of diabetes disease in the literature. 768 randomly selected data are used for training 78.33 classification accuracy was 71.1used Attribute Weighted Artificial Immune System with 10- fold cross validation method and obtained a classification accuracy of 75.87many other methods used for the classification of diabetes dataset with accuracy between 5976.5area of employing Computer Aided Diagnostic systems (CAD) to improve diagnostic decisions.

The Support Vector Machine (SVM) is a novel learning machine introduced first by Vapnik and has been applied in several financial applications recently, mainly in the area of time series prediction and classification.

A. *Data Collection*

Since classification of diabetes by machine learning approaches are highly relied on the datasets implemented, selecting an appropriate dataset has then become one of the most critical processes in training the model ko. In recent studies, most existing data-driven diabetes detection models are trained using a publicly available diabetes dataset for machine learning and deep learning purposes, named Pima Indians Diabetes Database (PIDD) . Released in 1988, this dataset records nine features of 768 female instances aged at least 21 years old. Recorded features are as follows: Age, Body-Mass-Index (BMI), number of pregnancies, blood pressure, triceps skin fold thickness ,insulin ,Glucose ,Diabetes Pedigree Function value and presence of diabetes in sample's body.

B. *Data Pre-processing*

Despite there are numerous kinds of datasets implemented in solving this task, it is still undeniable that most of them does not meet the quality constraint in training the machine learning and deep learning models, due to reasons as follows: First, it is unavoidable that many datasets contain missing or wrongly data when being constructed. Existence of such data points may affect the models' performance to some extent, and this must be avoided especially when dealing with medical or healthcare related tasks. Second, features recorded in the dataset may not be correlated to the target diabetic outcome, involving these features to train the classification models will not only drag the models' performance, but also increase the computational cost and time required. Other than these, various scales, units and distributions may be different in datasets, and this may cause domination of certain feature in the learning process, leading to incorrect and unfair comparisons between different features. Class balancing issue is also a concerning task in constructing the perfect dataset, as it can prevent biasing of model towards majority class. Therefore, before feeding the dataset to train the model, data preprocessing procedures such as data imputation, feature selection, data normalization and class balancing has to be performed accordingly to solve the stated issues. On top of that, depending on nature of the dataset, encoding of data has to be performed in order to allow the model in processing and understanding the categorical information effectively.

II. DISEASE CLASSIFICATION USING SVM

A. *Diabetes Disease Dataset*

Diabetes Disease Dataset The dataset for the suggested system includes both numerical and characteristic data. All patients can have the prediction made, regardless of their age or gender. The dataset, which has more than two lakh patient records, may be utilized for testing as well as training."0" or "1" are the possible values for the binary target variable. "0" denotes a negative test result for diabetes, while "1" indicates a positive result. Maximum of cases are in class "0" while some of the cases are in class "1". By fine-tuning parameters, the relevance of the automatically selected collection of variables was assessed further by hand. The variables that performed the best in terms of discrimination were those that made the final cut Eight variables—numeric and characteristic—are present are glucose level, and Blood pressure ,body, mass index(BMI), no. of pregnencies, skin thickness, insulin level, dibetes pedigree function and age (years).

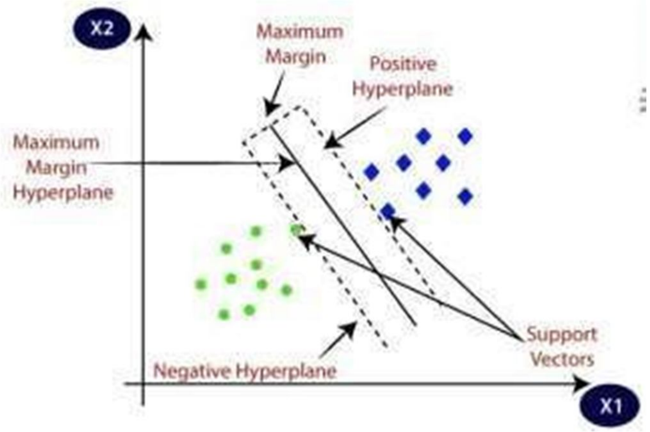


Fig. 1. SVM Algorithm

Even when there are missing values in the data set, these can be handled by other]Processes .One of the characteristic which accepts inputs in the following three formats: never, no information, and current.

B. Training and Test Dataset Evaluation

In the training data set, a 10-fold cross-validation was carried out to assess the SVM models’ resilience. Ten equalsized subgroups are initially created from the training data set. A model trained on all cases and an equal number of non-cases randomly chosen from the remaining nine datasets was trained using each subset as a test data set. Ten iterations of this cross-validation procedure were conducted, with one test data set serving as each subset. Test data sets evaluate the models’ performance.

III. SUPPORT VECTOR MACHINE

SVM Model Generation SVMs are also employed because they can identify intricate associations in your data without requiring you to perform a lot of manual modifications. When working with smaller datasets that contain tens to hundreds of thousands of characteristics, it’s an excellent choice. Because they can handle small, complex information better than other algorithms, they usually find more accurate findings. SVM is a group of related supervised learning techniques used in regression and classification diagnostics in medicine [1,16]. SVM maximizes the geometric margin while also minimizing the empirical classification error. Therefore, SVM stands for Maximum Margin Classifiers. The Structural Risk Minimization Principle, or assured risk boundaries in statistical learning theory, forms the foundation of SVM, a general algorithm. Using an implicit mapping of their inputs into high dimensional feature spaces, known as the kernel technique, support vector machines

(SVMs) may effectively execute nonlinear classification. The classifier may be constructed without explicitly knowing the feature space thanks to the kernel trick .An SVM model is a mapping of the instances as points

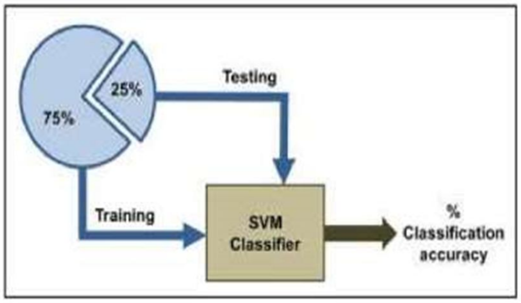
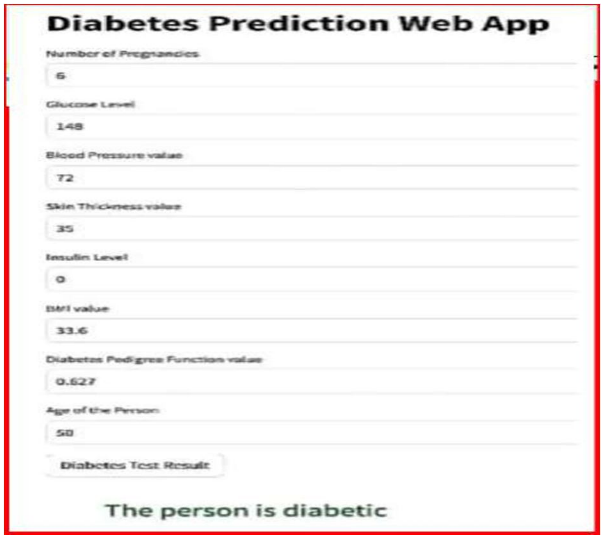


Fig. 2. SVM Classifier



Diabetes Prediction Web App

Number of Pregnancies
6

Glucose Level
148

Blood Pressure value
72

Skin Thickness value
35

Insulin Level
0

BMI value
33.6

Diabetes Pedigree Function value
0.627

Age of the Person
50

The person is diabetic

Fig 4 . output

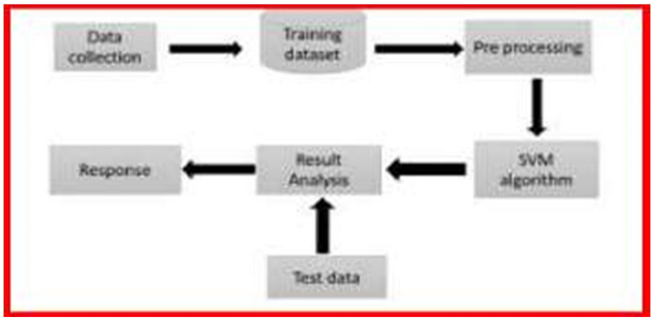


Fig.3.Proposed System Architecture.

space, so that the examples belong to distinct categories and are separated by as large a distance as feasible [1, 8]. An SVM, for instance, locates a hyperplane with the highest proportion of points from the same class on the same plane given a set of points that belong to either of the two classes. The optimal separating hyperplane (OSH) is a separating hyperplane that minimizes the possibility of misclassifying test dataset samples while optimizing the distance between the two parallel hyper planes. The SVM classifier performs classification using an appropriate threshold value after first converting the input vectors into a decision value. The hyperplane is divided (or separated) in order to display the training data. The data points that are closest to the hyperplane are called support vectors, and it is these that establish the margin. When the support vectors are evenly spaced and properly categorized, the margin is maximized relative to the hyperplane. Because SVMs seek to identify a hyperplane that optimizes the separation between classes, they are an extremely strong tool for binary classification. This approach also makes SVMs resistant to outliers and very effective in a wide range of real world scenarios.

IV. PROPOSED SYSTEM

We use a single algorithm in the proposed system, which lowers the time complexity .SVM (Support Vector Machine) is a machine learning technique used to predict diabetes. We are able to take into account patient data regardless of age or gender. The suggested system is an interactive application that asks the user to enter data in order to generate a

prediction. The updated dataset under consideration includes the following attributes: age, Number of Pregnancies, Glucose level, Skin Thickness, Insulin level, BMI, Dibetes pedigree Function value, and outcome. The proposed system takes into account patients who are younger than 21.

V. CHALLENGES

In general, future challenges in diabetes classification using Machine learning techniques can be concluded into three aspects: Availability of large and good quality datasets, me analyzation of features using data driven solution, and ethical issues in implementing this solution to the public. A huge dataset must be used to improve training and attain greater accuracy and performance. Building an accurate diabetes classification model requires a sizable database for model training. However, creating such a database involves a significant investment of time and money, causing that obtaining the perfect dataset comes with great number of issues to overcome. Researchers suggest exploring non-lab-invasive lab measurement databases, but it is crucial to analyze target features and diabetic outcomes before hand to prevent excessive costs associated with building the database. For instance, the availability of current diabetes datasets presents another difficulty because they are frequently not globally representative and may create racial or regional biases during model training. This problem becomes important since datasets are crucial to machine learning models. Additionally, because classification models frequently operate as "black boxes", it is challenging for medical professionals to understand the results adequately. For diabetes classification to advance and patient treatment to improve, these obstacles must be overcome. Medical personnel must interpret the outcome produced by a diabetes classification machine learning-based model for several reasons. First, machine learning learning classification models are a supportive tool aiding treatment decision .While the model’s outcome can provide insights into patients developing diabetes, the people need to interpret the information generated by the model when making treatment decisions, such as prescribing medication or suggesting lifestyle modifications and further diagnostic tests. This solution is less persuasive to the patients without considering the outcome. This is crucial since it is the patients’ constitutional right to comprehend the care being given. Therefore, knowing the model’s results enables medical providers to inform patients about their risk of developing diabetes or their diagnosis effectively. It gives patients the power to decide what is best for their health, including following treatment programs, forming healthy behaviours, and actively managing their disease.

VI. CONCLUSIONS

In a time when technology and healthcare are increasingly combining, the creation of a diabetic risk level prediction system has shown promise for improving patient care and overall health. Diabetes is a worldwide health issue that has farreaching effects. It frequently results in serious complications such as kidney disease, blindness, and heart failure. Patients have a significant time and financial burden because frequent trips to diagnostic centres become essential. The creation of this SVM-based diabetes risk prediction system is a major step in the direction of more efficient and proactive healthcare. One way to lessen the overall healthcare burden associated with complications related to diabetes is to be able to provide patients with timely information and interventions. The project’s success highlights machine learning’s potential in the healthcare industry and the beneficial effects it can have on people’s lives all over the world. The feature subset selection procedure can be used in the future to enhance the SVM classifier’s performance.

REFERENCES

[1] Research on Diabetes Prediction Method Based on Machine Learning.2020.

[2] Diabetes Prediction using Machine Learning Algorithms.2019(Institute of Technology, Chennai, India

[3] Machine Learning Tools for Long-Term Type 2 Diabetes Risk Prediction.2021Corresponding author: Nikos Fazakis (fazakis@ece.upatras.gr).