One-Stop Diabetes Disease Solution Using Machine Learning

Prof. Pallavi Bharambe[1] Prabhat R. Yadav[2] Hiralal S. Sonawane[3] Bhausaheb S. Vagare[4]
[1,2,3,4]Department of Computer Engineering
[1,2,3,4]Shivajirao S. Jondhale College of Engineering, Thane, Maharashtra, India

Abstract— Machine Learning (ML) algorithms can be used in various fields of Knowledge and one of which is medical field. The medical datasets can be used to predict early stage of any disease. Diabetes Disease is one the chronic diseases and is among the leading cause of death in the world. According to International Diabetes federation 382 million people are living with diabetes across the whole world. By 2035, this will be doubled as 592 million. Prediction of diabetes at early stage can lead to improved treatments. However, early prediction of diabetes is quite challenging task for medical practitioners due to complex interdependence on various factors as diabetes affects different human organs. In this paper, analysis and research on one the ML algorithms i.e., Random Forest Classifier (RFC) is done and how we can use the RFC to predict diabetes disease at its early stage.

Keywords: Machine Learning, Random Forest Classifier, Diabetes Disease

I. Introduction

Diabetes Disease is one the deadliest disease in the world. It is not only a disease but also a creator of different kinds of diseases like heart attack, blindness, kidney diseases, etc. In understanding diabetes and how it develops, we need to understand what happens in the body without diabetes. Sugar (Glucose) comes from the foods that we eat, specifically carbohydrate foods. Carbohydrate foods provide our body with its main energy source – everybody, even those people with diabetes, needs carbohydrate. Carbohydrate foods include bread, cereal, pasta, rice, fruit, dairy products and vegetables. When we eat these foods, the body breaks them down into glucose. The glucose moves around the bloodstream. Some of the glucose is taken to our brain to help us to think clearly and function [2].

The normal identifying process is that patients need to visit a diagnostic centre, consult their doctor, and sit tight for a day or more to get their reports. The disease or condition which is continual or whose effects are permanent is a chronic condition. These types of diseases affected quality of life, which is major adverse effect. Diabetes is one of the most acute diseases, and is present worldwide. A major reason of deaths in adults across the globe includes this chronic condition [1].

Diagnosis of diabetes is considered a challenging problem for quantitative research. Some parameters like A1c, fructosamine, white blood cell count, fibrinogen and hematological indices were shown to be ineffective due to some limitations [4]. A blood sample will be taken at random time. Regardless of when you last ate, a blood sugar level of 200 milligrams per decilitre (mg/dL) ---11.1 millimoles per litre (mmol/L) — or higher suggests diabetes [5]. A blood sugar level less than 140 mg/dL (7.8 mmol/L) is normal. A reading of more than 200 mg/dL (11.1 mmol/L) after two hours indicates diabetes. A reading between 140 and 199 mg/dL (7.8 mmol/L and 11.0 mmol/L) indicates prediabetes [5].

A. Types of Diabetes

Particularly there are three types of Diabetes –

1) Type - 1 Diabetes:

- Around 10% of the people with diabetes have type 1 diabetes. It is caused by an autoimmune reaction where the body's defense system attacks the cells that produce insulin.
- As a result, the body produces very little or no insulin. The exact causes of this are not yet known, but are linked to a combination of genetic and environmental conditions.
- It can happen at any age, but usually develops in children or young adults. People with type 1 diabetes need daily injections of insulin to control their blood sugar levels.

2) Type - 2 Diabetes:

- It is the most common type of diabetes, accounting for around 90% of all diabetes cases. It is generally characterized by insulin resistance, where the body does not fully respond to insulin.
- Because insulin cannot work properly, blood glucose levels keep rising, releasing more insulin.
- This type is most commonly diagnosed in older adults, but is increasingly seen in children, adolescents and younger adults due to rising levels of obesity, physical inactivity and poor diet.

3) Gestational Diabetes:

It's basically a type of diabetes which consists of high blood glucose during pregnancy and is associated with complications to both, the mother and the child. It usually disappears after pregnancy but women affected and their children are at a risk of developing Type – 2 diabetes in future.

II. Problem Definition

As of now in the current situation of Home isolation and Pandemic all over the globe, people are not able to move with utmost precautions and lack of confidence. Here, Machine Learning technique can be helpful to develop a model/system that can help in detecting diabetes disease with the help of its corresponding symptoms and show the possibility that a person may be diagnosed with the disease. And this can help as an early warning to a disease to be led attention to and take precautions in such a way that it helps curing the disease. Also medical field is mostly reliant on Machine Learning techniques in some domains of the field and it can also be helpful in verifying the result.

III. Literature Review

Naveen Kishore G. et al presented different algorithms in ML that can be used for prediction purposes and accordingly for good accuracy score as well. They basically used three different algorithms, particularly SVM (Support Vector

Machine), Decision Tree & Random Forest Classifier, it uses different parameters to detect the result and has a better accuracy.[1]

Priyanka Indoria et al presented a set of algorithms that can be used in particular, Levenberg-Marquardt Learning network and Naïve Bayesian algorithm. It balances the system with both, higher accuracy rates and also with smallest accuracy rate.[2]

Tejas N. Joshi et al presented studied algorithms particularly, SVM (Support Vector Machine), Logistic Regression and ANN (Artificial Neural Network) which is suitable for binary classification task and objective being to improve the classification accuracy.[3]

Minyechil Alehegn et al presented a project to detect diabetes disease using algorithms, KNN(K-Nearest-Means) and Niave Bayes and found out that it's useful for storage space which is minimal, and space complexity is also less but the accuracy level of the model is less and it's the only drawback that is seen. Ensemble learning approach is performed on the same dataset.[4]

From the above four papers that we studied and analysed made us come to a conclusion that Random Forest Classifier algorithm can be very helpful in our project that we'll be proceeding with and a better explanation for deciding it is discussed further. It will really support our design and architecture that we are going to make. So, in our project we have used RFC to predict and detect diabetes disease at early stage.

IV. ALGORITHM

A. *Random Forest Classifier*

The random forest method is a flexible, fast, and simple machine learning algorithm which is a combination of tree predictors. Random forest produces satisfactory results most of the time. It is a supervised system getting to know set of rules and difficult to improve on its performance, and it can also handle different types of data including numerical, binary, and nominal. Random forest builds multiple decision trees and aggregates them to achieve more suitable and accurate results. It has been used for both classification and regression. Classification is a major task of machine learning. It has the same hyper parameters as the decision tree or bagging classifier. The fact behind random forest is the overlapping of random trees, and it can be analysed easily. Suppose if seven random trees have provided the information related to some variable, among them four trees agree and the remaining three disagree. On the basis of majority voting, the machine learning model is constructed based on probabilities. This algorithm also solves the overfitting issue. It's also used to remedy classification and regression additionally. In this algorithm it consists of the trees. The number of tree structures present in the data is directly proportional to the accuracy of the result. Random Forests has a variety of applications, such as recommendation engines, image classification, and feature selection.[6]

We compared three different algorithms (SVM, KNN, RFC) that are primarily used for creating prediction models and concluded that RFC can work well considering the fact that it's nothing more than a bunch of decision trees combined and it does binary classification really well. They can handle categorical feature as well. This algorithm can also handle dimensional spaces as well as large number of training examples. It can work out of the box and that is one reason why they are very popular.[7]

B. *How the algorithm does works?*

It works in four steps:

- Select random samples from a given dataset.
- Construct a decision tree for each sample and get a prediction result from each decision tree.
- Perform a vote for each predicted result.
- Select the prediction result with most votes as the final prediction.

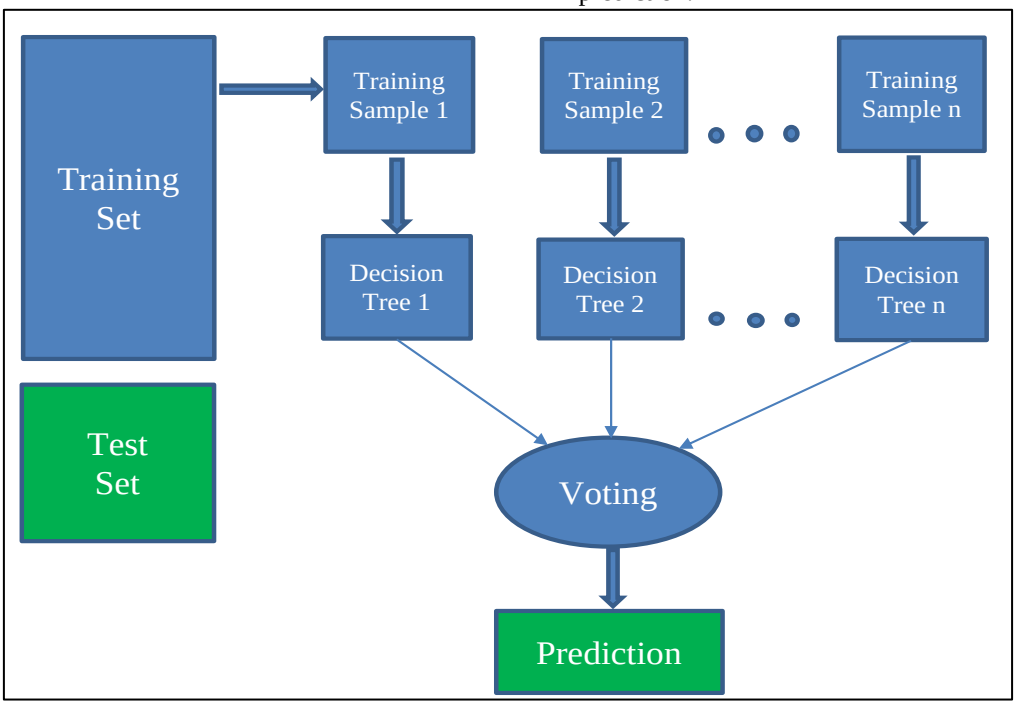

Fig. 1: Algorithm Flow.

V. METHODOLOGY

A. *Proposed System*

1) Stage – I:

– The user who needs to check on the basis of symptoms will provide answers to the questions that system asks to predict if there are chances of user getting diagnosed with the disease.
– Then according to the input, the system predicts and gives a message "You might be suffering with diabetes". If you want further check then proceed with further stage" or "You don't have any symptoms of diabetes, cheers!".
– If user have diabetes like symptoms then he/she may proceed to further stage to get clarity else he/she might exit.

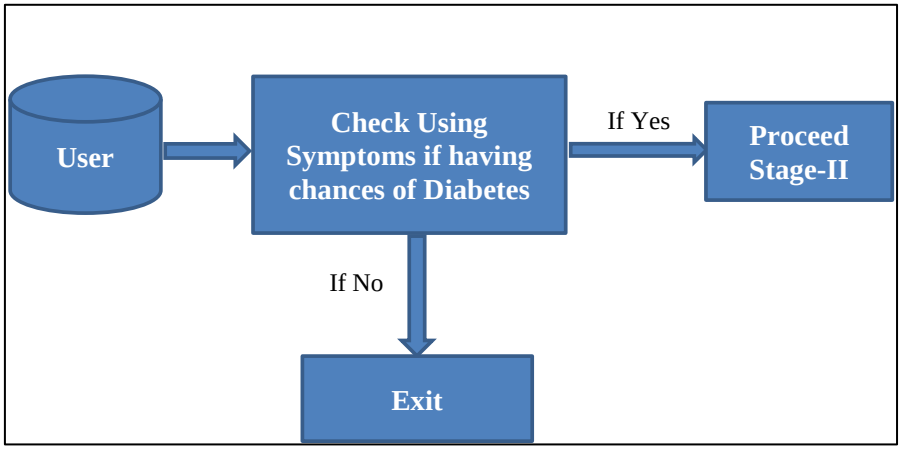

Fig. 2: Stage-I.

2) Stage II:

– In this stage, the user is asked to provide the medical data to detect which type of diabetes he/she might be suffering with.
– (Data such as – glucose, insulin, BMI, age, DFT, pregnancy, blood pressure)
– According to the input provided, the system will predict the type of diabetes and also will provide some precautions and a diet chart to be followed, that can help curing the disease. The model accuracy score is also displayed.
– If user wants any further information on his/her result then he/she might contact the doctors that are available via, "Contact with Doctor" portal by which an appointment will be scheduled with one doctor for solving user queries.

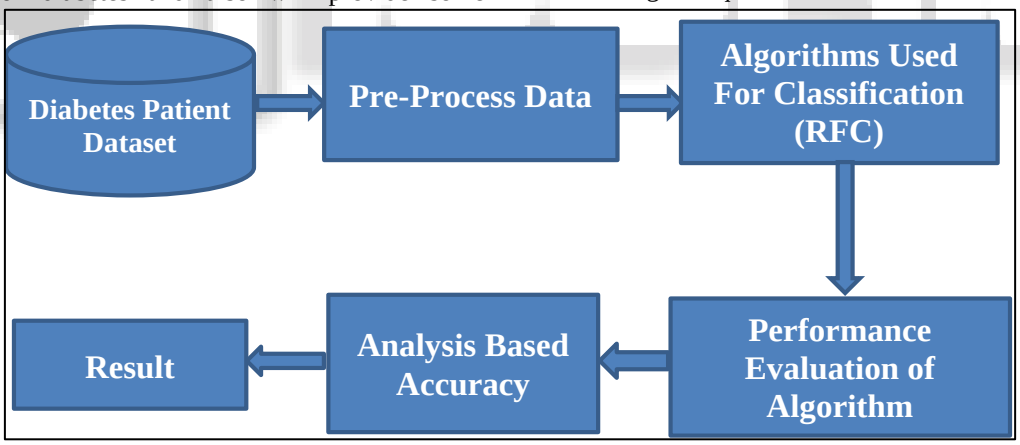

Fig. 3: Stage-II.

B. System Design

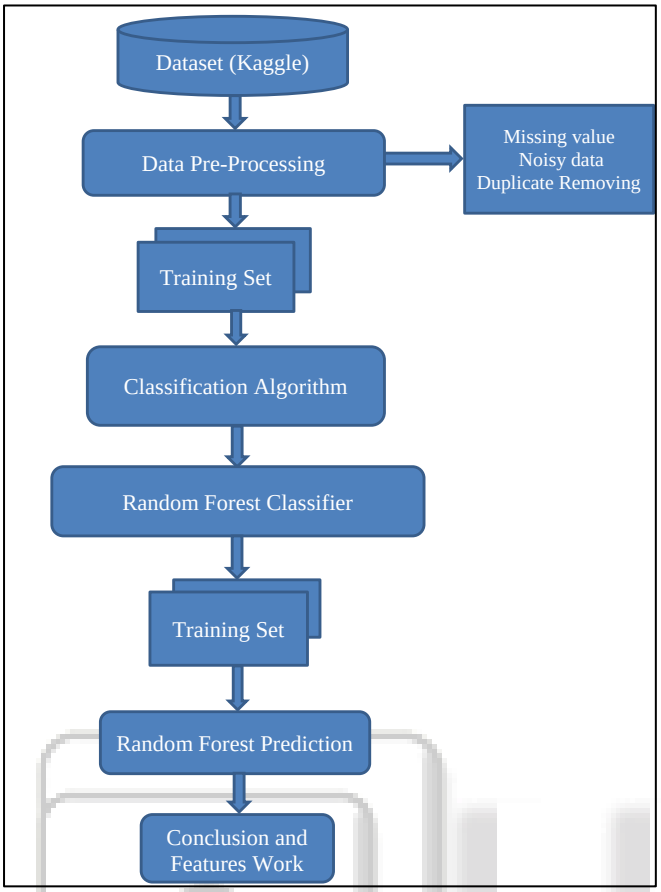

Fig. 4: Architecture Design.

VI. APPLICATION

This work has described a machine learning approach to predicting diabetes levels. The technique may also help researchers to develop an accurate and effective tool that will reach at the table of clinicians to help them make better decision about the disease status.

This system can be helpful basically in two ways:

- As a Pre-Diabetic measure this can be helpful for the general public who want clarity on if the symptoms that they are having are of diabetes or not and can also further check for the type of disease and their respective precautions.
- Also this system can be helpful in the medical field for verification purpose, the medical data that the clinics must be having for detecting diabetes can use that in the system and get to know the exact type of diabetes.
- In future, we can increase the transparency factor of the system by making the "Contact Doctor" portal more user-friendly. We can create a chatting and video calling option available there itself by which the user can clear it's queries at that point of time by the doctor available. Also, an ensemble learning model can be created using different algorithms by the model will be more hybrid and the efficiency will also be good.

VII. CONCLUSION

Machine learning and data mining techniques are valuable in disease diagnosis. The capability to predict diabetes early, assumes a vital role for the patient's appropriate treatment procedure and Machine Learning has the great ability to revolutionize the diabetes risk prediction with the help of advanced computational methods and availability of large amount of epidemiological and genetic diabetes risk dataset. Detection of diabetes in its early stages is the key for treatment. This work has described a machine learning approach to predicting diabetes levels. The Random Forest Algorithm had measured different parameters in the dataset and we have come across a better accuracy rate of 85%. The technique may also help researchers to develop an accurate and effective tool that will reach at the table of clinicians to help them make better decision about the disease status.

ACKNOWLEDGMENT

We sincerely wish to thank our project guide Prof. Pallavi Bharambe for her ever encouraging and inspiring guidance which helped us to make our project success. Our project guide made us endure with her expert guidance, kind advice and time motivation which helps us to determine about our project.

REFERENCES

[1] Naveen Kishore G, V. Rajesh, A. Vamsi Akki Reddy, K. Sumedh, T. Rajesh Sai Reddy, "Prediction of Diabetes using ML Classification Algorithms", International Journal of Scientific & Technology Research, 2020, IISNA no. 2277-8216, pp. 1805-1808

[2] Priyanka Indoria, Yogesh Kumar Rathore, "A Survey : Analysis and Prediction of Diabetes using ML", International Journal of Engineering Research and Technology(IJERT), 2017, IISN no. 2278-0181, pp. 285-291

[3] Tejas N. Joshi, Prof. Pramila M. Chavan, "Diabetes Prediction using ML", International Journal of Engineering Research and Application, 2018, IISN no. 2248-9622, pp. 9-13

[4] Minyechil Alehegn, Rahul Joshi, "Detection and Prediction of Diabetes using ML Techniques ", International Research Journal of Engineering and Technology(IRJET), 2017, IISN no. 2395-0056, pp. 426-436

[5] https://www.idf.org/aboutdiabetes/type-1-diabetes.html

[6] https://www.datacamp.com/community/tutorials/random-forests-classifier-python?utm_source=adwords_ppc&utm_campaignid=1455363063&utm_adgroupid=65083631748&utm_device=c&utm_keyword=&utm_matchtype=b&utm_network=g&utm_adpostion=&utm_creative=278443377086&utm_targetid=aud-299261629574:dsa-429603003980&utm_loc_interest_ms=&utm_loc_physical_ms=9300510&gclid=CjwKCAiA-_L9BRBQEiwA-bm5fhW1wShMZs_WUijz_EdUIyB1CKIhWn9K5_8n-YhP5KLxSJr9EMVfVRoCUskQAvD_BwE

[7] https://discuss.analyticsvidhya.com/t/which-one-to-use-randomforest-vs-svm-vs-knn/2897/3

