IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

A COMPARATIVE STUDY FOR DIABETES PREDICTION USING DIFFERENT MACHINE LEARNING ALGORITHMS

Ansari Huzaif [1], Saket Swarndeep [2]
[1]Student in ME, Computer Engineering, LJ University, Ahmedabad, Gujarat, India
[2] Assistant Professor, LJ University, Ahmedabad, Gujarat, India
[1]ansarihuzaif1507@gmail.com, [2]sanket.swarndeep@ljinstitutes.edu.in

Abstract: *Diabetes develops when your blood sugar, also known as blood glucose, is too high. Blood glucose is your major source of energy, and it is obtained from the foods you consume. Diabetes has contributed to difficulties with heart disease, foot problems, nerve damage, and other ailments throughout time. Millions of people worldwide are affected by this type of illness. Early diabetes detection is critical to ensuring people's health. As a result, in order to foresee diabetes, a system for monitoring people's diabetes is necessary. Understanding the signs of the disease is critical for making predictions about it. We provide a framework for diabetes prediction utilizing a variety of machine learning algorithms in this study. The PIMA Indian Diabetic dataset is also used to evaluate machine learning algorithms such as Artificial Neural Networks (ANN), Decision Trees, Random forests, K-Nearest Neighbors (KNN), Support Vector Machines (SVM), and Logistic Regression. Following that, the findings are examined. Based on the patient's current medical information, we utilize a fused model to predict whether or not the patient has diabetes.*

Keywords: Diabetes, Disease, Worldwide, Machine Learning, PIMA Datasets

INTRODUCTION

Rapid urbanization and modernization have resulted in the emergence of several chronic illnesses, posing a massive danger to global public health. Diabetes mellitus (DM), often known as Diabetes is one such illness. It is currently one of the most frequent illnesses in all age groups and environments. The number of diabetes patients (aged over 18 years) has increased rapidly from 4.7 to 8.5% from 1980 to 2014 which imposes crucial challenges in both developed and developing nations [1]. According to a report by the International Diabetes Federation in 2017 there were 425 million diabetics in the world at the time, and it was also concluded that the number will increase to 625 million by 2045[2]. However, the pace of growth in Africa is anticipated to be 143%, compared to 15% in Europe, with China, India, and the United States of America being the most affected shown in Fig 1 [3].

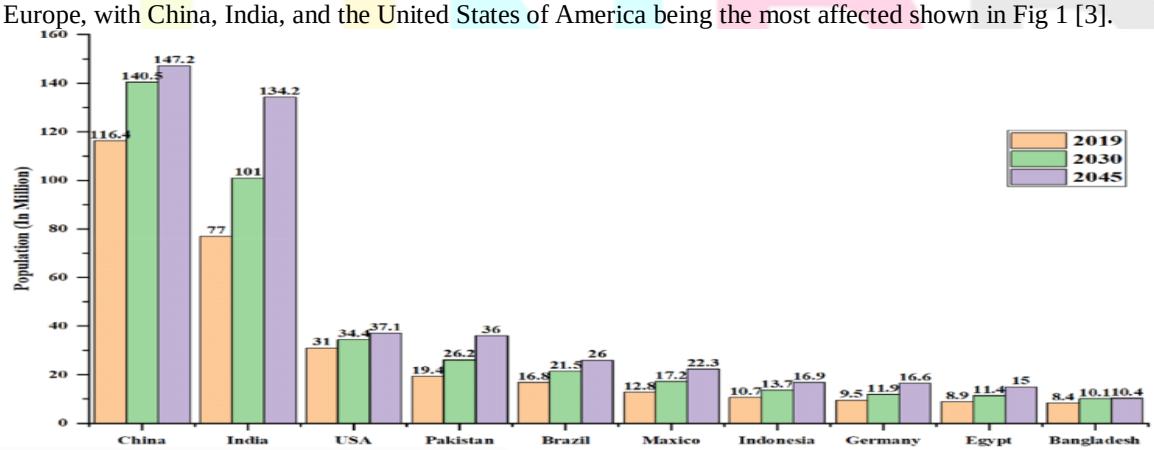

Fig 1. Top 10 most affected countries with diabetes [3]

Diabetes is a chronic condition that has become one of the most common lifestyle disorders, defined by sustained increased blood sugar levels. Diabetes causes organ failure such as liver, heart, kidneys, stomach, and so on in the long term.
Classification [4] of diabetes mellitus can be described as follows:

- Type I diabetes: The diabetic condition that depends on insulin occurs mainly in children and adolescents because of the genetic disorders.
- Type II diabetes: Generally, occurs in adults during the age of 40 years discernible by high blood sugar level.
- Pregnancy diabetes: The diabetes that occur during the pregnancy period.

- Diabetic retinopathy: This type of disorder leads to eye blindness.
- Diabetic neuropathy: Nerve disorder is the cause of this type of diabetes.

Factors Responsible for Diabetes:

- Combination of genetic susceptibility and environmental factor can cause diabetes.
- Overweight may lead to cause diabetes in the long run.
- If a parent or sibling has diabetes, then the risk is supposed to be increased.
- Aging increases risk of diabetes.
- More than 140/90 mm of Hg is linked to an increased risk of diabetes.
- Low levels of high-density lipoprotein (HDL) are also the cause of occurring the risk.

Complications Arise Due to Diabetes: The complications progress moderately. Possible complications those are included for arising:

- Cardiovascular disease: Diabetes vividly increases the risk of various cardiovascular problems;
- Damage in the nerves (Neuropathy);
- Damage in the kidneys (Nephropathy);
- Damage in eyes (Retinopathy);
- Damage in foot: Deficient blood flow to the feet increases the risk;
- Acute skin condition: Bacterial and fungal infections may happen;
- Impairment of hearing: The problems of hearing are common;
- Alzheimer's disease: Increases the chance of Alzheimer's disease.

Fig 2. Diabetes Symptoms[5]

Thus, if ignored, this ill situation depletes resources for people, families, and the entire country. The healthy life and overall well-being of people with prediabetes depend on early identification and symptomatic treatment. Disease identification and prevention will benefit from an intelligent medical diagnostic system based on symptoms, indicators, laboratory testing, and observations. Medical diagnosis systems have used artificial intelligence (AI) in a variety of intriguing methods for the identification of diseases. This study suggests a framework for combining machine learning and traditional methods to identify diabetes individuals early.

I. Techniques Used for Diabetes Detection

1.1 Artificial Neural Network (ANN)

In an artificial neural network, a neuron processing unit can represent different objects, such as features, letters, concepts, or some meaningful abstraction pattern. The type of processing unit in the network is divided into three categories: input unit, output unit and hidden unit. The input unit accepts signals and data from the outside world. The output unit realizes the output of the system processing result. The hidden unit is a unit that is located between the input and output units and cannot be observed outside the system [6].

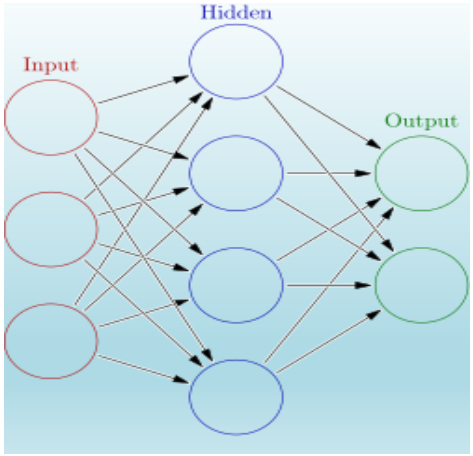

Fig 3. Artificial Neural Network [7]

1.2 Random Forest (RF)

A Random Forest Technique is a well-liked supervised machine learning algorithm that is utilized for Classification and Regression issues in machine learning. We are aware that a forest is made up of many different types of trees, and the more trees there are, the more robust the forest will be. Similar to this, the accuracy and problem-solving power of a Random Forest Algorithm increases with the number of trees in the algorithm. In order to increase the dataset's predicted accuracy, a classifier known as Random Forest uses many decision trees on different subsets of the input data. It is based on the idea of ensemble learning, which is the practice of integrating many classifiers to solve a challenging issue and enhance the performance of the model.

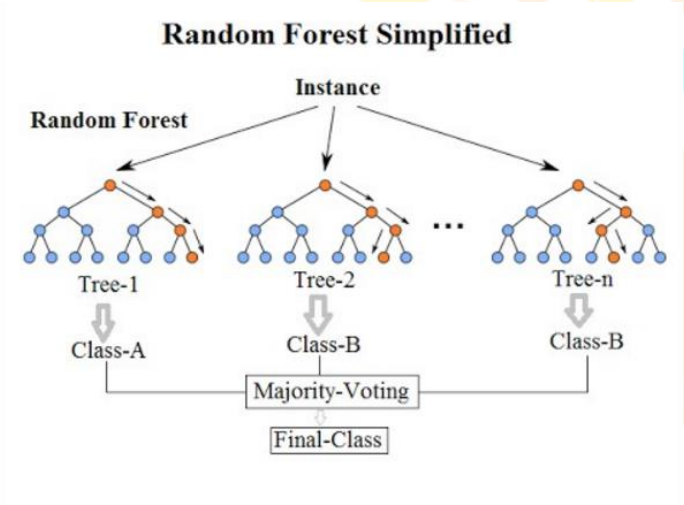

Fig 4. Random Forest [8]

1.3 K-Nearest Neighbors (KNN)

In this classification technique, the anonymous data points are discovered using the familiar data points which are known as nearest neighbors. k-Nearest neighbors (k-NN) is conceptually simple and is also called as lazy learning, where “k” is the nearest neighbor. In k-NN algorithm, the aim is to vigorously recognize k samples in the training dataset which are identical to a new sample [9].

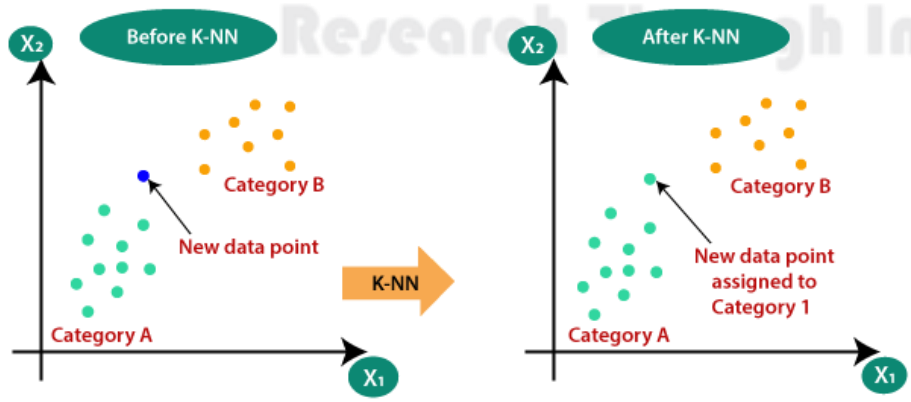

Fig 5. K-Nearest Neighbors [10]

1.4 Support Vetor Machine (SVM)

The occurrences of points in area are denoted by the SVM algorithm that are then plotted so that the classes are separated by strong gap. The goal is to determine the maximum-margin hyperplane which provides the greatest parting between the classes. The occurrences which is closest to the maximum-margin hyperplane are called support vectors. The vectors are chosen which are based on the part of the dataset that signifies the training set. Support vectors of two classes enable the creation of two parallel hyperplanes. Therefore, larger the periphery between the two hyperplanes, better will be the generalization error of the classifier. SVMs are implemented in a unique way as compared with other machine learning algorithms [11].

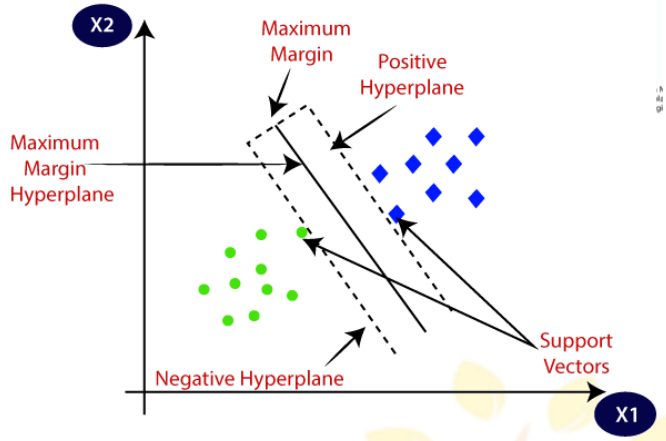

Fig 6. Support Vetor Machine [12]

1.5 Logistic Regression (LR)

Logistic Regression was mostly used in the biological research and applications in the early 20th century. Logistic Regression (LR) is one of the most used machine learning algorithms that is used where the target variable is categorical. Recently, LR is a popular method for binary classification problems. Moreover, it presents a discrete binary product between 0 and 1. Logistic Regression computes the relationship between the feature variables by assessing probabilities (p) using underlying logistic function [13].

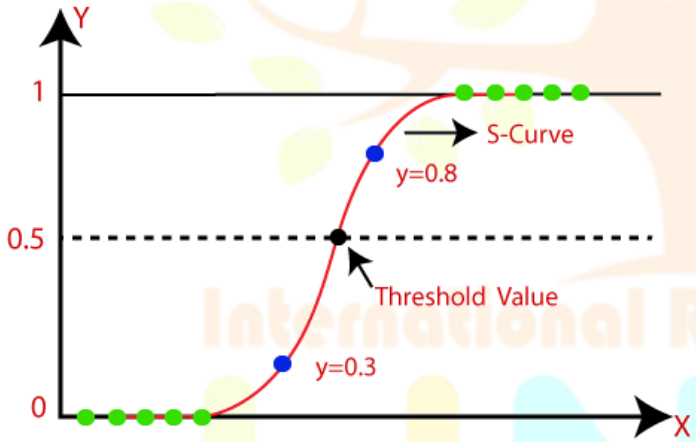

Fig 7. Logistic Regression [14]

II.LITERATURE SURVEY

Jobeda Jamal Khanam, Simon Y. Foo [15], In this study, they developed a system that can accurately predict diabetes, using the feature reduction method to predict diabetes, using DT, KNN, RF, NB, AB, LR, SVM, and ANN. They derived the result as: ANN = 88.57% accuracy max.

Harleen Kaur, Vinita Kumari [16], In this, the feature selection of the dataset is done with the help of the Boruta wrapper algorithm, which provides an unbiased selection of important features using linear kernel SVM, radial basis kernel SVM, k-NN, ANN, and MDR. Giving the result as K-NN = 92% maximum accuracy was obtained.

SMH Mahmud, MA Hossin, MR Ahmed, MR Ahmed, SRH Noori, MNI Islam Sarkar [17], In this, it has proposed a framework for diabetes prediction, monitoring, and application using the methods ANN, SVM, LR, DT, RF, and NB. The result was obtained as NB = 74% accuracy.

Kamrul Hasan, MD. Ashraful Alam, Dola Das, Eklas Hossain, Mahmudul Hasan [18], Diabetes prediction has been achieved in this literature. Utilizing the proposed ensemble model, where pre-processing is crucial for reliable and accurate prediction, where rejecting outliers and filling in missing values were key considerations. using different methods (K-NN, DT, RF, AB, NB, XB). in which XB has 95% accuracy.

Priyanka Sonar, K. Jaya Malini [19], The aim is to develop a system that might predict high levels of diabetes risk in a patient with better accuracy. using different methods: SVM, ANN, decision trees, and Naive Bayes. The outcomes were as follows: SVM = 82%, ANN = 82% accuracy as maximum.

Muhammad Azeem Sarwar, Nasir Kamal, Wajeeha Hamid, Munam Ali Shah [20], This paper aims to both help doctors and to be doctors. who are practicing early diabetes prediction utilizing ML techniques. DT, RF, SVM, LR, NB, and KNN are used. SVM = 77% and KNN = 77% accuracy were obtained.

Yuvaraj, K. R. Sri Preethaa [21], This study suggests a novel approach to machine learning algorithms for diabetes prediction using Hadoop-based clusters. The findings show that machine learning algorithms can be used to create extremely accurate diabetes predictive healthcare systems. NB, DT, and RF methods were used. I arrived at the conclusion that RF = 94% accuracy.

III. PROPOSED MODEL

The proposed system concentrates on combining algorithms shown in the Fig 8. ANN, Random Forest, Logistic Regression, KNN, Support Vector Machine for accuracy authentication are the fundamental classification algorithms.

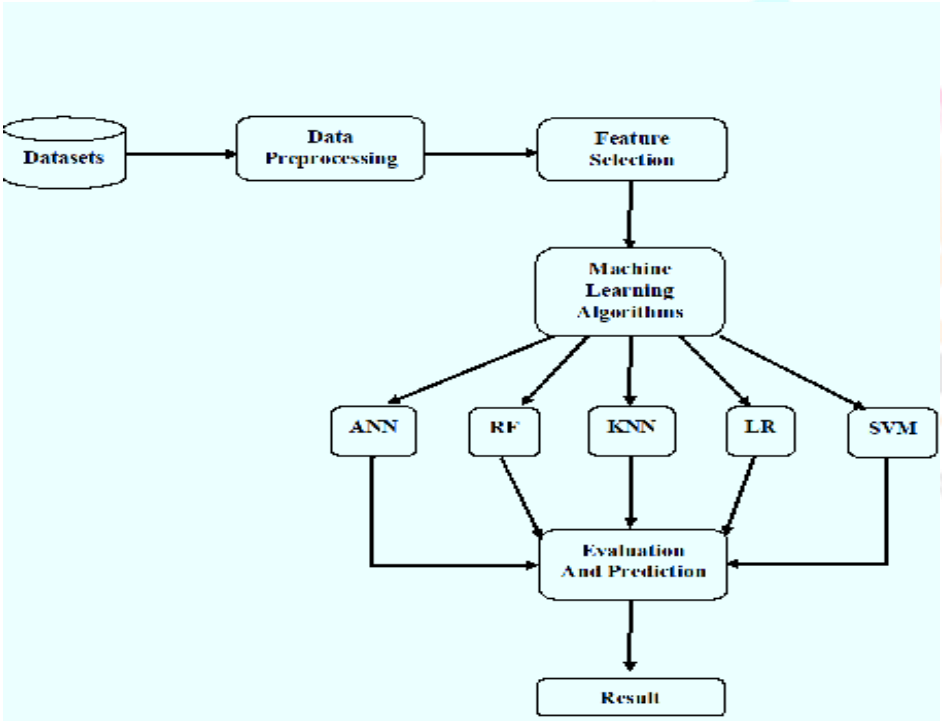

Fig 8. Proposed Model of Diabetes Prediction

Step 1: Initially, the first step in the diabetes dataset is taken. It is then applied to data pre-processing modules.

Step 2: In this, the Diabetes dataset's irrelevant features are removed by the pre-processing module, which then provides the pre-processed dataset with relevant features to the machine learning algorithms.

Step 3: In this step, feature selection is done, which is the procedure for selecting relevant features for your machine learning model to solve the problem.

Step 4: Then in this step, go through the evaluation and prediction of different machine learning models.

Step 5: The results computed and performance of each algorithm are evaluated to determine the best model for prediction.

IV.CONCLUSION

Diabetes develops when your blood sugar levels are abnormally high. Blood glucose is your main source of energy, and it comes from the foods you eat. Diabetes has been linked to a variety of diseases over time. This type of illness affects millions of people worldwide. Diabetes detection at an early stage is critical to people's health. As a result, in order to predict diabetes, a diabetes monitoring system is required. Machine learning techniques can help physicians diagnose and treat diabetic diseases. Machine learning is not limited to a single application; it has a broad range of applications for problem solving; we simply need to understand its techniques and identify areas where it can be used. Machine learning models, like humans, can be trained to behave well and learn from nature.

REFERENCES

[1] Hasan MK, Alam MA, Das D et al (2020) Diabetes prediction using ensembling of different machine learning classifiers. IEEE Access 8:76516–765 https://doi.org/10.1109/ACCESS.2020. 2989857

[2] Saeedi P, Petersohn I, Salpea P et al (2019) Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: Results from the International Diabetes Federation Diabetes Atlas, 9th edition. Diabetes Res Clin Pract 157:107843. https://doi.org/10.1016/j.diabres.2019.107843

[3] Gupta, H., Varshney, H., Sharma, T.K. *et al*. Comparative performance analysis of quantum machine learning with deep learning for diabetes prediction. *Complex Intell. Syst.* 8, 3073–3087 (2022). https://doi.org/10.1007/s40747-021-00398-7

[4] K. Sumangali, B. S. R. Geetika and H. Ambarkar, "A classifier-based approach for early detection of diabetes mellitus," *2016 International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT)*, 2016, pp. 389-392, Doi: 10.1109/ICCICCT.2016.7987979.

[5] Johnson, H. B., MD. (2016, October 11). *Diabetes Signs and Symptoms: What Are My Long-Term Risks?* University Health News. https://universityhealthnews.com/daily/diabetes/diabetes-signs-and-symptoms-what-are-my-long-term-risks/

[6] Wu, Yc., Feng, Jw. Development and Application of Artificial Neural Network. *Wireless Pers Commun* 102, 1645–1656 (2018). https://doi.org/10.1007/s11277-017-5224-x

[7] Agrawal, S. K. (2022, September 6). *Understanding the Basics Of Artificial Neural Network*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/07/understanding-the-basics-of-artificial-neural-network-ann/

[8] Wikipedia contributors. (2022, November 29). *Random forest*. Wikipedia. https://en.wikipedia.org/wiki/Random_forest

[9] Choudhury, A., & Gupta, D. (2019). A survey on medical diagnosis of diabetes using machine learning techniques. In *Recent developments in machine learning and data analytics* (pp. 67-78). Springer, Singapore.

[10] K-Nearest Neighbor(KNN) Algorithm for Machine Learning - Javatpoint. (n.d.). www.javatpoint.com. https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning

[11] Sonar, P., & JayaMalini, K. (2019, March). Diabetes prediction using different machine learning approaches. In *2019 3rd International Conference on Computing Methodologies and Communication (ICCMC)* (pp. 367-371). IEEE.

[12]] Support Vector Machine (SVM) Algorithm - Javatpoint. (n.d.). www.javatpoint.com. https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm

[13] Mahmud, S. H., Hossin, M. A., Ahmed, M. R., Noori, S. R. H., & Sarkar, M. N. I. (2018, August). Machine learning based unified framework for diabetes prediction. In *Proceedings of the 2018 International Conference on Big Data Engineering and Technology* (pp. 46-50).

[14]] *Logistic Regression in Machine Learning - Javatpoint*. (n.d.). www.javatpoint.com. https://www.javatpoint.com/logistic-regression-in-machine-learning

[15] Khanam, J. J., & Foo, S. Y. (2021). A comparison of machine learning algorithms for diabetes prediction. *ICT Express*, *7*(4), 432-439

[16] Kaur, H., & Kumari, V. (2020). Predictive modelling and analytics for diabetes using a machine learning approach. *Applied computing and informatics*.

[17] Mahmud, S. H., Hossin, M. A., Ahmed, M. R., Noori, S. R. H., & Sarkar, M. N. I. (2018, August). Machine learning based unified framework for diabetes prediction. In *Proceedings of the 2018 International Conference on Big Data Engineering and Technology* (pp. 46-50).

[18] Hasan, M. K., Alam, M. A., Das, D., Hossain, E., & Hasan, M. (2020). Diabetes prediction using ensembling of different machine learning classifiers. *IEEE Access*, *8*, 76516-76531

[19] Sonar, P., & JayaMalini, K. (2019, March). Diabetes prediction using different machine learning approaches. In *2019 3rd International Conference on Computing Methodologies and Communication (ICCMC)* (pp. 367-371). IEEE.

[20] Sarwar, M. A., Kamal, N., Hamid, W., & Shah, M. A. (2018, September). Prediction of diabetes using machine learning algorithms in healthcare. In *2018 24th international conference on automation and computing (ICAC)* (pp. 1-6). IEEE.

[21] Yuvaraj, N., & SriPreethaa, K. R. (2019). Diabetes prediction in healthcare systems using machine learning algorithms on Hadoop cluster. *Cluster Computing*, *22*(1), 1-9.