Implementation of a Web Application to Predict Diabetes Disease using Machine Learning Algorithm

Thirumalai teja B
Department of Electronics and Communication Engineering
Panimalar Institute Of Technology
Chennai, India
ronaldoteja273@gmail.com

Prasanth C
Department of Electronics and Communication Engineering
Panimalar Institute Of Technology
Chennai, India
cprasanth104@gmail.com

Peer Mohamed Afridi A
Department of Electronics and Communication Engineering
Panimalar Institute Of Technology
Chennai, India
peermohamedafridia@gmail.com

Dr S. Sathiya Priya
Department of Electronics and Communication Engineering
Panimalar Institute Of Technology
Chennai, India
ecehodpit@gmail.com

***Abstract*— Diabetes is a disease caused by the formation of a disproportionately large amount of sugar in the blood coagulating. Today, it is one of the fatal diseases in the world. This complex illness affects people worldwide both consciously and subconsciously. Diabetes is also the reason for diseases like heart attack, paralyzed, kidney disease, blindness etc. Various type of computer based detection works are proposed in literatures for prediction and simulation of diabetes. Identifying process for diabetic patients usual required much more time and money. However, machine learning is the key to creating a solution to this agitated problem. So we have create another architecture where we have ability to predict whether the patient has diabetes or not. The main objective of this analysis is to create a web application for predicting diabetes of a user through the prediction of a high accuracy by some high power machine learning algorithm. A benchmark dataset namely Pima Indian has been used for prediction of the use diabetes based on a diagnosis. With an accuracy of 82.35%of t he predicted speed of artificial neural networks (ANNs), a signific ant improvement in accuracy has been shown to developinteractiv e web applications for predicting diabetes.**

Keywords— Diabetes, SVM, ANN, Naive Bayes, Min Max Scaling

I. INTRODUCTION

Diabetes is a rapidly growing disease among young people and among people themselves. Increased sugar levels (glucose) in the blood can cause diabetes. Diabetes can be divided into two categories. B. Type-1-disaccharide and Type-2-disaccharide. Type-1 diabetes is an autoimmune disease. In this case, the body destroys the cells that are important to produce insulin, absorbing sugar and producing energy. And this type of diabetes can cause obesity. Obesity is an increase in the body mass index (BMI) as an individual's normal BMI level [1]. -Type 1 diabetes can occur in childhood or adolescence. -Type 2 diabetes usually affects obese adults. In this way, the body opposes observations of insulin or does not produce insulin. Type 2 usually occurs in the middle or age group. Additionally, there are other causes of diabetes, such as bacterial or viral infections, toxic or chemical content in food, autoimmune reactions, obesity, nutritional changes, lifestyle changes, feeding habits, and contamination. Complications, kidney problems, retinopathy, foot ulcers, etc. Machine learning algorithms usually find hidden patterns in large data records and find approximate final results. Machine learning is an AI under field, and machine learning algorithms can be divided into three categories: B. Study for monitored learning, unattended learning, and reinforcement. The system uses monitored learning algorithms to test the accuracy of popular machine learning algorithms (machine learning). Monitoring Learning Algorithms Learn patterns from existing data and attempt to predict new outcomes based on previous learning. ML algorithms are intended to identify existing data such as probability-based, function-based, typically tree-based, and instance-based. Various algorithms have been introduced for machine learning to support medical professionals using a variety of data mining algorithms. The effectiveness of a decision support system is recognized by its accuracy. Therefore, the main purpose of the structure of decision support systems is to predict specific diseases with maximum levels of accuracy. The system uses existing data records called Pima Indians to train and evaluate models, which are open source data records.

The deficits of this study were codified as follows: Section II discusses some related works that have been performed previously. The proposed architecture was demonstrated in Section III. Section IV contains isolated descriptions of methodologies and various algorithms. Applications and implementation design are discussed in Section V. Discussions of experimental procedures and outcome analysis are presented in Section VI. Complete the paper by presenting future work in Section VII.

II. RELAVENT WORK

This section presents some of the previous research work, which are answered along with the proposed work. In relation to, the authors proposed that the information provided to separate all the metadata would be pre-processed. KNN is used to find the next neighbour for the specified data record. If the desired level is found, the algorithm will stop execution if it is not used by the system classifier. According to reference, Naiver Bayes (NB) uses precise algorithms for machine learning to use Arabic web documents. The K-nearest neighbour classification was examined to estimate economic situation. Economics and

bankruptcy maintain intentional removal can be used to estimate the use of KNN technology. The amount of bankruptcy organizations is increasing in the event of a global emergency. The percentage of economic burden expectations draws a wide range of concerns about academic liability, not just economic and financial institutions. Related authors suggested that insulin withdrawal from the human body, leading to blood cells and glucose to diabetes, is a lifelong illness. Diabetes also affects difficulties such as stroke, heart disease, blindness, weight loss for kidney failure. Essentially, the human neuronal structures performed are learning, generalization and mathematical.

III. PROPOSED ARCHITECTURE

Monitoring Learning Algorithms Learn patterns from existing data and attempt to predict new outcomes based on previous learning. The ML algorithm is used to identify existing data, such as probability-based, feature based, rule-based, tree-based, and instance-based, according to Figure 1. View the broad architecture of the proposed model. According to our model, patients need to provide medical data to successfully diagnose diabetes
testing.

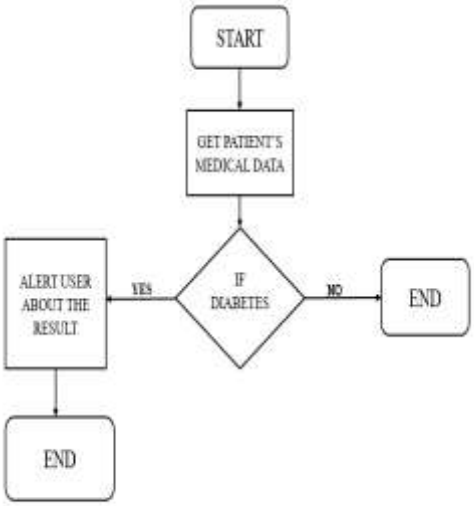

Fig. 1. Flowchart of Diabetes prediction Model

IV. METHODOLOGIES

A. Support Vector Machine (SVM)

The Support Vector Machine is a monitored learning algorithm that is primarily used for classification problems. Overcharging incorrectly derives ML models from a particular data record. In this case, SVM can prevent the nature of the test data from being overwhelmed and producing better accuracy. SVM has linear hyperbones with edges that divide the data records into positive and negative samples.

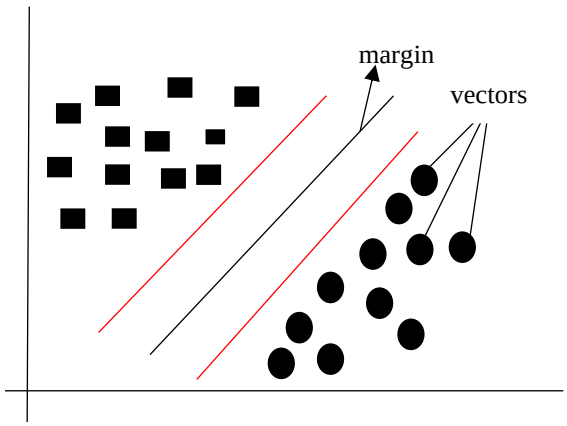

Fig. 2. Graphical representation of SVM working procedure

The SVM selects the hyperlevel that allows maximum removal. SVM decision limits for mathematical equations as follows:

$$min\frac{1}{2}\sum_{j=1}^{n}\theta_j^2$$

Where,

$$\theta^T x^{(i)} \geq 1 \quad if\ y^{(i)} = 1$$
$$\theta^T x^{(i)} \leq -1 \quad if\ y^{(i)} = 0$$

B. K- Nearest Neighbors(KNN)

K-Nearest Neighbors is a monitored learning algorithm and is a lot of vectors in K. The KNN working method is quite simple, with predictions being based on the values of the K parameter. The following graphical representation of the number of neighbors is shown in Figure 3 below.

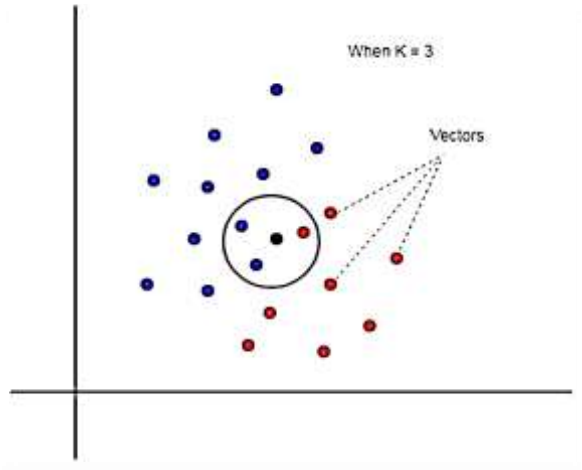

Fig. 3. Graphical representation of KNN working Procedure

From the above diagram, for k is 3, we try to meet the following three adjacent vectors: In the next step, KNN finds the highest ratio of adjacent vectors and finally creates the result of the entrance vector.

C. Naive Bayes Algorithm

The Naive Bayes algorithm is another machine learning algorithm for classification problems. Naive Bayes is an efficient classification algorithm for data mining, which can handle missing values during classification. Naive Bayes algorithms are machine learning and efficient models. Essentially, this model is used for text classification. Some known examples are spam filtration, sentimental analysis, and classification of new articles. Naive names are a kind of distinctive feature that is different from another distinctive event. Bayes refers to the theorem of statistician and philosopher Thomas Bayes. The NB theorem can be expressed mathematically as follows:

$$P(A \mid B) = \frac{P(B \mid A)\, P(A)}{P(B)}$$

- P (A | B): Probability of occurrence of event A given the event B is true.
- P (A) and P (B): Probabilities of the occurrence of event A and B respectively.
- P (B | A): Probability of the occurrence of event B given the event A is true.

Bayes Theorem for Naive Bayes Algorithm state the following relationship

$$P(C_i|x_1, x_2, \ldots\ldots, x_n) = \frac{P(x_1, x_2, \ldots\ldots, x_n \mid C_i)}{P(x_1, x_2, \ldots\ldots, x_n)} \quad for\ 1 \leq i \leq k$$

This relationship can be simplified as follows

$$P(C_i|x_1, x_1, \ldots\ldots, x_n) = \left(\prod_{j=1}^{j=n} P(x_j \mid C_i)\right) \cdot \frac{P(C_i)}{P(x_1, x_1, \ldots\ldots, x_n)} \quad for\ 1 \leq i \leq k$$

D. ARTIFICIAL NEURAL NETWORK (ANN)

Artificial Neuron Networks (ANNs) are considered a way of presenting human neurons in a mathematical way by reflecting their learning and generalization skills. The ANN model can expand strongly nonlinear systems where the relationships between variables are unknown or complex.

A neuronal network is made up of different neurons, followed by several layers. Knot represents the structure of human neurons such as Dendrit and Axon. This allows weighted connections to be connected between nodes as axons. The overall structure of a neural network is shaped by the input layer, one or more hidden layers, and the starting layer where the I th-neuron represents the connection to the J th neuron of the entire structure, and Wij as the strength of the connection. Between neurons. The nodes in the structure of ANN take in (functions) and then process and calculate the weight total of a kind, and send the bias term () to the next hidden layer by node data to send to JTH - Node [11]. The mathematical representation of the above argument can be expressed as follows:

$$net_j = \sum_{i=1}^{m} x_i * w_{ij} + \theta_j \ (j = 1, 2, \ldots, n)$$

V. DESIGN AND IMPLEMENTATION

As mentioned previously, our goal is to develop web-based applications that can predict diabetes based on patient health data. We tested some powerful models of machine learning, such as SVM, KNN, Naive Bayes, and ANN, to find the most appropriate ML algorithms that can be predicted more accurately. I successfully evaluated these models using several libraries for machine learning such as Scikit-Learn, Numpy, Matplotlib, Pandas. To remove the over-adjustment problem, split the data record into two subsections. One is for testing and the other is for training. Based on the exchange size of different training data, the accuracy speed of each defined model was reached. However, pre-processing of PIMA data in India could lead to greater prediction accuracy. Therefore, to improve prediction accuracy, we also calculate the accuracy of the defined model. Dates are an effective way to increase the accuracy of a particular model for machine learning, and machine learning models do not work well without data. The proposed model used Min Max Scaler (MMS) as the normalization model. MMS basically extends data to the range [0, 1] or [-1, 1]. The mathematical formula for maximum scaling in Min Max can be displayed as follows:

$$x' = \frac{x - min(x)}{max(x) - min(x)}$$

By using the MIN MAX scaler (MMS) processing method for each defined model, we achieved higher prediction accuracy than previous calculations. To implement the designed architecture, I developed a web application using Streamlit based on the highest prediction accuracy. The accuracy calculations for the defined algorithm are shown in Figure 4. The method is called getScores.

```
def getScores(X_train, y_train, X_test, y_test):
    SVM_Score = svm_clf.fit(X_train, y_train).score(X_test, y_test)
    KNN_Score = knn.fit(X_train, y_train).score(X_test, y_test)
    GNB_Score = gnb.fit(X_train, y_train).score(X_test, y_test)
    ANN_Score = ann.fit(X_train, y_train).score(X_test, y_test)
```

Fig. 4. Score Calculation code snippet of different algorithms

VI. RESULTS

In this section we will discuss regarding our results which we have achieved after experimental design. Following TABLE I. represents an insight description of our Pima Indian Dataset. This dataset is mainly based on the females those were living at Pima Indian heritage. Following 8 features ***(a-h)*** of Pima Indian dataset helps us to predict the diabetes of any Individuals with the help of our proposed methodologies.

a) Numbers of time Pregnant
b) Glucose Test
c) Blood Pressure
d) Triceps skinfold thickness
e) 2-Hour Serum Insulin
f) Body Mass Index
g) Diabetes Pedigree function
h) Age

Class	Attribute Number
Pregnancy Count	1
Glucose concentration in plasma	2
Blood pressure (diastolic, mm Hg)	3
Thickness of triceps skin fold (mm)	4
2-Hour serum insulin (μ U/ml)	5
Body mass index	6
Pedigree function of diabetes	7
Years of age	8

TABLE I. DIFFERENT ATTRIBUTES OF INDIAN PIMA DATASET

Following TABLE II depicts the average accuracy rate varies with training dataset size. We have experimented with different size of training data for SVM, KNN, GNB and ANN. Without Min Max Scaling (MMS) it shows an average accuracy of 76.25% with Gaussian Naïve Bayes Algorithm.

Training Dataset Size	SVM	KNN	GNB	ANN
368	63	64	76	63
468	63	68	77	63
568	63	68	76	63
668	63	66	76	63
Average	**63**	**66.5**	**76.25**	**63**

TABLE II. ACCURACY OF DIFFERENT MACHINE LEARNING ALGORITHMS BASED ON PIMA INDIAN DATASET

In order to improve the detection accuracy we have performed some data pre-processing by using Min Max Scaler Method. Following TABLE III. shows that by using Min Max Scaling Method we have achieved more higher accuracy than previous calculation in TABLE II. According to TABLE III. Artificial Neural Network achieved 82.35% detection accuracy based on the diagnosis nature of Indian Pima Dataset. Therefore we have built a web application using ANN Model which is capable of predicting whether a patient has diabetes or not.

Training Dataset Size	SVM + MMS	KNN + MMS	GNB + MMS	ANN + MMS
368	78	75.3	79.1	81.8
468	78	76	79.3	82.3
568	78.2	75.1	79.3	82.9
668	78	75.6	79.5	82.4
Average	**78.05**	**75.5**	**79.3**	**82.35**

TABLE III. ACCURACY OF ALGORITHMS WITH MIN MAX SCALER METHOD

We used the python programming language as the backend development and streamlit as the frontend for the machine learning model. Our proposed architecture typically collects data records values from the model of artificial neural networks from SQL databases and schools during training sessions. During the forecast period, the user must provide some information. Developed web applications allow you to predict whether your test results are positive or not. To test diabetic users, users must provide the following information in their web application. You will need some of the much-needed information such as blood pressure, body mass index (BMI), serum insulin, or oral glucose resistance tests.

To assess the model's identification accuracy, we compared our work with several state-ART research tasks. Table IV below is driven using different methods for prediction of different medical data. From the following, it is clear that the proposed model has the highest accuracy than other research work. The graphical representations of Tables II and III can be found in the following diagram. 5 and Figure 6.

Model Name	Accuracy
Gaussian Naive Bayes (GNB) [13]	76.52%
General Regression Network (GRNN) [14]	80.21%
Backpropagation Genetic Algorithm (BGA) [15]	74.80%
Fuzzy Min Max (FMM) [16]	69.28%
Our Proposed Model (ANN with MMS)	**82.35%**

TABLE IV. COMPARISON OF OUR PROPOSED MODEL WITH OTHERS

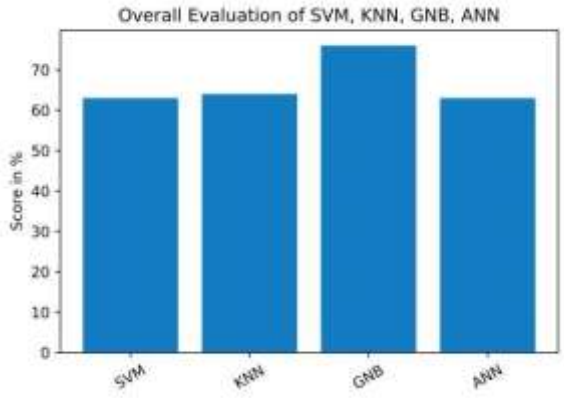

Fig. 5. Graphical representation of Accuracy of different machine learning algorithm

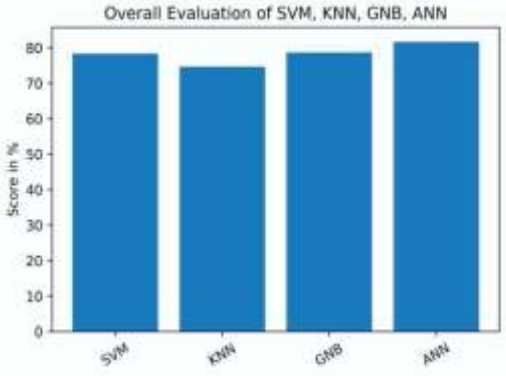

Fig. 6. Graphical representation of Accuracy of different machine learning algorithm with Min Max Scaler Method

VII. CONCLUSION AND FUTURE SCOPE

This paper highlights the development of a web-based application aimed at the accurate prediction of diabetes diseases. Through a comparative analysis of various machine learning algorithms, the study identifies Artificial Neural Network (ANN) as the most effective approach, achieving the highest accuracy when combined with the Min-Max Scaling Method on the Indian Pima Dataset. This combination demonstrates the potential to enhance diagnostic accuracy, offering valuable insights into diabetes prediction.

The proposed approach represents a significant step forward in utilizing machine learning within the medical field, showcasing the ability to process and analyze medical data

with precision. By advancing diabetes prediction methods, this application could aid in early detection, personalized care, and improved health outcomes for patients.

Looking ahead, the focus is on extending this work by exploring deep learning models, which are known for their capacity to capture complex patterns in large datasets. Additionally, the development of a location-based dataset using real-world medical data is envisioned. This future direction aims to refine the prediction capabilities further and adapt the model to diverse populations, paving the way for more robust and universally applicable diabetes prediction tools.

VII. REFERENCES

[1] Muni Balaji Thumu; N. Balajiraja; Muhammed Yousoof, "Predictive Modelling For Diabetes Mellitus And Cardiovascular Disease Detection Using Artificial Neural Network" 2024 4th International Conference on Data Engineering and Communication Systems (ICDECS)

[2] Benjamin Lobo , Leon Farhy, Mahdi Shafiei, and Boris Kovatchev, "A Data-Driven Approach to Classifying Daily Continuous Glucose Monitoring (CGM) Time Series", IEEE Transactions on Biomedical Engineering, Vol. 69, No.2, February 2022

[3] Sumeet Kalia , Olli Saarela, Tao Chen, Braden O'Neill, Christopher Meaney, Jessica Gronsbell, "Marginal Structural Models Using Calibrated Weights With Super Learner: Application to TypeII Diabetes Cohort", IEEE Journal of Biomedical and Health Informatics, Vol.26, No.8, August 2022.

[4] C. Charitha, A. Devi Chaitrasree, P. C. Varma and C. Lakshmi, "Type-II Diabetes Prediction Using Machine Learning Algorithms," 2022 International Conference on Computer Communication and Informatics(ICCCI), Coimbatore, India, 2022, pp. 1-5, doi: 10.1109/ICCCI54379.2022.9740844.

[5] H. Kaur and V. Kumari, "Predictive modelling and analytics for diabetes using the machine learning approach", Applied Computing and Informatics, vol. 18, no. 1, pp. 90-100, 2022.

[6] R. Jain, "Introduction to Naive Bayes Classification Algorithm in Python and R," February 2, 2017.

[7] D.k. Chouby, S. Paul, S. Kumar, and S. Kumar, "Classification of Pima indian diabetes dataset using naive bayes with genetic algorithm as an attribute selection," The International Conference on Communication and Computing Systems, ICCCS-2016.

[8] H. Shee, K. W Cheruiyot and S. Kimani, "Application of k-Nearest Neighbour Classification in Medical Data Mining," International Journal of Information and Communication Technology Research, Volume 4, No. 4, April 2014.

[9] V. Vijayan, and A. Ravikumar, "Study of Data Mining algorithms for Prediction and Diagnosis of Diabetes Mellitus," International Journal of Computer Application, Vol 94, pp .12-16, June 2014.

[10] C. Kalaiselvi and G. M. Nasira, "A New Approach for Diagnosis of Diabetes and Prediction of Cancer Using ANFIS," 2014 World Congress on Computing and Communication Technologies, Trichirappalli, 2014, pp. 188-190.

[11] V. Chandra S.S, and A. Hareendran S, "Artificial Intelligence and machine learning," PHI learning Private Limited, Delhi 110092, 2014.

[12] S. Kumari and A. Singh, "A data mining approach for the diagnosis of diabetes mellitus," 2013 7th International Conference on Intelligent Systems and Control (ISCO), Coimbatore, 2013, pp. 373-375.

[13] R. Motka, V. Parmarl, B. Kumar and A. R. Verma, "Diabetes mellitus forecast using different data mining techniques," 2013 4th International Conference on Computer and Communication Technology (ICCCT), Allahabad, 2013, pp. 99-103.

[14] F. Amato, A. López, E. M. Peña-Méndez, P. Vaňhara, A. Hampl and J. Havel, " Artificial neural networks in medical diagnosis," 2013.

[15] F.S Panchal, A. Ganatra, Y. Kosta, and M. Panchal, "Review on Methods of Selecting Number of Hidden Nodes in Artificial Neural Network," International Journal of Computer Theory and Engineering, vol. 3, no. 2, pp. 332–337, 2011.

[16] H. Hasan Örkcü, and H. Bal, "Comparing Performances of Backpropagation and Genetic Algorithms in the Data Classification," Expert Systems with Applications, vol. 38, 2011, pp. 3703-3709.

[17] P. Tan, M. Steinbach, V. Kumar, "Introduction to Data Mining," Pearson Education, Inc,2006.

[18] K. Kayaer, and T. Yildirim, "Medical Diagnosis on Pima IndianDiabetes Using General Regression Neural Networks," Proceedings of the International Conference on Artificial Neural Networks and Neural Information Processing, pp. 181–184, 2003.

[19] M. Elkourdi, A. Bensaid, and T. Rachidi, "Automatic Arabic Document Categorization Based on the Naïve Bayes Algorithm," Alakhawayn University, 2001.

[20] I. Aleksander, and H. Morton, "An introduction to neural computing," Int Thomson Computer Press, London 1995.

[21] M. Seera, and C.P. Lim, "A hybrid intelligent system for medical data classification," Expert Systems with Applications, Vol. 41 pp. 2239– 2249.