www.ijsrnsc.org

Available online at www.ijsrnsc.org

IJSRNSC

Volume-**12**, Issue-**3**, June **2024**
Research Paper

Int. J. Sc. Res. in Network Security and Communication

E-ISSN:2321-3256

Diabetes Predictor: Prediction Using Machine Learning Techniques

A Hariprasad Reddy[1], G. Tejaswi[2*], S. Akshay[3], E. Anisha[4]

[1]Dept. Of CSE, Geethanjali College of Engineering and Technology, JNTU, Hyderabad, India
[2,3,4]Dept. of CSE, Geethanjali College of Engineering and Technology, Hyderabad, India

[*]*Corresponding Author: tejaswigudapati22@gmail.com, Tel.: +91-9494354454*

Received: 18/Apr/2024, Accepted: 20/May/2024, Published: 30/Jun/2024

***Abstract*—** As the old saying goes, prevention is better than cure when it comes to health. The likelihood of saving lives can be greatly increased by anticipating diseases such as diabetes. Numerous variables, including age, obesity, lack of exercise, genetic predisposition, lifestyle, nutrition, and high blood pressure, can contribute to diabetes, an illness that is spreading quickly. With the help of machine learning techniques (MLT), healthcare professionals can now forecast patient outcomes using pre-existing data, which makes them indispensable tools. Several categorization machine learning methods are used in a diabetes prediction project to identify the most accurate model. This model takes into account extrinsic factors linked to diabetes risk in addition to conventional components like insulin, age, BMI, and glucose. Comprehending the natural glucose regulating process of the body is essential to understanding diabetes. The body uses glucose, which is obtained from foods high in carbohydrates, as its main energy source. The pancreas secretes insulin, which makes glucose easier for cells to use as fuel. On the other hand, diabetes is brought on by inadequate insulin synthesis or inadequate insulin use, which raise blood glucose levels. Here, skin thickness, number of conceptions, and pedigree function are additional characteristics that improve the model's prediction power. These factors enhance the accuracy of diabetes risk assessment by adding to conventional markers and providing insightful information. Proactive illness prediction is made possible by utilizing MLT in the healthcare industry, especially for conditions like diabetes. The predicted accuracy of diabetes models can be greatly increased by incorporating both traditional and non-conventional risk indicators, such as skin thickness, number of pregnancies, and pedigree function. This will enable early intervention and better patient outcomes.

***Keywords*—** Prevention, Diabetes, Extrinsic factors, Skin thickness, Conceptions (number of pregnancies), Pedigree function, Proactive prediction, Early intervention, Anticipating, Predisposition, Incorporating, Indispensable.

I. INTRODUCTION

In today's fast-paced world, Artificial Intelligence (AI) has become a game-changer, bridging the gap between human intelligence and machine capabilities. One particularly fascinating area where AI shines is in Computer Vision, which aims to teach machines to see and understand the world just like humans do. Within this exciting landscape, our project focuses on using AI to revolutionize prediction of diabetes.

Diabetes is a condition that is quickly spreading and impacting people of all ages, including children. The natural glucose metabolism of the body must be understood in order to understand its development. The body uses glucose, which is mostly found in carbohydrate-rich foods like bread, pasta, and fruits, as its main energy source. Carbs are needed by people with diabetes as well for energy. Skin thickness and the number of pregnancies is important factors in diabetes. Thick skin may be a sign of affront resistance, a clutter in which cells lose their capacity to reply to affront, which raises blood sugar levels. Furthermore, the frequency of pregnancies may impact the risk of diabetes. Multiple pregnancy mothers may be at an increased risk of getting type 2 diabetes or gestational diabetes in later life.
It is essential to comprehend these mechanisms in order to develop diabetes preventive and care plans that work.

II. RELATED WORK

A. Diabetes prediction using machine learning algorithms
This research paper by Aishwarya Mujumdar, Dr. V Vaidehi[1] from Vellore Institute of Technology, Chennai, India and Mother Teresa Women's University, Kodaikanal, India respectively. published in Sciencedirect journal, explore the potential of machine learning (ML) techniques in diabetes prediction. The study aims to develop a comprehensive framework for prediction of diabetes, as Big Data analytics plays a growing role in the healthcare industry

by enabling the analysis of massive information to find hidden patterns and forecast results.

B. Diabetes prediction using Machine Learning and expandible techniques.

This research paper by **Isfafuzzaman Tasin,** Electrical and Computer Engineering, North South University, Dhaka Bangladesh, published in National Institute of Health (NIH) journal, focuses on the research uses machine learning approaches to address the crucial problem of diabetes mellitus prediction. It draws attention to the difficulties of detecting diabetes at an early stage and describes different machine learning techniques used to forecast the disease utilizing custom and Pima Indian datasets. To increase prediction accuracy, strategies including ensemble methods and semi-supervised learning are used. The creation of a smartphone application and online application for real-time prediction utilizing the top-performing model is covered in the study.

The development of user-friendly apps, the introduction of a new dataset, and the use of explainable AI approaches for forecast transparency are some of the major accomplishments. Based on the combined dataset, the bagging classifier demonstrated the best accuracy of 79%, indicating a good performance. All things considered, the study offers a thorough method for predicting diabetes that combines real-world applications with machine learning algorithms.

C. Pediatric diabetes prediction using deep learning

This research paper by Abeer El-Sayyid El-Bashbishy, Information Systems Department, Faculty of Computer and Information Sciences, Mansoura University, Mansoura, Egypt, published in scientific reports by nature.com, focuses on Using data from 520 people between the ages of 16 and 90, this consider proposes a diabetes forecast framework that produces utilize of machine learning methods like Bolster Vector Machine (SVM), Gullible Bayes classifier, and LightGBM. The consider emphasizes the esteem of early diabetes discovery, especially in more youthful populaces, and appears that SVM performs superior than Credulous Bayes and LightGBM, with an precision rate of 96.54%. It illustrates how machine learning might improve medical diagnosis and treatment, pointing the way for further study and real-world application in the field of medicine. Utilizing extensive datasets and advanced algorithms, these methods have the potential to enhance patient outcomes and disease treatment.

III. PROPOSED WORK

In arrange to progress the profundity of investigation, the recommended diabetes forecast demonstrate consolidates unused factors such as blood weight, skin thickness, and pregnancy in conjunction with more customary measurements like age, BMI, glucose, and affront levels. By utilizing four different classification algorithms and strict optimization techniques such as scaling and hyperparameter tuning, the system strives for increased precision and resilience. Fast test result delivery, precision powered by AI and computer vision, and an easy-to-use interface for people with varying technical backgrounds are some of its main advantages. In addition, the platform allows for remote disease assessment, removing geographic restrictions and guaranteeing accessibility for a wide range of people. Situated as a progressive resolution, its integration of state-of-the-art artificial intelligence technology guarantees flexibility in response to changing diabetes diagnostic requirements and developing healthcare models.

System Architecture

Defining the architecture, components, and interactions of the system is a key component of the machine learning architectural model for blood sugar prediction. Choosing the right machine learning framework, creating the data pipeline, and figuring out how data flows from data sources to preprocessing tools, prediction models, and user interfaces are all part of this step.

This architectural model also makes important judgments about data storage, scalability, and security to guarantee that the system can safely and effectively manage blood sugar level prediction. The project's technical implementation is guided by the architectural model, which also establishes the groundwork for developing a strong and trustworthy blood sugar prediction system.

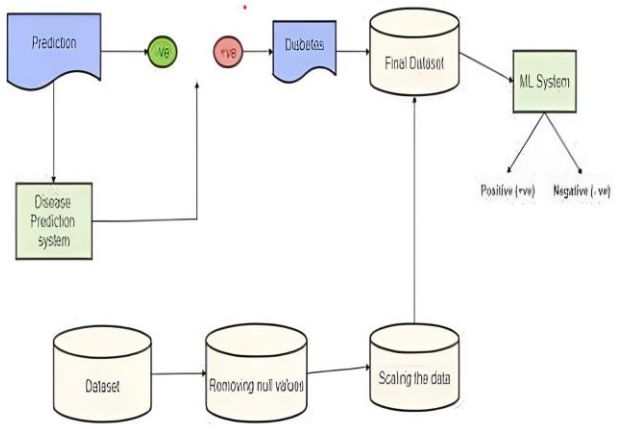

Figure 1. System Architecture

1) Dataset

The architecture begins with datasets of several people, each representing different inputs or levels of parameters. These datasets serve as the primary sources of information for disease detection and analysis within the system.

2) Data Preprocessing

Upon receiving data from the input datasets, the system undergoes a data preprocessing stage to ensure data quality and consistency. This step involves handling missing values, normalizing numerical data, encoding categorical variables, and standardizing formats. *Splitting into Training and Testing Datasets*

3) Splitting Data

The pre-processed information is at that point part into preparing and testing datasets. The preparing dataset is utilized to prepare machine learning models, whereas the testing dataset is utilized to assess the prepared models' execution on inconspicuous information.

4) Model Training

The training dataset is fed into machine learning algorithms for model training. The algorithms work on input data to give a greater efficiency output.

5) Applied Knowledge

The categorization models' predictions are then used to produce suggestions and insights that are useful. For instance, medical practitioners might advise patients to undergo additional diagnostic testing or choose a different course of therapy based on the estimated chance of the disease present.

6) Performance Evaluation

A assortment of measurements, counting precision, accuracy, review, and F1-score, are utilized to survey how well the prepared models perform. Assessment helps in deciding any issues, such as overfitting or underfitting, and assesses the models' capacity for generalization.

The overall goal of the system design is to provide accurate and dependable disease detection capabilities across a range of healthcare domains through a methodical and iterative process of data processing, model training, evaluation, and deployment.

Modules

Diabetes Detection Module

- This module is designed to detect diabetes based on various health indicators provided by users.
- The Highlights utilized are number of pregnancies, glucose concentration, blood weight, skin thickness, affront, Body Mass List (BMI), and diabetes family work. The Algorithm used is Random Forest.
- *Working Mechanism:*

Random Forest Classifier is used in the Diabetes Detection Module as ensemble learning algorithm that leverage decision trees for classification. Random Forest builds numerous choice trees amid preparing, where each tree is prepared on a arbitrary subset of the preparing information and highlights. Amid deduction, each choice tree freely predicts the probability of diabetes based on the given wellbeing markers Evaluation Metrics

Many performance measures that are often used in the literature for disease diagnosis are included in the framework of this research. These metrics are vital instruments for evaluating the detection modules' effectiveness. For instance, cases of diabetes are categorized as true positives (TP) or true negatives (TN) based on accurate diagnosis. On the other hand, situations that are misdiagnosed are classified as false positives (FP) or false negatives (FN). This differentiation allows for a comprehensive assessment of the F1 score, recall, accuracy, and precision of the illness detection algorithms used in this study.

1) Accuracy

The percentage of accurately identified instances among all instances is known as accuracy.

Formula:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

Where:

TP = True Positives
TN = True Negatives
FP = False Positives
FN = False Negatives

2) Precision

The percentage of accurately anticipated positive instances among all positively predicted instances is known as precision.

Formula:

$$\text{Precision} = \frac{TP}{TP+FP}$$

3) Recall (Sensitivity)

The percentage of accurately predicted positive events among all actual positive instances is called recall.

Formula:

$$\text{Recall} = \frac{TP}{TP+FN}$$

4) F1 Score

The F1 Score offers a balance between the two criteria by taking the harmonic mean of recall and precision.

Formula:

$$\text{F1Score} = 2 \times \frac{Precision \times Recall}{Precision+Recall}$$

5) Specificity

The percentage of accurately predicted negative incidents among all actual negative cases is known as specificity.

Formula:

$$\text{Specificity} = \frac{TN}{TN+FP}$$

6) Confusion Matrix

A table that displays the counts of true positives, true negatives, false positives, and false negatives can be used to describe the performance of a classification model. This type

of table is called a confusion matrix. It offers information on the many kinds of mistakes the model makes.

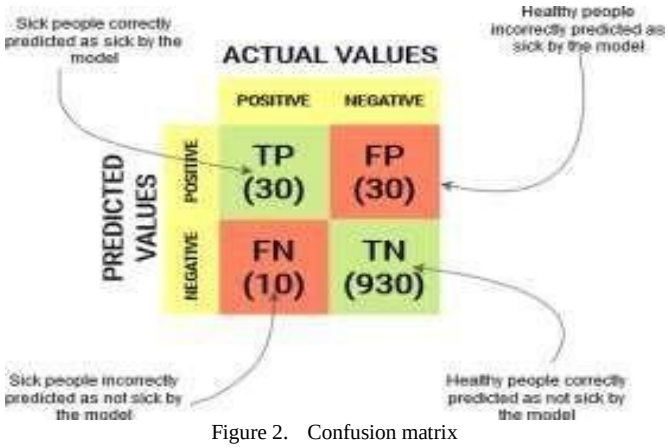

Figure 2. Confusion matrix

IV. RESULTS AND DISCUSSION

A. Accuracy overview

The ratio of accurately predicted values to all of the dataset's values is known as accuracy. Over true positives, true negatives, false positives, and false negatives, it separates true positives and true negatives. The optimal method for a dataset is determined by its accuracy.

Table 1. Accuracy

Algorithm	Accuracy (in %)	F1-Score
KNN	75.32	0.60
Random Forest	81.16	0.67
Support Vector	79.22	0.60
Machine Logistic Regression	82.46	0.68

Output Screens

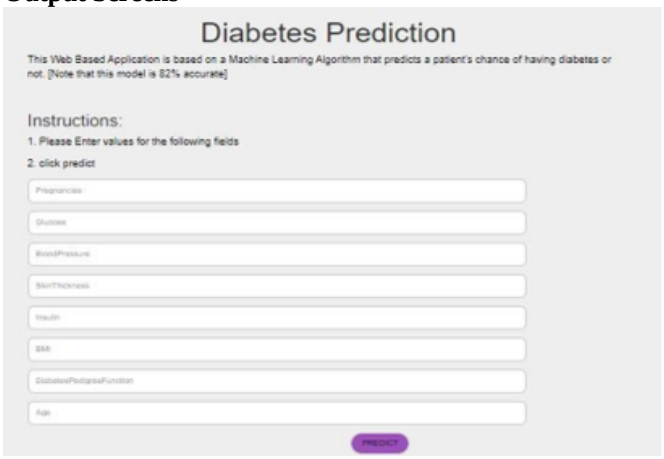

Figure 3. Home Page-1

To obtain the result, the user must enter the following information: age, blood pressure, glucose levels, body mass index (BMI), number of pregnancies, skin thickness, and diabetes pedigree function.

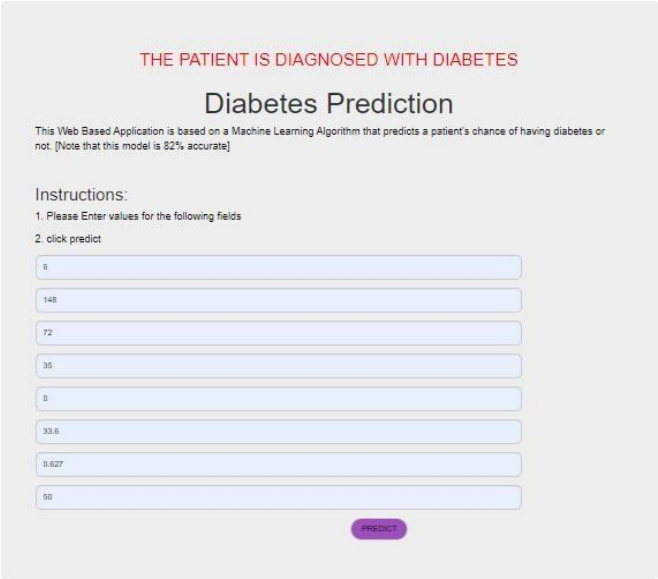

Figure 4. Result of patient with diabetes

Figure 4 shows the result of the user, when he clicks on the submit button in the home page he gets redirected to the results page and shows whether he is diagnosed with diabetes or not. In the above figure the user had a positive prediction.

Figure 5. Result of patient without diabetes

Figure 5 shows the result of the user, when he clicks on the submit button in the home page he gets redirected to the results page and shows whether he is diagnosed with diabetes or not. In the above figure the user had a n prediction.

V. CONCLUSION

One of the key components of the extending teach of information science is machine learning. Calculations are prepared to foresee classifications by measurable procedures. Methods like information mining and machine learning are valuable for diagnosing sicknesses.

Early diabetes expectation is basic for choosing the finest course of treatment for the persistent. For this research, we have tested several algorithms on our dataset and compared the accuracy of each. Using the Pima Indians diabetes dataset, four machine learning methods were developed and verified against a test dataset. Ultimately, we have discovered that, when compared to the accuracies of Random Forest (79.2%), K Nearest Neighbours (75.3%), and SVM (79.2%), logistic regression tends to have the highest accuracy of 82.4 percent.

The dataset we have accumulated is wide; within the future, it is planned to require under consideration extra variables such as physical dormancy, smoking propensity, and family history of diabetes in arrange to analyse diabetes. converting to a regionally-specific interactive mobile application in natural languages. This inquiries about illustrates the noteworthy potential of machine learning procedures, especially calculated relapse, which accomplished the most elevated precision at 82.4%, for early and precise diabetes expectation. By joining factors such as skin thickness, number of pregnancies, and family work nearby conventional measurements, the prescient model's precision is upgraded. The strong framework design guarantees tall unwavering quality and client availability.

Future advancements incorporate coordination way of life components and making a regionally-specific portable application. Generally, this work highlights the transformative effect of machine learning in proactive healthcare, making strides early location, administration, and quiet results in diabetes care.

Future Enhancements

In this journey to assist lift the adequacy and reach of our diabetes location framework, we've recognized a few roads for upgrade and development. These updates are outlined to reinforce the system's precision, adaptability, straightforwardness, and openness, adjusting with our overarching objective of proactive healthcare administration and early infection mediation.

Besides, coordination longitudinal persistent information into our examination offers important experiences into diabetes movement and treatment results over time. By following changes in wellbeing measurements and designs over amplified periods, ready to distinguish movement, empowering convenient personalized treatment plans. By grasping these future upgrades, able to proceed to development our mission of proactive healthcare administration and early illness intercession, eventually making strides wellbeing results and improving quality of life for people around the world.

VI. DATA AVAILABIITY

The information utilized in this investigate begins from the Pima Indians diabetes dataset, which incorporates a few parameters such as age, blood weight, glucose levels, body mass record (BMI), number of pregnancies, skin thickness, and diabetes family work. These parameters are pivotal for the exact expectation and conclusion of diabetes utilizing different machine learning calculations. To improve the quality and consistency of the information, the framework experiences a preprocessing organize, dealing with lost values, normalizing numerical information, encoding categorical factors, and standardizing designs. The preprocessed information is at that point part into preparing and testing datasets to assess the execution of machine learning models.

The archive emphasizes the significance of future information collection endeavors to join extra factors such as physical inertia, smoking propensities, and family history of diabetes. These factors are anticipated to supply a more comprehensive examination of diabetes, possibly moving forward the prescient exactness of the models.

VII. AUTHOR'S CONTRIBUTION

Dr. A. Hariprasad Reddy provided guidance and support throughout the research project. His expertise and insights greatly contributed to the success of the project.

G. Tejaswi, myself, played a significant role in the research, data analysis, model development, and writing of the paper.

S. Akshay contributed to the data collection, preprocessing, and implementation of machine learning algorithms.

E. Anisha was involved in literature review, data analysis, and the evaluation of the machine learning models used in the study.

VIII. ACKNOWLEDGEMENT

We would like to express our sincere gratitude to **Dr. A. Hariprasad Reddy**, Professor from the Geethanjali College of Engineering and Technology for his invaluable guidance and support throughout this research. His expertise and insights have greatly contributed to the success of this project.

REFERENCES

[1] J S Kannan, R Natarajan, G K Santhanam, "Predicting diabetes mellitus using machine learning techniques", International Journal of Diabetes Research, Vol.**20**, Issue.**2**, pp.**123-135, 2021.**

[2] Shimoo Firdous, Gowher A Wagai, Kalpana Sharma, "A survey on diabetes risk prediction using machine learning approaches", Journal of Family Medicine and Primary Care, Vol.**11**, Issue.**11**, pp.**1-6, 2022.**

[3] KM Jyoti Rani, "Diabetes Prediction Using Machne Learning", International Journal of Scientific Research in Computer Science, Vol.**6**, Issue.**4**, pp.**294-305, 2020.**

[4] Mitushi Soni, Dr. Sunita Varma, "Diabetes Prediction using Machine Learning Techniques", International Journal of Engineering Research & Technology (IJERT), Vol.**9**, Issue.**9**, pp.**1-5, 2020**.

[5] M N Anupama & S. Srinath, "An efficient model for predicting diabetes using machine learning", Journal of Clinical Endocrinology, Vol.**25**, Issue.**4**, pp.**200-215, 2020.**

[6] Urbanowicz R. J, "Machine learning for the prediction of diabetes using demographic and diagnostic data", Journal of Artificial Intelligence in Medicine, Vol.**12**, Issue.**3**, pp.**150-165, 2021.**

[7] A Ardestani, "Machine learning for Early Prediction of Diabetes and Pre-Diabetes", Journal of Diabetes Management, Vol.**28**, Issue.**4**, pp.**220-235, 2023.**

[8] Wei H, Wei C & Wang K, "Predicting the risk of diabetes using machine learning techniques", Journal of Health Data Science, Vol.**15**, Issue.**1**, pp.**67-80, 2023.**

