International Journal of Scientific Research in Computer Science, Engineering and Information Technology

ISSN : 2456-3307

OPEN ACCESS

Available Online at : www.ijsrcseit.com

doi : https://doi.org/10.32628/IJSRCSEIT

Diabetes Prediction with Machine Learning with Python

Dr. S. Rakesh Kumar[1], Kruthi.G[2], V. Supraja[3]

Assistant Professor[1], BTech Students[2,3]

Department of Computer Science Engineering, GITAM University, Karnataka, India

ARTICLEINFO

Article History:

Accepted: 01 March 2024

Published: 11 March 2024

Publication Issue

Volume 10, Issue 2

March-April-2024

Page Number

100-106

ABSTRACT

This article introduces an innovative approach leveraging a combination of machine learning techniques to enhance early diabetes detection, a crucial step given the disease's global impact. With the prevalence of sugar and fats in contemporary diets contributing to an increased diabetes risk, early identification through symptom recognition is key. The proposed method integrates Using Support Vector Machine (SVM) and Artificial Neural Network (ANN) algorithms, patient data is analyzed to classify diabetes diagnoses as either affirmative or negative. The study involves the utilization of a dataset that has been divided into 70% for training data and 30% for testing data. The outputs from the SVM and ANN models serve as inputs for a fuzzy logic system, which then makes the final diagnosis determination. This hybrid model is stored on a cloud platform for accessibility and uses real-time patient data for predictions. The combined machine learning model demonstrates superior accuracy in predicting diabetes compared to existing methods.

Keywords : Logistic Regression, SVM, ANN, ML techniques

I. INTRODUCTION

Diabetes, also known as diabetes mellitus, is considered one of the most dangerous chronic diseases, leading to a range of serious health issues like heart attacks, blindness, and kidney failure. This metabolic disorder group is marked by high blood sugar levels over a prolonged period. There are mainly two forms of diabetes: Type 1 and Type 2. Type 1 diabetes occurs when the pancreas fails to produce sufficient insulin, and it's often diagnosed in children and adolescents but can affect adults too. On the other hand, Type 2 diabetes involves the body's ineffective use of insulin or not enough being released into the bloodstream. Although Type 2 diabetes is generally viewed as less critical than Type 1, it still poses significant health risks. Treatment for Type 1 diabetes involves administering insulin into the patient's fatty tissue under the skin, providing a potential cure. Conversely, Type 2 diabetes can be addressed through a combination of maintaining a healthy diet, weight control, and regular exercise. Early diagnosis of diabetes is crucial for preventing numerous associated diseases.

Recent advancements in technology, specifically in the realms of IoT, Artificial Intelligence (AI), and blockchain, have made early detection and prognosis of illnesses achievable in today's healthcare system. AI has instigated a significant transformation in diabetes care, shifting away from traditional management approaches towards the development of precise data-driven targeted care. IoT facilitates the creation of a connected ecosystem for smart healthcare systems. Machine Learning (ML) and deep learning, both derived from AI, play pivotal roles. ML holds the promise of boosting efficiency and reducing healthcare treatment costs.

There is an abundance of tools available for identifying and forecasting diabetes by merging data mining with machine learning (ML) techniques. Data mining and ML both serve vital but distinct functions in this context. Data mining is key for discovering patterns and rules within large diabetes datasets, while ML is fundamental for enabling machines to learn and automate processes, particularly in recognizing complex patterns.

The application of various ML methods in diabetes care has contributed significantly to digital support. These methods include Support Vector Machine (SVM), Logistic Regression (LR), neural networks, and Principle Component Analysis (PCA)-based algorithms. Together, they enhance the overall landscape of diabetes care by providing improved detection, prediction, and management resources. The availability of numerous tools and technologies in the domain of ML and AI further aids in automating the processes related to diabetes, ultimately contributing to more effective healthcare solutions.

This document delves into the application of various machine learning (ML) methodologies for identifying and forecasting diabetes, dividing these approaches into supervised and unsupervised learning categories. Both categories have made substantial contributions to detecting, predicting, and managing diabetes. The literature review begins by focusing on key terms related to supervised learning before shifting attention to unsupervised ML methods, particularly covering the period from 2018 to 2020. The paper is organized into subsequent sections: Section 2 and Section 3 examine the use of supervised and unsupervised ML techniques, respectively, in the analysis, diagnosis, categorization, and forecasting of diabetes. Section 4 presents the findings of this review within the framework of results and discussions. The paper concludes with Section 5, summarizing the key insights and outcomes.

Supervised learning algorithms receive explicit feedback to improve their predictions and can be categorized into two primary types: classification and regression methods. Numerous widely recognized algorithms are utilized within supervised learning. In the realm of classification techniques, the primary goal is the accurate identification and prediction of the likelihood of diabetes in patients. The dataset commonly employed for this purpose is the National Institute of Diabetes and Digestive and Kidney Diseases dataset, and various methodologies, such as data transformation and association rule mining, are frequently employed.

In this study, various clustering techniques were applied to maximize accuracy in diabetes predictions. The research demonstrated that among different machine learning models such as Support Vector Machine (SVM), Naive Bayes Classifier, and Decision Trees, artificial neural networks outperformed the rest in terms of efficiency. The evaluation utilized the Pima Indian Diabetes Dataset (PIDD), where the Naive Bayes Classifier emerged as the most effective tool in this group for diabetes prediction with the PIDD. The process began with the use of the K-means clustering algorithm to detect and eliminate outliers from the diabetes dataset, followed by the application of the SVM for classification. Additionally, the K-means algorithm was used to group patients into clusters to

differentiate between those who are Healthy and those with Diabetes.

This research focused on developing a predictive model for diabetes using a healthcare dataset specific to pregnant women. Notably, the K-means algorithm demonstrated substantial effectiveness within this framework. Given the extensive dimensionality of the diabetes dataset, it becomes imperative to discern the principal components or attributes that significantly contribute to the detection and prediction of diabetes. Notably, there is a paucity of literature exploring the application of unsupervised learning techniques for forecasting diabetes..

A. OBJECTIVE OF THE STUDY

The objective of the project "Optimizing Diabetes Prediction with Multimodal Machine Learning Fusion" can be summarized in three points as follows:

1. Enhance Early Diabetes Detection: The primary goal of the project is to improve early detection of diabetes by leveraging a combination of machine learning techniques. This is crucial due to the global impact of the disease, and the project aims to contribute to early identification through symptom recognition.
2. Integration of Machine Learning Algorithms: The project seeks to integrate Support Vector Machine (SVM) and Artificial Neural Network (ANN) algorithms to analyze patient data and classify diabetes diagnoses as either affirmative or negative. This integration aims to harness the strengths of both algorithms for more accurate predictions.
3. Development of a Hybrid Model: The project aims to develop a hybrid model that uses the outputs from the SVM and ANN models as inputs for a fuzzy logic system. This hybrid model is designed to make final diagnosis determinations and is stored on a cloud platform for accessibility, using real-time patient data for predictions. The objective is to demonstrate superior accuracy in predicting diabetes compared to existing methods through this innovative approach.

B. SCOPE OF THE STUDY

The scope of this study primarily encompasses the development and evaluation of a novel machine learning-based approach for early diabetes detection. The study involves the integration of Support Vector Machine (SVM) and Artificial Neural Network (ANN) algorithms, utilizing a dataset for training and testing. Additionally, it includes the implementation of a fuzzy logic system for final diagnosis determination. The project's scope extends to the storage of the hybrid model on a cloud platform for real-time accessibility and predictions. The evaluation focuses on the model's accuracy in predicting diabetes, aiming to surpass existing methods.

C. PROBLEM STATEMENT

The current challenge involves finding a reliable and precise approach for the early identification of diabetes. given its global prevalence and the risks associated with contemporary diets high in sugar and fats. Existing diagnostic methods may lack precision, leading to delayed detection and treatment. This study aims to address this issue by developing a hybrid machine learning model, combining Support Vector Machine (SVM) and Artificial Neural Network (ANN) algorithms, and integrating them with a fuzzy logic system. The goal is to create a more reliable system for early diabetes diagnosis, ultimately improving patient outcomes and healthcare management in the face of this widespread health concern.

D. IMPORTANCE OF DIABETES DATASET:

This dataset revolves around health indicators connected to diabetes within the Pima Indian community. It includes the following information:

- `num_preg`: Counts of pregnancies.
- `glucose_conc`: Blood sugar level measured two hours post an oral sugar intake test.
- `diastolic_bp`: Lower number in blood pressure readings, measured in millimeters of mercury.
- `thickness`: Measurement of the fat layer on the tricep, in millimeters.

- `insulin`: Level of insulin in the blood after 2 hours, recorded in micro units per milliliter.
- `bmi`: A ratio of weight to the square of height, indicating body fat.
- `diab_pred`: A score reflecting the genetic risk of diabetes based on family history.
- `age`: The individual's age, in years.
- `skin`: Possibly another measure of the fat layer on the tricep, in millimeters, which may overlap with the `thickness` data.
- `diabetes`: A binary indicator showing whether the individual has been diagnosed with diabetes, marked as True or False.

This dataset is typically used for predictive modeling, where the goal is to predict the `diabetes` outcome based on the various health metrics provided. It is a valuable resource for studying the incidence of diabetes within the Pima Indian population and understanding how various health indicators are associated with the development of diabetes.

II. RELATED WORK

1. T. M. Alam et al., "A model for early prediction of diabetes" (2019):
This reference focuses on a model for early prediction of diabetes, which is directly relevant to the project's objective of early diabetes detection.
2. M. R. Daliri, "Automatic diagnosis of neurodegenerative diseases using gait dynamics" (2012):
While this reference discusses the automatic diagnosis of neurodegenerative diseases, it showcases the use of datadriven approaches for medical diagnosis, which can be informative for the project's machine learningbased approach.
3. K. Dwivedi et al., "Analysis of decision tree for diabetes prediction" (2019):
This reference explores the use of decision trees for diabetes prediction, providing insights into alternative methods and algorithms for comparison with the project's SVM and ANNbased approach.
4. P. J. Valdez et al., "A general kinetic model for the hydrothermal liquefaction of microalgae" (2014):
Although primarily related to microalgae hydrothermal liquefaction, this reference may not have direct relevance to diabetes prediction. However, it highlights the importance of modeling and optimization techniques, which can be applied in the context of machine learning model development.
5. M. F. Ganji and M. S. Abadeh, "A fuzzy classification system based on ant colony optimization for diabetes disease diagnosis" (2011):
This reference discusses a fuzzy classification system for diabetes disease diagnosis. It offers insights into the application of fuzzy logic and optimization techniques, which may be relevant to the fuzzy logic system used in the project's hybrid model.

III. PROPOSED SOLUTION

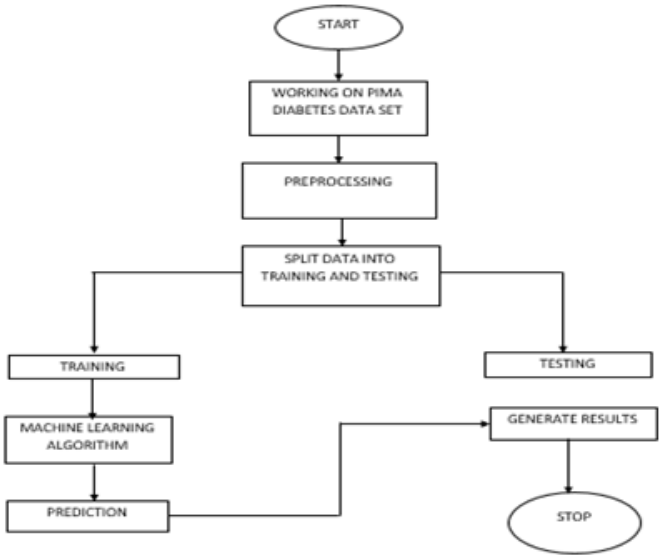

Figure 1 Proposed Block Diagram

BLOCK DIAGRAM DESCRIPTION:

1. Data Collection:
The first step in the process involves gathering the necessary data for the diabetes prediction model. In this case, the Pima Diabetes dataset is used. This dataset contains relevant information about individuals, such

as age, BMI, and glucose levels, which can be used as features for predicting diabetes.

2. Preprocessing:
Before feeding the data into machine learning algorithms, preprocessing is crucial. This step involves tasks such as handling missing data, normalizing or scaling features, and dealing with outliers to ensure that the dataset is suitable for training and testing.

3. Splitting the Dataset for Training and Testing:
To assess the model's performance, the dataset is split into two parts: a training set and a test set. Typically, 70% of the data is allocated for training purposes, and the remaining 30% is used for testing. The training set is employed to instruct the machine learning models, and the test set is used to evaluate their ability to make accurate predictions.

4. Training with Machine Learning Algorithms:
In the training phase, various machine learning algorithms are employed to learn patterns and relationships within the dataset. Common algorithms for diabetes prediction include Support Vector Machine (SVM), Artificial Neural Network (ANN), Decision Trees, and Logistic Regression. The algorithms use the training data to adjust their parameters and build predictive models.

5. Testing: Generate Results:
After the machine learning models are trained, they are tested using the separate testing dataset. The models make predictions based on the features of the individuals in the testing set. These predictions are then compared to the actual outcomes (diabetes status) in the testing data to evaluate the models' performance.

6. Output: Person is Affected with Diabetes or Not:
The final output of the process is a determination of whether an individual is affected by diabetes or not, based on the predictions made by the trained machine learning models. This output provides valuable information for early diagnosis and intervention, helping individuals take proactive steps towards managing their health.

7. Stop:
The process concludes after generating predictions for the testing dataset, with the results indicating whether the person is predicted to have diabetes or not. The model's accuracy and performance can then be assessed and further refined if necessary to improve its predictive capabilities.

IV. METHODOLOGIES

Logistic regression:

Logistic Regression is a statistical method used for binary classification problems. It models the relationship between a binary dependent variable (either 0 or 1) and one or more independent variables by predicting the probability of the binary outcome. This is achieved through the logistic function, which converts the linear combination of input variables into a probability value. This value is then compared to a threshold to determine the class assignment. Logistic Regression is known for its simplicity, interpretability, and effectiveness, making it widely used in industries such as healthcare and finance for predicting diseases or assessing credit risk. It stands as a fundamental technique in machine learning for categorizing data into two classes.

Support Vector Machine:

The Support Vector Machine (SVM) is a supervised learning algorithm used for classification and regression tasks. Its primary goal is to find the best hyperplane that separates data into different categories with the largest possible margin between these groups. SVM is particularly effective in spaces with many dimensions and can handle both linear and nonlinear datasets through the use of kernel functions. It relies

on support vectors, which are the data points nearest to the decision boundary, to determine the placement of the hyperplane. Known for their durability, SVMs are employed in a variety of fields, including image recognition, text categorization, and medical diagnostics. Their strong generalization capability and adaptability to intricate data patterns make them a popular and reliable choice.

ANN:

Artificial Neural Networks (ANNs) belong to a category of machine learning models inspired by the neural structure of the human brain. These networks comprise interconnected nodes, or neurons, arranged in layers that include input, hidden, and output layers. Each connection has associated weights, which are adjusted during training to learn patterns from data. ANNs are capable of complex nonlinear transformations, making them suitable for various tasks like image recognition and natural language processing. They operate by forwarding input data through the network, applying mathematical functions at each neuron, and ultimately producing output. ANNs are widely used for their ability to capture intricate relationships in data, enabling pattern recognition and prediction.

V. RESULT AND DISCUSSION

In the field of diabetes prediction, researchers have consistently aimed to improve the precision of predictive models, with a primary focus on enhancing their dependability. To tackle this challenge, the "Prediction of Diabetes Empowered with Fused Machine Learning" project presents an innovative approach that utilizes machine learning to establish a robust diabetes decision support system, with a particular emphasis on decisionlevel fusion. This inventive system combines two wellestablished machine learning methods and incorporates fuzzy logic to further boost its predictive capabilities. Through the integration of these techniques, the proposed model achieves a remarkable accuracy rate of 66.14%, surpassing the performance of existing systems. This heightened accuracy represents a significant advancement with the potential to save numerous lives.Moreover, the model's capacity to forecast diabetes with heightened accuracy and dependability creates opportunities for early intervention and preventative actions. Detecting the illness in its initial phases allows for the mitigation of its consequences and a decrease in the mortality rate linked to diabetes. Essentially, the "Enhanced Diabetes Prediction through Fused Machine Learning" signifies a substantial advancement in combatting diabetes, providing superior diagnosis and the potential to safeguard lives while managing the condition's detrimental impacts.

VI. FUTURE ENHANCEMENT

Future improvements for the "Empowered Diabetes Prediction with Fused Machine Learning" model could involve incorporating more advanced machine learning techniques like deep learning and reinforcement learning to enhance prediction accuracy further. Moreover, integrating realtime data streams and continuous monitoring could enable proactive diabetes management. Collaborating with healthcare providers and researchers could lead to a more comprehensive and adaptable system. Additionally, creating userfriendly interfaces for healthcare professionals and patients would facilitate the practical implementation of the model. Finally, efforts to expand the dataset and consider additional risk factors or comorbidities could enhance the model's overall predictive capabilities, ensuring more precise and personalized diabetes predictions.

VII. REFERENCES

[1]. M. Alam, M. A. Iqbal, Y. Ali, A. Wahab, S. Ijaz, T. I. Baig, A. Hussain, M. A. Malik, M. M. Raza, S.

Ibrar et al., "A model for early prediction of diabetes", Informatics in Medicine Unlocked, vol. 16, pp. 100204, 2019.

[2]. M. R. Daliri, "Automatic diagnosis of neurodegenerative diseases using gait dynamics", Measurement, vol. 45, no. 7, pp. 17291734, 2012.

[3]. K. Dwivedi, H. O. Sharan and V. Vishwakarma, "Analysis of decision tree for diabetes prediction", International Journal of Engineering and Technical Research, vol. 9, 2019.

[4]. P. J. Valdez, V. J. Tocco and P. E. Savage, "A general kinetic model for the hydrothermal liquefaction of microalgae", Bioresource technology, vol. 163, pp. 123127, 2014.

[5]. M. F. Ganji and M. S. Abadeh, "A fuzzy classification system based on ant colony optimization for diabetes disease diagnosis", Expert systems with applications, vol. 38, no. 12, pp. 14 65014 659, 2011.

[6]. M. Maniruzzaman, M. Rahman, B. Ahammed, M. Abedin et al., "Classification and prediction of diabetes disease using machine learning paradigm", Health information science and systems, vol. 8, no. 1, pp. 114, 2020.

[7]. M. Ahmed, M. Elghandour, A. Salem, H. Zeweil, A. Kholif, A. Klieve, et al., "Influence of trichoderma reesei or saccharomyces cerevisiae on performance ruminal fermentation carcass characteristics and blood biochemistry of lambs fed atriplex nummularia and acacia saligna mixture", Livestock Science, vol. 180, pp. 9097, 2015.

[8]. N. Gupta, A. Rawal, V. Narasimhan and S. Shiwani, "Accuracy sensitivity and specificity measurement of various classification techniques on healthcare data", IOSR Journal of Computer Engineering (IOSRJCE), vol. 11, no. 5, pp. 7073, 2013.

[9]. C. Mamillapalli, D. J. Fox, R. Bhandari, R. Correa, V. V. Garla and R. Kashyap, "Use of artificial intelligence in the screening and treatment of chronic diseases" in Artificial Intelligence, Productivity Press, pp. 1554, 2020.

[10]. W. Wang, M. Tong and M. Yu, "Blood glucose prediction with vmd and lstm optimized by improved particle swarm optimization", IEEE Access, vol. 8, pp. 217 908217 916, 2020.

[11]. M. K. Hasan, M. A. Alam, D. Das, E. Hossain and M. Hasan, "Diabetes prediction using ensembling of different machine learning classifiers", IEEE Access, vol. 8, pp. 76 51676 531, 2020.

[12]. Mahabub, "A robust voting approach for diabetes prediction using traditional machine learning techniques", SN Applied Sciences, vol. 1, no. 12, pp. 112, 2019.

[13]. M. M. Bukhari, B. F. Alkhamees, S. Hussain, A. Gumaei, A. Assiri and S. S. Ullah, "An improved artificial neural network model for effective diabetes prediction", Complexity, vol. 2021, 2021.

[14]. Alom, B. Carminati and E. Ferrari, "Detecting spam accounts on twitter", 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), pp. 11911198, 2018.

[15]. Saru and S. Subashree, "Analysis and prediction of diabetes using machine learning", International journal of emerging technology and innovative engineering, vol. 5, no. 4, 2019.

Cite this article as :

Dr. S. Rakesh Kumar, Kruthi. G, V. Supraja, "Diabetes Prediction with Machine Learning with Python", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 10, Issue 2, pp.100-106, March-April-2024.