2024 3rd International Conference for Innovation in Technology (INOCON)
Karnataka, India. Mar 1-3, 2024

2024 3rd International Conference for Innovation in Technology (INOCON) | 979-8-3503-8193-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/INOCON60754.2024.10511558

Machine Learning Approach for Diabetes Prediction using Genetic Algorithm based Feature selection

Dr T.S.Ravi Kiran
Associate Professor,
Department of Computer Science
P.B. Siddhartha College of Arts & Science Technology
Vijayawada, A.P, India
tsravikiran@pbsiddhartha.ac.in

A. Srisaila
Assistant professor,
Department of Information Technology
V.R Siddhartha Engineering College
Vijayawada, A.P, India
a.srisaila@vrsiddhartha.ac.in

G. Siva Shankar
Associate Professor,
Department of AI&ML
Aditya Engineering College
Surampalem,India
sivacse517@gmail.com

Bodasingi Sowjanya
Assistant professor,
Department of CSE
Centurion University of technology and Management
Vizianagaram,Andhra Pradesh
sowjanyabodasingi@gmail.com

A.Lakshmanarao
Associate professor,
Department of IT
Aditya Engineering College
Surampalem,India
laxman1216@gmail.com

***Abstract*— Diabetes, a prevalent and complex medical condition, demands accurate predictive models for early detection and effective management. This paper introduces a novel approach for diabetes prediction by combining genetic algorithm-based feature selection with ML classification. By combining the genetic algorithm's capability to optimize feature selection and the predictive power of ML classifier, this work offers a promising avenue for improving diabetes prediction accuracy. Two datasets from Kaggle were collected. Initially, RF applied on both datasets. Later, datasets ae balanced using oversampling technique "ADASYN". Later, genetic algorithm is employed to optimize feature selection, with the fitness function minimizing the negative accuracy of the model. The selected features are then used to train a final model, and the accuracy is evaluated on the test set. The results showcase the effectiveness of the proposed approach in enhancing diabetes prediction accuracy when compared to base model. Results from both datasets shown accuracy enhancement with GA feature selection. The selected features provide valuable insights into the influential factors contributing to diabetes outcomes.**

Keywords*— *Diabetes prediction, Genetic Algorithm, Machine Learning, Kaggle.

I. Introduction

Diabetes is a global health concern with a significant impact on individuals, healthcare systems, and society as a whole. Diabetes is a chronic metabolic disorder that reasons high glucose levels. This condition arises either due to the insufficient production of insulin by the pancreas or the body's inability to effectively use the insulin it produces. There are primarily two categories of diabetes, type 1 Diabetes results from the immune system mistakenly attacking and destroying the insulin-producing beta cells in the pancreas. Management of blood sugar requires insulin shots for Type 1 diabetes. Type 2 Diabetes, more frequent, arises when the body becomes insulin-resistant or the pancreas doesn't generate enough insulin. Poor nutrition, inactivity, and obesity may cause type 2 diabetes.

The significances of uncontrolled diabetes can be dangerous and comprise various health difficulties, such as cardiovascular diseases, kidney problems, nerve damage, and eye issues. Early detection and effective management of diabetes are critical in preventing difficulties and refining the worth of life for persons affected by this condition. Diabetes has reached epidemic proportions worldwide, affecting millions of people across diverse population.

Diabetes is a growing health concern in India, with an increasing number of reported cases. The prevalence of diabetes in the country has risen significantly over the years, posing a major public health challenge. Factors contributing to this rise include changes in lifestyle, dietary habits, and an overall increase in sedentary behavior. The adoption of urbanized living, characterized by less physical activity and the consumption of processed foods, has been linked to the escalating diabetes epidemic. In India, 77 million individuals who are 18 years of age or older have type 2 diabetes, while 25 million people are prediabetic. This suggests that the prevalence of diabetes is high. It is concerning that more than 50% of people with diabetes are ignorant about their illness since this increases the risk of complications. Cardiovascular problems are more likely to occur in people with diabetes, sometimes twice or three times as often. Heart attacks and strokes are among these problems. Diabetic retinopathy, the primary cause of blindness, and neuropathy, which may result in foot sores and even amputation, are among the side effects. Moreover, hyperglycemia is a primary cause of renal failure. It is important to identify and treat this prevalent health issue in India as soon as possible. The impact of diabetes extends beyond the immediate health implications, affecting individuals economically and burdening the healthcare system. Uncontrolled diabetes can move to various difficulties such as cardiovascular diseases, kidney failure, neuropathy, and vision impairment. Moreover, a substantial portion of the diabetes population in India remains undiagnosed, emphasizing the need for awareness.

Traditional approaches to diabetes prediction often rely on clinical markers and risk factors, but advancements in machine learning (ML) provide an opportunity to harness the power of data-driven predictive models. In the context of this research, the focus is on leveraging advanced machine learning techniques, specifically genetic algorithm-based

Authorized licensed use limited to: NIIT University Alwar. Downloaded on January 08,2025 at 13:55:12 UTC from IEEE Xplore. Restrictions apply.

feature selection and classification models, to enhance the prediction of diabetes. By harnessing the power of data and innovative algorithms, the goal is to contribute to more accurate and personalized approaches to diabetic risk assessment and management. In this work, genetic algorithm approach used for feature selection. Later, ML classifier applied for the selected features.

A. Genetic algorithm approach

The key idea of GA is towards creating a population of potential solutions, evolve them over multiple generations, and iteratively improve the solutions through genetic operators such as selection, crossover, and mutation. Genetic Algorithms (GAs) begin by creating a population of potential solutions, where each solution is represented as a set of parameters. This initial population is randomly generated. Genetic Algorithms (GAs) start by creating a population of potential solutions with diverse parameter sets. These solutions are randomly generated to form the initial population. During the evaluation phase, each solution's performance is assessed using a fitness function, assigning scores based on problem-solving effectiveness. In the selection phase, solutions with higher fitness scores are preferentially chosen for reproduction, following the principle of "survival of the fittest." Crossover, the next step, involves exchanging portions of genetic information between selected solutions, simulating genetic recombination and generating new offspring. Mutation introduces random changes to the genetic information of some solutions, fostering diversity and exploring alternative possibilities. Less fit solutions are then replaced with new offspring, ensuring a higher persistence chance for better solutions in the next generation. The algorithm iterates through these steps for a defined number of generations or until specific termination criteria are met, such as achieving a target fitness level or reaching a maximum number of generations. Upon completion, the best solution or a set of high-performing solutions is extracted from the final population, representing the optimal or near-optimal solution to the problem. GAs excels in optimization problems, seeking the ideal combination of parameters or features to maximize or minimize a specific objective function.

B. Need for genetic algorithm approach in feature selection

Feature selection is an important step in ML model development, aiming to identify the most relevant features that contribute to accurate predictions. In the context of diabetes prediction using a Random Forest model, the dataset likely contains numerous features, and not all of them may be equally informative. The Genetic Algorithm (GA) is employed in this approach to address specific challenges and enhance the overall effectiveness of the feature selection process. The dataset may have a high-dimensional feature space, making it computationally expensive and time-consuming to exhaustively search through all possible feature combinations. The GA efficiently explores this large solution space, considering various combinations of features, and converges towards solutions that are likely to contribute to accurate predictions. Feature selection involves choosing subsets of features, resulting in a combinatorial optimization problem. The number of possible feature combinations grows exponentially with the number of features. The GA's ability to work with binary variables allows it to handle combinatorial aspects effectively, exploring different combinations of features in a systematic and adaptive manner. The complexity of feature selection varies across datasets. The GA's adaptability allows it to handle diverse problem complexities, providing a robust and scalable approach for identifying relevant features in the context of diabetes prediction.

II. LITERATURE SURVEY

Applying ML techniques to health area is not new era [1]. Several authors applied ML models for diabetes detection. In Boon Feng Wee et al. [2] evaluated the impact of using various approaches to oversampling and feature selection on the precision of diabetes diagnosis. Their work provided ideas for potential future directions to enhance accuracy and reliability through the use of more complex feature selection algorithms. The authors in [3] used DT, SVMs, RF, Logistic Regression, KNN techniques to diabetes prediction. The recommended approach was trained and tested on all classification models, although the ADASYN-trained XGBoost classifier performed best. This classifier has 82% accuracy, 0.82, F1 coefficient, and 0.85 AUC. To demonstrate system adaptability, a domain adaptation approach was developed. Explainable AI using the LIME and SHAP frameworks was introduced to better understand how the model predicts. Finally, a website framework and Android smartphone app for rapid diabetes prediction using several input factors were created.

The primary goal of [4] was to use several ML techniques to predict the potential early development of diabetes, particularly in females. Early detection is essential for both preventing the development of the condition and mitigating its severe consequences, such as heart and kidney issues. The significance of having a tool to assist doctors in identifying diabetes in its early stages was emphasized. The model obtained an accuracy of 83% using RF model. Five ML models were tested to predict diabetes in patients by [5]. Pregnancies, glucose concentration, blood pressure, skinfold thinness, insulin levels, BMI, genetic background, family history, age, and result (with or without diabetes) were used from the Kaggle Pima Indian dataset. 768 diabetic and non-diabetic individuals participated. The K-NN and BNB models fared best. K-NN had 80% diabetes diagnostic accuracy, outperforming BNB's 77.3%. The research suggests machine learning systems may detect diabetes early. In [6], patients' diabetes is predicted using logistic regression based on diagnostic metrics from databases. The Vanderbilt and PIMA Indians Diabetes databases are consulted. Performance is enhanced via stacking and maximum voting, which are the two methods used to choose features. With highest Voting on dataset 1 and Maximum Voting with stacking on dataset 2, the highest accuracy is 78% and 93%, respectively. The research demonstrates how feature selection, normalization, and logistic regression may enhance model accuracy and runtime.

After comparing the models in [7], LightGBM was selected as the principal model because of its excellent accuracy. After that, Hyper settings were modified for maximum performance. The findings show that machine learning algorithms can effectively diagnose diabetes and uncover its causes. This work also reveals disease progression variables. This study provides the groundwork for future ML based diabetes prediction. Several ML models have been implemented in [8] to identify

Authorized licensed use limited to: NIIT University Alwar. Downloaded on January 08,2025 at 13:55:12 UTC from IEEE Xplore. Restrictions apply.

patterns and risk factors in diabetes dataset with the use of a Python data manipulation tool. Six ML techniques have been applied to categorize the patient as either diabetes or non-diabetic: SVM, KNN, gradient boosting, DT, RF and logistic regression. The authors of [9] examined healthcare prediction analytics and used four different ML approaches to address issues. This study made use of binary and early Detection databases. These datasets were used to calculate the accuracy, precision, and recall of KNNs and RF approaches. Their findings would have been beneficial to stakeholders, medical professionals, researchers working on diabetes prediction research. In their work, SVM beat KNN and Logistic Regression. The model described in [10] turned the diabetes prediction challenge into a classification issue by hiding layers of a deep NN model with dropouts to avoid overfitting. Through parameter tuning and the use of binary cross-entropy as the error function, a DNN prediction model with exceptional accuracy was produced. A training accuracy of 98.07% was achieved using the Pima Indians diabetes dataset, demonstrating the efficacy and appropriateness of recommended method, according to experiments. Comprehensive testing on the diabetes and diabetic type databases of the Pima Indians shows that the proposed model is better than current methods.

III. METHODOLOGY

The proposed approach is signified by the model shown in Figure. 1. It employs a hybrid approach, combining machine learning with a genetic algorithm to optimize feature selection and enhance the accuracy of diabetes prediction. Initially, two diabetes datasets are collected from Kaggle. The datasets are then split into training and testing sets to facilitate model evaluation.

To ensure uniformity in feature representation, the features are standardized using the StandardScaler. The genetic algorithm (GA) is set up, utilizing the Random Forest classifier as the machine learning model. A fitness function is designed for the Genetic Algorithm, aiming to minimize the negative accuracy by selecting pertinent features. Binary variables are employed to represent feature selection in the Genetic Algorithm.

The Genetic Algorithm is executed to identify the optimal set of features for enhanced prediction. The selected features are then extracted based on the results of the Genetic Algorithm. Subsequently, the final model is trained using the chosen features, and its accuracy is evaluated on the test set. The results of the model, including the final accuracy, are displayed. Additionally, the list of selected features contributing to the improved accuracy is presented. This integrated approach provides a comprehensive strategy for refining diabetes prediction models, emphasizing feature relevance and interpretability. The proposed method applied on two datasets and achieved good results.

IV. EXPERIMENTS AND RESULTS

A. Details of the dataset

Two datasets are gathered from Kaggle. The features of two datasets are shown in Table I. There are 8 input features in both the datasets. The target variable in DD-1(Diabetic Dataset-1) is “Outcome”. The target variable in DD-2(Diabetic Dataset-2) is “diabetes”.

B. Checking Imbalance in Datasets

Total number of rows in DD-1 is 768. Out of this, 500 rows belong to “no diabetes” class and 268 belongs to “yes diabetes” class. The number of rows in DD-2 is 1,00000. Out of this, 8500 represents “yes diabetes” class and 91,500 rows represent “no diabetes” class. So, both the datasets ae imbalanced. The class distribution of DD-1 and DD-2 are shown in Fig.2 and Fig.3 respectively. So, there is a need to handle imbalance issue. To handle imbalance, we applied “ADASYN” over sampling technique.

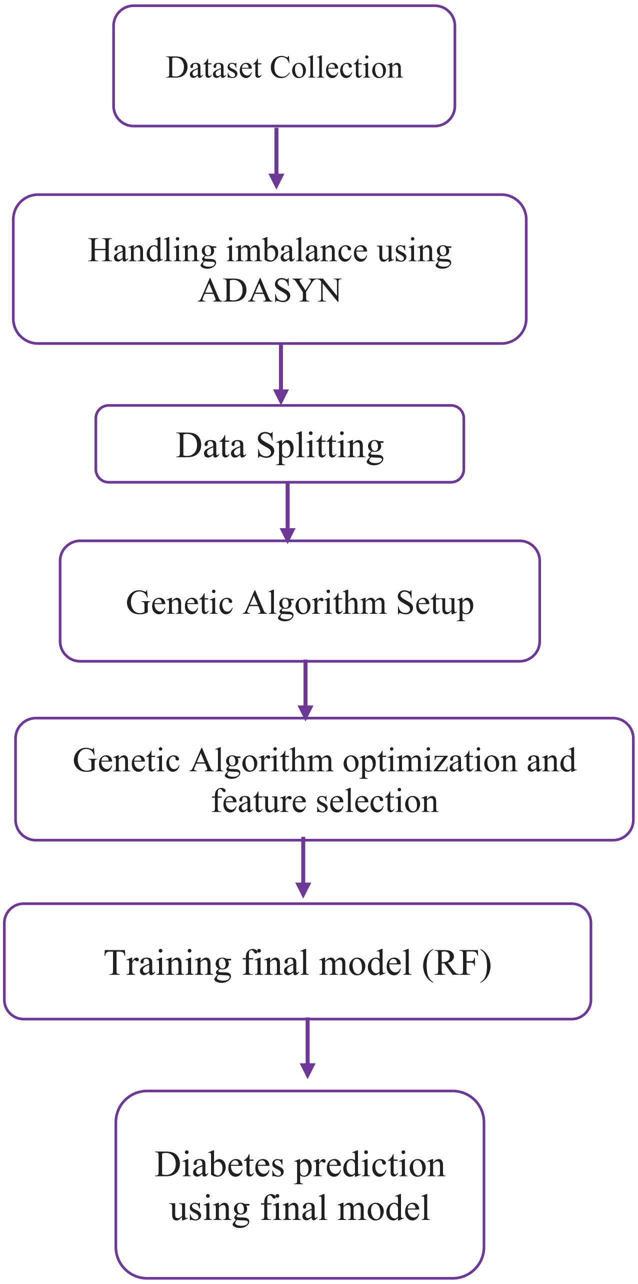
Dataset Collection
Handling imbalance using ADASYN
Data Splitting
Genetic Algorithm Setup
Genetic Algorithm optimization and feature selection
Training final model (RF)
Diabetes prediction using final model

Fig. 1. Proposed Model

Authorized licensed use limited to: NIIT University Alwar. Downloaded on January 08,2025 at 13:55:12 UTC from IEEE Xplore. Restrictions apply.

TABLE I. DATASET FEATURES

DD-1 Features	DD-2 Features
Pregnancies	age
Glucose	Gender
BloodPressure	BMI
DiabetesPedigreeFunction	Hypertension
Skin Thickness	heart disease
Insulin	smoking history
BMI	blood_glucose
Age	HbA1c_level
Outcome	diabetes

“CNN is a powerful classifier for images. So, we first applied Convnets for skin cancer detection. The proposed CNN consists of the dataset

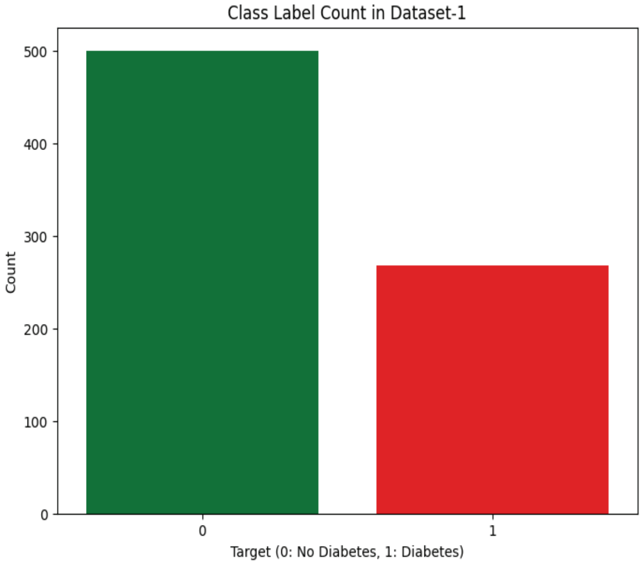

Fig. 2. Class distribution in dataset-1

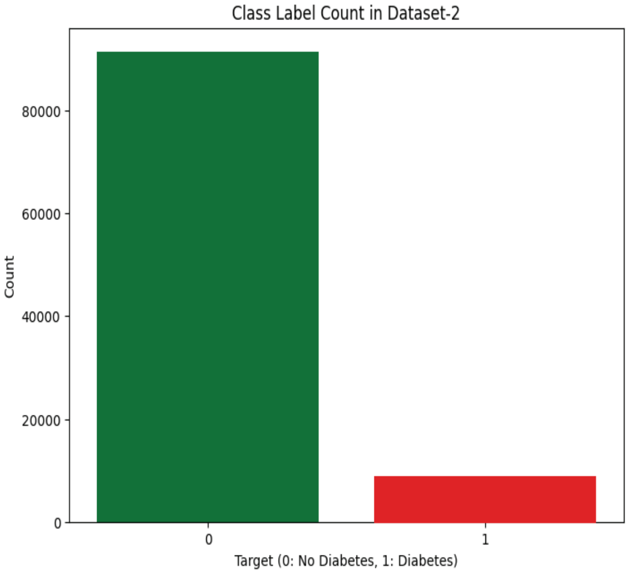

Fig. 3. Class distribution in dataset-2

C. Apply ADASYN for resolving data imbalance issue

In this step, both DD-1 and DD-2 are experimented with ADASYN technique. It is a technique to balance number of samples for both classes. After applying this technique, the number of class 0 samples in DD-1 are 500 and number of class1 samples is 474. Similarly, the number of class 0 samples in DD-2 is 91,500 and the number of class1 samples is 92,193. Now the dataset is balanced.

D. Data Splitting

In this step, both DD-1 and DD-2 are divided as two parts training and testing samples. The number of training and testing samples information is given in Table II.

TABLE II. DATASET DIVISION

Dataset	Train samples	Test samples
DD-1	779	195
DD-2	1,83,693	36,739

E. Genetic Algorithm setup

In the initial setup for the Genetic Algorithm (GA), a ML model, specifically Random Forest, is defined as the base model for subsequent feature selection. A fitness function is e The heart of the Genetic Algorithm (GA) lies in the definition of the fitness function. This function acts as a guide for the GA, providing a measure of the quality of a particular set of features. The function takes a set of binary weights as input, indicating which features are selected (1) or not selected (0). Inside the function, the selected features are used to train the Random Forest model on the training set. The negative accuracy on the test set is then computed. The negative accuracy is employed because the GA aims to minimize this fitness function, and a lower accuracy (negative) corresponds to better feature selection. To facilitate the Genetic Algorithm, certain configurations are specified. It defines the bounds for each variable (feature selection) as binary values (0 or 1). Additionally, the algorithm variable is initialized using the geneticalgorithm library. It incorporates the defined fitness function, the dimensionality of the problem (number of features), the variable type (binary), and the variable boundaries. These configurations set the stage for the GA to explore the solution space effectively.

F. Genetic Algorithm Optimization

The Genetic Algorithm is executed by calling run(). This triggers the iterative optimization process where the GA explores various combinations of features to find the set that minimizes the negative accuracy. The algorithm iteratively refines its solutions over multiple generations, mimicking the principles of natural selection. The goal is to converge towards a set of features that enhances the Random Forest model's predictive performance for diabetes.

G. Feature Selection

Later, the algorithm is executed to identify the optimal set of features. Later, selected features are extracted the based on the GA results. The genetic algorithm can be applied through python “geneticalgorithm” package [13]. .The features selected in the two datasets is shown in Table III.

Authorized licensed use limited to: NIIT University Alwar. Downloaded on January 08,2025 at 13:55:12 UTC from IEEE Xplore. Restrictions apply.

TABLE III. FEATURE SELECTION THROUGH GA

Dataset	Features Selected
DD-1	Pregnancies, Glucose, BloodPressure, Insulin, Age
DD-2	Age,, Gender, Hypertension heart disease, smoking history, blood_glucose

H. *Model Training and Evaluation*

The Random Forest model is trained the features nominated by the Genetic Algorithm. The accuracy of the model is assessed, and the results are recorded, including the final model accuracy and the selected features. The accuracy achieved with the proposed model was shown in the Table IV and Fig.4. To compare the performance of the model, initially RF is applied without feature selection process.

TABLE IV. ACCURACY COMPARISON

Dataset	Accuracy with RF without GA	Accuracy with RF Feature Selection by GA
DD-1	84.5%	90.3%
DD-2	94.5%	97.6%

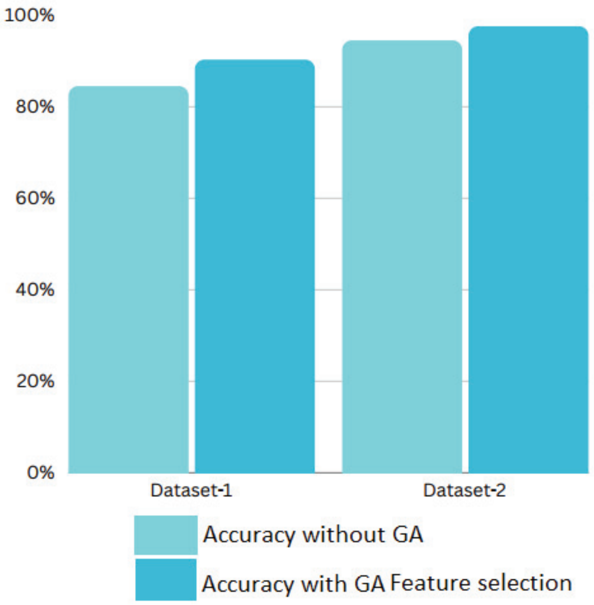

Fig. 4. Accuracy comparison with and without GA feature seletion

V. CONCLUSION

This paper proposed a strategy for diabetes prediction, combining genetic algorithm-based feature selection with machine learning classification. This fusion of genetic algorithm optimization and machine learning modeling presents a promising avenue for advancing the accuracy of diabetic prediction. Two Kaggle datasets are collected. Later As number of class labels in both datasets are imbalance, an oversampling technique “ADASYN” applied to make class labels equal. Later data is divided into training and testing sets. Initially, random forest applied for both datasets. Later, a genetic algorithm applied for both the datasets and best features extracted. By using these best features, a random forest applied and achieved good accuracy when compared to base model. Beyond accuracy enhancement, the model identifies key features crucial for diabetic prediction, contributing to a comprehensive understanding of the disease's determinants.

REFERENCES

[1] A. Lakshmanarao et al., ”Heart Disease Prediction using Feature Selection and Ensemble Learning Techniques,” International Conference on Intelligent Communication Technologies and Virtual Mobile Networks, Tirunelveli, India, 2021.

[2] J. Avanija et al., “Skin Cancer Detection using Ensemble Learning,” ICSCSS, Coimbatore, India, 2023, pp. 184-189.

[3] I. Tasin et al., “Diabetes prediction using machine learning and explainable AI techniques,” Healthcare Technology Letters, vol. 10, no. 1–2. Institution of Engineering and Technology, Dec, 2022.

[4] N. Abdulhadi et al., “Diabetes Detection Using Machine Learning Classification Methods,” International Conference on Information Technology , Amman, Jordan, 2021.

[5] O. Iparraguirre-Villanueva et al., “Application of Machine Learning Models for Early Detection and Accurate Classification of Type 2 Diabetes,” Diagnostics, vol. 13, no. 14. MDPI AG, p. 2383, Jul. 15, 2023.

[6] V, Viswanatha et al., “Diabetes Prediction Using Machine Learning Approach,” Strad Research, Vol 10, Issue 8, August 7,2023.

[7] E. Daniel et al., “An Efficient Diabetes Prediction Model using Machine Learning,”vInternational Conference on Electronics and Sustainable Communication Systems, Coimbatore, India, 2023.

[8] R. Katarya and S. Jain, “Comparison of Different Machine Learning Models for diabetes detection,” International Conference on Advances and Developments in Electrical and Electronics Engineering”, Coimbatore, India, 2020.

[9] H.N.Lakshmi et al.,”Analysis of Diabetic Prediction Using Machine Learning Algorithms on BRFSS Dataset,” International Conference:ICTEI,Tirunelveli, India, 2023.

[10] R. Bhargava and J. Dinesh, “Deep Learning based System Design for Diabetes Prediction,” International Conference on Smart Generation Computing, Communication and Networking, Pune, India, 2021.

[11] https://www.kaggle.com/datasets/mathchi/diabetes-data-set.

[12] https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset/data

[13] https://pypi.org/project/geneticalgorithm/

Authorized licensed use limited to: NIIT University Alwar. Downloaded on January 08,2025 at 13:55:12 UTC from IEEE Xplore. Restrictions apply.