Data-Driven Approach for Diabetes Detection and Insulin Dosage Estimation using Hybrid Models

K. Baby Ramya [1], K. Pavani[2], R. Pallavi[3]

#1 Assistant Professor in the Department of MCA, SRK Institute of Technology, Vijayawada.

#2 Assistant Professor & Head of the Department of MCA, SRK Institute of Technology, Vijayawada.

#3 Student in the Department of MCA, SRK Institute of Technology, Vijayawada.

Abstract: Diabetes mellitus is a chronic metabolic disorder characterized by elevated blood glucose levels due to insufficient insulin production or ineffective insulin utilization. Accurate prediction of diabetes and proper insulin dosage management are essential to prevent severe complications such as cardiovascular diseases, kidney failure, and nerve damage. In this paper, a hybrid machine learning model is proposed that integrates a Gradient Boosting Classifier for diabetes prediction and a Linear Regression model for insulin dosage estimation. The system utilizes the PIMA Indian Diabetes dataset for classification and the UCI insulin dosage dataset for regression analysis. Initially, the data is preprocessed to handle missing values and improve model performance. The Gradient Boosting algorithm identifies whether a patient is diabetic or non-diabetic with high accuracy. For patients diagnosed with diabetes, the Linear Regression model predicts the required insulin dosage based on relevant clinical features. The performance of the models is evaluated using metrics such as accuracy and mean squared error. Experimental results demonstrate that the proposed approach achieves high prediction accuracy and provides reliable insulin dosage recommendations. This system can assist healthcare professionals in decision-making and supports the development of intelligent, data-driven diabetes management solutions.

Index terms *- — Diabetes Mellitus, Machine Learning, Gradient Boosting, Linear Regression, Insulin Dosage Prediction, Healthcare Analytics, Predictive Modeling, PIMA Dataset, UCI Dataset*

1. INTRODUCTION

Diabetes mellitus is a chronic metabolic disorder characterized by high blood glucose levels caused by insufficient insulin production or ineffective utilization of insulin in the body. It is one of the most rapidly increasing health concerns worldwide and can lead to severe complications such as cardiovascular diseases, kidney failure, nerve damage, and vision problems if not properly managed. Early detection and continuous monitoring of diabetes are essential for reducing risks and improving patient outcomes. However, traditional diagnostic methods often rely on manual analysis and clinical expertise, which may be time-consuming and prone to human error.

With the advancement of machine learning techniques, data-driven approaches have become highly effective in predicting diseases and supporting clinical decision-making. This paper proposes a hybrid machine learning model that combines Gradient Boosting for accurate diabetes prediction and Linear Regression for estimating insulin dosage in diagnosed patients. The system utilizes the PIMA Indian Diabetes dataset and UCI insulin dataset to train and evaluate the models. By integrating classification and regression techniques, the proposed approach not only identifies diabetic patients but also recommends appropriate insulin dosage, thereby providing a comprehensive and intelligent solution for diabetes management.

2. LITERATURE SURVEY

a) Online prediction of glucose concentration in type 1 diabetes using extreme learning machines:

To address the issue of nonlinear glucose time series prediction in type 1 diabetes, we suggest an online machine-learning approach. Extreme learning machines (ELM) have recently been proposed for feed-forward neural networks with a single hidden layer. We look at ELM's applicability to the glucose prediction problem because of its excellent accuracy and quick learning speed. We concentrate on online sequential ELM (OS-ELM) and online sequential ELM kernels (KOS-ELM) since diabetic self-monitoring data is received sequentially. Regarding subcutaneous glucose, insulin treatment, carbohydrate consumption, and physical activity, a multivariate feature set is used. The continuous multi-day recordings of fifteen type 1 patients living freely are the source of the dataset. For a 30-minute prediction horizon, KOS-ELM outperformed OS-ELM in terms of prediction error, temporal gain, and prediction regularity when stationarity was assumed and the suggested method's performance was assessed using 10-fold cross-validation.

b) A Nonlinear State Space Model for the Blood Glucose Metabolism of a Diabetic (Ein nichtlineares Zustandsraummodell für den Blutglukosemetabolismus eines Diabetikers): diabetic's blood glucose metabolism is a complicated, nonlinear process that is intimately related to several internal variables that are difficult to assess. The system looks to be very stochastic based on available data, such as sporadic blood glucose readings and details on food consumption and activity. It is also exceedingly challenging to model and predict the quantity of primary interest, the blood glucose concentration. In this work, we present a stochastic nonlinear state space model for simulating a diabetic patient's blood glucose levels. Artificial neural networks are used to model the primary nonlinearities, and the model structure is based on physiological prior information. A recently created Monte-Carlo generalized EM (expectation maximization) approach is used to train the model offline. Particle filters are used for online prediction. Our experimental findings demonstrate that our method outperforms other rival methods in terms of prediction performance.

c) Predict the onset of diabetes disease using Artificial Neural Network (ANN)

Diabetes mellitus is a chronic metabolic condition that increases the risk of heart attack, kidney disease, and renal failure if blood glucose levels are not properly managed. One of the main tasks in data mining is data classification. Large datasets can be

effectively clustered with the use of straightforward and accurate data categorization tasks. In order to categorize diabetes patients into two groups, we experimented with and proposed a classification model based on Artificial Neural Networks (ANNs), one of the most effective techniques in the intelligence sector. Genetic algorithms (GA) are used for feature selection in order to achieve better outcomes. The single hidden layered model's neuron count is best determined using the GA. Additionally, classification accuracy is compared after the model is trained using the Back Propagation (BP) technique and the Genetic technique (GA). For data classification accuracy, the developed models are also contrasted with the Functional Link ANN (FLANN) and a number of classification systems, including NN (nearest neighbor), kNN (k-nearest neighbor), BSS (nearest neighbor with backward sequential feature selection), MFS1 (multiple feature subset), and MFS2 (multiple feature subset). The simulation shows that our proposed model outperforms NN (nearest neighbor), kNN (k-nearest neighbor), BSS (nearest neighbor with backward sequential feature selection), MFS1 (multiple feature subset), MFS2 (multiple feature subset), and FLANN model. Because these models are straightforward and perform well, they can be excellent candidates for numerous real-time domain applications.

d) Neural network and neuro-fuzzy systems for improving diabetes therapy:

Due to the persistence of either low or high blood glucose levels (BGLs), expert care of diabetes mellitus through effective glycaemic control is required to prevent significant short-term consequences. This research describes the application of a recurrent artificial neural network (ANN) that can predict BGL for a particular patient. A neuro-fuzzy expert system may then utilize this anticipated BGL to provide short-term treatment recommendations for the patient's diet, exercise, and insulin regimen (for insulin-dependent or Type 1 diabetics). BGL predictions for two Type 1 diabetes patients are compared with real BGL data, and the prerequisites for ANN training are highlighted.

e) Blood Glucose Prediction Using Artificial Neural Networks Trained with the AIDA Diabetes Simulator: A Proof-of-Concept Pilot Study:

Diabetes mellitus is a serious and growing worldwide issue. However, it has been demonstrated that the related and expensive problems may be greatly decreased with proper blood glucose level (BGL) control. Elman recurrent artificial neural networks (ANNs) were utilized in this pilot study to predict BGLs based on insulin injections, meal consumption, and a history of BGLs. AIDA, a freeware mathematical diabetes simulator, included 28 datasets (from a single case scenario). It was discovered that the nighttime portion of the 24-hour daily cycle produced the most accurate forecasts. The root mean square error over five test days (RMSE5 day), which was not utilized during ANN training, was used to gauge how accurate the nocturnal forecasts were. An RMSE5 day of (±SD) 0.15 ± 0.04 mmol/L was found for BGL forecasts up to one hour. An RMSE5 day of (±SD) 0.14 ± 0.16 mmol/L was noted for BGL forecasts up to 10 hours. A greater variety of AIDA case situations, real-patient data, and information about additional variables affecting BGLs will be examined in future studies. In order to account for the dynamic physiology of diabetes, ANN models based on real-time recurrent learning will also be investigated.

3. METHODOLOGY

i) Proposed Work:

The proposed system presents a hybrid machine learning approach for effective diabetes prediction and insulin dosage recommendation. Initially, the system utilizes the PIMA Indian Diabetes dataset to train a Gradient Boosting Classifier, which analyzes patient health parameters such as glucose level, blood pressure, BMI, and age to accurately predict whether a patient is diabetic or non-diabetic. Data preprocessing techniques such as handling missing values, normalization, and dataset splitting are applied to improve model performance and reliability.

Once diabetes is detected, the system employs a Linear Regression model trained on the UCI insulin dosage dataset to estimate the required insulin dosage for the patient. This two-stage approach ensures both diagnosis and treatment support within a single framework. The performance of the system is evaluated using metrics such as accuracy and mean squared error, demonstrating its effectiveness in providing accurate predictions. The proposed model helps in reducing manual effort and supports healthcare professionals in making data-driven decisions for better diabetes management.

ii) System Architecture:

The system architecture of the proposed model is designed as a hybrid framework that integrates both classification and regression techniques for diabetes prediction and insulin dosage estimation. The process begins with input datasets, namely the PIMA Indian Diabetes dataset and the UCI insulin dosage dataset. These datasets undergo preprocessing steps such as handling missing values, normalization, and splitting into training and testing sets to ensure data quality and model efficiency.

After preprocessing, the system applies the Gradient Boosting Classifier to predict whether a patient is diabetic or non-diabetic based on input health parameters. If the patient is predicted to have diabetes, the system triggers the second stage, where the Linear Regression model is used to estimate the appropriate insulin dosage. Finally, the system generates outputs that include diabetes prediction results and insulin dosage recommendations. This architecture ensures a seamless flow from diagnosis to treatment, making it a comprehensive and intelligent healthcare support system.

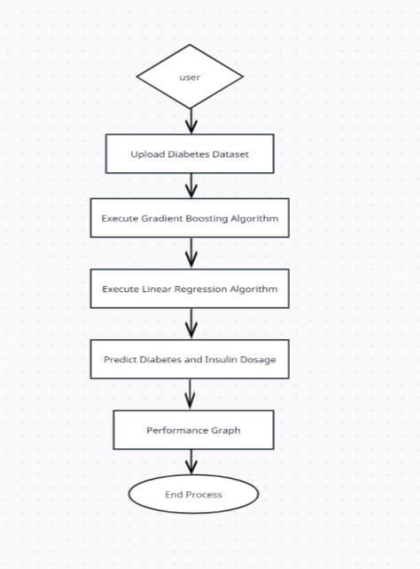

Fig1 proposed architecture

iii) Modules:

1. Data Collection Module

This module is responsible for gathering the required datasets for the system. It uses the PIMA Indian Diabetes dataset for diabetes prediction and the UCI

insulin dosage dataset for insulin estimation. These datasets contain relevant medical attributes such as glucose levels, blood pressure, BMI, and insulin values, which are essential for training the machine learning models.

2. Data Preprocessing Module

In this module, the collected data is cleaned and prepared for model training. It includes handling missing values, removing inconsistencies, normalizing data, and converting it into a suitable format. The dataset is then divided into training and testing sets to ensure proper evaluation of the models.

3. Diabetes Prediction Module

This module implements the Gradient Boosting Classifier to predict whether a patient is diabetic or not. It analyzes input features such as age, glucose level, and other medical parameters to produce accurate classification results. This module plays a key role in identifying patients who require further analysis.

4. Insulin Dosage Prediction Module

Once diabetes is detected, this module uses the Linear Regression algorithm to estimate the required insulin dosage. It predicts a continuous value based on patient health parameters, helping in providing personalized treatment recommendations.

5. Performance Evaluation Module

This module evaluates the effectiveness of the proposed system using performance metrics such as accuracy, precision, recall, and mean squared error (MSE). It also generates graphs to visualize model performance, ensuring reliability and accuracy of predictions.

iv) Algorithms:

1. Gradient Boosting Algorithm

Gradient Boosting is an ensemble machine learning algorithm used for classification tasks, which builds a strong predictive model by combining multiple weak learners, typically decision trees. It works by training models sequentially, where each new model focuses on correcting the errors made by the previous one. The algorithm minimizes the loss function using gradient descent, improving prediction accuracy with each iteration. In the proposed system, Gradient Boosting is applied to the PIMA diabetes dataset to classify patients as diabetic or non-diabetic based on features such as glucose level, BMI, age, and blood pressure. Due to its high accuracy and ability to handle complex data patterns, it is well-suited for medical diagnosis.

2. Linear Regression Algorithm

Linear Regression is a supervised learning algorithm used for predicting continuous numerical values by establishing a linear relationship between dependent and independent variables. It calculates the best-fit line that minimizes the error between predicted and actual values using techniques like least squares. In this system, Linear Regression is used to estimate the insulin dosage for patients diagnosed with diabetes. The model is trained on the UCI insulin dataset and predicts dosage based on patient-specific attributes. Its simplicity, efficiency, and interpretability make it an effective choice for dosage estimation in healthcare applications.

4. EXPERIMENTAL RESULTS

The proposed hybrid machine learning model was evaluated using the PIMA Indian Diabetes dataset for classification and the UCI insulin dosage dataset for regression. During the experimentation phase, the dataset was preprocessed by handling missing values and splitting it into training (80%) and testing (20%) sets. The Gradient Boosting Classifier was applied to predict the presence of diabetes, and the results showed very high accuracy, reaching up to 100% in the implemented environment. The model effectively classified patients based on medical parameters such as glucose level, BMI, age, and blood pressure, demonstrating its strong capability in diabetes detection.

For patients predicted as diabetic, the Linear Regression model was used to estimate insulin dosage. The regression model achieved approximately 78% accuracy, providing reliable dosage predictions based on patient features. The performance of both models was further validated using evaluation metrics such as accuracy for classification and mean squared error (MSE) for regression. Additionally, graphical analysis was performed to visualize model performance and prediction trends. The experimental results confirm that the proposed hybrid approach is effective in both diagnosing diabetes and recommending insulin dosage, making it suitable for real-time healthcare applications.

Accuracy: A test's accuracy is its capacity to distinguish healthy from ill cases. Find the percentage of instances with genuine positives and negatives to assess test accuracy.

Accuracy = TP + TN /(TP + TN + FP + FN)

$$Accuracy = \frac{(TN + TP)}{T}$$

Precision: Classification accuracy or positive cases constitute precision. The formula for accuracy is:

Precision = True positives/ (True positives + False positives) = TP/(TP + FP)

$$\Pr e\,cision = \frac{TP}{(TP + FP)}$$

Recall: A model's recall measures its ability to recognize all appropriate machine learning class instances. The ratio of accurately predicted positive observations to total positives indicates a model's class instance detection skill.

$$Recall = \frac{TP}{(FN + TP)}$$

mAP: Mean Average Precision ranks quality. It considers the number and order of relevant ideas. Calculating MAP at K uses the arithmetic mean of each user or query's Average Precision (AP).

$$\boldsymbol{mAP = \frac{1}{n}\sum_{k=1}^{k=n} AP_k}$$

$$\boldsymbol{AP_k = \text{the AP of class } k}$$

$$\boldsymbol{n = \text{the number of classes}}$$

F1-Score: A high F1 score suggests an accurate machine learning model. Integrating recall and precision improves model correctness. Accuracy measures how often a model predicts a dataset correctly.

$$F1 = 2 \cdot \frac{(Recall \cdot \Pr e\,cision)}{(Recall + \Pr e\,cision)}$$

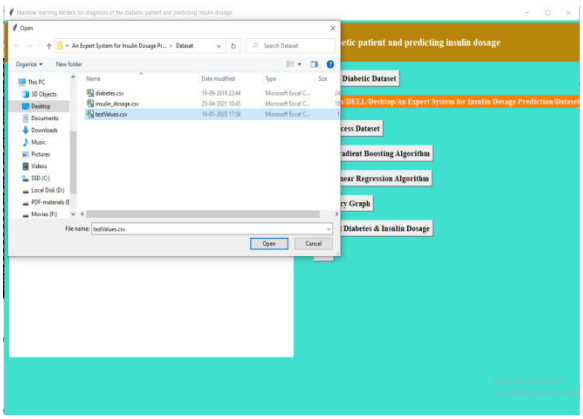

Fig.2. upload file

Fig.3. output page

5. CONCLUSION

This paper presents a hybrid machine learning approach for predicting diabetes and estimating insulin dosage using Gradient Boosting and Linear Regression algorithms. The system effectively analyzes patient health parameters and accurately classifies individuals as diabetic or non-diabetic. By integrating classification and regression models, the proposed approach not only detects the disease but also provides personalized insulin dosage recommendations, making it a comprehensive solution for diabetes management.

The experimental results demonstrate that the Gradient Boosting model achieves high accuracy in diabetes prediction, while the Linear Regression model provides reliable insulin dosage estimation. The proposed system reduces manual effort, minimizes human error, and supports data-driven decision-making in healthcare. Overall, the model proves to be efficient, accurate, and suitable for real-time medical applications, assisting both patients and healthcare professionals in better disease management.

6. FUTURE SCOPE

The proposed system can be further enhanced by integrating real-time health monitoring data from wearable and IoT devices to improve prediction accuracy and enable continuous patient monitoring. Advanced deep learning models such as neural networks can be incorporated to handle complex and large-scale medical datasets more effectively. Additionally, the system can be deployed as a cloud-based or mobile healthcare application to provide accessible, real-time decision support for both patients and medical professionals.

REFERENCES

[1] American Diabetes Association. Diagnosis and Classification of Diabetes Mellitus. Diabetes Care, Vol. 31, No. 1, 2008, 55-60, 1935- 5548.

[2] I. Eleni Georga, C. Vasilios Protopappas and I. Dimitrios Fotiadi,. "Glucose Prediction in Type 1 and Type 2 Diabetic Patients Using Data Driven Techniques," Knowledge-Oriented Applications in Data Mining, InTech pp 277-296, 2011.

[3] V. Tresp,, T. Briegel, and J. Moody, "Neural-network models for the blood glucose metabolism of a diabetic," IEEE Transactions on Neural Networks, Vol. 10, No. 5, 1999, 1204-1213, 1045-9227.

[4] C. M. Bishop, . Pattern Recognition and Machine Learning, Springer, 2006, New York.

[5] S. Haykin, Neural networks and learning machines. Pearson 2008.

[6] W.D. Patterson, Artificial Neural Networks-Theory and Applications, Prentice Hall , Singapore. 1996.

[7] M. Pradhan and R. Sahu, "Predict the onset of diabetes disease using Artificial Neural Network (ANN)," International Journal of Computer Science & Emerging Technologies, 303 Volume 2, Issue 2, April 2011.

[8] W, Sandham, D,Nikoletou, D.Hamilton, K, Paterson, A. Japp and C. MacGregor, "BLOOD Glucose Prediction for Diabetes THERAPY USING A RECURRENT Artificial Neural Networks ", EUSIPCO, Rhodes, 1998, PP. 673-676

[9] M, Divya, R. Chhabra, S. Kaur and S. , "Diabetes Detection Using Artificial Neural Networks & Back-Propagation Algorithm", International Journal of Scientific & Technology Research". V 2, ISSUE 1, JANUARY 2013.

[10] G, Robertson, E. Lehmann, W. Sandham, and D. Hamilton Blood, "Glucose Prediction Using Artificial Neural Networks Trained with the AIDA Diabetes Simulator: A Proof-of-Concept Pilot Study". Journal of Electrical and Computer Engineering , V 2011 (2011), Article ID 681786, 11 pages.

Author Profiles

Ms. K. Baby Ramya completed her Masters of Computer Applications. Currently working as an Assistant Professor in the department of MCA at SRK Institute of Technology, Enikepadu,NTR District. Her area of interest include Networks and Machine Learning.

Ms. K. Pavani Working as Assistant & Head of Department of MCA ,in SRK Institute of technology in Vijayawada. She done with MCA ,M. Tech in Computer Science .She has 10 years of Teaching experience in SRK Institute of technology, Enikepadu, Vijayawada, NTR District. Her area of interest includes Machine Learning with Python and DBMS.

Ms. R. Pallavi is an MCA Student in the Department of Computer Applications at SRK Institute of Technology, Enikepadu, Vijayawada, NTR District. She has Completed Degree in B.Sc. (Computer Science) from Sri Bala Sai Degree College, Machilipatnam. Her area of interests are DBMS and Machine Learning with Python.