iJETRM
INTERNATIONAL JOURNAL OF ENGINEERING TECHNOLOGY RESEARCH & MANAGEMENT (IJETRM)
Peer-Reviewed | Open Access | International Journal
https://ijetrm.com/

DIABETES PREDICTION USING MACHINE LEARNING ALGORITHMS AND ONTOLOGY

Rayavarapu Venkata Chaitanya (MCA Student),
Dr A Mary Sowjanya (Associate Professor),
Varsha Kumari (Research Scholar)
Department of Information Technology and Computer Applications Andhra University College of Engineering, Visakhapatnam, AP.

Corresponding Author: Rayavarapu Venkata Chaitanya,
rayavarapuchaitanyaz@gmail.com , kvarsha162@gmail.com .

ABSTRACT

Diabetes mellitus is a common chronic metabolic disease which, if not diagnosed, can cause a host of serious diseases like cardiovascular diseases, kidney failure, nerve damage, etc. Timely and accurate diagnosis can help to save treatment costs and enhance treatment outcomes for patients with diabetes. In this project, a Machine Learning based diabetes prediction system is developed on the basis of the Pima Indian Diabetes dataset which contains clinical parameters like Glucose level, Blood pressure, BMI, Insulin level, Age, Heart disease, Smoking history. The pre-processed data is used to implement, train and test three supervised classification algorithms: Decision Tree, Random Forest, and Gradient Boosting. In order to compare the models, their accuracy is chosen as the primary performance measurement. The results of the experiments indicate that the Gradient Boosting model has the highest accuracy of 97%, followed by Random Forest with 90% accuracy and Decision Tree with 86% accuracy. The results show that ensemble and boosting based methods significantly outperform simple tree-based classifiers for this prediction task, which makes Gradient Boosting the most appropriate model to be used with a real-world diabetes risk assessment tool.

Keywords:
Diabetes Prediction, Machine learning, Ontology, Gradient Boosting, Decision Tree, Random Forest, SWRL, Clinical Decision Support System

INTRODUCTION

Diabetes mellitus is a common chronic disease that affects millions of people worldwide. It's when the body does not make enough insulin or the insulin it does make doesn't work properly, causing blood glucose to rise. Diabetes can cause heart disease, kidney damage and failure, nerve damage, and vision loss if it isn't caught early. Hence, diabetes diagnosis is crucial to enhance the health of the patient and minimize health issues associated with diabetes. As healthcare data continues to expand, machine learning has proven to be a valuable tool in healthcare predictions. With the help of machine learning algorithms, patient information can be analysed, patterns can be identified and the risk of diabetes can be predicted fairly accurately. But many prediction systems only deliver the result and they don't show you how to make the prediction, so it's hard for the healthcare professional to understand the result. An ontology-based Diabetes Prediction system based on machine learning algorithm is proposed in this work. The system is comparing three different classification algorithms: Decision Tree, Random Forest and Gradient Boosting. Of these, Gradient Boosting has the best prediction performance; Decision Tree is adopted to provide rules which are translated into SWRL rules, and are stored in an OWL ontology. The ontology provides an explanation of the prediction by doing semantic reasoning on patient data. The proposed system is developed as a web app using Flask which allows users to input clinical data and get diabetes risk predictions along with ontology-based explanations. This would integrate the predictiveness of machine learning with the interpretability of ontology, thereby being a helpful decision support system for early diabetes screening.

INTERNATIONAL JOURNAL OF ENGINEERING TECHNOLOGY RESEARCH & MANAGEMENT (IJETRM)
Peer-Reviewed | Open Access | International Journal
https://ijetrm.com/

LITERATURE SURVEY

Diabetes mellitus is a one of the most prevalent chronic diseases globally and should be diagnosed in a timely manner to prevent serious adverse health effects. As medical information develops, machine learning has proven to be a useful tool to support early prediction of diabetes. Many studies have used the PIMA Indian Diabetes Dataset to evaluate different classification algorithms. The previous researches primarily have targeted the accuracy of the predictions by employing appropriate preprocessing methods, feature engineering and complex machine learning models.

A. Diabetes Prediction Using Machine Learning

The use of machine learning has been effective in predicting diabetes by recognizing patterns of patient health information. Logistic Regression, Support Vector Machine, Naïve Bayes, Decision Tree, Random Forest and Gradient Boosting are some of the popular algorithms used for disease classification. The methods are reported to be more accurate and rapid than the traditional statistical methods, and could be useful in clinical decision support.

B. Decision Tree Based Diabetes Prediction

Decision Tree is a simple supervised learning algorithm and it interprets the decisions in a tree-like structure. It is widely used in healthcare as the decision rules generated are easy to understand. Decision Trees give satisfactory classification results, but are prone to overfitting and are sensitive to noisy data. Despite these restrictions, they are still significant as a baseline model for medical prediction tasks.

C. Diabetes Prediction by Using Random Forest.

Random Forest is an ensemble learning algorithm which uses several Decision Trees to increase the prediction accuracy. It resulting more reliable and stable predictions. Random Forest is among the most popular algorithm for healthcare applications as several studies have reported its higher accuracy compared to individual Decision Trees.

D. Gradient Boosting Based Diabetes Prediction

Gradient Boosting is a strong ensemble method in which each new tree in the ensemble corrects the errors from the previous tree in a sequence. It can detect complicated interactions between clinical features and typically has high prediction accuracy. Although it takes longer to compute, Gradient Boosting has proven to be an effective technique for prediction in many disease prediction systems due to its high predictive power.

Research Gap

There are numerous machine learning models that have been proposed for diabetes prediction, but there are still some challenges. The majority of the current works in the literature rely on one algorithm while few provide a comparison of several algorithms under the same conditions. Moreover, certain approaches focus exclusively on predictive power rather than model interpretability, computational efficiency, and/or applicability in clinical settings. These constraints emphasize the need for a proper comparative assessment of the reliability of machine-learning algorithms processed through a standardised pre-processing and evaluation approach.

OBJECTIVES

This is the primary objective which aims at creating a machine learning based diabetes prediction system. The pre-processed PIMA Indian Diabetes Dataset is finally fed into three classification algorithms, namely the Decision Tree, Random Forest and Gradient Boosting. These models are tested and compared to find out which model would yield the most accurate prediction of diabetes.

METHODOLOGY

In this study, diabetes prediction utilizing machine learning and ontology for early prediction of Diabetes. The proposed framework classifies the patients as diabetic and non-diabetic by using supervised machine learning techniques based on patient's clinical data. The methodology has six major steps: data preprocessing, feature preparation, model training, integration of ontology, prediction and performance evaluation. The algorithms used include Decision Tree, Random Forest, and Gradient Boosting, which are trained and evaluated to find out which one is the best. The Decision Tree rules are translated to SWRL rules and combined with an OWL ontology for semantic reasoning and interpretation of prediction. The final model is implemented as a web application written in Flask, which returns diabetes predictions along with explanations in the form of an ontology.

A. Overview of the Proposed System

To build an accurate and interpretable diabetes prediction framework, the proposed system combines the use of machine learning algorithms with ontology-based reasoning. Supervised learning techniques are used to analyse

iJETRM
INTERNATIONAL JOURNAL OF ENGINEERING TECHNOLOGY RESEARCH & MANAGEMENT (IJETRM)
Peer-Reviewed | Open Access | International Journal
https://ijetrm.com/

patient clinical information to identify if a patient is diabetic or not. The framework can be divided into six phases: data collection, data preprocessing, model development, integration of ontology, prediction generation and performance evaluation. Three algorithms: Decision Tree, Random Forest, and Gradient Boosting are trained and evaluated to find the best model. Decision Tree rules are further translated into SWRL rules and are added in an OWL ontology for better interpretability of the prediction results. Finally, it was deployed to flask web application.

B. Data Collection and Preprocessing

1) Data Collection

The diabetes prediction system proposed is based on a public set of 100,000 patient records. There are 8 clinical attributes in each record: Gender, Age, Hypertension, Heart Disease, Smoking History, BMI, HbA1c Level, Blood Glucose Level. Target variable: Diabetic or Non-Diabetic. These clinical features are used as input features to train and test the machine learning algorithms.

2) Data Preprocessing

Before training the model, the data is preprocessed to improve the quality of the data and to make sure the predictions are accurate. The dataset is loaded from a CSV file through the Pandas library and analyzed to check the structure, data types and class distribution. Label Encoding: Numerical values are created from categorical features – Gender, Smoking History. The chosen clinical characteristics are then used as feature input to building the model. Finally, the data is split into 80% training data and 20% testing data by stratified sampling, which will maintain the same class distribution. This processed data set is then fed into a Decision Tree, Random Forest, and Gradient Boosting model for training and testing

C. Machine Learning Model Development

In this work three supervised machine learning algorithms have been implemented and compared. **Decision Tree (DT):** A rule-based classification algorithm which makes a prediction based on important clinical features using recursive partitioning of the dataset. It is also applied to produce rules for integration of ontologies.

Random Forest (RF): An ensemble learning technique that picks several decision trees and classifies in the manner of majority voting to increase the stability of the prediction and decrease the overfitting.

Gradient Boosting (GB): It is an ensemble method which constructs decision trees one by one, with the aim of learning from the mistakes of the previous tree. This method can enhance the prediction accuracy and overall model performance.

All the models are trained and tested on the same set of data and with the same experimental setup to have an unbiased comparison.

D. Ontology Design and Rule-Based Reasoning

To improve the interpretation of the machine learning predictions and to convey clinical knowledge, a diabetes ontology was created in Protégé. The ontology represents the relationships between patient information, clinical parameters and diabetes diagnosis. It offers a semantic basis for reasoning and explaining the prediction results by using the rules.

The ontology is designed to represent the key concepts involved in diabetes prediction. It consists of classes such as Patient, Diabetic, Non-Diabetic, and Risk Level, which describe the main entities in the system. Data properties including Age, Gender, BMI, HbA1c Level, Blood Glucose Level, Hypertension, Heart Disease, and Smoking History are used to capture each patient's clinical information. Individual patient records are then created using these attributes, allowing the ontology to represent real-world clinical data in a structured manner.

1) SWRL Rule Generation

The Decision Tree model is used to derive classification rules from the training data. The rules are then translated into Semantic Web Rule Language (SWRL) rules and added to the ontology. The ontology can be used for semantic reasoning and consistent diabetes risk classification by allowing the representation of the decision logic in a human-readable format. Below are given some example rules written in SWRL.

- If patient has HbA1c Level ≥ 6.5 and Blood Glucose Level ≥ 126, then that means the patient is classified as Diabetic.
- If the BMI is ≥ 30 and the Blood Glucose Level is ≥ 126, then categorise the patient as High Risk.
- If HbA1c level < 5.7 and Blood Glucose level < 100, then the patient is classified as Nondiabetic.2) Ontology Reasoning

2) Ontology Reasoning

iJETRM
INTERNATIONAL JOURNAL OF ENGINEERING TECHNOLOGY RESEARCH & MANAGEMENT (IJETRM)
Peer-Reviewed | Open Access | International Journal
https://ijetrm.com/

The ontology is implemented with a semantic reasoner to run the SWRL rules and draw conclusions about the patient's category. Once data about the patient is loaded into the ontology, the reasoning process will access the rules defined and decide the status of the patient's diabetes and the risk level. The inferred results are then shown in the Flask application in addition to the machine learning prediction, thus creating an explainable and interpretable decision supporting system.

E. Evaluation

Each model is evaluated on the held-out test set using accuracy as the primary metric. The comparative results are summarized below:

Algorithm	Accuracy (%)	Key Characteristic	Remarks
Decision Tree	85%	Rule-based hierarchical splitting	Simple, interpretable, moderate accuracy
Random Forest	90%	Ensemble of multiple decision trees	Reduces overfitting, improves stability
Gradient Boosting	97%	Sequential boosting of weak learners	Best performing algorithm in this study

(Evaluation Results)

Gradient boosting is clearly better than both random forest and decision tree, as evident from the results, which indicates that error correction sequential procedures in boosting methods are more effective on this diabetes classification problem than using both random forest and decision tree methods. Random Forest is an ensemble of independent trees that is better than the single Decision Tree, but is not as accurate as boosting

RESULTS AND DISCUSSION

The experimental evaluation shows that all the three machine learning algorithms are able to predict the diabetes. Decision Tree performed well with an accuracy of 86% and gave easy to understand decision rules. Random Forest enhanced the predictability performance to 95%, as it was able to combine several decision trees and prevent overfitting. Gradient Boosting had the highest accuracy (97%) among all the algorithms in this study; thus it is the most effective algorithm in predicting diabetes. It is an effective sequential learning strategy that reduces the prediction errors and enhances the generalization. So, Gradient Boosting is chosen as the final prediction model due to its better classification performance.

Confusion Matrix Analysis

The Gradient Boosting model has the classification accuracy in the confusion matrix depicted in Figure. It accurately identified 18268 non-diabetic patients (True Negatives) and 1179 diabetic patients (True Positives). There were 32 non-diabetic patients who were incorrectly classified as diabetic (False Positive) and 521 diabetic patients who were incorrectly classified as non-diabetic (False Negative).

The few misclassifications show that the proposed model has a good predictive ability. The confusion matrix also shows that the Gradient Boosting algorithm is reliable and accurate for predicting Diabetes and has very few misclassifications.

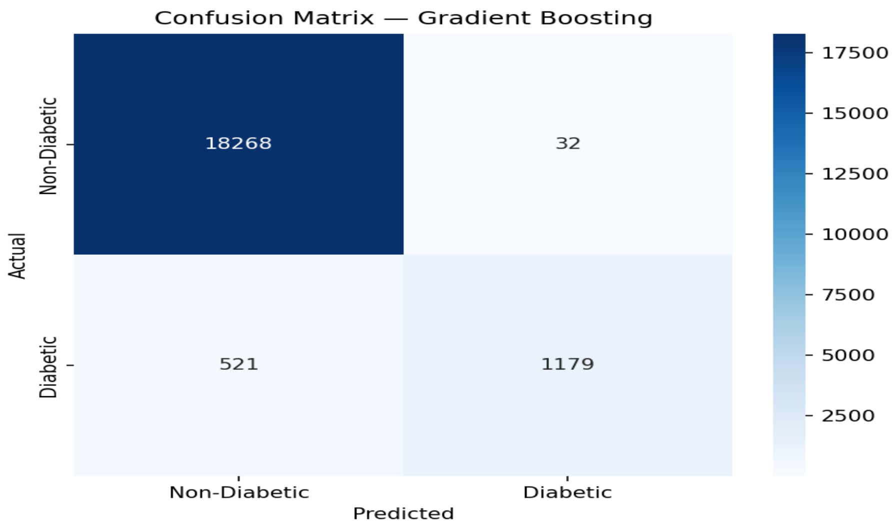

Figure 1: Confusion Matrix of the Model

ISSN: 2456-9348 (Online)
SJIF Impact Factor: 8.232

iJETRM
INTERNATIONAL JOURNAL OF ENGINEERING TECHNOLOGY RESEARCH & MANAGEMENT (IJETRM)
Peer-Reviewed | Open Access | International Journal
https://ijetrm.com/

Performance Comparison

Performance comparison of the prediction accuracy of the three machine learning algorithms used in this study is shown in the Figure. Decision Tree had an accuracy of 85%, Random Forest had an accuracy of 90% and Gradient Boosting has the highest accuracy at 97%. The predictive ability of the Gradient Boosting algorithm was clearly better than the other algorithms due to its sequential learning mechanism which is most useful in reducing the prediction error.

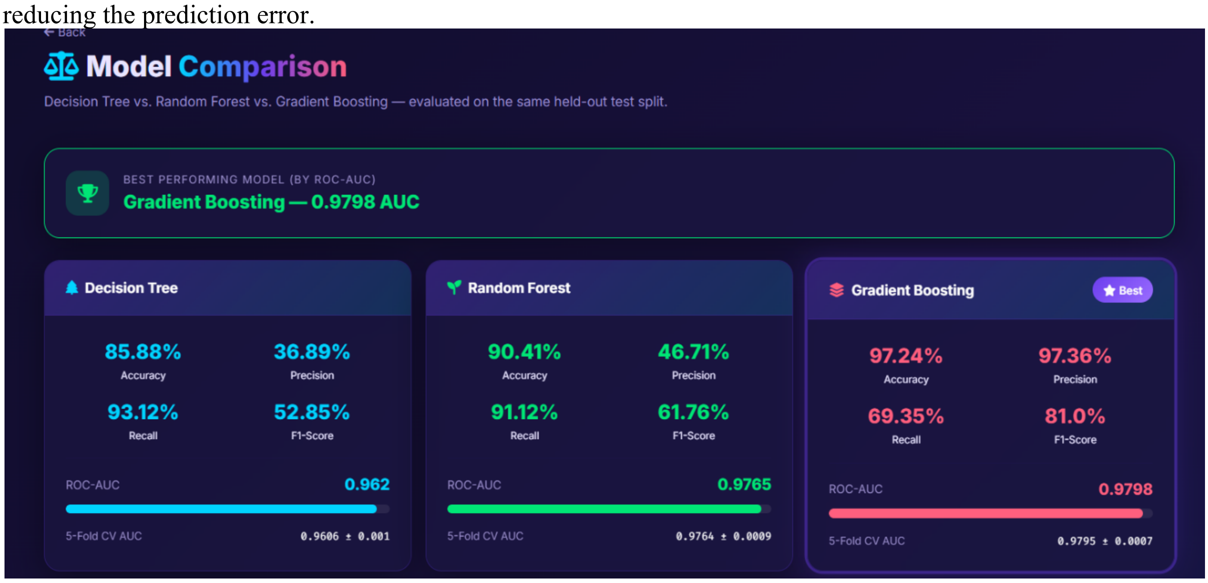

Figure 2: Performance Comparison

Home Page of the Proposed System

This is the initial screen of the proposed system. This is the Home Page of the Proposed System**.** Home page of the Diabetes Prediction System is shown. The web application offers a simple and user-friendly interface with which the users can navigate to the module of predictions. The initial screen explains the functionality of the system and leads to the diabetes prediction section.

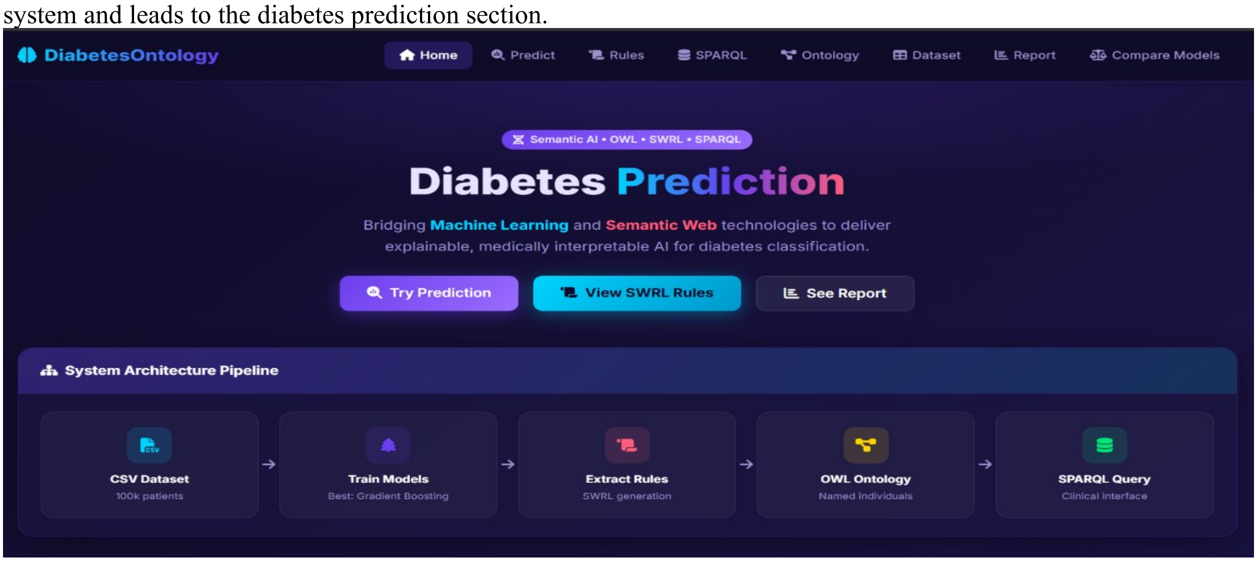

Figure 3: Home Page

Patient Details Input Page

Figure shows the clinical parameters page for patients to enter clinical data including glucose level, HbA1c level, BMI, age, hypertension, history of heart disease, gender and smoking history. The machine learning model uses these features as inputs to make a decision about the diabetes or not..

iJETRM
INTERNATIONAL JOURNAL OF ENGINEERING TECHNOLOGY RESEARCH & MANAGEMENT (IJETRM)
Peer-Reviewed | Open Access | International Journal
https://ijetrm.com/

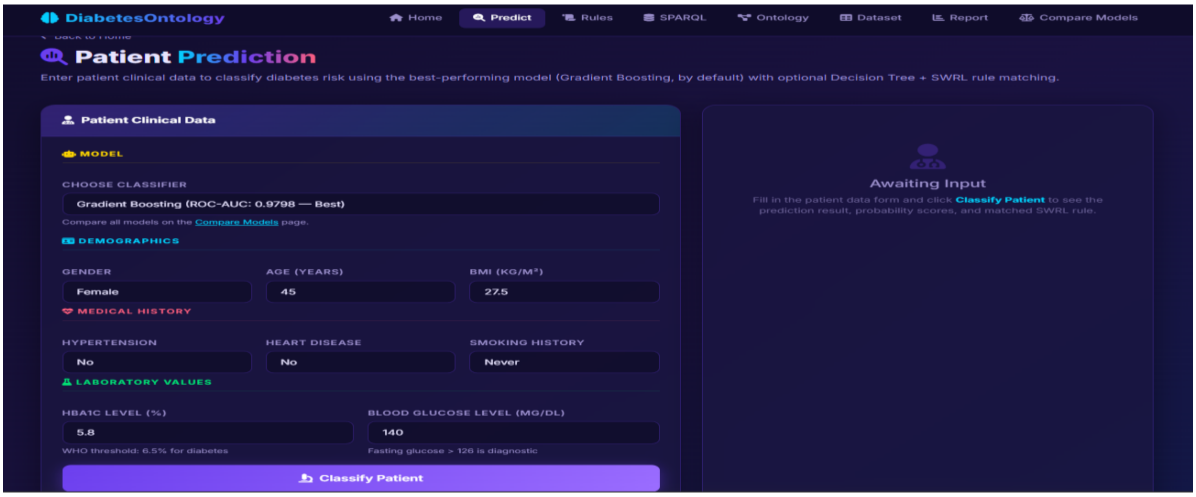

Figure 4: Input Page

Prediction Result

Non-Diabetic Page

The figure shows the prediction result for a non-diabetic patient (as defined by the proposed Diabetes Prediction System). The trained Gradient Boosting model predicts that the patient has a 0.99 probability of not having diabetes, given the clinical parameters the user inputs, such as the level of HbA1c, blood glucose, BMI, age, hypertension, history of heart disease, gender and smoking. It is predicted in real-time and presented on the Flask Web application.

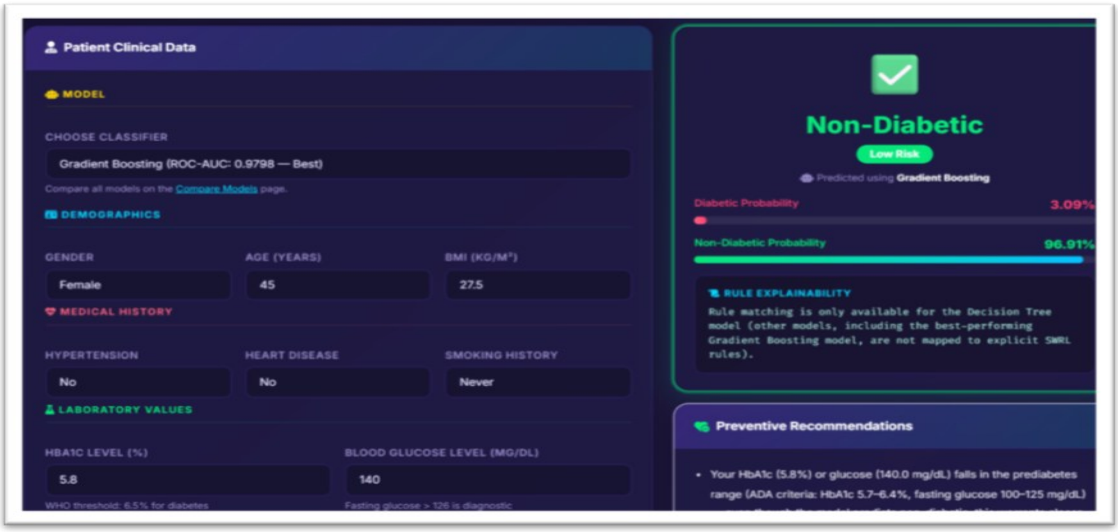

Figure 5: Non-Diabetic

Diabetic Page

The prediction result for Diabetic Classified by the proposed Diabetes Prediction System is given below in Figure. Once the user provides the clinical data, the input is analysed by the trained Gradient Boosting model and the model predicts that the patient is diabetic. This result is immediately presented via the Web application and enables the user to get an initial evaluation. The ontology-based reasoning mechanism will be able to classify the patient as a Diabetic Patient by applying the rules generated with SWRL, simultaneously. The use of machine learning and ontology improves the accuracy and explainability of the prediction process.

iJETRM
INTERNATIONAL JOURNAL OF ENGINEERING TECHNOLOGY RESEARCH & MANAGEMENT (IJETRM)
Peer-Reviewed | Open Access | International Journal
https://ijetrm.com/

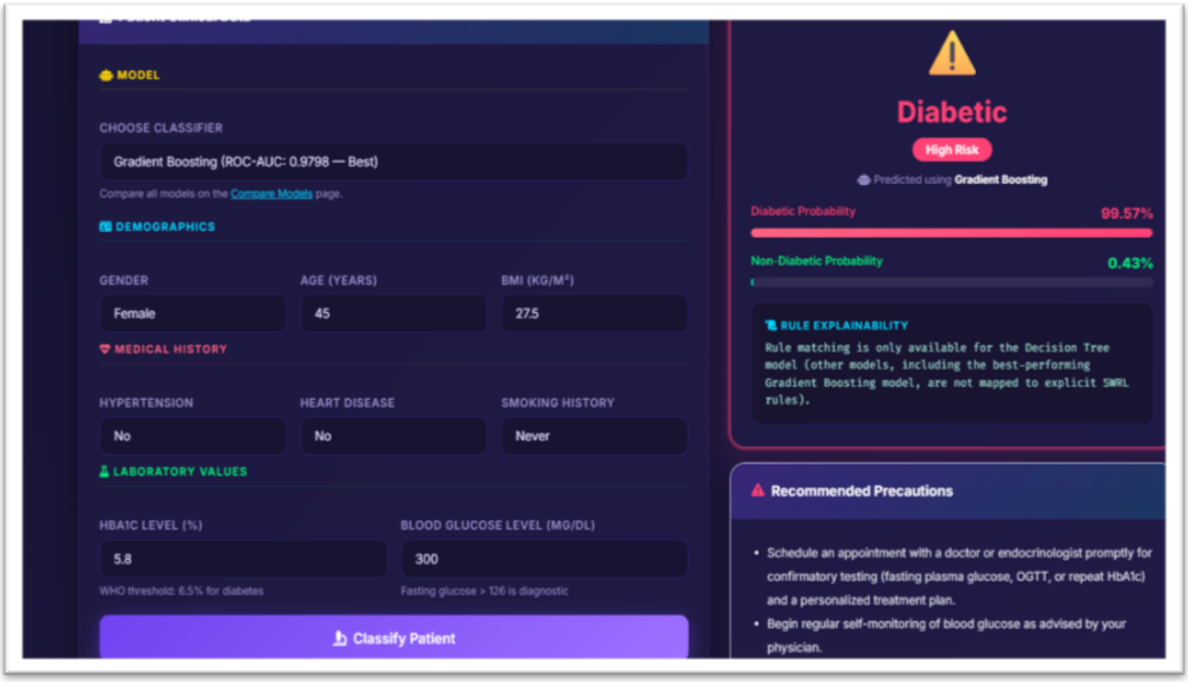

Figure 6: Diabetic

CONCLUSION

In this project, the authors designed an intelligent system for prediction of diabetes by applying three supervised machine learning algorithms: Decision Tree, Random Forest, and Gradient Boosting. This dataset consisted of diabetes data, which were used to train and test the models to find out which technique will be more accurate to predict the disease. The experimental results showed that the accuracy of Decision Tree, Random Forest and Gradient Boosting were 86%, 90% and 97%, respectively. In all the models Gradient Boosting always performed best in terms of prediction performance by reducing classification errors and the overall accuracy. The proposed system can help the healthcare professionals to detect diabetes as early as possible and diagnose and treat the diabetes at early stage. The system provides a rapid, accurate and economical decision-making aid for diabetes prediction through using machine learning techniques. Evaluation methodology is implemented using the common metrics like accuracy, precision, recall, F1-score and confusion matrix. In conclusion, the project successfully showed that machine learning can be a valuable tool in the prediction of diabetes with Gradient Boosting being the best algorithm for this specific dataset.

REFERENCES

1) H. El Massari, Z. Sabouri, S. Mhammedi, and N. Gherabi, "Diabetes Prediction Using Machine Learning Algorithms and Ontology," Journal of ICT Standardization, vol. 10, no. 2, pp. 319–338, 2022.
2) J. J. Khanam and S. Y. Foo, "A Comparison of Machine Learning Algorithms for Diabetes Prediction," ICT Express, vol. 7, no. 4, pp. 432–439, 2021.
3) A. Mujumdar and V. Vaidehi, "Diabetes Prediction Using Machine Learning Algorithms," Procedia Computer Science, vol. 165, pp. 292–299, 2020.
4) P. Cihan and H. Coşkun, "Performance Comparison of Machine Learning Models for Diabetes Prediction," in Proceedings of the 29th Signal Processing and Communications Applications Conference (SIU), 2021.
5) F. Alaa Khaleel and A. M. Al-Bakry, "Diagnosis of Diabetes Using Machine Learning Algorithms," Materials Today: Proceedings, 2021.
6) M. A. Sarwar, N. Kamal, W. Hamid, and M. A. Shah, "Prediction of Diabetes Using Machine Learning Algorithms in Healthcare," in Proceedings of the 24th International Conference on Automation and Computing (ICAC), 2018.

INTERNATIONAL JOURNAL OF ENGINEERING TECHNOLOGY RESEARCH & MANAGEMENT (IJETRM)
Peer-Reviewed | Open Access | International Journal
https://ijetrm.com/

7) G. Tripathi and R. Kumar, "Early Prediction of Diabetes Mellitus Using Machine Learning," in Proceedings of the International Conference on Reliability, Infocom Technologies and Optimization (ICRITO), 2020.
8) M. Rady, K. Moussa, M. Mostafa, A. Elbasry, Z. Ezzat, and W. Medhat, "Diabetes Prediction Using Machine Learning: A Comparative Study," in Proceedings of NILES Conference, 2021.
9) M. U. Emon et al., "Primary Stage of Diabetes Prediction Using Machine Learning Approaches," in Proceedings of ICAIS, 2021.
10) T. Mahboob Alam et al., "A Model for Early Prediction of Diabetes," Informatics in Medicine Unlocked, vol. 16, 2019.
11) L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.
12) J. H. Friedman, "Greedy Function Approximation: A Gradient Boosting Machine," The Annals of Statistics, vol. 29, no. 5, pp. 1189–1232, 2001.
13) J. R. Quinlan, C4.5: Programs for Machine Learning. Morgan Kaufmann, 1993.
14) C. M. Bishop, Pattern Recognition and Machine Learning. Springer, 2006.
15) T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning, 2nd ed. Springer, 2009.
16) A. Géron, Hands-On Machine Learning with Scikit-Learn, Keras and TensorFlow, 3rd ed. O'Reilly Media, 2022.
17) F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.
18) I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, Data Mining: Practical Machine Learning Tools and Techniques, 4th ed. Morgan Kaufmann, 2016.
19) S. Raschka and V. Mirjalili, Python Machine Learning, 3rd ed. Packt Publishing, 2019.
20) T. M. Mitchell, Machine Learning. McGraw-Hill, 1997.

