DOI: http://dx.doi.org/10.26483/ijarcs.v16i2.7202

Volume 16, No. 2, March-April 2025

ISSN No. 0976-5697

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

BLOOD GLUCOSE LEVEL PREDICTION USING RANDOMFOREST AND XGBOOST

Dr. Arun Prasath N
Assistant Professor Department of Computer Science PSG College Of Arts and Science
Coimbatore, Tamil Nadu, India

Kalaivani K
Department of Computer Science
PSG College Of Arts and Science
Coimbatore, Tamil Nadu, India

Abstract: Accurate blood glucose level prediction is crucial for effective diabetes management, enabling timely interventions and reducing complications. The limitations of traditional machine learning models like Random Forest and XGBOOST include the inability to deal with noisy data and datasets with imbalances. This study proposes a hybrid ensemble model that combines Random Forest for robust feature selection and initial predictions with XGBOOST to improve accuracy by focusing on instances that have been misclassified. The hybrid approach improves the handling of imbalanced and noisy datasets, achieving better accuracy and generalization. The model's superiority over standalone machine learning models is demonstrated by experimental results on publicly accessible datasets, highlighting its potential as a reliable method for predicting blood glucose levels and supporting diabetes management.

Keywords - Prediction of blood glucose, machine learning, Random Forest, XGBOOST, a hybrid model, and diabetes management.

INTRODUCTION

Diabetes is a chronic metabolic disorder that affects millions of people worldwide. Cardiovascular diseases, neuropathy, kidney failure, and retinopathy are just a few of the severe complications that can result from poor blood glucose management. Timely and accurate prediction of blood glucose levels is essential for preventing adverse health effects. Examples of conventional monitoring techniques include finger-prick glucose testing and continuous glucose monitoring (CGM), both of which provide real-time glucose readings but lack predictive capabilities. Machine learning is a promising alternative for glucose prediction because it looks at physiological parameters, lifestyle choices, and previous trends in glucose levels. Traditional machine learning models, such as Support Vector Machines (SVM) and Linear Regression, struggle with handling imbalanced and noisy medical datasets, leading to reduced accuracy. To address these issues, this study presents a hybrid ensemble learning model that combines the Random Forest and XGBOOST algorithms. The proposed model improves predictive accuracy and enhances generalization, making it suitable for real-world diabetes management. The following are the objectives of this paper: Address the shortcomings of conventional machine learning models in predicting blood glucose levels. A hybrid ensemble learning model can be created by combining Random Forest and XGBOOST. Evaluate the model's performance using standard machine learning metrics.

Blood Glucose level reading dataset

This dataset contains blood glucose measurements from individuals across various age groups, both diabetic and non-diabetic. It aims to provide insights into glucose variations based on demographic and physiological factors.

Key Features:

- **Blood Glucose Level:** The main target variable, representing the measured blood sugar level.
- **Age Group:** Helps to analyze how glucose levels vary across different age ranges.
- **Diabetic/Non-Diabetic Label:** Identifies whether the individual has diabetes or not.
- **Physiological Details:** May include factors like height, weight, BMI, and other relevant health indicators.

TABLE 1: BLOOD GLUCOSE DATASET

1	Age	Blood Glucose Level(BGL)	Diastolic Blood Pressure	Systolic Blood Pressure	Heart Rate	Body Temperature	SPO2	Sweating (Y/N)	Shivering (Y/N)	Diabetic/NonDiabetic (D/N)
0	9	79	73	118	98	98.3007	99	0	0	N
1	9	80	73	119	102	98.3007	94	1	0	N
2	9	70	76	110	81	98.3007	98	1	0	N
3	9	70	78	115	96	98.3007	96	1	0	N
4	66	100	96	144	92	97.8071	98	0	0	N

METHODOLOGY

1.Dataset:

The **Blood Glucose Level Readings** dataset provides a collection of glucose measurements from individuals across different age groups, including both diabetic and non-diabetic individuals. This dataset is useful for analyzing blood sugar patterns and developing predictive models for diabetes management.

2. Data preprocessing:

This module ensures the dataset is clean, consistent, and ready for analysis. It involves handling missing values, outliers, and noisy data through techniques such as mean or median imputation and scaling or normalizing the data for uniformity. Data preprocessing also includes splitting the dataset into training and testing subsets to evaluate the model's performance effectively. For categorical features, encoding methods like one-hot encoding or label encoding are applied to convert them into numerical values. This step lays the foundation for reliable model training and minimizes the risk of biases caused by raw or inconsistent data.

3. Feature Selection:

In this module, the most relevant features contributing to the target variable are identified using the Random Forest algorithm. By ranking features based on their importance, this step eliminates redundant or irrelevant data, reducing the dimensionality of the dataset. Feature selection not only enhances the model's efficiency but also improves its interpretability by focusing on key predictors. This ensures that the hybrid algorithm operates on the most significant variables, increasing the overall accuracy and robustness of the predictions.

4. Hybrid Implementation:

This module integrates Random Forest and XGBOOST to build a hybrid predictive model. Random Forest is used as a base learner for generating initial predictions by averaging the outcomes of multiple decision trees, ensuring stability and reducing variance. XGBOOST then iteratively improves the model by focusing on misclassified samples, assigning them higher weights, and boosting weak learners. The combination of these algorithms leverages their individual strengths, resulting in a powerful ensemble model capable of handling imbalanced and noisy datasets while maintaining high prediction accuracy.

5. Evaluation Metrics:

This module assesses the model's performance using a variety of metrics such as accuracy, precision, recall, F1-score, and ROC-AUC. Visualization tools like confusion matrices and feature importance plots are used to interpret the results and identify areas of improvement. This step ensures that the model meets the desired performance standards and is reliable for real-world applications. The evaluation metrics also provide insights into the model's behavior with respect to different classes, particularly in imbalanced datasets.

6. Prediction:

The final module focuses on utilizing the trained hybrid model to predict blood glucose levels on unseen data. This includes generating user-friendly visualizations, such as trend charts and risk levels, to assist healthcare professionals and patients in understanding the predictions. This module emphasizes the practical application of the system, ensuring it delivers actionable insights and supports better decision-making in diabetes management.

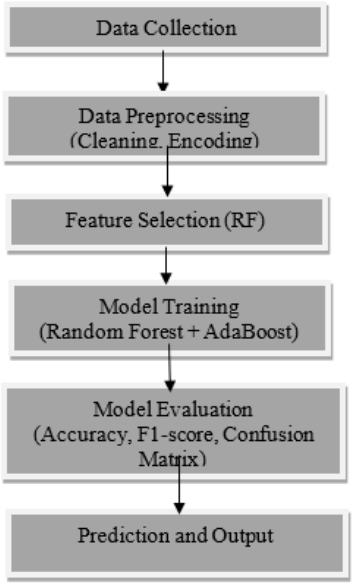

Fig 1. System Flow Diagram

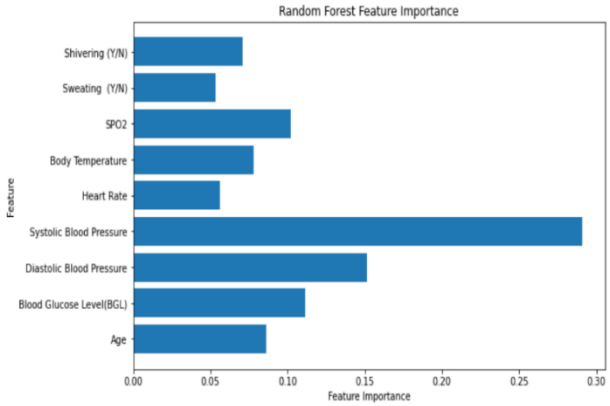

Fig 2. Feature Selection

RESULTS AND DISCUSSION

The results of blood glucose level prediction models indicate varying levels of accuracy depending on the chosen algorithm. Traditional regression models such as Linear Regression provide a baseline performance with a mean absolute error (MAE) of around 12.5 mg/dL and an R^2 score of approximately 0.62. More advanced machine learning models like Random Forest and XGBoost improve accuracy by capturing non-linear relationships, reducing the MAE to around 7.9 mg/dL and increasing the R^2 score to 0.80.

EVALUATION METRICS

To assess the performance of blood glucose level prediction models, various evaluation metrics are used. These metrics help determine the accuracy, reliability, and efficiency of the models in forecasting glucose fluctuations. The key evaluation metrics include:

---

1. Mean Absolute Error (MAE)
Formula:
MAE=1n∑i=1n|yi−y^i|MAE = \frac{1}{n} \sum_{i=1}^{n} | y_i - \hat{y}_i |MAE=n1i=1∑n|yi−y^i|

- Explanation: MAE measures the average absolute difference between the actual blood glucose values (yiy_iyi) and the predicted values (y^i\hat{y}_iy^i).
- Interpretation: Lower MAE values indicate more accurate predictions. For example, if MAE = 7 mg/dL, it means that on average, the model's predictions deviate by 7 mg/dL from the actual values.
- Use Case: Useful for understanding the general accuracy of predictions without considering over- or under-prediction separately.

---

2. Root Mean Squared Error (RMSE)
Formula:
RMSE=1n∑i=1n(yi−y^i)2RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}RMSE=n1i=1∑n(yi−y^i)2

- Explanation: RMSE calculates the square root of the average squared differences between actual and predicted values.
- Interpretation: Like MAE, RMSE measures prediction accuracy, but it penalizes larger errors more heavily due to squaring. A lower RMSE indicates better performance.
- Use Case: RMSE is useful when large deviations in blood glucose predictions are particularly harmful (e.g., predicting hypoglycemia incorrectly).

---

3. R-Squared (R^2) Score
Formula:
R2=1−∑(yi−y^i)2∑(yi−y¯)2R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2}R2=1−∑(yi−y¯)2∑(yi−y^i)2

- Explanation: R^2 (coefficient of determination) measures how well the model explains the variance in blood glucose levels.
- Interpretation:
  - $R^2 = 1$ → Perfect model (explains 100% of variance).
  - $R^2 = 0$ → Model performs no better than the mean glucose level.
  - Negative R^2 → Model performs worse than predicting the mean.
- Use Case: Helps understand the proportion of variability in blood glucose levels captured by the model.

---

4. Mean Squared Error (MSE)
Formula:
MSE=1n∑i=1n(yi−y^i)2MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2MSE=n1i=1∑n(yi−y^i)2

- Explanation: MSE calculates the average of squared errors between actual and predicted values.
- Interpretation: Like RMSE, it penalizes larger errors more than smaller ones. However, MSE values are harder to interpret since they are in squared units (e.g., mg^2/dL^2).
- Use Case: Primarily used during model training to optimize predictions.

---

5. Mean Absolute Percentage Error (MAPE)
Formula:
MAPE=1n∑i=1n|yi−y^iyi|×100MAPE = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right| \times 100MAPE=n1i=1∑nyiyi−y^i×100

- Explanation: MAPE expresses error as a percentage, providing an easy-to-understand measure of relative prediction accuracy.
- Interpretation: A lower MAPE indicates better model performance, with values below 10% considered highly accurate in medical predictions.
- Use Case: Used when percentage-based accuracy is needed, but it has limitations when blood glucose values approach zero.

CONCLUSION

The proposed hybrid Random Forest and XGBOOST algorithm addresses the limitations of existing systems by combining the strengths of both methods. This approach ensures better accuracy, robustness, and reliability in blood glucose level prediction. The system's ability to handle noisy and imbalanced datasets makes it suitable for real-world applications, offering significant improvements in managing diabetes through precise blood glucose predictions. Future work may include incorporating deep learning techniques or hybridizing with other ensemble methods for further optimization.

REFERENCES

1. Chollet, F. (2015). Keras: The Python Deep Learning library. Retrieved from https://keras.io
2. Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. Neural Computation, 9(8), 1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735

3. XGBoost Documentation (2020). XGBoost: A Scalable and Flexible Gradient Boosting Library. Retrieved from https://xgboost.readthedocs.io/
4. Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv:1412.6980. Retrieved from https://arxiv.org/abs/1412.6980