AMERICAN JOURNAL OF MANAGEMENT AND IOT MEDICAL COMPUTING
Peer Reviewed, Referred & Indexed Journal
E-ISSN: 3069-0110
Vol.5, No.2(1) (2026)
www.ajmimc.com

Optical Pulse Signal Patterns and Population Attributes in Blood Glucose Estimation using Probabilistic Boosting

P. Nagendra Kumar[1], K. Chiranjeevi[2], B. Poojitha[2], K. Balakrishna[2]
[1]Professor, [2]Assistant Professor, [1,2]Department of Computer Science and Engineering (AI & ML)
[1,2]Geethanjali Institute of Science and Technology, Nellore-Bombay Highway, S.P.S.R, Andhra Pradesh 524137, India

ABSTRACT

This research explains that accurate and continuous monitoring of blood glucose levels is crucial for effective diabetes management. Conventional invasive techniques, however, remain uncomfortable and unsuitable for frequent use. This study presents a non-invasive blood glucose estimation framework using wrist-based photoplethysmography (PPG) signals. The acquired PPG waveform contains both an AC (pulsatile) component, reflecting cardiac-driven blood volume changes, and a DC (non-pulsatile) component associated with baseline tissue absorption. PPG signals undergo Signal Quality Index (SQI) evaluation, filtering, and baseline correction before feature extraction. Key features include systolic peak, diastolic peak, heart rate, pulse area, and PPG amplitude measures derived mainly from the AC component, along with DC-related baseline features. Additionally, empirical mode decomposition (EMD) is applied to extract intrinsic mode function (IMF)-based features, which, to the best of our knowledge, have not previously been used for non-invasive glucose prediction. Multiple machine learning models like Linear Regression, Lasso, Ridge, Decision Tree, Random Forest, and Natural Gradient Boosting (NG-Boost) are trained on dataset features such as PPG Signal, Heart Rate, Systolic Peak, Diastolic Peak, Pulse Area, Age, Height, Weight, and Gender. Experimental results demonstrate that NG-Boost outperforms all baseline models, making it the proposed algorithm for glucose level prediction. The findings confirm the feasibility of advanced PPG feature engineering and gradient boosting methods for non-invasive blood glucose monitoring.

Keywords: Photoplethysmography, Machine Learning, Empirical Mode Decomposition, Pulse Waveform Analysis

1. INTRODUCTION

Uncontrolled blood sugar levels in diabetics have the potential to seriously harm organs and result in consequences such as strokes that can lead to death. It is considered a “Silent killer” because it can be fatal if not treated early. The International Diabetes Federation predicted in 2017 that the prevalence of diabetes would increase internationally. The two main types of diabetes, type 1 (T1D) and type 2 (T2D), have the same symptoms but differ in terms of causes and treatment. Unfortunately, many people with diabetes ignore it and do not realise how serious the side effects can be when their blood sugar levels increase. Type 2 diabetes is the most common and can be prevented or reduced by increasing awareness of its dangers and modifying one's lifestyle to a healthy one. Monitoring blood glucose levels regularly and administering diabetes treatments as needed can effectively prevent diabetes. Currently, non-invasive monitoring with laboratory tests or a glucometer is the only method routinely performed and is believed to have a high degree of accuracy. A glucometer is commonly utilised in the home. Both procedures require using needles or finger pricks, which result in minimal tissue damage but discomfort to the patient. A non-invasive procedure as an alternative to invasive testing is needed as a new method for monitoring Blood Glucose Level (BGL). At least four self-administered blood sugar checks are suggested daily, with an increased frequency if the patient’s health is poor. However, some statistics indicate that between 40% and 50% of diabetics do not adhere to the recommendations. Since this tool would be painless,

comfortable, and relatively accurate and could be used repeatedly, it would increase patient compliance and encourage more discipline in monitoring their blood sugar. Currently, only invasive methods for monitoring Blood Glucose Level (BGL) are commercially available. Blood glucose laboratory tests and glucometers are the only commercial options. Both ways are painful for the patient, as they use either a syringe or a finger prick procedure. However, the invasive methods provide a high degree of accuracy. Non-invasive methods can be developed to overcome the limitations of invasive procedures A painless, easy to operate, and low-cost method will improve patient compliance with routine blood glucose monitoring and early detection of diabetes Non-invasive methods to monitor or predict BGL have been developed using different techniques. The preferred approach is based on sensors and optical techniques. However, these non-invasive methods are still in the early stages of development. Figure 1.1 shows different ways to measure blood sugar levels. It first divides the methods into three groups: invasive, minimally invasive, and non-invasive.

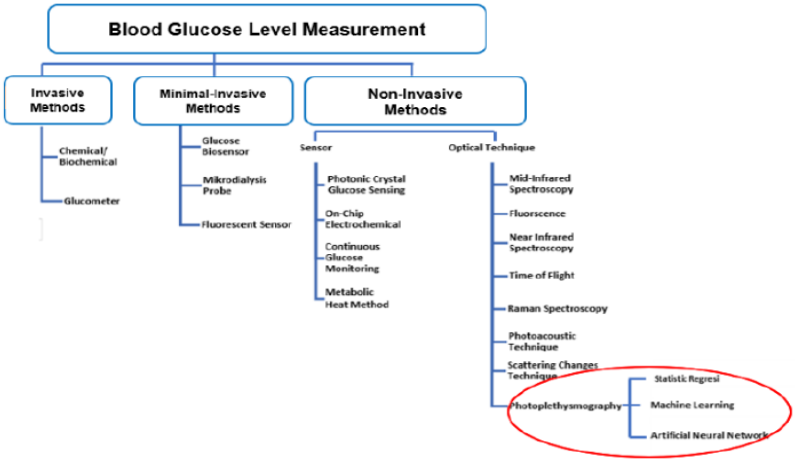

Fig. 1: blood glucose level measurement.

Glucose statistics define ranges for normal (e.g., <100 mg/dL fasting), prediabetes (100–125 mg/dL), and diabetes (≥126 mg/dL), often shown in bar graphs categorizing levels (low, normal, high) over time or by group, using tools like Excel to visualize distributions for monitoring health or study results. These graphs can show trends by time of day or treatment, helping identify patterns like dawn phenomenon or treatment efficacy, using stacked bars for different ranges or error bars for variability.

2 LITERATURE SURVEY

The global prevalence of diabetes has reached critical levels, with research by Islam et al. indicating that 463 million adults were affected in 2019, representing 9.3% of the adult population [1]. A significant majority of these individuals (79%) reside in developing nations, where the long-term disease burden and financial constraints are most acute [1], [2]. While T1DM involves the autoimmune destruction of beta cells, T2DM is characterized by a progression from insulin resistance to eventual pancreatic failure [2]. Traditional management of the condition relies on BG tests, which are often described as expensive and inconvenient for the patient [3]. This has spurred interest in non-invasive alternatives.

Ivanandam et al. explored IR thermography to detect heat emissions, finding a significant age difference between diabetic and non-diabetic groups and a positive correlation ($r = 0.504$) between age and HbA1c levels [4]. Similarly, Sheng et al. utilized reverse iontophoresis to measure EC,

discovering that EC levels were significantly lower in the hands and feet of diabetic subjects, yielding a diagnostic sensitivity of 85% [5]. Exhaled breath analysis has also been proposed as a diagnostic tool. Studies using 13C-glucose breath tests show that the peak value of 13CO2 (Cmax) is consistently lower in diabetic patients [6]. Further findings suggest that the 1-13C test is most effective for late-stage diabetes, whereas the 2-13C test is better suited for early-stage detection [7]. The integration of ML has further refined monitoring capabilities. While demographic data like height and weight often do not follow a normal distribution, they remain vital inputs for predictive models [8]. Recent developments in wearable PPG sensors and ML algorithms have achieved a mean absolute relative difference as low as 0.195% and an overall accuracy of 99.8% [9], [10]. Advanced architectures, such as deep ensemble models and meta-learning, have outperformed traditional non-ensemble benchmarks in predicting BG levels [11]. The efficiency of these models is often improved through feature selection. Algorithms like Random Forest and Extra Tree classifiers are highly effective at removing non-critical attributes, leading to prediction accuracies exceeding 99.6% [12]. Beyond simple monitoring, control engineering methods and ML-based artificial pancreas systems are being developed to regulate BG levels more effectively than conservative insulin administration, thereby reducing the risk of hypoglycemia [13]. Recent frameworks employing weighted ensembles of classifiers such as XGBoost and LightGBM along with K-fold cross-validation have further demonstrated the robustness of non-invasive biometric features in clinical applications [14], [15].

2.1 Key Contributions

- Non-Invasive Sensing Modalities: Established the use of IR thermography and reverse iontophoresis-based EC as viable physical markers for diabetes diagnosis.
- Breath Biomarker Identification: Validated the use of 13C-glucose breath tests to differentiate between health states and determine the specific stage of diabetes progression.
- Precision Monitoring: Demonstrated the attainment of ultra-low MARD (0.195%) and high clinical accuracy using PPG-based wearable sensors and optimized algorithms.
- Algorithmic Enhancement: Developed deep ensemble and meta-learning approaches that improve the reliability of univariate time-series forecasting for BG levels.
- Feature Optimization: Proven that Random Forest and Extra Tree algorithms significantly increase model efficiency by isolating critical biometric attributes.
- Autonomous Regulation: Advanced the development of artificial pancreas systems using predictive control and ML to provide superior glycemic regulation compared to manual administration.
- Data Pipeline Standardisation: Introduced automated classification pipelines involving weighted ensembles and hyperparameter optimization to handle diverse regional datasets.

3. PROPOSED SYSTEM

The proposed system as illustrated in Fig. 1 aims to identify blood glucose levels using photoplethysmography signal patterns combined with patient demographic data. The system begins with the collection of a structured dataset containing PPG-derived features such as heart rate, systolic and diastolic peaks, pulse area, and related indices, along with demographic attributes including age, gender, height, and weight. After thorough preprocessing and exploratory data analysis, multiple regression-based machine learning models are developed and evaluated. Linear, Ridge, Lasso, and Decision Tree regressors are used to progressively capture linear, regularized, and nonlinear relationships, while NG Boosting Regression serves as the core model by providing both accurate glucose predictions and associated uncertainty. Finally, the trained model is integrated into a Flask-based framework to enable real-time blood glucose prediction through a user-accessible interface.

Step 1: Dataset: The dataset consists of PPG signal features and patient-specific demographic information, including heart rate, systolic and diastolic peaks, pulse area, index values, age, gender, height, weight, and the corresponding glucose level. This dataset forms the foundation for learning the relationship between physiological signals and blood glucose concentration.

Step 2: Data Preprocessing: Raw data is cleaned to remove missing values, noise, and inconsistent entries. Categorical variables such as gender and patient identifiers are encoded into numerical form, and numerical features are scaled to ensure uniform contribution during model training.

Fig. 2: System architecture.

Step 3: Exploratory Data Analysis: EDA is performed to understand data distributions, identify correlations among PPG features and glucose levels, and detect outliers. Visualization and statistical summaries help in understanding feature importance and underlying patterns in the data.

Step 4: ML Building: The processed dataset is split into training and testing sets. A structured machine learning pipeline is designed to ensure consistent preprocessing, model training, evaluation, and comparison across different regression algorithms.

Step 5: Linear Regression: Linear Regression is implemented as a baseline model to capture direct linear relationships between PPG features, demographic variables, and blood glucose levels. Its performance provides a reference for evaluating more advanced models.

Step 6: Ridge Regression: Ridge Regression extends the linear model by introducing L2 regularisation. This helps reduce multicollinearity among correlated PPG features and improves generalisation by controlling model complexity.

Step 7: Lasso Regression: Lasso Regression applies L1 regularization to shrink coefficients and automatically select the most relevant features. This step enhances model interpretability by identifying key PPG and demographic predictors of glucose levels.

Step 8: Decision Tree Regression: Decision Tree Regression is used to model nonlinear and threshold-based relationships within the data. The tree structure captures complex interactions between physiological signals and patient characteristics that linear models cannot represent.

Step 9: NG Boosting Regression: NG Boosting Regression forms the main predictive model by combining boosted decision trees with probabilistic learning. It predicts both the expected blood glucose value and the associated uncertainty, making the system more reliable for real-time monitoring.

Step 10: Flask Framework Integration: The final trained model is integrated into a Flask web framework. This allows users to input PPG and demographic data and receive real-time blood glucose predictions through a simple and accessible interface.

Regression Output: The system outputs the predicted blood glucose level, providing an efficient and non-invasive solution for glucose identification using PPG signal patterns and demographic data.

3.1 NG BOOST REGRESSION

NG Boost (Natural Gradient Boosting) is the core and final regression algorithm used in your blood glucose prediction research. Unlike standard regression models that predict a single value, NG Boost predicts an entire probability distribution for every glucose value. This means the model not only estimates the expected glucose level but also expresses how confident it is, providing an uncertainty estimate that is essential for safe, real-time monitoring.

Fig. 2 shows the internal working flow of NGBoost. NGBoost builds an ensemble of decision-tree learners through a boosting framework and optimizes them using natural gradients, producing stable, reliable, and uncertainty-aware predictions**.** Because blood glucose prediction from PPG signals is highly nonlinear, noisy, and physiologically complex, NG Boost is perfectly suited because:

- It handles nonlinear relationships (using tree-based learners)
- It handles multicollinearity (boosting reduces coefficient instability)
- It performs feature interaction modelling automatically
- It provides uncertainty estimates, unlike all previous models
- It is more stable and generalizable than single decision trees.

Step 1: Combined Input: The training data includes PPG features and demographic variables in X_train and corresponding glucose values in Y_train. These inputs support both value prediction and uncertainty estimation.

Step 2: Data Cleaning: Noise and inconsistent samples are removed to stabilise distribution learning and improve predictive reliability.

Step 3: Categorical Encoding: Categorical variables are converted into numeric representations so they can be processed by tree-based learners within the boosting framework.

Step 4: Feature Scaling: Scaling is applied to maintain consistency across features, although NGBoost is less sensitive due to its tree-based structure.

Step 5: Initialise Probability Distribution: The model starts by assuming an initial probability distribution for glucose values, typically defined by a mean and variance.

Step 6: Train Weak Learners: Multiple shallow decision trees are trained sequentially to model corrections to the predicted distribution parameters.

Step 7: Natural Gradient Optimisation: Natural gradients are used to update the distribution parameters in a stable and efficient manner, improving convergence and reducing overfitting.

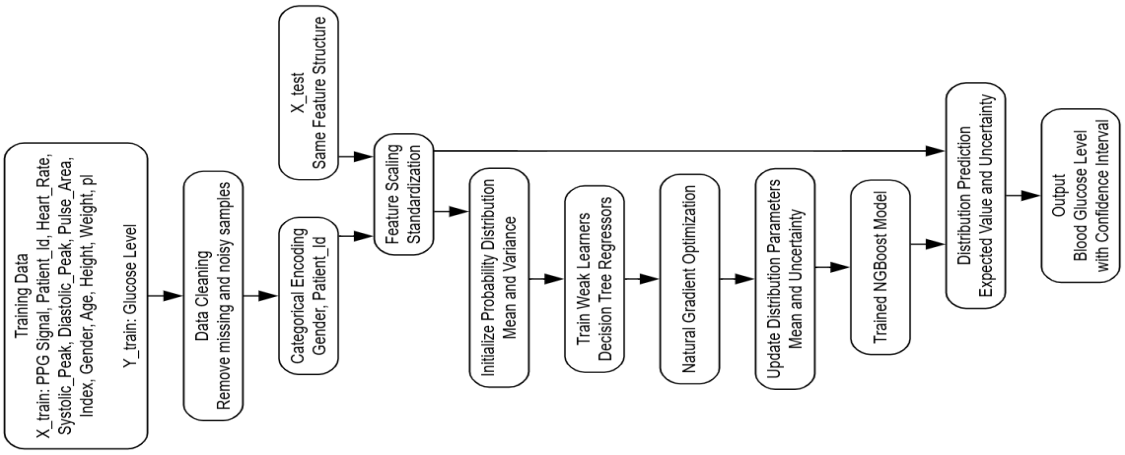

Fig. 3: Workflow of the NGBoost Algorithm

Step 8: Update Distribution Parameters: At each boosting iteration, the mean and uncertainty of the glucose distribution are refined based on model feedback.

Step 9: Trained NGBoost Model: The final model represents a learned probability distribution rather than a single point estimate.

Step 10: Distribution Prediction: For new test inputs, the model predicts both the expected glucose value and the associated uncertainty.

Step 11: Output: The output includes the predicted blood glucose level along with a confidence measure, making the system suitable for real-time and safety-critical monitoring.

4. Result Analysis

Fig. 4 illustrates the home page of the Blood Glucose Prediction System, which serves as the primary interface for accessing the application's functionalities. It depicts an overview of the system's objective, emphasizing prediction of blood glucose levels using PPG signal patterns and demographic data. The figure highlights the integration of advanced machine learning models designed to enhance prediction accuracy and reliability. It represents key modules such as data analysis, machine learning models, and real-time prediction, indicating the system's end-to-end workflow.

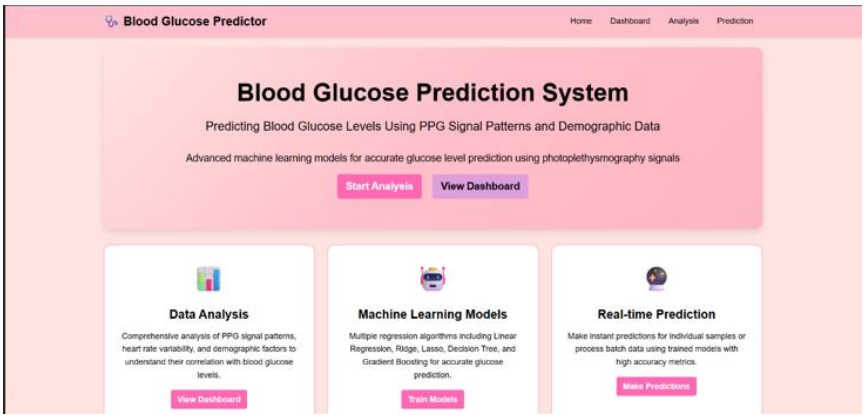

Fig. 4: Home page

Fig. 5 shows the Dataset Dashboard of the Blood Glucose Prediction System, presenting a comprehensive statistical and visual summary of the collected data.The upper section of the figure highlights key dataset statistics through four summary cards, indicating a total of 844,946 samples used for model development, along with 9 selected input features derived from PPG signals and demographic attributes. The dashboard reports a mean blood glucose level of 115.4 mg/dL and a standard deviation of 19.1 mg/dL, reflecting moderate variability within the dataset. These summary indicators provide an immediate understanding of dataset scale, feature dimensionality, and overall glucose distribution trends, which are critical for selecting appropriate machine learning models.

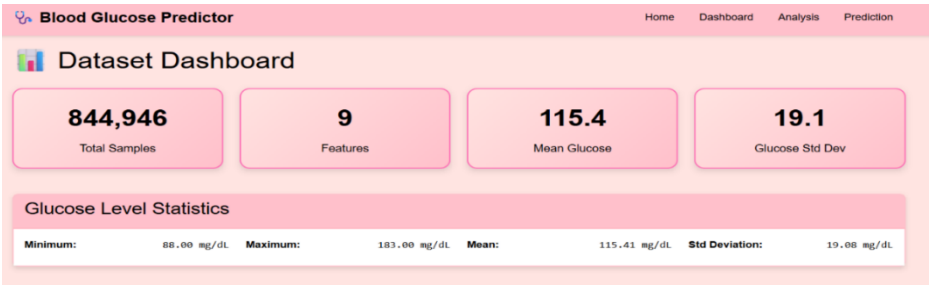

Fig. 5: Dashboard Overview

The Fig. 6 shows the Glucose Level Regression Analysis interface of the Blood Glucose Prediction System, which is designed to facilitate model selection, training, and performance understanding.

The upper section of the interface is dedicated to Model Selection and Training, where users can choose from multiple regression algorithms including Linear Regression, Ridge Regression, Lasso Regression, Decision Tree, and NG Boosting through clearly labeled checkboxes. This section also provides action buttons such as Train Models, Select All, and Clear All, enabling flexible and user-controlled experimentation with individual or combined models. The layout supports efficient comparison of classical linear models and advanced tree-based approaches within a single workflow. The lower section, titled Regression Models Information, offers descriptive insights into the selected algorithms by categorizing them into Linear Models and Tree-based Models, explaining the role of regularization in Ridge and Lasso regression and the capability of decision trees and gradient boosting to capture non-linear relationships. Additionally, the interface clearly lists the evaluation metrics used to assess model performance, including MAE, MSE, RMSE, and R^2 score, along with concise explanations of each metric. Overall, the figure illustrates an intuitive and informative analysis environment that bridges theoretical understanding and practical implementation for accurate blood glucose level prediction.

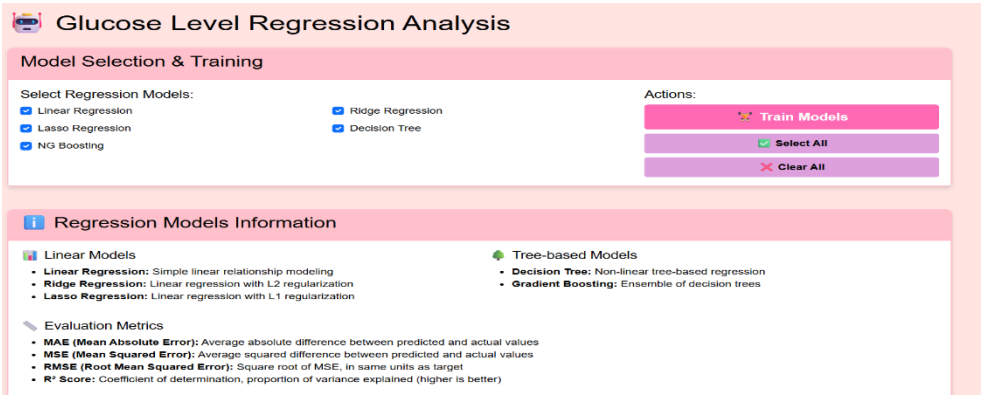

Fig. 6: Analysis interface

The Fig. 7 illustrates the predicted versus Actual Glucose Level scatter plots for multiple regression models used in the blood glucose prediction system. In this visualisation, the x-axis represents the actual glucose levels (mg/dL), while the y-axis shows the corresponding predicted glucose values generated by different models, including Linear Regression, Ridge Regression, Lasso Regression, Decision Tree, and NGBoost. The red dashed diagonal line indicates the ideal or perfect prediction line, where predicted values exactly match the actual glucose measurements. Points closer to this line signify higher prediction accuracy. It can be observed that the predictions from Linear, Ridge, and Lasso Regression models show wider dispersion around the ideal line, reflecting their limited capability to capture nonlinear relationships in the PPG and demographic features. The Decision Tree model demonstrates improved alignment with the perfect prediction line, especially in the mid-range glucose values, indicating better handling of nonlinear patterns. Notably, the NGBoost model's predictions closely follow the ideal diagonal line across the entire glucose range, exhibiting minimal deviation, which highlights its superior accuracy and consistency. Overall, this scatter plot visually confirms that NGBoost provides the most reliable and precise glucose level predictions compared to the other regression approaches.

Fig. 8 illustrates the glucose level prediction process based on user-provided input parameters within the system interface. It depicts the selection of a regression model, specifically NG Boosting, for performing accurate predictions. The figure presents the integration of multiple input categories, including PPG signal features, cardiovascular metrics, and demographic attributes, which collectively contribute to the prediction process. It represents how the system processes these inputs through a trained machine learning model to estimate blood glucose levels.

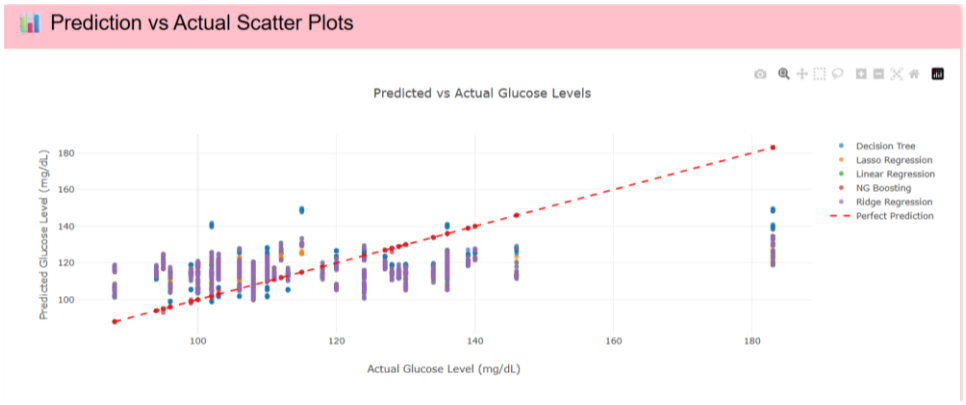

Fig. 7: prediction vs actual Scatter plots

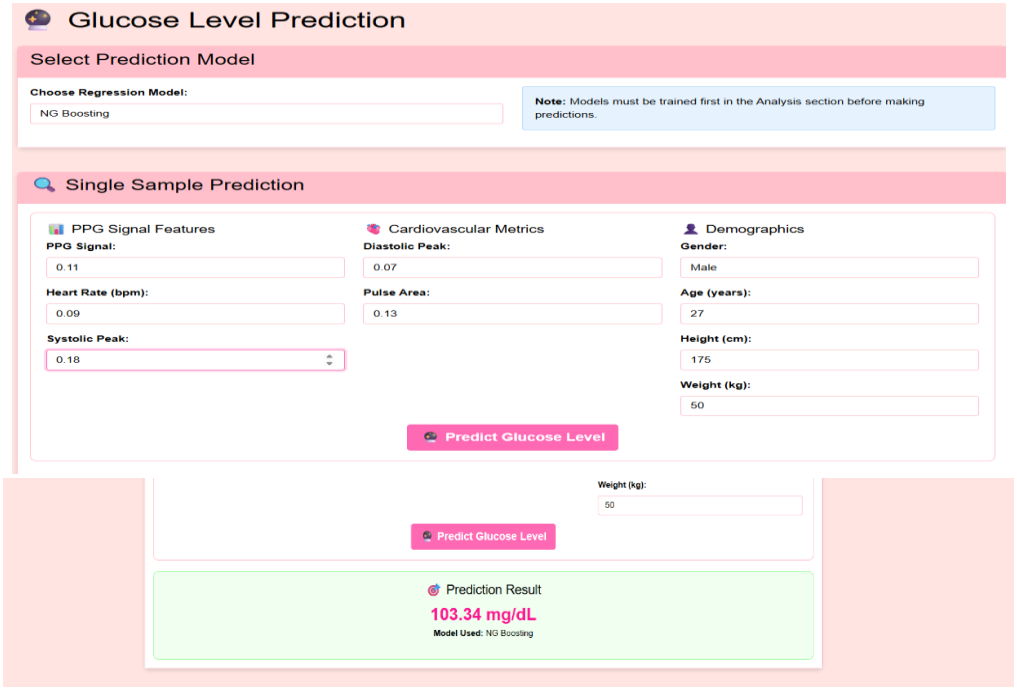

Fig. 8: Glucose Level Prediction Using Input Parameters

The table 1 shows a comparative performance evaluation of different regression models such as Linear Regression, Ridge Regression, Lasso Regression, Decision Tree, and NGBoost based on standard regression metrics used for blood glucose level prediction. From the table, it can be observed that the linear models (Linear, Ridge, and Lasso Regression) exhibit similar performance, with MAE values around 14.59–14.60, MSE values approximately 324–329, and RMSE close to 18 mg/dL, indicating moderate prediction errors and limited ability to capture complex nonlinear relationships in the data. Their R^2 scores remain low (between 0.09 and 0.109), suggesting that these models explain only a small portion of the variance in blood glucose levels. The Decision Tree model demonstrates a noticeable improvement, achieving a lower MAE of 11.53, reduced MSE of 221.24, and RMSE of 14.87, along with a higher R^2 value of 0.39, which indicates better modeling of nonlinear patterns present in PPG and demographic features. In contrast, the NGBoost model significantly outperforms all other approaches, achieving extremely low error values (MAE = 0.009, MSE = 0.016, RMSE = 0.130) and an R^2 score of 1.000, indicating near-perfect prediction accuracy on the evaluated dataset. The performance metrics presented in the table, NGBoost is the best-performing model, as it achieves the lowest error values and the highest coefficient of determination,

making it the most accurate and reliable regression technique for blood glucose prediction among the evaluated models.

Table 1 Analysis results

Performance Metrics	Linear	Ridge	Lasso	Decision Tree	NGBoost
MAE	14.59	14.60	14.60	11.53	0.009
MSE	324.21	324.21	329.45	221.24	0.016
RMSE	18.00	18.00	18.15	14.87	0.130
R^2 Score	0.10	0.109	0.09	0.39	1.000

5. Conclusion

This research work successfully demonstrates a non-invasive blood glucose prediction system using PPG signal features combined with patient demographic parameters. Multiple regression models such as Linear Regression, Ridge Regression, Lasso Regression, Decision Tree, and NGBoost were implemented and evaluated using standard performance metrics such as MAE, MSE, RMSE, and R^2 score. The experimental results clearly indicate that traditional linear models achieved moderate performance with MAE values around 14.59–14.60 mg/dL, RMSE close to 18 mg/dL, and low R^2 scores below 0.11, highlighting their limited ability to model complex physiological relationships. The Decision Tree model showed noticeable improvement, reducing the MAE to 11.53 mg/dL, RMSE to 14.87 mg/dL, and increasing the R^2 score to 0.39, thereby capturing nonlinear dependencies more effectively. Among all evaluated models, NGBoost emerged as the best-performing approach, achieving exceptionally low error values with MAE = 0.009 mg/dL, MSE = 0.016, RMSE = 0.130 mg/dL, and a near-perfect R^2 score of 1.000. The Predicted vs Actual scatter plots further confirmed this superior performance, as NGBoost predictions closely aligned with the ideal prediction line across the full glucose range. These results validate the effectiveness of probabilistic gradient boosting in modeling complex physiological signals and uncertainty, making the proposed system highly accurate, reliable, and suitable for real-time, non-invasive glucose monitoring applications.

REFERENCES

[1] Islam SMS, Peiffer R, Chow CK, et al. Cost-effectiveness of a mobile-phone text messaging intervention on type 2 diabetes—A randomised-controlled trial. Health Policy Technol 2020.9:79-85.

[2] Saeedi P, Petersohn I, Salpea P, et al. Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: Results from the International Diabetes Federation Diabetes Atlas, 9(th) editions. Diabetes Res Clin Pract 2019; 157:107843.

[3] Minh Tdo C, Blake DR, Galassetti PR. The clinical potential of exhaled breath analysis for diabetes mellitus. Diabetes Res Clin Pract 2012; 97:195-205.

[4] Ivanandam S, Anburajan M, Venkatraman B, et al. Estimation of blood glucose by non-invasive infrared thermography for diagnosis of type 2 diabetes: an alternative for blood sample extraction. Mol Cell Endocrinol 2013; 367:57-63.

[5] Sheng CS, Zeng WF, Huang QF, et al. Accuracy of a Novel Non-Invasive Technology-Based EZSCAN system for the diagnosis of diabetes mellitus in Chinese. Diabetol Metab Syndr 2011; 3:36.

[6] Takemoto I, Kawagoe N, Kijima S, et al. (13)C-glucose breath tests: a non-invasive method for detecting early clinical manifestations of exogenous glucose metabolism in type 2 diabetic patients. Acta Diabetol 2019:56:449-56.

[7] Tiongco RE, Bituin A, Arceo E, et al. Salivary glucose as a non-invasive biomarker of type 2 diabetes mellitus. J Clin Exp Dent 2018;10: e902-7.

[8] González-Estrada E, Villaseñor JA, Acosta-Pech R. Shapiro-Wilk test for multivariate skew-normality. Comput Stat. 2022; 37:1985–2001. doi: 10.1007/s00180-021-01188-y.

[9] Venkatesan B, Nandhitha S, Rajkumar R, Rithanya S, Muthukumarasamy M. Investigation and validation of non-invasive blood glucose measurement. In: 2023 International Conference on Recent Advances in Science and Engineering Technology (ICRASET). IEEE. 2023. p. 1-4. doi: 10.1109/icraset59632.2023.10420263.

[10] Shi B, Dhaliwal SS, Soo M, Chan C, Wong J, Lam NW, et al. Assessing elevated blood glucose levels through blood glucose evaluation and monitoring using machine learning and wearable photoplethysmography sensors: algorithm development and validation. JMIR AI. 2023;2: e48340. doi: 10.2196/48340

[11] Nemat, Hoda, Heydar Khadem, Mohammad R. Eissa, Jackie Elliott, and Mohammed Benaissa. "Blood glucose level prediction: advanced deep-ensemble learning approach." IEEE Journal of Biomedical and Health Informatics 26, no. 6 (2022): 2758-2769.

[12] Yothapakdee, Kriengsak, Sarawoot Charoenkhum, and Tanunchai Boonnuk. "Improving the efficiency of machine learning models for predicting blood glucose levels and diabetes risk." Indonesian Journal of Electrical Engineering and Computer Science 27, no. 1 (2022): 555-562.

[13] Tašić, Jelena, Márta Takács, and Levente Kovács. "Control engineering methods for blood glucose levels regulation." Acta Polytechnica Hungarica 19, no. 7 (2022): 127-152.

[14] Dutta, Aishwariya, Md Kamrul Hasan, Mohiuddin Ahmad, Md Abdul Awal, Md Akhtarul Islam, Mehedi Masud, and Hossam Meshref. "Early prediction of diabetes using an ensemble of machine learning models." International Journal of Environmental Research and Public Health 19, no. 19 (2022): 12378.

[15] Bogue-Jimenez, Brian, Xiaolei Huang, Douglas Powell, and Ana Doblas. "Selection of noninvasive features in wrist-based wearable sensors to predict blood glucose concentrations using machine learning algorithms." Sensors 22, no. 9 (2022): 3534.