The Bioscan AN INTERNATIONAL QUARTERLY JOURNAL OF LIFE SCIENCES 21(2): S.I (2), 764-779, 2026 www.thebioscan.com

Diabetes Prediction using PPG Signal and Clinical data with XGBoost Classification

J. Jannathul Firthous[1,2], Dr. G. Murugeswari[3]

[1]Research Scholar, Reg No: 18221192162012, Department of Computer Science and Engineering, Manonmaniam Sundaranar University, Abishekapatti, Tirunelveli-627012, Tamil Nadu, India

[2]Assistant Professor, Department of Computer Science, Sadakathullah Appa College, Affiliated of Manonmaniam Sundaranar University, Tirunelveli-627011, Tamil Nadu, India

[3]Associate Professor, Department of Computer Science and Engineering, Manonmaniam Sundaranar University, Abishekapatti, Tirunelveli-627012, Tamil Nadu, India

DOI: https://doi.org/10.63001/tbs.2026.v21.i02.S.I(2).pp764-779

KEYWORDS
Diabetes Prediction, PPG Signal, Machine Learning, PapaGei Model, Clinical Data, Hand-crafted Features, Multimodal Learning.

Received on: 14-04-2026

Accepted on: 04-05-2026

Published on: 13-05-2026

Abstract

Diabetes mellitus (DM) is a long-term metabolic disease marked by high blood sugar levels brought on by either inadequate insulin synthesis or inefficient insulin usage. It is a significant worldwide health issue linked to serious consequences like neuropathy, renal failure, and cardiovascular disorders. For diabetes to be well managed and long-term problems to be avoided, early detection and precise prediction are essential. This work combines advanced machine learning techniques with physiological signal analysis to propose an effective model named as PPG-XGei for the prediction of diabetes mellitus. The main input of the proposed model is Photoplethysmography (PPG) digitized signals, which are processed using a deep representation model (PapaGei) to produce PPG embeddings. To provide a complete fused representation, these learnt features are combined with 86 hand-crafted PPG features and 13 clinical information elements. The model can capture both physiological patterns and clinical signs related to diabetes due to this multimodal feature fusion. For the final prediction, the merged feature set is fed to XGBoost classifier. With an accuracy of 96.34%, precision of 94.12%, recall of 97.79%, F1-Score of 95.92%, and an AUC of 98.3%, the experimental findings show excellent performance. Overall, the work shows how well hand crafted features and clinical data can be combined with deep learning based embeddings to provide scalable and accurate diabetes prediction.

1. INTRODUCTION

One of the most prevalent chronic illnesses in the world, diabetes mellitus (DM) is characterized by consistently elevated blood glucose levels brought on by reduced insulin secretion or activity.

Since diabetes is linked to serious problems such as kidney damage, eye loss, and cardiovascular issues, its rising prevalence has become a significant public health concern. To lower these risks and improve patients' health conditions, early detection and ongoing observation are crucial.

Non-invasive methods of measuring human physiological parameters are common nowadays.

For this reason, the PPG system is becoming more popular in the medical research community and medical care. Adoption of the PPG system is primarily motivated by its wearability, ease of construction, and minimal hardware requirements. Heart rate (HR), breathing rate (BR), and the percentage of oxygen in blood (SpO2) are consistently measured using the PPG signals; however, additional clinical measurements, such as blood pressure, blood glucose, and others, are still in the research stage and are rarely commercially available. Additionally, motion artifacts can affect PPG devices. However, PPG signals presents a number of difficulties, including inter-subject variability, motion artifacts, and noise susceptibility, making precise feature extraction and analysis. Furthermore, physiological data may contain intricate and hidden patterns that cannot be captured by relying just on manufactured features. Managing high-dimensional feature spaces, guaranteeing model robustness and generalization, and successfully integrating diverse data sources such as PPG signals and clinical metadata present another significant difficulty.

This work investigates the combination of cutting-edge deep learning and machine learning approaches, driven by the need for a non-invasive, accurate, and scalable diabetes prediction system. Combining the benefits of multimodal learning with deep learning models' capacity to automatically extract meaningful representations from unprocessed data offers a solid basis for enhancing prediction accuracy. Specifically, a more thorough knowledge of the underlying physiological and medical issues is made possible by utilizing PPG signals in conjunction with clinical and handcrafted aspects. The PapaGei model is used in this work as a crucial element for deep feature extraction from unprocessed PPG signals. The main benefit of employing the PapaGei model is its capacity to carry out end-to-end representation learning, automatically producing high-dimensional embeddings that capture intricate temporal correlations and minute physiological traits linked to diabetes mellitus. In contrast to conventional handcrafted methods, it minimizes human bias, preserves important information through compact yet expressive feature representations, and lessens reliance on manual feature engineering.

The study uses the XGBoost (Extreme Gradient Boosting) algorithm, a potent and effective ensemble machine learning method based on decision trees, to carry out the final classification. XGBoost builds several weak learners one after the other, with each new tree concentrating on fixing the mistakes of the preceding ones.

It makes use of parallel processing for quicker calculation, regularization strategies to avoid overfitting, and gradient boosting optimization. XGBoost is especially well-suited for this multimodal framework because of its capacity to manage complicated feature interactions and high-dimensional data. It makes good use of the fused characteristics to increase prediction robustness, accuracy, and precision.

Overall, the suggested system addresses important issues in diabetes prediction by combining deep learning with sophisticated ensemble classification. The primary contributions of this research include the use of the PapaGei model for automated feature extraction from physiological signals, followed by the integration of handcrafted features to capture domain-specific patterns. Additionally, clinical metadata features are incorporated to enhance the contextual understanding of patient health. These combined features are then optimized using Principal Component Analysis (PCA) for dimensionality reduction and improved efficiency. Finally, the XGBoost model is employed for accurate classification, resulting in a scalable, reliable, and non-invasive framework for early Diabetes Mellitus prediction and improved healthcare decision-making.

This paper is organized as follows: Section II presents a comprehensive review of existing state-of-the-art research related to non-invasive diabetes prediction and PPG-based analysis. Section III describes the proposed methodology, including the PapaGei-based PPG embedding extraction, multimodal feature fusion, and the XGBoost classification framework. Section IV discusses the experimental results, and discussion of the proposed system. Finally, Section V concludes the paper by summarizing the key findings and outlining potential directions for future research.

2. RELATED WORK

Ali et al. [1] combined explainable AI methods like SHAP and LIME with handmade features, deep learning One-Dimensional Statistical Convolutional Neural Network model, and XGBoost to develop a non-invasive diabetes screening method. For better prediction, the technique combines feature fusion with multi-channel waveform representations (PPG, VPG, and APG). Better performance is evidenced by the experimental results. In order to increase sample quantity and model performance, M.X. Xiao et al.

[2] applied the instantaneous frequency of maximal energy frequency (f_{Emax}), calculated from toe PPG signals to improve diabetes and neuropathy prediction. Artificial Neural Networks, Logistic Regression, and Fisher discriminant analysis provided the improved accuracy of 93.07%, 86.30% respectively for type-2 diabetic prediction. Islam, T.T et al. [3] implemented a non-invasive technique for glucose measurement by smartphone videos based PPG signals. Machine learning models like Principal Component Regression (PCR), Partial Least Square Regression (PLS), Support Vector Regression (SVR), and Random Forest Regression (RFR) are used for feature extraction and prediction. The PLS outperformed other techniques for diabetic prediction. Angielevska. A et al. [4] focused on classification of HbA1c levels using heart rate variability (HRV) features such as SDNN, RMSSD, NN50, and PNN50, combined with multiple machine learning models. A preprocessing pipeline and patient-wise data evaluation ensure reliable analysis. The machine learning model achieved an F1 score of 93.20% demonstrating strong effectiveness and robustness in HbA1c classification. Using PPG signals and subject metadata a two-stage hybrid machine learning approach for non-invasive blood glucose prediction was developed [5]. XGBoost and CatBoost models are used for diabetic prediction pre-processing and feature extraction. This method is claimed as an appropriate method for wearable-based continuous glucose monitoring since it guarantees both interpretability and reliability.

Chandel, A.k et al. [6] deals with Fourier-Bessel Transform (FBT) properties taken from PPG signals to present a non-invasive glucose monitoring method.

Further Feature selections, machine learning models were used for diabetic prediction. Random Forest produced the highest accuracy of 94.9%. A low-cost, non-invasive glucose monitoring device that makes use of machine learning and PPG signals is suggested by Ali, J. et al. [7]. Feature extraction and real-time glucose level prediction are carried out by processing data obtained through an ESP32 and MAX30105 sensor. Arumugam, P. et al. [8] presented a hybrid WOA-XGBoost model to improve Type-2 Diabetes Mellitus prediction by optimizing hyperparameters The model was evaluated on PIMA and Diabetes risk prediction datasets and it is observed that an accuracy of 95.4% was achieved. Kartikasari, P. et al. [9] developed a XGBoost based model for early diabetes diagnosis. With an accuracy of 97.12%, the model proved to be useful for accurate forecasting.

Using XGBoost in conjunction with SMOTE for class balancing and Bayesian Optimization for hyperparameter tuning, Ulum, M. N. I. et al. [10] created a diabetes classification model. The model was tested on Pima Indians dataset. Using XGBoost and Bayesian Optimization for effective hyperparameter tuning, Muqtadir, A. et al. [11] created the diabetes detection model that utilizes the XGBoost and Bayesian Optimization and it outperformed well among the traditional models such as LihtGBM, Decision Tree, Logistic Regression, KNN, SVM and Navie Bayes.

In order to improve feature selection [12] assessed Random Forest and XGBoost for diabetes prediction. It attained the accuracy of 83.5% with XGBoost and 86% with Random Forest. Using the XGBoost classifier and key clinical characteristics like blood pressure, insulin, and glucose, Jayadhas, S. A. et al. [13] suggested a diabetes prediction model. To improve model performance, data preprocessing methods such as handling missing values and standardization were used. Metrics like accuracy, precision, recall, F1-score, and ROC-AUC were used to assess the model. It outperformed conventional classifiers with an accuracy of 81.25% and an AUC of 84%. Saha, S, et al. [14] suggested a non-invasive diabetes diagnostic approach using photoplethysmogram (PPG) signals. Spectral entropy, Shannon energy, and other time-frequency domain properties are retrieved from the PPG data.

Machine learning methods such as Random Forest and extreme gradient boosting are used to evaluate the significance of each feature in diabetes prediction. Using these features several machine-learning models such as Random Forest and extreme gradient boosting are used to classify the PPG signals into normal or diabetes groups. Monte-Moreno et al. [15] built datasets from the MSLR database and used Random Forest model with 33 PPG waveform features and BMI, age, and weight data for the prediction of diabetics. Using concatenated infrared PPG and ECG signals Padmavilochanan, D. et al. [16] suggested a Gated Convolutional Neural Network with light-weight dimension reduction for capturing both spatial and temporal dependencies for improved diabetes detection. Using paired PPG

signals and the MSLR database, Liu WR et al. [17] developed a unique deductive learning technique that provided an accuracy of 93.5%. Li J et al. [18] suggested a non-invasive blood glucose monitoring system that combines ECG and PPG signals using spatiotemporal feature fusion and a Choquet integral–based multimodel approach to achieve high clinical accuracy of 93.4%. C. Sonawane et al.

[19] uses machine learning algorithms such as Support Vector Machine (SVM) and Decision Tree, on the Pima Indian Diabetes dataset to predict diabetes based on features like glucose, insulin, and BMI. The models were evaluated using accuracy, precision, and recall. SVM produced 76.6% of accuracy, 75% for Decision Tree and making it a more effective method for diabetes prediction. Papri, N. J. et al. [20] developed a non-invasive IoT-based system using PPG signals and machine learning models such as XGBoost, Random Forest, SVM, Logistic Regression for diabetes detection. Using ESP32 hardware and AWS cloud for real-time monitoring, the system achieved 87.88% and 90% accuracy, showing an effective solution for early detection.

The majority of current research focuses on sophisticated feature extraction or high-accuracy models, but these approaches fall short of providing a comprehensive, useful solution. Many studies use sophisticated deep learning models that are challenging to apply in low-cost settings, rely on single datasets, or lack real-time implementation.

Our approach aims to close these gaps by integrating machine learning and non-invasive PPG readings into a single, cohesive solution. Since research indicates that no single model consistently performs best across all datasets, the need to improve reliability drives the use of the XGBoost classifier. Our method is also closes the main gaps found in the literature by offering a real-time diabetes detection system that is both accurate and useful for practical healthcare applications.

3. METHODOLOGY

The proposed methodology of PPG-XGei model is shown in Figure 1 which follows a structured workflow of Diabetes Mellitus (DM) prediction using multimodal data integration. The methodology is presented in the following sections.

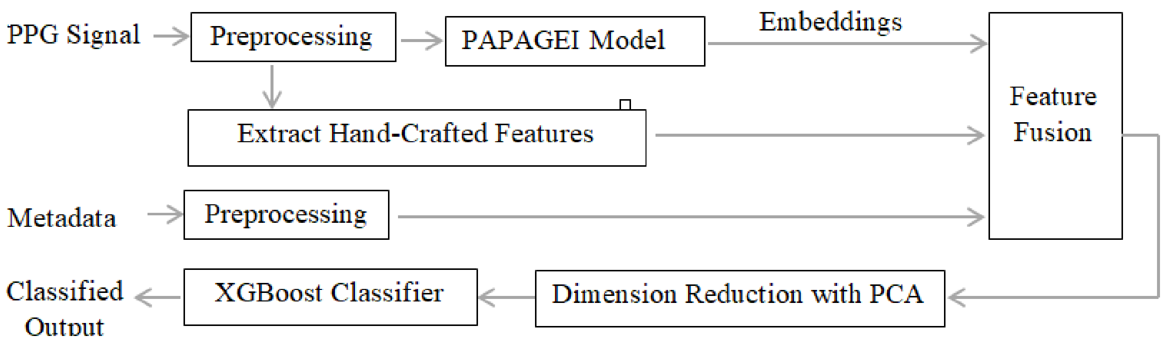

Figure 1- Architecture of proposed PPG-XGei model

3.1 Data Loading

The proposed model begins with the data loading stage, where the dataset is collected and systematically organized for further processing. The input primarily consists of digitized Photoplethysmography (PPG) signals along with associated metadata such as class labels and patient information, which are directly read as time-series numerical data from CSV file formats.

To ensure data quality and reliability, the dataset is carefully examined for noise, inconsistencies, and missing values, and appropriate preprocessing techniques such as noise filtering and missing value handling are applied where necessary. The metadata, including class labels and patient-specific details, is accurately aligned with the corresponding PPG signals using unique identifiers to maintain data integrity.

Finally, data is fed into the subsequent stage like model training, thereby enabling efficient and consistent processing by the proposed model.

3.2 PaPaGei Model

The PaPaGei Model was developed by Pillai et al. [21] as an open foundation model for photoplethysmography (PPG) signals pretrained on over 57000 hours of data comprising 20 million PPG segments from publically available datasets. PaPaGei uses a self-supervised representation learning approach that takes advantage of differences in PPG signal shape to learn significant physiological properties. From these preprocessed data, this technique generates 512 embeddings using ResNet-style CNN encoder in the PaPaGei model. In parallel, domain-specific PPG morphological features, such as inflection point area and reflection index (RI), are extracted separately and fused with the learned representations along with clinical metadata to form a comprehensive feature set. The morphology-aware variant, PaPaGei-S, further enhances the learning process by capturing physiologically relevant characteristics associated with metabolic health indicators such as diabetes.

This hybrid representation enables the model to generalize effectively across domain shifts and diverse subject populations, producing robust and physiologically meaningful embeddings suitable for downstream tasks such as diabetic classification and regression.

3.3 Hand Crafted Features

Handmade features extracted from digitized photoplethysmographic (PPG) signal values are used in machine learning models. Time-domain features capture basic statistical properties of the discrete samples, including mean, standard deviation, skewness, kurtosis, and percentiles that describe the overall variability and distribution. Morphological features focus on the shape of the waveform derived from the sampled values like peaks and troughs, peak intervals, amplitudes, and rise or decay patterns, which are important for analyzing pulse characteristics. Frequency-domain features analyse the digitized samples in terms of its frequency components, including power in different frequency bands, dominant frequency, and spectral entropy, helps to identify hidden periodic patterns. Hjorth parameters namely activity, mobility, and complexity provide a compact representation of signal dynamics and irregularity. Additionally, other features like autocorrelation and differences between successive samples capture temporal dependencies and trends. Together, these features create a comprehensive representation of the digitized signal, improving the performance and accuracy of machine learning models.

3.4 Feature Fusion of PPG-XGei

Feature fusion is the process of combining multiple sources of information to form a unified and more informative representation for prediction. In the proposed architecture, three different types of features are integrated:

(i) embeddings obtained from the PaPaGei model,

(ii) handcrafted PPG features that capture domain-specific physiological characteristics, and

(iii) clinical metadata containing patient-related information.

The PaPaGei model extracts embeddings automatically from digitized PPG signals, while handcrafted features provide interpretable physiological insights, and metadata contributes contextual clinical information. These complementary feature sets are combined at the feature level to enhance the overall representation. This fusion enables the model to leverage data-driven learning and domain knowledge, thereby improving robustness, interpretability, and predictive performance for downstream tasks such as diabetes classification. The fusion process is mathematically represented as:

$$F = F_{papgei} || F_{Hand-craft} || F_{Metadata} \quad \text{........ (1)}$$

Where $F_{papagei}$ denotes the deep feature embeddings learned from the PaPaGei model, $F_{Hand\text{-}craft}$ represents the manually extracted PPG features, and $F_{metadata}$ corresponds to the clinical or patient-specific attributes. The symbol || indicates feature concatenation, meaning that all feature vectors are joined together to form a single unified feature vector F. This combined representation is then used as input for subsequent steps such as dimensionality reduction (PCA) and classification.

3.5 Dimensionality Reduction using PCA

Principal Component Analysis (PCA) is used to preserve the most crucial information in the feature space while reducing dimensionality. Following feature fusion, the dataset includes a large number of features from both manual techniques and deep learning embeddings, which may result in high computational costs, redundancy, and noise. Only the most important features are retained after PCA reduces them to a smaller collection of uncorrelated principal components that capture the greatest variation. This makes the dataset more condensed and useful by lowering noise and crossovers. As a result, PCA increases efficiency by reducing memory consumption, accelerating training, and minimising overfitting while maintaining crucial patterns for precise prediction. Models like XGBoost are subsequently trained using the smaller feature set for final classification.

3.6 XGBOOST Classifier Evaluation

The final model is created with XGBoost (Extreme Gradient Boosting), a very efficient and scalable machine learning technique built on the gradient boosting framework. XGBoost works by generating numerous decision trees successively, with each new tree trained to rectify the faults of the preceding ones. It employs a gradient descent optimization technique to reduce the loss function and enhance forecast accuracy. One of XGBoost's primary features is its ability to efficiently handle high-dimensional data and complicated feature interactions. It also provides built-in regularization algorithms to control overfitting and increase the model's generalizability. Because of its speed, accuracy, and robustness, XGBoost is commonly utilized in structured dataset classification and prediction applications.

In the present work, XGBoost is chosen as the major classification model due to its superior performance and ability to operate efficiently with the fused feature set obtained from deep learning and handcrafted feature extraction. The smaller feature set obtained from PCA is utilized to train the XGBoost model, allowing it to acquire meaningful patterns and correlations for accurate prediction. In addition to XGBoost, with other classifiers it seems to best and so serves as the final model in the proposed system and, its algorithm is shown in Table 1.

Table 1 Algorithm for our proposed PPG-XGei work

```
Input : Raw PPG signals S, Metadata M
Output : 0 or 1 (0 – Non-diabetic and 1 – diabetic)

Begin

   Load(S, M)
   Preprocess_Data()
   Em ← Extract Metadata

   Initialize PaPaGei_Model()
   Ed ← Extract_Deep_Features(S)

   Ft ← Extract_Time_Domain_Features(S)
   Fm ← Extract_Morphological_Features(S)
   Ff ← Extract_Frequency_Domain_Features(S)
   Fh ← Compute_Hjorth_Parameters(S)

   Fhc ← Combine(Ft, Fm, Ff, Fh)

   Fcombined ← Fuse(Em,Ed, Fhc)

   Freduced ← Apply_PCA(Fcombined)

   Model ← Initialize_XGBoost()
   Train(Model, Freduced)
   Test(Model,Freduced)
   Prediction ← Predict(Model, Freduced)

End
```

4. RESULTS AND DISCUSSIONS

4.1 Dataset

This study utilizes the publicly available Guilin People's Hospital PPG dataset hosted on Figshare. The dataset contains 219 samples, each representing a physiological instance produced from Photoplethysmography (PPG) signals. Along with the digitized PPG signals, the dataset includes accompanying physiological information such as systolic and diastolic blood pressure, pulse pressure, body mass index (BMI), and heart rate. These measurements, paired with clinical labels indicating diabetic or non-diabetic status, provide a reliable foundation for supervised model development.

4.2 Model Execution

The model execution phase entails training and assessing the XGBoost classifier on the processed dataset using a 5-fold cross-validation approach to ensure consistent results. In this method, the dataset is separated into five equal subgroups, and the model is trained on four of them while being tested on the fifth. This procedure is repeated five times, so that each subset is tested once. This strategy reduces overfitting and allows for a more broad evaluation of the model. During execution, the XGBoost algorithm learns patterns from input features and iteratively improves predictions by decreasing error using gradient boosting. After completing all folds, the final results are calculated by averaging performance indicators such as accuracy, precision, recall, F1-score, and AUC. This methodical execution guarantees that the model is strong, dependable, and capable of operating effectively with unknown data.

4.3 Experimental Results

The performance of the proposed model is evaluated over 100 epochs using metrics, such as accuracy, precision, recall, F1-score, and Area Under the Curve (AUC), with the dataset split into 80% for training and 20% for testing. The experimental finding in Table 2 compares different models with proposed model for diabetes prediction. The basic CNN model shows the lowest performance of 68.02% accuracy.

Adding PapaGei features improves results of 79.93%, and XGBoost further increases accuracy as 83.35%. Incorporating PCA significantly boosts performance 89.45%. The proposed PPG-XGei model achieves the best results with 96.34% accuracy, 94.12% precision, 97.79% recall, and 95.92% F1-score, and 98.3% of AUC making it the most effective and reliable approach.

Table 2 Performance of our Proposed PPG-XGei for Diabetics Prediction

Model	Feature Type	Accuracy (%)	Precision (%)	Recall (%)	F1Score (%)
CNN	PPG + Derivative	68.02	47.73	49.25	48.47
PapaGei + CNN	Embeddings	79.93	64.77	29.03	40.10
PapaGei + XGBoost	Embeddings	83.35	67.89	66.78	67.33
PapaGei + PCA + XGBoost	PCA-Reduced Embeddings	89.45	90.12	92.23	91.17
Proposed (PPG-XGei)	PapaGei + Handcrafted + PCA	96.34	94.12	97.79	95.92

4.4 DISCUSSION:

The table highlights a clear improvement in model performance with the integration of advanced techniques. The CNN model performs poorly, indicating limited effectiveness with raw data. Adding PapaGei features improves results, proving the importance of deep feature extraction. Using XGBoost further enhances performance due to better classification capability. A significant boost is achieved by combining handcrafted features and PCA, highlighting the value of feature engineering and dimensionality reduction.

The proposed PPG-XGei model delivers the best performance across all metrics, showing that a hybrid approach (deep features + handcrafted features + PCA + XGBoost) provides the most accurate and reliable diabetes prediction.

5. CONCLUSION

This paper suggests an effective methodology for diabetes prediction utilising PPG data that combines deep feature extraction (PaPaGei), handcrafted features, PCA-based dimensionality reduction, and sophisticated machine learning.

While PCA increases efficiency and generalisation by eliminating redundancy, feature fusion offers a rich representation. According to experimental data, the suggested method works better than conventional models like CNN, PapaGei+CNN, PapaGei+XGBoost, PapaGei+Handcrafted Features+PCA+XGBoost, our proposed PPG-XGei achieves the best performance of 96.34% accuracy accompanied with high precision, recall, and AUC, guaranteeing accurate diabetic case diagnosis. Overall, the approach is reliable and appropriate for practical healthcare applications; more advancements could be made with bigger datasets, more effective class imbalance management, and sophisticated deep learning models.

REFERNCES

1. Ali, Mubashir, Jingzhen Li, and Zedong Nie. "PPG Based Digital Biomarker for Diabetes Detection with Multiset Spatiotemporal Feature Fusion and XAI." Computer Modeling in Engineering & Sciences 145.3 (2025): 4153.
2. Xiao, M. X., Lu, C. H., Ta, N., Wei, H. C., Yang, C. C., & Wu, H. T. (2022). Toe PPG sample extension for supervised machine learning approaches to simultaneously predict type 2 diabetes and peripheral neuropathy. Biomedical Signal Processing and Control, 71, 103236.
3. Islam, T. T., Ahmed, M. S., Hassanuzzaman, M., Bin Amir, S. A., & Rahman, T. (2021). Blood glucose level regression for smartphone ppg signals using machine learning. Applied Sciences, 11(2), 618.
4. Angjelevska, A. & Gusev, M. & Gjorgjieva, S.. (2024). Classification of Hemoglobin A1c from Long and Extra-long Term Heart Rate Variability. 1187-1192. 10.1109/MIPRO60963.2024.10569581.
5. Sinha, S. K., Singh, O. N., & Singh, A. A Hybrid Machine Learning Pipeline for Noninvasive Blood Glucose Estimation Using PPG Signals.
6. Chandel, A., Nayak, A., & Gupta, S. (2026). Advancing diabetes management: Machine learning-based non-invasive glucose monitoring using wearable PPG sensors. In Deep Learning for Cardiac Signal Analysis in Robotic Applications (pp. 147-158). Academic Press.
7. Ali, J., Hyder, S. S., Sobia, S., Furqan, E. M., Moorat, S., & Shams, S. (2026). NON-INVASIVE GLUCOSE MONITORING DEVICE USING PHOTOPLETHYSMOGRAPHY SIGNAL AND MACHINE LEARNING

MODEL. Spectrum of Engineering Sciences, 4(3), 415-421.

8. Arumugam, P., Srinivasan, A. S., Mohan, D. B., Kumar, S., Mehta, M., & Haque, M. (2026). A Hybrid Whale Optimization and XGBoost Framework for Accurate Prediction of Type 2 Diabetes Mellitus. Bangladesh Journal of Medical Science, 25(1), 78-90.
9. Kartikasari, P., Warsito, B., Rochayani, M. Y., Arisandi, R., & Rahman, S. D. F. (2026, March). Improving diabetes mellitus disease patients classification using XGBoost method. In AIP Conference Proceedings (Vol. 3411, No. 1, p. 040012). AIP Publishing LLC.
10. Ulum, M. N. I., & Unjung, J. (2026). Enhancing diabetes classification performance using XGBoost integrated with SMOTE and bayesian hyperparameter optimization. Journal of Soft Computing Exploration, 7(1), 9-18.
11. Muqtadir, A., Indah, A. S., & Aufar, A. (2026, March). Bayesian optimization method in XGBoost hyperparameter tuning: a novel approach for robust diabetes disease classification. In Third International Conference on Emerging Trends in AI and Computational Technologies (ICONEST 2025) (Vol. 14141, pp. 194-203). SPIE.
12. Jawza, D. N., Mazdadi, M. I., Farmadi, A., Saragih, T. H., Kartini, D., & Abdullayev, V. (2025). The enhancing diabetes prediction accuracy using random forest and XGBoost with PSO and GA-based feature selection. Journal of Electronics, Electromedical Engineering, and Medical Informatics, 7(2), 295-306.
13. Jayadhas, S. A., & Roslin, S. E. (2025, August). Predicting Diabetes Using XGBoost Classifiers and Multivariate Clinical Factors. In 2025 9th International Conference on Inventive Systems and Control (ICISC) (pp. 669-673). IEEE.
14. Saha, S., Saha, U., Saha, S., & Fattah, S. A. (2024, February). PPG-based feature extraction for type ii diabetes prediction: A Machine Learning Approach. In 2024 2nd International Conference on Computer, Communication and Control (IC4) (pp. 1-6). IEEE.
15. E. Monte-Moreno, “Non-invasive estimate of blood glucose and blood pressure from a photoplethysmograph by means of machine learning techniques,” *Artif. Intell. Med.*, vol. 53, no. 2, pp. 127–138, Oct. 2011.
16. Padmavilochanan, D., Pathinarupothi, R. K., Menon, K. A. U., Kumar, H., Guntha, R., Ramesh, M. V., & Rangan, P. V. (2023). *Diabetes mellitus detection using intelligent convolutional neural network–gated recurrent unit model*.

17. W.-R. Lu, W.-T. Yang, J. Chu, T.-H. Hsieh, and F.-L. Yang, “Deduction learning for precise noninvasive measurements of blood glucose with a dozen rounds of data for model training,” Sci. Rep., vol. 12, no. 1, p. 6506, Apr. 2022.
18. J. Li, J. Ma, O. M. Omisore, Y. Liu, H. Tang, P. Ao, Y. Yan, L. Wang, and Z. Nie, “Noninvasive blood glucose monitoring using spatiotemporal ECG and PPG feature fusion and weight-based choquet integral multimodel approach,” IEEE Trans. Neural Netw. Learn. Syst., vol. 35, no. 10, pp. 1–15, Oct. 2023.
19. C. Sonawane, K. Somwanshi, R. Patil and R. Raut, "Diabetic Prediction Using Machine Algorithm SVM and Decision Tree," *2023 International Conference on Applied Intelligence and Sustainable Computing (ICAISC)*, Dharwad, India, 2023, pp. 1-7, doi: 10.1109/ICAISC58445.2023.10199539.
20. Papri, N. J., Ahmed, A., & Chowdhury, A. (2025). IoT and cloud-based non-invasive diabetes detection system from photoplethysmogram. Discover Internet of Things, 5(1), 57.
21. A. Pillai, D. Spathis, F. Kawsar, and M. Malekzadeh, “PaPaGei: Open Foundation Models for Optical Physiological Signals,” arXiv:2410.20542, 2024.