IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Mathematical Modelling of Engineering Problems
Vol. 12, No. 11, November, 2025, pp. 3829-3840
Journal homepage: http://iieta.org/journals/mmep

Multimodal Approach for Non-Invasive Blood Glucose Estimation Using Fingertip Video

Asawari Chinchanikar[1*], Manisha Dale[2]

[1] Department of Electronics and Telecommunication Engineering, AISSMS Institute of Information Technology, Savitribai Phule Pune University, Pune 411001, India
[2] Department of Electronics and Telecommunication Engineering, MES's Wadia College of Engineering, Pune 411001, India

Corresponding Author Email: chinchanikar.a@gmail.com

https://doi.org/10.18280/mmep.121108

Received: 13 August 2025
Revised: 9 October 2025
Accepted: 16 October 2025
Available online: 30 November 2025

Keywords:
fingertip video analysis, deep learning, smartphone-based health monitoring, ResNet-50, feature importance

ABSTRACT

Conventional blood glucose monitoring methods, such as finger-prick tests and intravenous sampling, are invasive and often cause discomfort, leading to poor adherence and psychological stress. Non-invasive prediction offers a more user-friendly alternative. This study proposes a non-invasive approach for blood glucose estimation via fingertip videos captured by a mobile camera under near-infrared illumination. Three regression models were trained using (i) handcrafted photoplethysmography (PPG) features, (ii) ResNet-50 deep learning features, and (iii) a hybrid feature set. Feature importance analysis guided the selection of the most informative features to reduce redundancy and enhance prediction accuracy. The hybrid approach consistently outperformed single-feature-based models, achieving a coefficient of determination (R^2) of 0.88 and a Mean Absolute Error (MAE) of 14.50 mg/dL. Five-fold cross-validation verified robustness with an average R^2 of 0.89 and MAE of 15.13 mg/dL, while Bland–Altman analysis demonstrated over 90% agreement with reference measurements. These findings demonstrate that integrating ResNet-derived features with handcrafted PPG features significantly enhances predictive performance, validating fingertip video analysis as a feasible, accurate, and low-cost alternative to invasive glucose monitoring.

1. INTRODUCTION

Diabetes mellitus is a globally prevalent metabolic disorder characterized by an impaired ability to regulate blood glucose levels (BGLs). Type 1 diabetes develops due to the autoimmune destruction of pancreatic β-cells, leading to inadequate insulin production [1], while Type 2 diabetes is associated with both reduced insulin sensitivity and impaired insulin secretion. Insulin is essential for glucose metabolism, facilitating the uptake and utilization of glucose by cells. Persistent hyperglycemia is the defining feature of diabetes.

Type 2 diabetes has emerged as a growing public health challenge, with its global prevalence rising steadily over recent decades. As of 2017, it impacted around 462 million people (6.28% of the global population) and contributed to more than 1 million fatalities yearly, ranked as the ninth leading cause of death worldwide. The condition affects both men and women equally, with incidence peaking around the age of 55, and is strongly associated with aging, sedentary lifestyles, and poor dietary habits. Its prevalence is highest among older adults and is projected to rise to 7,079 cases per 100,000 by 2030, highlighting the urgent need for better prevention and management strategies [2].

Current glucose monitoring technologies, such as continuous glucose monitors (CGMs), rely on subcutaneous sensor insertion, making them invasive and prone to complications, including skin irritation, infection, and the need for periodic recalibration. These limitations highlight the critical need for a non-invasive (NI), needle and pain-free glucose monitoring approach. Such a method would improve convenience, safety, and user compliance, offering a transformative shift in diabetes management by reducing reliance on invasive techniques.

In recent years, numerous studies have explored the development of NI and minimally invasive devices for glucose monitoring. Several approaches have focused on sensors that analyze alternative biological fluids, such as saliva [3], tears [4], and sweat [5], as well as optical techniques including mid-infrared spectroscopy, photoacoustic detection [6], and near-infrared (NIR) spectroscopy [7].

Among optical techniques, photoplethysmography (PPG) has gained widespread attention as an NI, low-cost approach for detecting blood volume changes in peripheral circulation. It involves projecting light onto the skin, where a photodetector (or camera) captures the reflected or absorbed light to assess volumetric changes in blood flow. During systole, increased blood volume absorbs more light, while during diastole, reduced volume leads to greater light reflection. This physiological cycle produces fluctuations in light intensity, which are captured as the PPG signal. Previous research has demonstrated that glucose exhibits measurable absorption characteristics in the NIR spectrum [8], making

NIR-based PPG a promising candidate for NI BGL estimation.

Sensor-driven approaches have employed NIR optocoupler pairs to acquire PPG signals, forming the basis for wearable systems aimed at continuous, NI BGL monitoring [9, 10]. In parallel, modern smartphones have emerged as powerful platforms for health monitoring due to their integrated cameras, sensors, and processing capabilities. Their versatility enables remote, real-time, and NI diagnostics through mobile healthcare applications. In this context, the PPG signal can be effectively captured using a smartphone camera in conjunction with an NIR illumination source [11]. Positioning the fingertip on the camera enables video capture, and averaging the pixel intensities within the Region of Interest (ROI) across frames yields a one-dimensional PPG signal.

The process of extracting PPG signals from fingertip video involves converting high-dimensional spatiotemporal data into a one-dimensional physiological waveform. Hence, these temporal variations in light intensity that represent the physiological blood volume changes can also be exploited for deep feature learning. Specifically, each video frame represents spatial light absorption patterns, which are processed by a pre-trained ResNet-50 model to extract deep features per video sample. Deep features are combined with handcrafted PPG features to form a hybrid feature dataset, which is then used to train the model for robust BGL estimation.

ResNet-based feature extraction has been employed in the NI estimation of hemoglobin levels from fingertip videos [12], demonstrating its capability to capture meaningful representations of underlying blood-related biomarkers.

In the proposed study, three experimental models were developed to evaluate the predictive capability of different feature representations. The study evaluates three feature sets: handcrafted features derived from PPG waveforms, deep features extracted using ResNet-50, and a combined multimodal feature set that integrates both PPG and ResNet-50 features. Each feature set was trained and tested using different machine learning models.

Key contributions of this study include:

(1). Development of a novel fingertip video dataset comprising recordings from 243 subjects.

(2). Design of a hybrid framework that integrates handcrafted PPG features with ResNet-50-based deep features for robust NI-BGL estimation.

(3). Comprehensive evaluation of individual and combined feature sets using multiple machine learning models, demonstrating the superiority of the multimodal approach.

2. RELATED WORK

Prior work has focused on deriving BGL through non-invasive means, driven by the growing demand for painless and user-friendly monitoring solutions. These techniques were developed as alternatives to traditional finger-prick methods, which are invasive and often uncomfortable for users. Broadly, NI glucose estimation techniques can be categorized into two primary groups: sensor-driven methods and spectroscopy-based approaches.

Prasad et al. [13] introduced an IoT-enabled system for random blood glucose estimation using Photoacoustic Spectroscopy (PAS) signals analyzed with a Shallow Dense Neural Network (SDNN). With 105 subjects, their system achieved Root Mean Square Error (RMSE) = 2.86 mg/dL and Mean Absolute Relative Difference (MARD) = 8.49%. Kumar et al. [14] employed NIR spectroscopy with a 940 nm LED sensor, where ensemble regressors achieved a coefficient of determination (R^2) = 0.921 and RMSE = 27.36 mg/dL from 611 measurements. Song et al. [15] proposed Multi-Scale Fusion (RBANet), a deep learning model integrating wearable physiological data (blood volume pulse, exploratory data analysis, heart rate, accelerometry) with nutrition information, reporting Mean Square Error (MSE) = 0.22 mmol/L and 96.75% accuracy for hyperglycemia detection. Chellamani et al. [16] used PPG signals with a Deep Sparse Capsule Network (DSCNet), obtaining $R^2 = 0.98$ and Mean Absolute Percentage Error (MAPE) = 3.02%. Further, Piao et al. [17] combined Graph Attention Networks with Gated Recurrent Units (GRUs) to model multivariate signals from Empatica devices, achieving RMSE ≈ 19.86 mg/dL. Mazgouti et al. [18] presented a hybrid Long Short-Term Memory (LSTM)–XGBoost model for Type 1 diabetes prediction, where CGM data from 12 patients yielded RMSEs of 7.97–10.93 mg/dL with R^2 up to 0.98. Fathimal et al. [19] designed a dual-wavelength NIR optical setup (940 nm and 1050 nm) with polynomial regression, reporting MAPE = 5.99%. Jian et al. [20] employed dual-wavelength PPG signals (660 nm, 880 nm), where Random Forest achieved MARD = 5.15% and R = 0.93.

Despite significant progress in sensor-based NI BGL estimation, many existing methods depend on bulky, costly, or specialized hardware and focus predominantly on signal-level processing. An alternative line of research has explored video-based methods, leveraging fingertip or facial recordings to derive PPG signals. Golap et al. [11] analyzed fingertip recordings with 850 nm LEDs and smartphone cameras, extracting 48 features; Multigene Genetic Programming achieved $R^2 = 0.881$. Nie et al. [8] proposed non-contact imaging photoplethysmography (IPPG) from facial NIR recordings, where Random Forest Regression reported MAE = 1.72 mmol/L. Sridevi et al. [21] explored Quantum Machine Learning with NIR-illuminated fingertip videos, achieving 89.30% accuracy. Chinchanikar and Dale [22] processed fingertip videos in Red, Green and Blue (RGB) and Hue, Saturation and Value (HSV) color spaces, where XGBoost yielded $R^2 = 0.89$ (RGB) and 0.84 (HSV). Table 1 summarizes representative NI glucose estimation studies, highlighting methods, devices, dataset sizes, and reported performance.

Fingertip videos capture changes in light intensity caused by volumetric variations in blood during systole and diastole. In most existing video-based approaches for NI BGL, these recordings were processed by averaging pixel intensity values over selected regions of interest to extract the PPG signal. While this reduction facilitates interpretation and physiological analysis, it inherently compresses complex spatial–temporal data (width × height × time) into a single time-series waveform, potentially discarding rich spatial and spatiotemporal information that may hold subtle yet valuable cues for accurate glucose estimation. Although the potential of deep learning to capture such complex patterns has been demonstrated in related applications, such as hemoglobin level estimation [12], its role in blood glucose prediction remains underexplored. This highlights the need for hybrid frameworks that integrate physiological signal features with deep visual representations to achieve robust, NI BGL prediction.

At the core of this study, PPG waveforms were extracted from fingertip videos, from which 46 handcrafted features

were computed. In parallel, ResNet-50 was applied to each video frame, and after a mean pooling operation, 2,048 deep features per video were obtained. Three experimental models were developed to evaluate the predictive capability of different feature representations: (i) handcrafted PPG features, (ii) ResNet-50–based deep features, and (iii) a combined multimodal feature set integrating both handcrafted and deep representations. Models trained on individual feature sets provided useful insights, but the combined approach consistently outperformed the others, highlighting the complementary nature of handcrafted physiological descriptors and deep visual embeddings.

Table 1. Summary of NI BGL estimation studies

Study	Method	Device	Subjects	Performance
Prasad et al. [13]	SDNN	905 nm pulsed laser with photodiode detector	105	RMSE = 2.86 mg/dL, MAD = 8.77 mg/dL, MARD = 8.49%
Kumar et al. [14]	Ensemble learning (Voting Regressor: RF + ET + XGB)	940 nm IR LED sensor	611	R^2 = 0.921, RMSE = 27.36 mg/dL
Song et al. [15]	Multi-scale fusion (RBANet)	Multi-signal wearable	PhysioNet dataset	MSE = 0.22 mmol/L, 96.75% accuracy
Chellamani et al. [16]	DSCNet	IR + Red light PPG sensor	835	R^2 = 0.98, MAPE = 3.02%, RMSE = 0.062
Piao et al. [17]	GARNNs (GAT + GRU, GATv2 + GRU)	Graph-based PPG fusion architecture	136	RMSE ≈ 19.86 mg/dL, MAE ≈ 13.81 mg/dL
Mazgouti et al. [18]	LSTM + XGBoost Fusion	Continuous Glucose Monitoring (CGM) data	12	RMSE = 7.97–10.93 mg/dL, R^2 up to 0.98
Fathimal et al. [19]	Polynomial Regression (Linear, Quadratic, Cubic)	Dual NIR LED system (940 and 1050 nm)	45	MAPE = 5.99%
Jian et al. [20]	Random Forest and XGBoost	MAX86150 sensor (Red + IR PPG)	1 over 13 days	MARD = 5.15%, R = 0.93
Golap et al. [11]	Video-based, MGGP (Multigene Genetic Programming)	850 nm 6 NIR LEDs + 1 flash LED board	111	R^2 = 0.881, MAE ±0.324
Nie et al. [8]	Video-based, PCR, PLS, SVR, RFR (best)	Facial video via 940 nm NIR camera	8 over 15 days	MAE = 1.72 mmol/L
Sridevi et al. [21]	Video-based, QSVM	10-second fingertip video using smartphone camera and 850 nm and 940 nm LEDs	136	QSVM accuracy = 89.30%, CV = 92.50%
Chinchanikar and Dale [22]	Video-based, XGBoost, CatBoost, RFR, GBR	6 NIR LEDs,1 flash LED, Smartphone camera	234	R^2 = 0.89 (RGB), 0.84 (HSV)

Note: RF: Random Forest, ET: Extra Trees, GRANN: Graph Attentive Recurrent Neural Network, GAT: Graph Attention Network, GRU: Gated Recurrent Unit, MGGP: Multi-Gene Genetic Programming, PCR: Principal Component Regression, PLS: Partial Least Squares, SVR: Support Vector Regression, RFR: Random Forest Regression, QSVM: Quantum Support Vector Machine, GBR: Gradient Boost Regression

3. SYSTEM FRAMEWORK

The proposed system layout, as shown in Figure 1, illustrates the complete operational workflow of the setup. Data acquisition was performed using a smartphone camera to record fingertip videos. From these videos, PPG signals were extracted, and handcrafted features capturing key temporal and frequency characteristics were computed. In parallel, ResNet-50 was applied to each video frame to extract deep features. Three experimental models were then developed and trained separately using (i) handcrafted features, (ii) ResNet-50 features, and (iii) a combined multimodal feature set. To enhance robustness and minimize redundancy, an importance-based feature selection process was applied to retain the most informative attributes. The optimized feature set was then used to train predictive models for accurate BGL estimation.

3.1 Experimental arrangement

The experimental platform configuration for the BGL estimation system consisted of an NIR illumination unit, as shown in Figure 2, and a smartphone camera. The NIR illumination unit [23] incorporated six peripheral NIR LEDs along with a central flash LED, which was used to boost the overall illumination intensity. In the present study, a 940 nm NIR illumination board was selected, as glucose exhibits notable absorption characteristics in the NIR spectrum, particularly within the 940–1000 nm range. The video data were acquired using a Samsung A51 mobile, powered by Android 10 and equipped with a 48-megapixel camera. The camera recorded footage at a frame rate of 30 FPS, with a screen resolution of 1080 × 2400 pixels, ensuring high-quality image capture for accurate signal extraction.

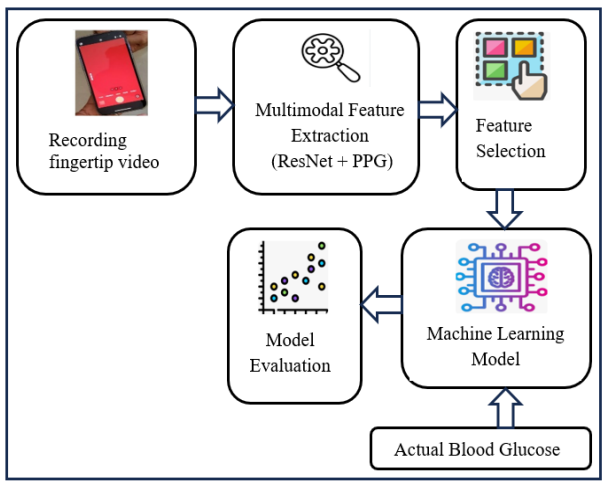


Figure 1. Model architecture overview

Figure 2. Hardware setup

During data collection, subjects placed their right-hand fingertip on the NIR illumination board for video acquisition. Initial trials revealed instability due to board displacement, finger motion, and minor camera shifts from breathing. To ensure consistency, a custom stabilization enclosure was designed to securely fix both the NIR illumination unit and the smartphone [22].

Despite this, occasional quality issues (e.g., sneezing, coughing, or finger shifts) were observed. A video-quality screening step was therefore implemented, where only recordings yielding clear PPG signals were retained; otherwise, re-acquisition was performed. Reference BGL values were obtained using an Accu-Chek® Instant device for validation of NI estimates.

3.2 Data acquisition

In the present study, fingertip videos in *.mp4 format were recorded from 243 subjects (121 females and 122 males), aged 18–88 years, with varying weights and health conditions, including both diabetic and non-diabetic individuals. Each subject placed their right index fingertip on an NIR illumination unit, and a 15-second video was recorded using a smartphone camera. Subjects were recruited from various local institutions in Pune, representing diverse socio-economic and lifestyle backgrounds. Before recording, all subjects followed hygiene protocols and provided informed consent. Figure 3 depicts the structured workflow of the data collection method.

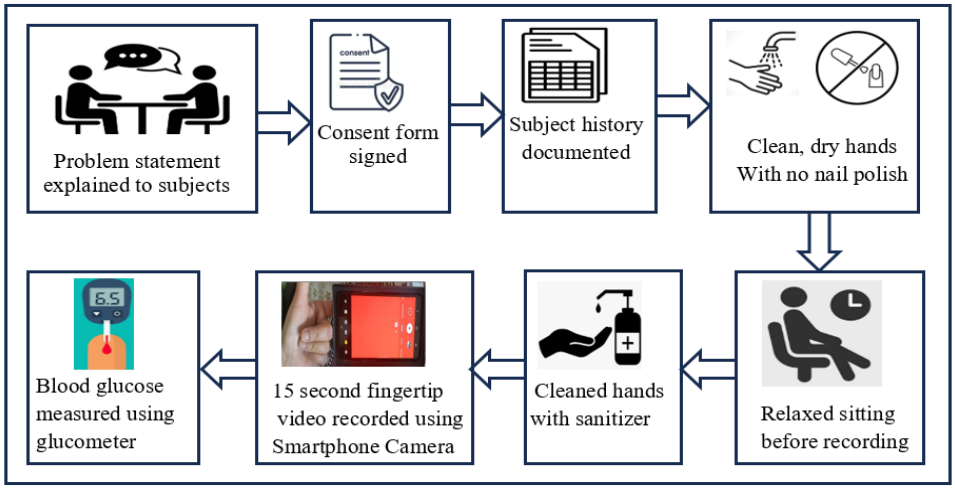


Figure 3. Step-by-step process for subject data collection

4. SIGNAL PROCESSING AND MULTI-SCALE FEATURE EXTRACTION

To establish a robust framework for NI BGL estimation, two primary feature sets were generated from fingertip videos: handcrafted features derived from PPG signals, and deep features extracted using the pre-trained ResNet-50 model. Additionally, a hybrid feature set was created by combining these two feature sets, leveraging the complementary strengths of physiological descriptors and deep visual representations for improved predictive performance.

4.1 PPG signal acquisition and characteristic feature extraction

Figure 4 depicts the overall framework used for clean PPG signal extraction from fingertip videos. For the extraction of the PPG signal, the red channel was selected as it exhibited the highest pixel intensity among the RGB channels.

To ensure signal quality and remove potential distortions, the initial 3 seconds and concluding 2 seconds of each video were excluded. This resulted in 300 usable frames per subject. To identify the ROI, K-means clustering was employed on the video frames, segmenting pixel values into separate groups.

Based on the clustering outcome, a 500 × 500 pixel area spanning rows 750 to 1250 and columns 0 to 500 was selected as the ROI for computing the mean intensity of the red channel. The mean red channel values across the defined region were used to derive the unprocessed PPG signal. This process is described by Eq. (1).

$$PPG(t) = \frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N} I_{red}(i,j,t) \quad (1)$$

where, M and N indicate the number of rows and columns in the region of focus, and $I_{red}(i, j, t)$ corresponds to the intensity of the pixel located at (i, j) at time t. To ensure precise PPG

signal analysis, preprocessing was applied to 10-second videos (300 frames). A Butterworth band-pass filter (0.5–4 Hz) was applied to remove motion artifacts and baseline drift while preserving the physiological frequency range of heart rates (30–240 bpm). The Butterworth design was chosen over other filters (e.g., Chebyshev, elliptic) because it provides a maximally flat frequency response in the passband, ensuring minimal signal distortion—an important factor for maintaining the integrity of the PPG waveform morphology. Once the raw PPG signal was filtered, peaks were identified using a peak detection method. A single PPG cycle with the most distinct systolic peak was identified.

Figure 5 shows the block diagram of the PPG signal processing and feature derivation applied to the filtered PPG signal. The process began with the application of a peak detection algorithm to identify individual PPG pulses within the signal. Among these, the PPG cycle corresponding to the maximum peak amplitude was selected as the representative waveform, as it was assumed to be the least affected by noise and the most physiologically relevant.

The first and second derivatives of the selected PPG waveform were then computed to capture the rate of change and acceleration in signal morphology. Additionally, the Fast Fourier Transform (FFT) was applied to extract the frequency-domain characteristics of the waveform. From the original waveform, its derivatives, and its frequency representation, a total of 46 features were extracted as shown in Table 2. Let h denote the derived handcrafted feature vector, as given in Eq. (2).

$$h = [h_1, h_2, .. h_{46}] \in R^{46} \quad (2)$$

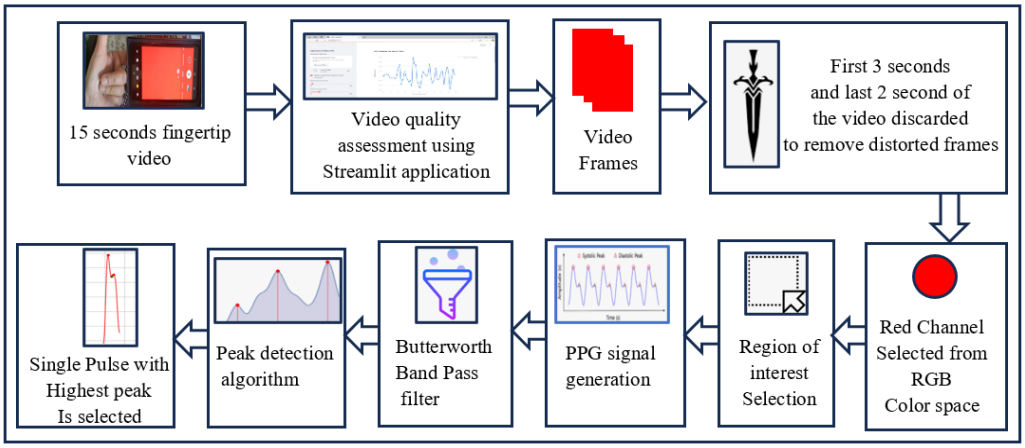
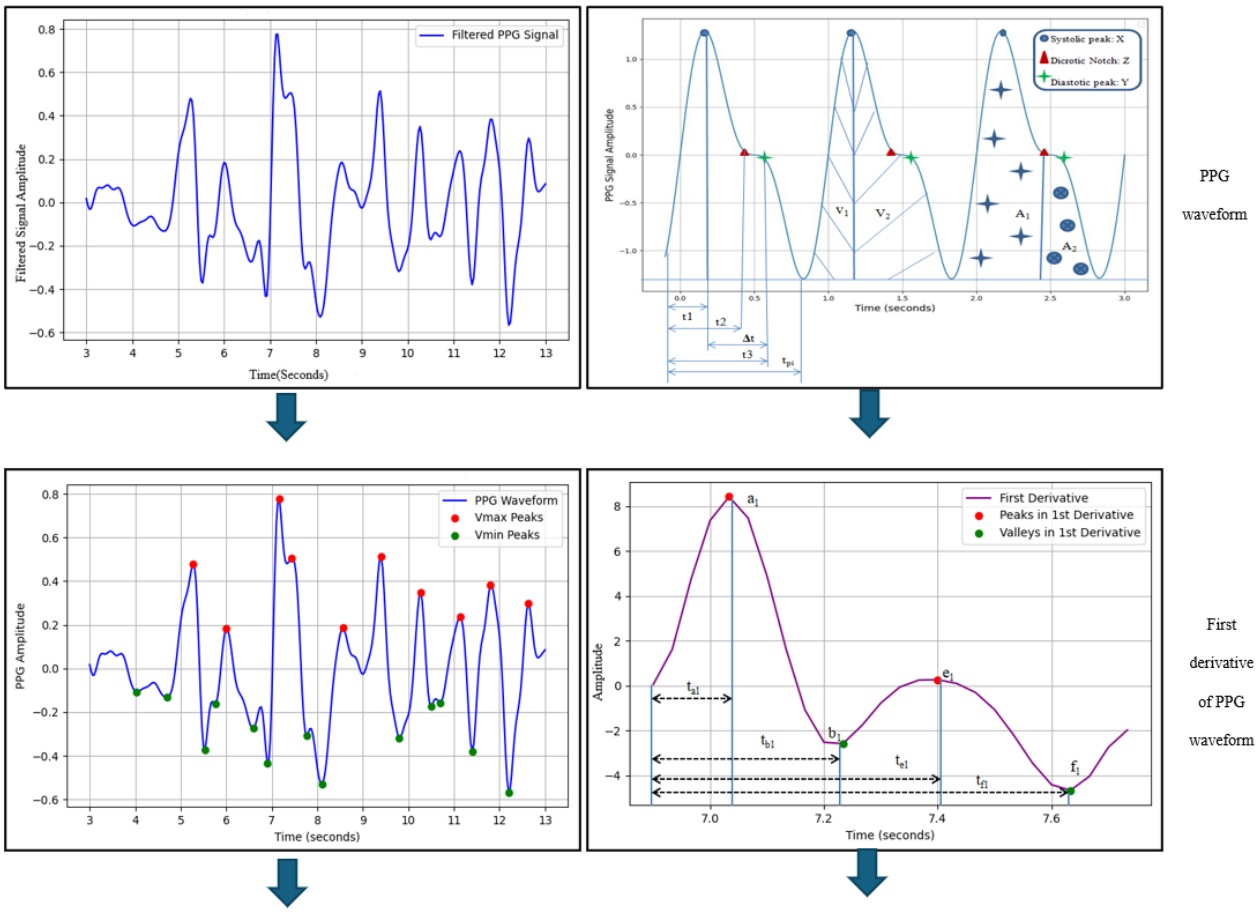


Figure 4. Overview of the PPG signal processing pipeline



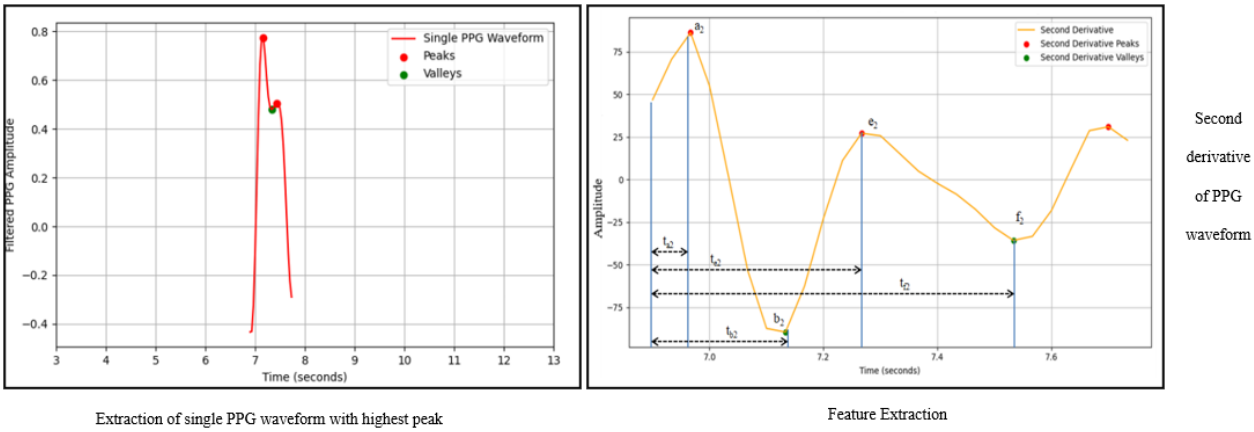


Figure 5. Schematic representation of the PPG signal processing and feature extraction steps [22]

Table 2. Handcrafted features

Feature	Description	Feature	Description	Feature	Description
f1	x	f17	t_1/t_{pi}	f33	t_{f1}/t_{pi}
f2	y	f28	t_2/t_{pi}	f34	t_{a2}/t_{pi}
f3	z	f19	t_3/t_{pi}	f35	t_{b2}/t_{pi}
f4	T_{pi}	f20	$\Delta t/t_{pi}$	f36	$(t_{a1}+t_{a2})/t_{pi}$
f5	y/x	f21	t_{a1}	f37	$(t_{b1}+t_{b2})/t_{pi}$
f6	(x-y)/x	f22	t_{b1}	f38	$(t_{e1}+t_2)/t_{pi}$
f7	z/x	f23	t_{e1}	f39	$(T_{f1}+t_3)/t_{pi}$
f8	(y-x)/x	f24	t_{f1}	f40	$X(f_0)$
f9	t_1	f25	b_2/a_2	f41	$\|X(f_0)\|$
f10	t_2	f26	e_2/a_2	f42	$X(f_1)$
f11	t_3	f27	$(b_2+e_2)/a_2$	f43	$\|X(f_1)\|$
f12	Δt	f28	t_{a2}	f44	$X(f_2)$
f13	$t_1/2$	f29	t_{b2}	f45	$\|X(f_2)\|$
f14	A_2/A_1	f30	t_{a1}/t_{pi}	f46	v_2/v_1
f15	t_1/x	f31	t_{b1}/t_{pi}		
f16	$y/(t_{pi}-t_3)$	f32	t_{e1}/t_{pi}		

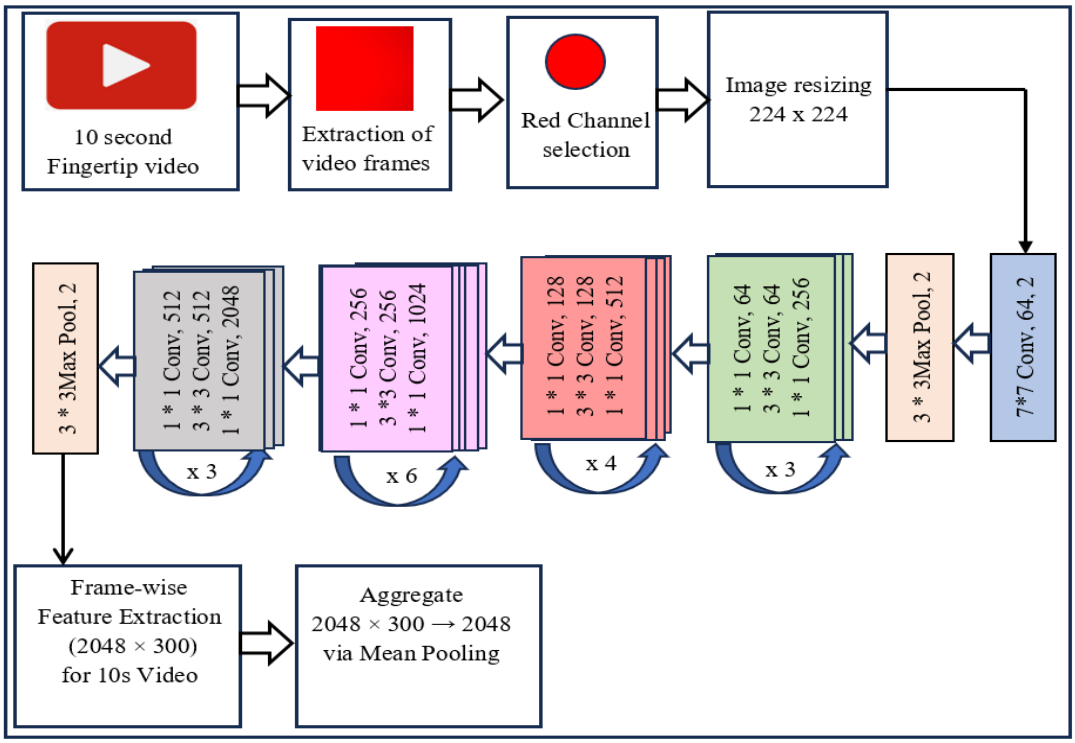


Figure 6. Feature extraction pipeline from fingertip video via ResNet-50

4.2 Deep feature extraction from video frames using ResNet

To preserve the spatial and textural information that is often lost during the transformation of fingertip videos into a 1D PPG signal, pretrained ResNet-50 was employed as a deep feature extractor. Unlike handcrafted features that rely solely on signal morphology, ResNet-50 focuses on learning spatiotemporal variations in light absorption, which may be indicative of blood volume changes linked to glucose levels. Its ability to learn complex representations from raw visual data makes it a powerful tool for enhancing model performance in NI BGL estimation.

ResNet-50 is a deep CNN consisting of 50 layers, renowned for its use of residual connections that facilitate the training of very deep architectures without performance degradation. These shortcut connections address the vanishing gradient issue and enable efficient learning of both low- and high-level features. The architecture consists of multiple convolutional blocks, identity mappings, and batch normalization layers. ResNet-50 is widely used for feature extraction in image-based tasks due to its robustness and strong generalization capability. Figure 6 shows the block diagram illustrating the process of feature extraction from fingertip videos.

In the proposed work, fingertip video data for each subject was processed by extracting frames between the 3rd and 13th seconds, resulting in a consistent temporal segment of approximately 300 frames per video. The frames were scaled to 224 × 224 pixels and standardized using the predefined ImageNet mean and standard deviation values [0.485, 0.456, 0.406] and [0.229, 0.224, 0.225], respectively, to ensure alignment with the ResNet-50 framework.

To derive features, the final fully connected classification layer of the ResNet-50 framework was removed. Let f_{ResNet} be a function that represents the ResNet-50 model's feature extraction process and is given by Eq. (3).

$$f_{ResNet} = R^{224\times224\times1} = R^{2048} \tag{3}$$

For each frame I_t, deep feature vector is given by Eq. (4).

$$x_t = f_{ResNet}(I_t) = R^{2048} \tag{4}$$

For t = 1, 2, 3, ...T, where, T = 300, x_t represents features from one frame and I_t is the tth red frame of the video. This resulted in a feature matrix for the entire video and is given by Eq. (5).

$$X = [x_1, x_2, \ldots, x_T] \in R^{2048\times T} \tag{5}$$

This configuration enabled access to the 2048-dimensional feature vector from the penultimate layer. The model was executed in inference mode using PyTorch with GPU acceleration where available. Each processed frame was passed through the network, and the resulting features were stored in separate files for each video to facilitate modular analysis. Processing 243 fingertip videos through ResNet-50 for deep feature extraction required approximately 2 hours and 32 minutes.

To convert the frame-level feature matrix $X \in R^{2048\times T}$ into a single, fixed-length representation for each video, temporal mean pooling was applied across all frame-level features. This operation averaged the ResNet embeddings over 300 frames, resulting in a 2048-dimensional vector that compactly captured the aggregated spatiotemporal information of the entire video, as expressed in Eq. (6). By averaging out transient variations, this strategy reduces frame-level noise and motion artefacts while emphasizing stable spatial patterns that are more likely to reflect underlying physiological changes. Mean pooling was employed to obtain a compact and robust feature representation suitable for tree-based regression models. It reduces the influence of outlier frames that may adversely affect performance in max pooling. Furthermore, mean pooling lowers dimensionality and computational cost, improving the sample-to-parameter ratio and enhancing the robustness of subsequent regression modeling.

$$\bar{x} = \frac{1}{T}\sum_{t=1}^{T} x_t \in R^{2048} \tag{6}$$

The resulting pooled vector $\bar{x}$ captures the average spatiotemporal characteristics of the video, serving as its compact deep representation. The aggregated dataset was compiled into a unified CSV file, with each row corresponding to one video sample.

Thus, two distinct feature sets were derived: the first comprising 46 handcrafted PPG features and the second consisting of 2048 deep features extracted using ResNet-50. A third hybrid feature set was then constructed by concatenating the handcrafted and deep features, resulting in a total of 2094 features. These three feature sets were subsequently employed to train and evaluate three experimental models: one based solely on handcrafted PPG features, one utilizing ResNet-50 features, and one leveraging the combined hybrid feature set.

5. FEATURE OPTIMIZATION AND MODEL CONSTRUCTION

Feature importance-based optimization was utilized to identify and retain the most relevant features for NI BGL prediction. This method evaluated the significance of every feature to the model's predictive performance, enabling the selection of an informative subset while eliminating redundant or irrelevant variables.

The importance of each feature was quantified based on the average reduction in the loss function each time it was used to split a decision node within the ensemble, as shown in Eq. (7). The significance of feature f was evaluated based on its influence on the loss function.

$$I_f = \sum_{t\epsilon Tf} \triangle L_t \tag{7}$$

where, Tf is the set of all nodes where feature f is used for splitting. ΔL_t denotes the reduction in the loss function at node t.

Features were ranked according to their importance scores, and only the most informative ones were retained to improve generalization, reduce the risk of overfitting, and enhance computational efficiency. Feature importance analysis was performed independently for each ensemble model (CatBoost, Random Forest, XGBoost, and Gradient Boosting) and for each feature set.

After optimization, the final number of selected features used for model training was as follows: all 46 features were retained for the handcrafted PPG feature set; for the ResNet-50 deep features, 1153–2048 features were retained depending

on the model; and for the hybrid feature set, 793–2094 features were retained.

It should be noted that tree-based feature importance can be biased toward features with higher cardinality or larger numeric ranges, potentially overestimating their influence. Moreover, these importance scores do not explicitly capture complex feature interactions, which should be considered when interpreting the predictive contribution of individual features. Despite these limitations, this approach provides a practical and computationally efficient method for identifying informative features in moderate-sized datasets.

6. RESULTS AND DISCUSSION

A total of 243 subjects, including both diabetic and non-diabetic individuals, participated in this study, with ages ranging from 18 to 88 years. For each subject, a 15-second fingertip video was recorded using a mobile camera paired with a NIR illumination unit to improve signal fidelity. From each video, a PPG waveform was extracted, from which 46 handcrafted features were computed. Additionally, 2048 deep features were extracted from video frames using ResNet-50 CNN. A third hybrid feature set was then constructed by combining the handcrafted and deep features, resulting in a 2094-dimensional dataset per subject.

Feature importance-based selection was employed to prioritize the most informative variables within the feature space. This approach ensured that the most relevant information was preserved while reducing redundancy among correlated features. By retaining variables that contributed most strongly to predictive performance, the method provided a refined and balanced feature set for subsequent model training. R^2 quantifies the proportion of the target variable's variance explained by the predictions, thereby indicating its degree of fit. In contrast, MAE measures the mean of the absolute deviation among actual and predicted glucose values, serving as an intuitive metric of prediction error that remains stable in the presence of outliers. The corresponding mathematical definitions are provided in Eqs. (8) and (9).

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_{gact} - y_{gpred}\right| \tag{8}$$

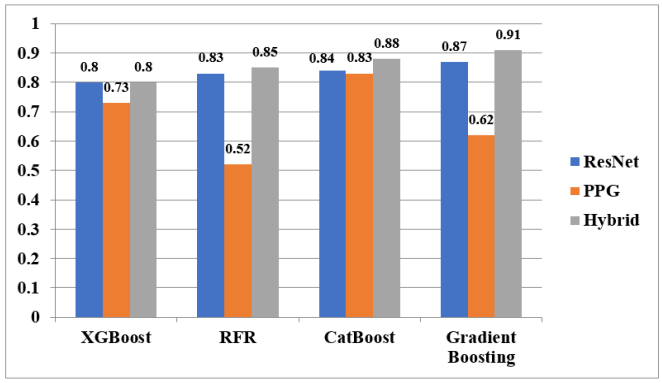
$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_{gact} - y_{gpred})^2}{\sum_{i=1}^{n}(y_{gact} - y_{mean\_of\_gact})^2} \tag{9}$$

where, y_{gact}: Measured BGL, y_{gpred}: Algorithm-generated BGL, $y_{mean\_of\_gact}$: Average measured BGL.

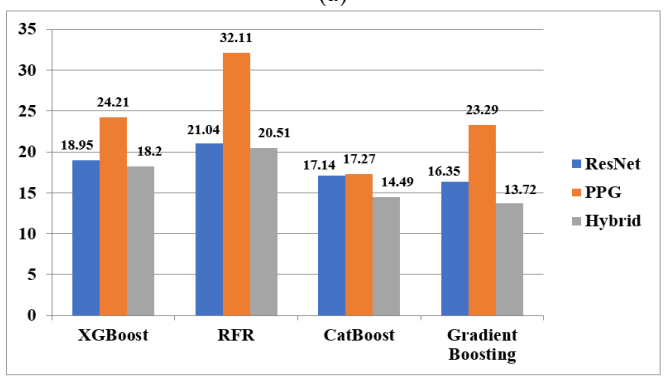
An initial 80:20 train-test split was carried out to evaluate model generalization on unseen data. A stratified 5-fold CV was subsequently applied to the training set by dividing it into five equal subsets. In each iteration, four folds were employed for training, while the remaining fold was used for validation. This process helped assess the model's robustness, reduce variance from individual splits, and ensure more reliable performance estimation.

To predict BGL, several ensemble regression models were evaluated, including RFR, GBR, XGBoost, and CatBoost. Each model was trained on an optimized subset of features, obtained by applying an importance-based feature selection method to the PPG dataset, the deep feature dataset, and the hybrid dataset. The number of selected features varied across datasets and regression models, reflecting the differing contributions of features to predictive performance, reducing redundancy, and the varying sensitivity of models to feature relevance.

Figures 7(a) and (b) present the prediction performance of the evaluated regression models on the test dataset in terms of R^2 and MAE, respectively. CatBoost and Gradient Boosting achieved the highest predictive accuracy when trained on hybrid features (R^2 = 0.88 and 0.91, MAE = 14.49 and 13.72, respectively), while XGBoost and Random Forest showed moderate performance (R^2 = 0.80 – 0.85). Models trained solely on handcrafted PPG features had lower predictive power (R^2 = 0.52 – 0.83), with ResNet-50 features yielding intermediate performance. These results highlight the complementary nature of hybrid features: deep embeddings provide rich abstract representations, whereas handcrafted descriptors preserve physiologically interpretable signal properties. 5-fold cross-validation further confirmed these observations. Hybrid features maintained superior generalization performance across all models. Figures 8(a) and (b) illustrate the R^2 and MAE values obtained after 5-fold cross-validation, highlighting the comparative performance of the three feature sets. CatBoost achieved the highest R2 of 0.89 and lowest MAE of 15.13 on hybrid features, followed by Gradient Boosting (R^2 = 0.79, MAE = 19.14). XGBoost and Random Forest achieved R^2 = 0.70 – 0.76 and MAE = 18.96 – 25.02. The superior performance of CatBoost can be attributed to its strong ability to model intricate nonlinear interactions among heterogeneous features, its inherent robustness to outliers, and its efficient handling of high-dimensional hybrid feature spaces, thereby making it particularly suitable for reliable blood glucose level prediction from fingertip video-derived data.

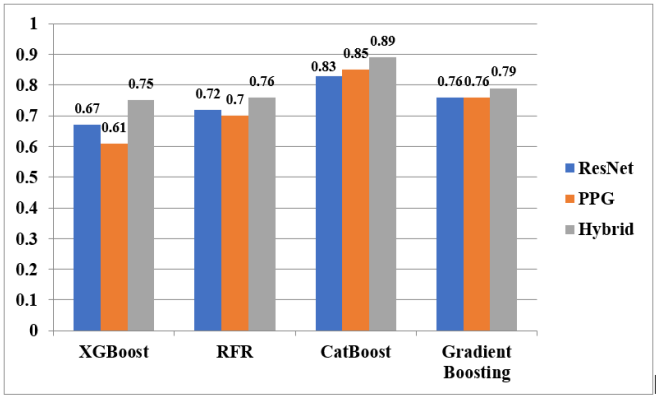


(a)

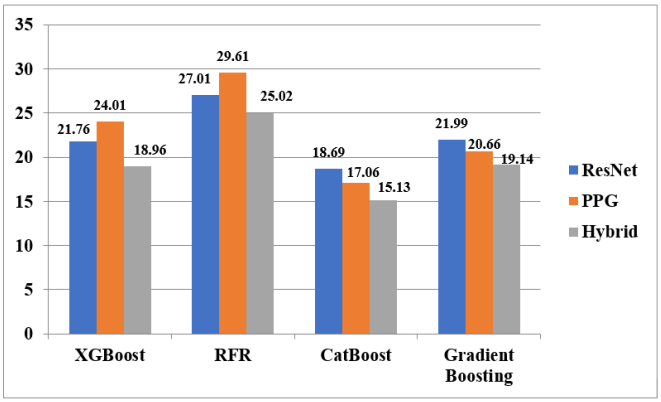


(b)

Figure 7. Comparison of (a) R^2 and (b) MAE values for different regression models across PPG features, ResNet-50 deep features, and the hybrid feature set



(a)



(b)

Figure 8. Comparative (a) R^2 and (b) MAE values of regression models trained on PPG, ResNet-50, and hybrid feature sets under 5-fold cross-validation

To enhance statistical rigor, all experiments were repeated ten times using different random seeds. The plots in Figures 7 and 8 illustrate the distribution of model performance across runs, showing mean values along with 95% confidence intervals. This approach mitigates the effect of random initialization and sampling variations on the reported results. Among all ensemble regressors, CatBoost consistently achieved the lowest MAE and the highest R2 across repeated trials, confirming its robustness and stability. A paired t-test was conducted to examine whether the observed improvement of CatBoost over other models was statistically significant. The analysis revealed that CatBoost's MAE was significantly lower than those of Gradient Boosting ($p < 0.05$), XGBoost ($p < 0.01$), Bagging Regressor ($p < 0.01$), and Random Forest ($p < 0.01$), confirming that its superior performance is statistically meaningful.

Error analysis revealed that the largest prediction errors were primarily associated with extreme glucose values and suboptimal acquisition conditions, such as low ambient lighting or slight finger motion. Variability in demographics, including age, BMI, and finger thickness, also contributed to error differences, with younger subjects with thinner fingers exhibiting slightly lower errors due to higher PPG signal quality. While the models demonstrated robust performance, several limitations may affect generalizability. Skin tone can influence PPG signal contrast, BMI and vascular health may affect peripheral blood flow, and environmental factors such as lighting and motion artifacts can introduce additional noise. Furthermore, tree-based feature importance may not fully capture complex feature interactions, and the study population was limited.

In addition to predictive performance, the computational efficiency of the evaluated models was assessed to consider real-time deployment feasibility. Inference time per sample was measured on a standard workstation with an Intel i7 CPU and 16 GB RAM. CatBoost and Gradient Boosting required approximately 5–7 ms per sample, whereas XGBoost and Random Forest required 6–9 ms per sample.

Overall, integrating handcrafted PPG and deep ResNet-50 features consistently reduced prediction error and enhanced generalization across all models. Figure 9 summarizes the final hyperparameter settings for each regression model.

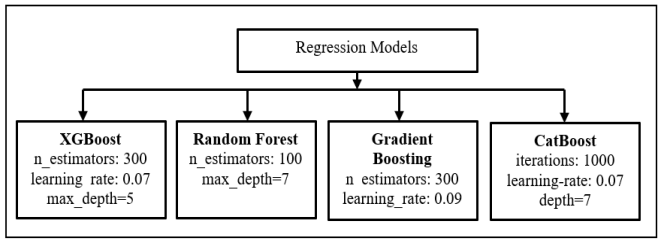
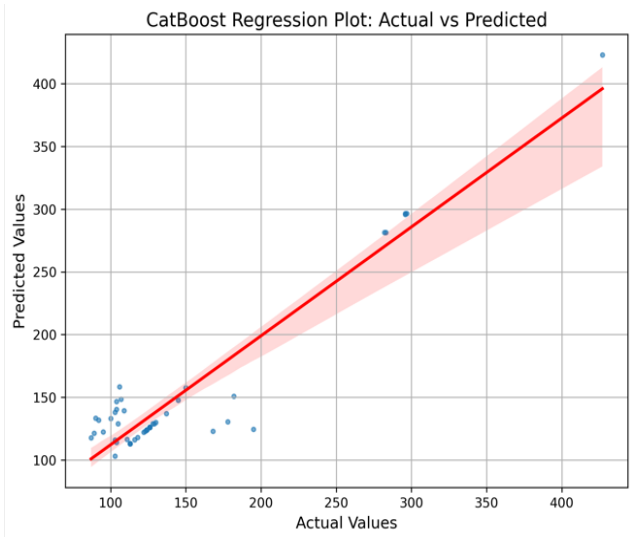
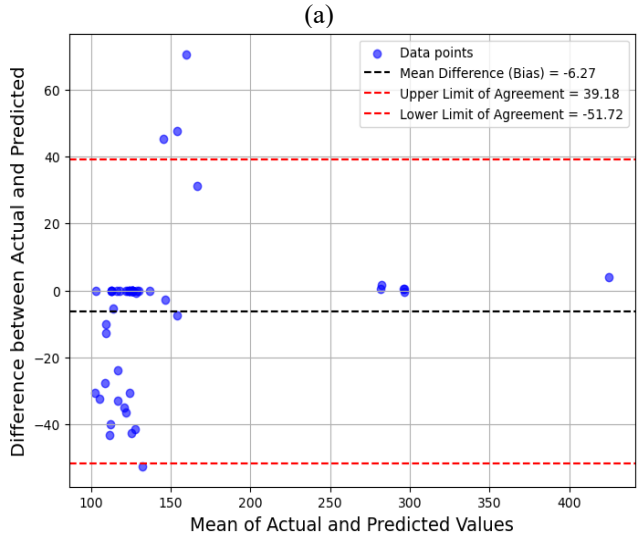


Figure 9. Hyperparameter configurations for evaluated regression models



(a)



(b)

Figure 10. (a) CatBoost regression and (b) Bland–Altman agreement analysis for the hybrid dataset

Table 3. Evaluation of the proposed method against current video-based BGL prediction techniques

Reference	Device	Feature Extraction	Subject Count	Algorithm	Results
Nie et al. [8]	industrial NIR camera	PPG	8	RFR	$R^2 = 0.60$
Golap et al. [11]	850 nm NIR LED Nexus- 6p smartphone	PPG	111	MGGP	$R^2 = 0.88$
Sridevi et al. [21]	Pixel-2 smartphone with 850 nm and 940 nm LEDs	PPG	136	Quantum Support Vector Machine	Accuracy = 89.30%
Haque et al. [24]	850 nm NIR LED, Nexus- 6p smartphone	PPG	93	Deep Neural Network (DNN)	$R^2 = 0.90$
Islam et al. [25]	OnePlus 6T	PPG	52	PLS	prediction errors = 17.02 mg/dL
Proposed work	940 nm NIR LED module, Samsung A51 smartphone	Handcrafted PPG + Deep Features (ResNet-50)	243	CatBoost	$R^2 = 0.88$ and avg. CV $R^2 = 0.89$

In terms of computational efficiency, all tree-based ensemble models demonstrated sufficiently low inference times for near real-time blood glucose estimation from fingertip videos. CatBoost combines high predictive accuracy with efficient inference, making it a strong candidate for practical deployment in mobile or embedded applications. Although training with hybrid features required a higher computational cost due to increased dimensionality, this is a one-time process and does not affect real-time prediction. Overall, the results highlight that integrating handcrafted PPG features with deep ResNet-50 representations enhances both predictive accuracy and error minimization, establishing hybrid features as a robust strategy for non-invasive BGL estimation from fingertip video data. Figure 9 presents the final hyperparameter settings used for each ensemble regression model.

Figure 10(a) presents the regression and (b) Bland–Altman plots for the CatBoost model trained on the hybrid feature set, evaluated on the 20% test dataset comprising 49 subjects. The regression plot shows a strong correlation between predicted and reference BGL, indicating high predictive accuracy. In the Bland–Altman analysis, 45 out of 49 points (91.84%) were located within the 95% limits of agreement (Region 1, considered acceptable), while 4 points (8.16%) fell outside these limits (Region 2), suggesting minor deviations. These findings further validate the clinical reliability of the model, with a large majority of predictions exhibiting acceptable agreement with reference values.

Table 3 presents a performance comparison between the proposed approach and multiple previously reported contact and non-contact methods for estimating BGL using video data.

All the methods listed above estimate BGL by extracting PPG signals from video data. Although Sridevi et al. [21] and Haque et al. [24] reported higher R^2 values (0.89 – 0.90), their studies involved relatively small subject counts (136 and 93, respectively). The proposed work achieved comparable predictive performance ($R^2 = 0.88$; average CV $R^2 = 0.89$) on a larger cohort of 243 subjects, demonstrating improved generalizability. Unlike these approaches, the proposed method incorporates deep features extracted directly from fingertip video frames using a pre-trained ResNet-50 CNN. By fusing handcrafted PPG features with high-level CNN-derived representations, the model leverages both physiological domain knowledge and abstract visual patterns. Specifically, ResNet-50 captures spatiotemporal variations in light absorption across video frames, reflecting subtle blood volume changes associated with glucose levels—information often lost during conventional signal extraction.

7. CONCLUSION AND FUTURE SCOPE

The proposed study presented an NI approach for BGL estimation using fingertip videos captured via a smartphone camera. A hybrid feature set was developed by combining 46 handcrafted PPG features with 2048 deep features derived from a pre-trained ResNet-50 model, resulting in a 2094-dimensional representation. Feature importance-based selection was applied to prioritize the most informative variables, preserving relevant information while minimizing redundancy. Among the evaluated machine learning models, CatBoost outperformed others, achieving a test R^2 score of 0.8801 and MAE of 14.50. Five-fold cross-validation further validated its robustness, yielding an average R^2 of 0.8943 and MAE of 15.13. Bland–Altman analysis showed that 91.84% of the predictions fell within the 95% limits of agreement, indicating strong alignment with reference glucose values.

While the proposed method demonstrated promising performance, the dataset used was relatively small for developing a highly generalizable model. Expanding the dataset to include a larger and more diverse population is essential to enhance model robustness and ensure broader applicability. Future studies should also consider physiological and environmental variations, including skin tone, BMI, ambient lighting, temperature, and vascular health, which may influence PPG signal quality and system performance.

To facilitate clinical translation, several steps are necessary before deployment: rigorous validation on diverse patient populations, integration with real-time smartphone applications, regulatory compliance, and assessment of usability in daily life. Moreover, incorporating adaptive modeling techniques that adjust to individual physiological differences can enhance prediction accuracy, computational efficiency, and user-friendliness. Overall, the proposed approach provides a foundation for scalable, real-time NI glucose monitoring, and further development could enable practical clinical and home-based applications for diabetes management.

REFERENCES

[1] Lambert, P., Bingley, P.J. (2002). What is Type 1 diabetes? Medicine, 30(1): 1-5. https://doi.org/10.1383/medc.30.1.1.28264

[2] Khan, M.A.B., Hashim, M.J., King, J.K., Govender, R.D., Mustafa, H., Al Kaabi, J. (2020). Epidemiology of Type 2 diabetes – global burden of disease and forecasted

trends. Journal of Epidemiology and Global Health, 10: 107-111. https://doi.org/10.2991/jegh.k.191028.001

[3] Zhang, W.J., Du, Y.Q., Wang, M.L. (2015). Noninvasive glucose monitoring using saliva nano-biosensor. Sensing and Bio-Sensing Research, 4: 23-29. https://doi.org/10.1016/j.sbsr.2015.02.002

[4] Badugu, R., Reece, E.A., Lakowicz, J.R. (2018). Glucose-sensitive silicone hydrogel contact lens toward tear glucose monitoring. Journal of Biomedical Optics, 23(5): 057005. https://doi.org/10.1117/1.JBO.23.5.057005

[5] Lee, H., Song, C., Hong, Y.S., Kim, M., et al. (2017). Wearable/disposable sweat-based glucose monitoring device with multistage transdermal drug delivery module. Science Advances, 3(3): e1601314. https://doi.org/10.1126/sciadv.1601314

[6] Kottmann, J., Rey, J.M., Sigrist, M.W. (2016). Mid-infrared photoacoustic detection of glucose in human skin: Towards non-invasive diagnostics. Sensors, 16(10): 1663. https://doi.org/10.3390/s16101663

[7] Xue, J.T., Ye, L.M., Li, C.Y., Zhang, M.X., Li, P. (2018). Rapid and nondestructive measurement of glucose in a skin tissue phantom by near-infrared spectroscopy. Optik, 170: 30-36. https://doi.org/10.1016/j.ijleo.2018.05.050

[8] Nie, Z.H., Rong, M., Li, K.Y. (2023). Blood glucose prediction based on imaging photoplethysmography in combination with machine learning. Biomedical Signal Processing and Control, 79(2): 104179. https://doi.org/10.1016/j.bspc.2022.104179

[9] Habbu, S., Dale, M., Ghongade, R. (2019). Estimation of blood glucose by non-invasive method using photoplethysmography. Sādhanā, 44: 135. https://doi.org/10.1007/s12046-019-1118-9

[10] Monte-Moreno, E. (2011). Non-invasive estimate of blood glucose and blood pressure from a photoplethysmograph by means of machine learning techniques. Artificial Intelligence in Medicine, 53(2): 127-138. https://doi.org/10.1016/j.artmed.2011.05.001

[11] Golap, M.A.U., Raju, S.M.T.U., Haque, M.R., Hashem, M.M.A. (2021). Hemoglobin and glucose level estimation from PPG characteristics features of fingertip video using MGGP-based model. Biomedical Signal Processing and Control, 67: 102478. https://doi.org/10.1016/j.bspc.2021.102478

[12] Sabir, H., Khan, K.U., Ishaq, O., Alazeb, A., Aljuaid, H., Algarni, A. (2024). Fingertip video dataset for non-invasive diagnosis of anemia using ResNet-18 classifier. IEEE Access, 12: 68880-68892. https://doi.org/10.1109/ACCESS.2024.3398353

[13] Prasad, P.N.S.B.S.V., Hussain, S.A., Singha, A.K., Jana, B., Mandal, P., Sanki, P.K. (2025). An advanced IoT-based non-invasive in vivo blood glucose estimation exploiting photoacoustic spectroscopy with SDNN architecture. Sensors and Actuators A: Physical, 387: 116391. https://doi.org/10.1016/j.sna.2025.116391

[14] Kumar, V., Divekar, A., Habbu, S., Joshi, S., Joshi, A., Dalvi, V.H. (2025). Non-invasive blood glucose estimation using a novel white-box model: An interpretable machine learning approach. Biomedical Signal Processing and Control, 105: 107520. https://doi.org/10.1016/j.bspc.2025.107520

[15] Song, Y., Yuan, Z.Y., Wu, Y.X. (2025). Multi-scale feature fusion model for real-time blood glucose monitoring and hyperglycemia prediction based on wearable devices. Medical Engineering & Physics, 138: 104312. https://doi.org/10.1016/j.medengphy.2025.104312

[16] Chellamani, N., Albelwi, S.A., Shanmuganathan, M., Amirthalingam, P., Alharbi, E.M., Alatawi, H.Q.S., Prabahar, K., Aljabri, J.B., Paul, A. (2025). A deep sparse capsule network for non-invasive blood glucose level estimation using a PPG sensor. Sensors, 25(6): 1868. https://doi.org/10.3390/s25061868

[17] Piao, C., Zhu, T., Baldeweg, S.E., Taylor, P., Georgiou, P., Sun, J., Wang, J., Li, K. (2025). GARNN: An interpretable graph attentive recurrent neural network for predicting blood glucose levels via multivariate time series. Neural Networks, 185: 107229. https://doi.org/10.1016/j.neunet.2025.107229

[18] Mazgouti, L., Laamiri, N., Ben Ali, J., El Idrissi, N.E.A., Di Costanzo, V., Naeck, R., Ginoux, J.M. (2025). Optimization of blood glucose prediction with LSTM-XGBoost fusion and integration of statistical features for enhanced accuracy. Biomedical Signal Processing and Control, 107: 107814. https://doi.org/10.1016/j.bspc.2025.107814

[19] Fathiaml, S.F., Kumar, J.S., Prabha, J.A., Selvaraj, J., F., F.J.S., S.P., A.K. (2024). Potential of near-infrared optical techniques for non-invasive blood glucose measurement: A pilot study. IRBM, 46(1): 100870. https://doi.org/10.1016/j.irbm.2024.100870

[20] Jian, S.E., Lo, Y.L., Chuang, Y.T., Kuo, S.H. (2025). Using machine learning to predict blood glucose level based on photoplethysmography. Measurement, 253: 117421. https://doi.org/10.1016/j.measurement.2025.117421

[21] Sridevi, P., Rabbani, M., Aziz, M.H., Upama, P.B., Mamun, S.M., Khan, R.A., Ahamed, S.I. (2025). Noninvasive estimation of blood glucose and HbA1c using quantum machine learning technique. Machine Learning with Applications, 19: 100626. https://doi.org/10.1016/j.mlwa.2025.100626

[22] Chinchanikar, A.K., Dale, M.P. (2025). Analyzing RGB and HSV color spaces for non-invasive blood glucose level estimation using fingertip imaging. International Journal of Advanced Computer Science and Applications (IJACSA), 16(4). https://doi.org/10.14569/IJACSA.2025.0160419

[23] Chinchanikar, A.K., Dale, M.P. (2024). Extraction of photoplethysmography signal for heart rate estimation using smartphone camera. In 2024 2nd International Conference on Emerging Trends in Engineering and Medical Sciences (ICETEMS), Nagpur, India, pp. 522-525. https://doi.org/10.1109/ICETEMS64039.2024.10965098

[24] Haque, M.R., Raju, S.M.T.U., Golap, M.A.U., Hashem, M.M.A. (2021). A novel technique for non-invasive measurement of human blood component levels from fingertip video using DNN based models. IEEE Access, 9: 19025-19042. https://doi.org/10.1109/ACCESS.2021.3054236

[25] 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. https://doi.org/10.3390/app11020618

NOMENCLATURE

K	Number of clusters
M	Number of rows
N	Number of columns
I_{red}	Intensity of red pixels
h	Handcrafted PPG feature
f_{ResNet}	ResNet feature set
X	ResNet feature matrix per frame
T	Total frames
$\bar{x}$	Agree gated feature
Tf	Set of all nodes
L_t	Total loss
y_{gact}	Actual blood glucose value
y_{gpred}	Predicted blood glucose value
$y_{mean\_of\_gact}$	Average measured blood glucose value