Development of a blood HbA1c level detection model based on Support Vector Regression (SVR) using microtest data

Nizam Ghazali[1,2*], *Yaya* Suryana[2], *Naufal Muharam* Nurdin[3,4], *Renan Prasta* Jenie[5,6], *Karlisa* Priandana[7], *Husin* Alatas[1], *Irzaman* Irzaman[1*], and *Muhammad Mahyiddin* Ramli[8]

[1]Department of Physics, Faculty of Mathematics and Natural Sciences, IPB University, Bogor, 16680, Indonesia
[2]Equipment Manufacturing Technology Research Centre, National Research and Innovation Agency (BRIN), Serpong, 15346, Indonesia
[3]Department of Community Nutrition, Faculty of Human Ecology, IPB University, Bogor, 16680, Indonesia
[4]Faculty of Medicine, IPB University, Bogor, 16680, Indonesia
[5]Master of Public Health Program, Faculty of Health Sciences and Technology, Binawan University, Jakarta, 13630, Indonesia
[6]Directorate of Research, Development, and Innovation, Indonesian Artificial Intelligence Society, Jakarta, 12930, Indonesia
[7]Computer Science Study Program, School of Data Science, Mathematics and Informatics, IPB University, Bogor, 16680, Indonesia
[8]Institute of Nano Electronic Engineering, Universiti Malaysia Perlis, Kangar, Perlis 01000, Malaysia

Abstract. Glycated hemoglobin (HbA1c) is a key indicator of long-term glycemic control and a marker of diabetes diagnosis. Rapid and cost-effective prediction from microtest data may support screening in resource-limited settings. This study developed and evaluated an HbA1c prediction model using Support Vector Regression (SVR) on small-scale primary microtest data (10 subjects, three repeated sessions) with strict procedures to prevent data leakage. Clinical and biometric numerical variables were standardized and modeled using an SVR with a Radial Basis Function (RBF) kernel. In 5-fold cross-validation, Spearman correlation was applied exclusively to the training data to select the top 10 features per fold, followed by hyperparameter optimization (C, epsilon, gamma) using grid search with cross-validation. Model performance was evaluated using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R^2. The SVR achieved MAE $\approx$ 0.705, RMSE $\approx$ 1.285, and $R^2 \approx -0.162$, indicating performance close to the mean baseline under leakage-free validation. Frequently selected predictors included HbA1c measurements at multiple time points and clinical indicators such as Impaired Glucose Tolerance (IGT). While predictive performance remains limited by sample size, the study establishes a methodologically robust framework for small-scale HbA1c modeling.

* Corresponding author: nizamghazali@apps.ipb.ac.id, irzaman@apps.ipb.ac.id

1 Introduction

Glycated hemoglobin (HbA1c) is a widely used primary indicator for assessing long-term glycemic control in patients with diabetes mellitus because it reflects the average blood glucose level over the preceding ~3 months. Conventional HbA1c testing typically requires well-equipped laboratory facilities, relatively long processing times, and higher costs, creating significant challenges in healthcare settings with limited resources [1-3]. In parallel, advances in data analysis technology−particularly machine learning−have enabled the development of predictive models for medical biomarkers that are faster and more efficient, with the potential to support rapid screening. Previous studies have shown that supervised learning algorithms such as linear regression, Random Forests, and XGBoost can achieve meaningful predictive performance for HbA1c; however, there remains a need to explore approaches that handle high-dimensional data with limited sample sizes and repeated measurements [4-7]. This study, therefore, employs Support Vector Regression (SVR) to develop an HbA1c prediction model using primary microtest data collected from 10 subjects, with three repetitions per subject. SVR was selected because it can capture non-linear relationships and offers strong regularization properties that help reduce overfitting in small datasets [8, 9]. The objective of this work is to develop and evaluate an SVR-based HbA1c prediction model using small-scale microtest data and to identify the most influential features. In terms of the current state of the art, HbA1c prediction research has commonly used linear regression for direct relationships, Random Forests, and XGBoost to capture nonlinearity and feature interactions, and mixed-effects models for longitudinal or panel data; yet, these approaches are typically applied to large datasets and have not been extensively studied for microtest settings with limited repetitions [10]. Consequently, there is limited evidence on SVR performance for HbA1c prediction in small, multidimensional primary clinical datasets, particularly within repeated-test designs that incorporate rapid microtest outcomes rather than relying mainly on conventional medical-record features [11]. This study contributes to the literature by presenting an early-stage, real-world application of SVR to a local HbA1c microtest dataset (10 subjects, three repetitions), highlighting both the methodological constraints and the modeling potential of SVR for high-dimensional, small-scale data, and providing a foundation for future improvements through embedded, leakage-aware feature selection and scalable data expansion.

2 Methodology

2.1 Inclusion criteria

The study included 10 adult participants aged 40 years or older with varying baseline metabolic profiles. Participants had no prior history of diagnosed diabetes mellitus or cardiovascular disease. They were not undergoing insulin therapy or receiving medication specifically prescribed for glycemic control at the time of recruitment. Individuals were required to be clinically stable and free from acute illness during the study period. Participants were able to provide written informed consent and comply with study procedures. They were willing to undergo repeated measurements at three scheduled time points (D1, D7, and D22). Only individuals with complete microtest and clinical measurement records across all sessions were included in the final analysis.

2.2 Exclusion criteria

Participants with severe comorbid conditions (e.g., advanced cardiovascular disease or uncontrolled hypertension) were excluded. Pregnant women or individuals with conditions that could affect the accuracy of microtest results were also excluded. Additionally, individuals taking medications known to affect glucose metabolism, such as corticosteroids or antidiabetic drugs, were excluded. This study involved 10 participants. The sample size was limited due to recruitment challenges and participant availability.

2.3 Study design

This study is a continuation of ongoing development. In this study, a pilot descriptive, analytical, and longitudinal design was implemented to assess physiological and microtest-based predictors of Impaired Glucose Tolerance (IGT) in a small cohort. The primary microtest dataset consisted of 10 subjects, each with three repeated measurements of clinical/biometric parameters, including blood glucose levels, blood pressure, body mass index, and other parameters. Data were collected at three different time points: Day 1 (D1), Day 7 (D7), and Day 22 (D22). The study aimed to identify biomarkers and physiological responses that could serve as early predictors of IGT, a precursor to type 2 diabetes (T2D). [12, 13]. Table 1 summarizes the physiological and biochemical parameters measured during the microtest, along with the corresponding time points. The goal is to provide a clear overview of the input variables used in SVR modeling.

Table 1. Summary of key parameters and measurement time points

Parameter	Measurement Time Points
HbA1c	Day 1, Day 7, Day 22
Fasting Glucose	Day 1, Day 7, Day 22
Hemoglobin (venous)	Day 1, Day 7, Day 22
Body Mass Index (BMI)	Day 1, Day 7, Day 22
Hemoglobin (fingerstick)	M0, M30, M60, M90, M120, M150
Postprandial Glucose	M0, M30, M60, M90, M120, M150
Systolic Blood Pressure	M0, M30, M60, M90, M120, M150
Diastolic Blood Pressure	M0, M30, M60, M90, M120, M150
Heart Rate	M0, M30, M60, M90, M120, M150
Body Temperature	M0, M30, M60, M90, M120, M150

2.4. Data collection

A comprehensive dataset comprising 214 columns was completed in February 2025. The dataset includes a variety of physiological and biochemical markers measured at multiple time points throughout the study. The main parameters recorded were:

- HbA1c (Hemoglobin A1c): To assess long-term glycemic control, using the measuring tool Hipro HbA1c Analyzer HP-AFS/1.

- Glucose Levels: Including fasting glucose and postprandial glucose at each measurement point, with measuring tools Hipro HbA1c Analyzer HP-AFS/1, and Inezco Test N'GO Vita-Voice Multi-Functional Monitoring System, including the strips.
- Blood Pressure: Systolic and diastolic BP measured at several intervals, with measuring tools Omron Automatic Blood Pressure Monitoring Model Hem-7156.
- Heart Rate: Recorded during different phases of the microtest, using the measuring tools Omron Automatic Blood Pressure Monitoring Model Hem-7156.
- Temperature: Infrared forehead thermometer (non-contact) BK8005
- Additional Microtest Responses: Physiological reactions measured under controlled stress, including temperature and biorhythm data, with measuring tools, Thermogun Life resources.
- Questionnaire Results: Self-reported information on lifestyle, dietary patterns, and other personal health factors.

2.5. Measurement Procedure

Before conducting the microtests, Ethical Approval was obtained (Approval No.: 1483/IT3.KEPMSM-IPB/SK/2024) [14].
This study involved repeated measurements conducted at three different time points:

- Day 1 (H1): Baseline measurements, including initial glucose and HbA1c levels.
- Day 7 (H7): Follow-up testing after 7 days to observe changes in glucose metabolism.
- Day 22 (H22): Final measurements to assess long-term physiological responses to the microtest intervention.

All subjects were required to fast for at least 8 hours before each microtest session (the night before). Upon arrival in the morning of H1, body weight, body temperature, systolic and diastolic blood pressure, and heart rate were recorded. For HbA1c testing, fasting blood glucose and hemoglobin were collected from venous blood, while fingertip samples were taken for postprandial glucose and hemoglobin at minute 0 (M0).
Subsequently, each subject consumed 250 mL of mineral water containing 50 g of glucose. Blood samples were then collected from the fingertip at 30-minute intervals (M30, M60, M90, M120, and M150) to measure glucose and hemoglobin levels. After the 150-minute measurement period, subjects were provided with protein, carbohydrates, and fruits. The same procedures were repeated on Day 7 and Day 22.

- Each measurement session thus consisted of multiple sampling intervals (M0‑M150) to capture detailed physiological responses. The intervals were defined as follows:
- M0: Baseline measurement (before any intervention).
- M30‑M150: Post-intervention intervals, where data were collected at each time point to monitor changes in physiological metrics over time.

This approach allowed data collection both at rest and under mild stress, enabling assessment of the body's dynamic response to stimuli and providing insights into glucose metabolism and cardiovascular responses. The dataset comprised 214 variables covering HbA1c, glucose (venous and fingerstick), blood pressure, heart rate, temperature, BMI, biorhythm indices, and questionnaire-derived indicators (e.g., KDia, IGT). Several measurements, such as body temperature before breaking the fast and the time of breaking the fast, become the 0th-minute values. Body weight upon arrival before breaking the fast and the questionnaire data for height are used to obtain the BMI (Body Mass Index) for that day, i.e.,

$$\text{BMI} = \text{Weight (kg)} / \text{Height}^2 \text{ (m}^2\text{)} \quad (1)$$

Meanwhile, the biorhythm values on D1, D7, and D22 were calculated, where *t* is the number of days between the microtest date and the subject's date of birth. The biorhythm values were obtained using the following formula:

$$f(t) = \sin(2\pi t / T) \quad (2)$$

where for Physical T=23, Emotional T=28 and Intellectual T=33. Values range from -1 to 1, where 0 is the critical point, 1 is the best condition, and -1 is the worst condition.
This microtest was conducted at the IPB University Health Laboratory in Dramaga, Bogor. All procedures were carried out in accordance with ethical clearance procedures. The Laboratory has laboratory personnel trained to perform all measurements required for this microtest.

2.6. Data Preprocessing

1. Removal of identifier columns.
2. Selection of numeric variables only.
3. Deletion of missing values in the target variable.
4. Exclusion of HbA1c variables other than the target to prevent information leakage.

2.6.1. Support Vector Regression Algorithm

Feature selection was performed in each fold using Spearman correlation.

$$\rho = 1 - [6\sum d_i^2] / [n(n^2 - 1)] \quad (3)$$

Only the 10 features with the most significant absolute correlation values ($|\rho|$) were selected from the training data.

2.6.2. Support Vector Regression Model

An SVR model with an RBF kernel was employed, defined as:

$$f(x) = \Sigma_i(\alpha_i - \alpha_i^*) K(x_i, x) + b \quad (4)$$

Hyperparameters C, ϵ, and γ were optimized using grid search with cross-validation on the training set within each fold.

2.6.3. Model Evaluation

Model performance was assessed using 5-fold cross-validation with the following metrics:

$$MAE = (1/n) \Sigma|y_i - \hat{y}_i| \quad (5)$$

$$RMSE = \sqrt{[(1/n) \Sigma(y_i - \hat{y}_i)^2]} \quad (6)$$

$$R^2 = 1 - [\Sigma(y_i - \hat{y}_i)^2] / [\Sigma(y_i - \bar{y})^2] \quad (7)$$

A comparative performance analysis between SVR and other models (narrative or tabular) was also conducted.

2.7. Transparency

The dataset used in this study is the same microtest dataset (10 subjects, 3 repetitions, Days 1–22) that has also been analyzed in another study by our group. However, the research objectives and analytical approaches differ substantially. While the other work focused on statistical modeling and ensemble machine learning methods (Linear Regression, Mixed-Effects, Random Forest), the present article focuses exclusively on the methodological evaluation of Support Vector Regression (SVR) with rigorous fold-wise feature selection and hyperparameter optimization.

3 Microtest results

In the microtest experiments, HbA1c values were obtained for all 10 subjects, with repeated measurements on Days 1, 7, and 22, as illustrated in Figure 1. The results of fasting glucose and postprandial glucose measured over 150 minutes are presented in Figure 2.

4 Analysis

Based on the collected data, the target variable was defined as HbA1cPVD22M150, representing the HbA1c measurement obtained at the third stage. Non-numeric columns were removed, and only numerical variables were retained for modeling.

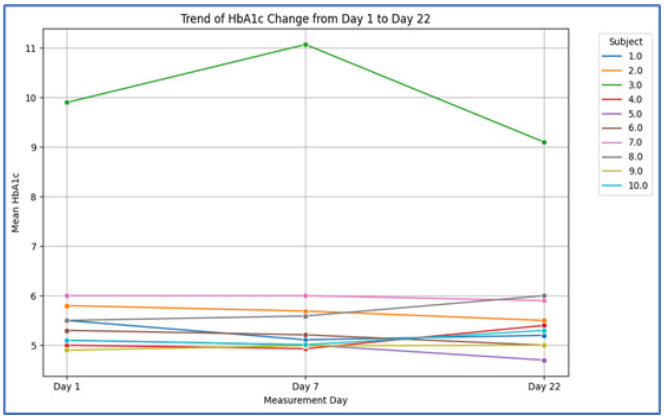

Fig. 1. Trend of HbA1c changes over 22 days from 10 subjects who took this microtest, from the first day, the seventh day, and ending with the microtest on the 22nd day. where each subject was reminded to continue their normal activities and maintain the same lifestyle during the microtest.

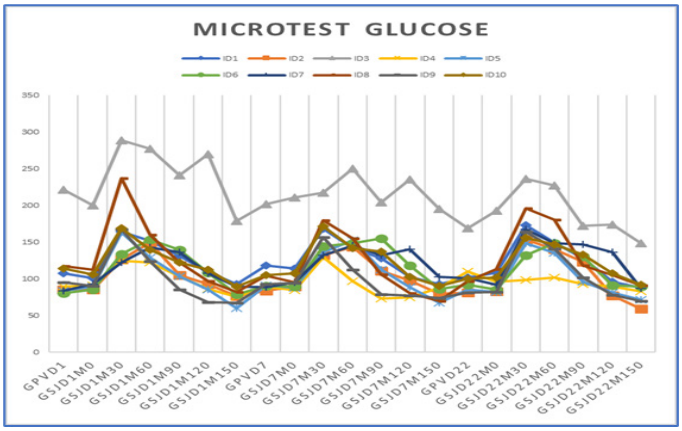

Fig. 2. Glucose trends of 10 subjects over 150 minutes across 3 Microtest Stages within 22 days.

For the first reference analysis (SVR – global feature selection), predictor ranking was performed once at the beginning using Spearman correlation. To minimize information leakage, HbA1c variables measured on the target day (D22) were excluded. The highest-ranked remaining predictors in Table 2 were then applied across all folds, and SVR hyperparameters were manually fixed as a baseline model without cross-validation tuning. This configuration serves as a comparison against the leakage-free fold-wise feature selection combined with grid search cross-validation used in the final model.

Table 2. Top Correlated Predictors after Excluding Target-Day HbA1c Variables

Col	Feature	spearman_rho	p_value	abs_rho
1	IGT	0.811503	0.004369	0.811503
138	PJanD22M60	0.741645	0.014075	0.741645
21	HbA1cPVD1M60	0.709483	0.021561	0.709483
3	HbA1cPVD1M0	0.709483	0.021561	0.709483
49	HbA1cPVD1M150	0.709483	0.021561	0.709483

In Figure 3, the red dashed line represents the identity line (prediction = actual), which serves only as a reference and not as a fitted model. Since features were selected globally before data splitting, there is a potential risk of data leakage. This may result in the model performing very well on the training set, but with reduced validity when generalized to unseen data.

After defining the SVR pipeline, hyperparameters were tuned within each fold using grid search cross-validation on the training split. Model performance was then evaluated on the corresponding held-out fold using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R^2, as summarized in Table 3. The optimal hyperparameter ranges selected across folds were $C = 1–100$, $\varepsilon = 0.01–0.10$, and $\gamma = 0.01$ or "scale", with the most frequently selected combination being $C = 10$, $\varepsilon = 0.01$, and $\gamma = 0.01$.

The SVR model yielded a Mean Absolute Error (MAE) of 0.436, a Root Mean Squared Error (RMSE) of 1.206, and an R^2 of −0.023.

These results indicate that the model's predictive performance was close to the mean baseline, with a negative R^2 suggesting limited explanatory power on unseen data.

Table 3. Performance comparison between global feature selection, fold-wise selection with grid search, and baseline

Analysis Version	MAE	RMSE	R^2
Initial (Global Top-10)	0.436	1.206	-0.023
Final (Fold-wise FS + grid search)	0.705	1.285	-0.162
Baseline (Mean)	0.774	1.192	0.000

Baseline (Mean) refers to predicting the training-fold mean HbA1c for all samples in the corresponding test fold.

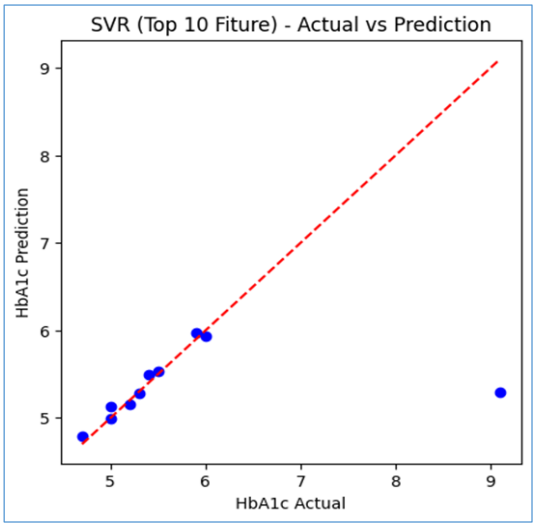

Fig. 3. SVR (Global Feature Selection) – Actual vs Prediction, the red dashed line represents the identity line, which serves only as a reference and not as a fitted model

After performing the initial SVR analysis, a second computation was conducted to generate the second plot (SVR – Fold-wise Feature Selection + grid search with cross-validation). In this approach, feature selection was performed within each cross-validation fold, using only that fold's training data to select the top 10 features. Subsequently, SVR hyperparameters were optimized using grid search with cross-validation on the training data of each fold.

In Figure 4, the blue line represents the identity line; however, the prediction results are more widely scattered around it, indicating that the model generalizes more realistically. This method is stricter and avoids data leakage, resulting in slightly lower performance but producing more accurate predictions for unseen data.

The detailed outcomes are presented in Table 4. Across the 5 folds (n_train=8; n_test=2 per fold), the selected feature sets were generally dominated by metabolic-response variables, especially IGT status and glucose measurements (plasma and fingerstick), with heart rate and temperature appearing intermittently. One fold showed a distinct pattern in which blood-pressure variables (systolic/diastolic) dominated the selected predictors, indicating instability in feature importance under small-sample conditions. Grid search with cross-validation most frequently converged to $C = 10$, $\varepsilon = 0.01$, and $\gamma = 0.01$ (occasionally $\gamma =$ 'scale' and $\varepsilon = 0.10$), suggesting a consistent regularization regime despite fold-to-fold variability in selected features. Most frequently selected predictors across folds ($K = 5$) based on fold-wise Spearman correlation ranking performed exclusively on the training split. The values indicate how many times each feature was selected among the top-ranked predictors within the five cross-validation folds. Variables selected in a greater number of folds (e.g., GSJD22M0 and

IGT, selected in 4/5 folds) suggest relatively stable and consistent contributions to HbA1c prediction under leakage-aware validation.

Table 4. Most frequently selected predictors across folds (K = 5)

Feature	Selected folds (out of 5)
GSJD22M0	4
IGT	4
GPVD1M0	2
GPVD1M30	2
GPVD1M90	2
GPVD1M150	2
GSJD1M0	2
PjanD22M0	2
PjanD22M60	2
TempD1M120	2

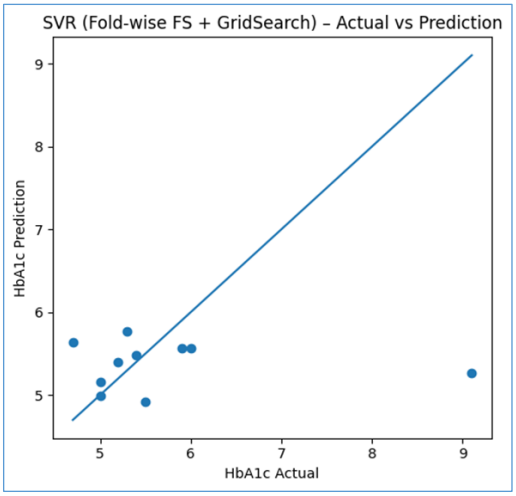

Fig. 4. Scatter plot of actual vs predicted values showing a relatively wide dispersion from the identity line, consistent with the low R^2.

5 Discussion

The SVR performance, which is close to the baseline, highlights the challenges of prediction in a small-n, large-p setting. The perfect correlations observed between several HbA1c variables across different time points and the target indicate extreme multicollinearity. For future research, redundant variables should be removed before modeling.

6 Conclusion

The first plot appeared overly optimistic because feature selection was performed once for the entire dataset. In contrast, the second plot was more conservative but methodologically ensured proper separation of training and testing within each fold, yielding more reliable results. SVR with Spearman correlation–based feature selection can be applied to small-scale HbA1c microtest data; however, its performance was not significantly better than the baseline.

Increasing the sample size, reducing redundant features, and integrating repeated-measure information are expected to enhance model performance. The findings suggest that SVR can be applied robustly under leakage-free evaluation in predictions despite limited data.

This research was supported by the Directorate General of Higher Education, Research, and Technology, Ministry of Education, Culture, Research, and Technology, under Grant No: 06/C3/DT.05.00/PL/2025, dated 9 June 2025.

The authors declare that there are no conflicts of interest related to the publication of this article. All research activities and findings presented were conducted independently, without any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. World Health Orgnization (WHO) Health Topics: Diabetes. Retrieved March 12, 2026, from https://www.who.int/health-topics/diabetes
2. M. David A. D'Alessio, American Diabetes Association (ADA), diabetes journals. Accessed: May 09, 2024. [Online]. Available: https://diabetesjournals.org/diabetes
3. R.A. Ajjan, T. Battelino, J. Seufert, P. Blin, G. de Pouvourville, E. Vicaut, L. Carcaillon-Bentata, F. Levrat-Guillen, E. Cosson, Do continuous glucose monitoring (CGM) metrics predict macrovascular and microvascular complications in diabetes? The FACULTY protocol of a retrospective real-world cohort study. BMJ Open. **15**, 1, e085961 (2025). https://doi.org/10.1136/bmjopen-2024-085961
4. K. Kumar, A. Gupta, N. Gupta, S. Srivastava, M. Asim, Non-Invasive Hemoglobin prediction: A machine learning approach, in Proceedings of the 5th International Conference on Advancement in Electronics & Communication Engineering (AECE), Ghaziabad, India, November 21-22, 2025 (2025), IEEE. https://doi.org/10.1109/AECE67531.2025.11386530
5. Irzaman, R.P. Jenie, Y. Suryana, S. Prambudi, T. Widayanti, D. Mariesta, I. Rahayu, A. Aridarma, S.K. Rahayu, T.S. Riadhie, H. Hardhienata, H. Alatas, Pre-clinical test for non-invasive (*in vitro*) blood glucose levels measuring at visible light wavelengths; in Proceedings of the International Conference and School on Physics in Medicine And Biosystem (ICSPMB): Physics Contribution in Medicine and Biomedical Applications, Depok, Indonesia, November 6-8, 2020 (2021). AIP Conference Proceedings, American Institute of Physics Inc. https://doi.org/10.1063/5.0048161
6. H. Alatas, Y. Suryana, S. Pambudi, T. Widayanti, R.P. Jenie, R. Zaheri, A. Aridarma, S.K. Rahayu, T.S. Riadhie, V. Rahmawaty, N.P. Har, M. Zuhri, T. Sumaryada, Irzaman, Fourier Transform Infra-Red spectrophotometry observation to find appropriate wavelength for non-invasive blood glucose level measurement optical device. Journal of Physics: Conference Series. **1882**, 012009 (2021). https://doi.org/10.1088/1742-6596/1882/1/012009
7. Y. Chen, X. Hu, Y. Zhu, F. Liu, Y. Li, Real-time non-invasive hemoglobin prediction using deep learning-enabled smartphone imaging. BMC Medical Informatics and Decision Making. **24**, 187 (2024). https://doi.org/10.1186/s12911-024-02585-1
8. Y. Zhang, H. Zhang, D. Wang, N. Li, H. Lv, G. Zhang, Development of a 5-year risk prediction model for transition from prediabetes to diabetes using machine learning: Retrospective cohort study. J. Med. Internet Res. **27**, e73190 (2025). https://doi.org/10.2196/73190
9. M. Shabestari, A. Mehrabbeik, S. Barbieri, P. Marques-Vidal, P. Heshmati-nasab, R. Azizi, Predictive factors of hypoglycemia in type 2 diabetes: a prospective study using machine learning. Sci. Rep. **15**, 18143 (2025). https://doi.org/10.1038/s41598-025-03030-7

10. B. Han, Y. Wang, H. Li, X. Sun, J. Zhou, X. Yu, A deep learning framework for HbA1c levels assessment using short-term continuous glucose monitoring data. Biotechnology and Bioprocess Engineering. **30**, 12-29 (2024). https://doi.org/10.1007/s12257-024-00161-y
11. F. Al-hussein, M. Abdollahian, L. Tafakori, K. Al-Shali, A hybrid approach to enhance HbA1c prediction accuracy while minimizing the number of associated predictors: A case-control study in Saudi Arabia. PLoS One. **20**, 6, e0326315 (2025). https://doi.org/10.1371/journal.pone.0326315
12. J. Niu, E. Rodriguez, T. Štambuk, I. Trbojević-Akmačić, N. Mraz, J. Seissler, T. Skurk, S. Schlesinger, A. Peters, G. Lauc, C. Gieger, H. Grallert, Longitudinal study reveals plasma glycans associations with prediabetes/type 2 diabetes in KORA study. Cardiovasc Diabetol. **24**, 321 (2025). https://doi.org/10.1186/s12933-025-02853-y
13. L. Wang, J. Xie, Z. Gu, X. Miao, L. Ma, S. Yan, Y. Gong, C. Li, B. Sun, Y. Ruan, Predicting isolated impaired glucose tolerance without oral glucose tolerance test using machine learning in Chinese Han men. Front Endocrinol (Lausanne). **16**, 1514397 (2025). https://doi.org/10.3389/fendo.2025.1514397
14. S. Hussain, Lab test guide. Accessed: Aug. 19, 2025. [Online]. Available: https://www.labtestsguide.com/