

HbA1c Estimation with Machine Learning and Implementation on Diabetes Self-Management Platform Based on CGM Measurements

Şaban Can Tiryaki

Dept. of Computer Engineering
Kocaeli University
İzmit, Türkiye
245112019@kocaeli.edu.tr

Adnan Kavak

Dept. of Computer Engineering
Kocaeli University
İzmit, Türkiye
akavak@kocaeli.edu.tr

Alpaslan Burak İinner

Dept. of Computer Engineering
Kocaeli University
İzmit, Türkiye
binner@kocaeli.edu.tr

Abstract— Diabetes self-management requires accurate monitoring of long-term blood glucose levels, typically measured by HbA1c. However, traditional laboratory tests provide delayed information and cannot capture real-time trends. This study presents a real-time approach based on Gradient Boosting based machine learning models to predict HbA1c from Continuous Glucose Monitoring (CGM) data to enable personalized diabetes management. CGM data set, i.e., a multimodal resource CGMacros dataset is used in this study. Data augmentation techniques including SMOTE for regression and jittering are employed to improve model learning. Simulation results show that among the models applied, CatBoost model resulted in the highest accuracy, robustness, and reliability in HbA1c prediction. This model is also implemented on a diabetes self-management mobile platform referred to as AIDCARE system via FastAPI, PostgreSQL, and Docker.

Keywords— CGM; CatBoost; Diabetes Self-management; HbA1c Estimation; Machine Learning; XGBoost

I. INTRODUCTION

Accurate assessment of long-term glycemic control, most commonly evaluated through glycated hemoglobin (HbA1c) levels, is fundamental to effective diabetes management. However, conventional laboratory assessments offer only intermittent in-vivo measurements and may fail to capture short-term glycemic variability. Continuous Glucose Monitoring (CGM) measurements based HbA1c estimation may overcome this limitation by delivering frequent, real-time glucose data, thus facilitating more individualized and precise diabetes care strategies. [1].

Recent research has employed machine learning approaches, including Random Forest, XGBoost, and Support Vector Machines (SVM), to predict HbA1c levels from Continuous Glucose Monitoring (CGM) data, demonstrating potential in capturing the nonlinear relationships between glucose dynamics and HbA1c outcomes [2]. Nonetheless, limitations persist, particularly with respect to small sample sizes, class imbalance, and the insufficient exploitation of temporal features.

In [3], a few-shot learning architecture is proposed for long-term HbA1c prediction [3]. While the model demonstrated reasonable predictive capability, it lacked real-time implementation and mobile platform deployment and did not employ data augmentation techniques such as SMOTE.

An explainable machine learning framework for real-time hypoglycemia and hyperglycemia prediction focusing on short-term glucose events rather than HbA1c prediction is presented [4]. A feature-based machine learning model is introduced in [5] for real-time hypoglycemia prediction. These studies did not incorporate mobile-friendly deployment or fast API-based services for real-time diabetes self-management. A deep transfer learning model is explored to improve blood glucose prediction in [6]. While this approach improved short-term glucose estimation, it did not specifically target HbA1c prediction based on short-term CGM data and also lacked mobile or real-time deployment capabilities.

This study proposes a real-time machine learning framework for predicting HbA1c from Continuous Glucose Monitoring (CGM) data, utilizing data augmentation strategies to mitigate class imbalance and extreme glucose values. Gradient boosting-based models, specifically XGBoost and CatBoost, are employed owing to their proven effectiveness in modeling nonlinear associations and handling high-dimensional, noisy biomedical data. The proposed gradient boosting framework is further integrated into AIDCARE, a mobile-enabled, AI-powered diabetes self-management platform, to facilitate real-time HbA1c prediction and personalized diabetes monitoring.

The remainder of this paper is structured as follows. Section II provides a detailed description of the dataset utilized in this study. Section III outlines the HbA1c estimation problem and reports the corresponding simulation results. Section IV presents the implementation of the HbA1c prediction framework within the AIDCARE platform. Finally, Section V provides concluding remarks.

II. DATASET DESCRIPTION

This study used the CGMacros dataset [7], a multimodal resource from the Sansum Diabetes Research Institute (2021–2024), which includes CGM data, meal information, physical activity, demographics, blood tests, and gut microbiome profiles. The dataset consists of 45 participants—15 healthy, 16 prediabetic, and 14 with type 2 diabetes—aged 18–69, each monitored for ~10 days using two CGM devices and a Fitbit Sense. From CGM data, over 30 statistical and time-based features were extracted, including mean, standard deviation, coefficient of variation, area under the curve, trend coefficient, time in range, and daily max–min averages. Missing values were imputed using Iterative Imputer to preserve multivariate relationships. To improve model training and account for extreme HbA1c values, the target variable was transformed using Power Transformer (Yeo-Johnson) and standardized. Additionally, the dataset was augmented via regression-adapted SMOTE and jittering by increasing the sample size and enhancing learning for high-risk individuals [8].

III. HbA1c ESTIMATION USING GRADIENT BOOSTING BASED MACHINE LEARNING MODELS

Let the input feature vector be $X=[x_1 \ x_2 \ \dots \ x_N]$ which is extracted from Libre Glucose data, the input feature data set is X_s after Power Transformer and Standard Scaler preprocessing. The target variable is

$$y = \log(1 + HbA1c). \quad (1)$$

The predicted HbA1c, $\hat{y}_t(i)$ for sample i at iteration t is given by

$$\hat{y}_t(i) = \widehat{y}_{t-1}(i) + \eta \cdot f_t(X_s(i)), \quad (2)$$

where η is the learning rate, $f_t(\cdot)$ is the decision tree output. The optimization objective at iteration t is

$$C_t = \sum_i L(y_t, \hat{y}_t(i)) + \Omega(f_t), \quad (3)$$

where L is the loss function (in this study, RMSE) and $\Omega(f_t)$ is the regularization term controlling the model complexity.

To estimate HbA1c from CGM-derived features, two nonlinear regression models namely XGBoost, and CatBoost were applied. These models were chosen for their ability to capture nonlinear relationships and handle high-dimensional, noisy biomedical data

XGBoost regressor is a gradient boosting method in which sequential trees iteratively correct errors from previous iterations. Each new tree is fitted to the residuals of the previous ensemble to improve accuracy while regularization and learning rate tuning help reduce overfitting. XGBoost builds each tree sequentially, using sampling strategies to improve diversity and generalization. Specifically, the subsample parameter controls the fraction of rows (observations) used to build each tree, while colsample_bytree specifies the fraction of features (columns)

sampled for each tree. These strategies increase tree diversity, reduce overfitting, and make the model effective for complex, high-dimensional datasets.

CatBoost Regressor is a gradient boosting method that handles categorical variables automatically and reduces overfitting through ordered boosting, early stopping, and L2 regularization. It supports fast and stable learning on mixed data types.

After M iterations, the final predicted log-transformed HbA1c value is converted back to the original scale,

$$HbA1c_i = \exp(\widehat{y}_M(i)) - 1 \quad (4)$$

The simulations were conducted in a Python environment using the Scikit-learn, XGBoost, and CatBoost libraries applied on CGMacros dataset. Model parameters were tuned by optimizing the learning rate, tree depth, and regularization terms. Data preprocessing steps are applied using Power Transformer and Standard Scaler. Data augmentation with regression-adapted SMOTE and Jittering expanded the dataset in order to improve model learning capability, especially for high HbA1c values.

In the model training, the RMSE loss function and gradient boosting techniques were applied. The XGBoost and CatBoost models were tuned to optimize prediction performance. XGBoost model was adjusted for learning rate, tree depth, and row/column sampling ratios to balance accuracy and generalization. CatBoost achieved the best performance with 300 iterations, a depth of 7, a learning rate of 0.05, and an L2 regularization term (L2_leaf_reg) of 1. XGBoost and CatBoost regression models were evaluated on the log-transformed HbA1c values using RMSE, MAE, and R^2 . The evaluation results are summarized in Table 1.

CatBoost achieved the smaller RMSE and MAE, and the highest R^2 , demonstrating superior predictive accuracy and robustness for HbA1c estimation from CGM-derived features. These results highlight the effectiveness of combining data augmentation techniques with gradient boosting models for real-time HbA1c prediction, supporting their potential integration into personalized diabetes management systems.

TABLE I. EVALUATION RESULTS FOR HbA1c ESTIMATION

Model	Measures		
	RMSE	MAE	R^2
CatBoost	0.1540	0.1183	0.9806
XGBoost	0.1811	0.1438	0.9732

IV. IMPLEMENTATION ON AIDCARE-DIABETES SELF MANAGEMENT PLATFORM

The findings of our theoretical investigation into the effectiveness of Gradient Boosting Models motivated their implementation within AIDCARE platform, a real-time, AI-powered, and mobile-assisted diabetes self-management platform currently under development [9]. The system consists of seven main modules: Risk, Measurement, Recommendations, Exercise, Diet, Reminders, and Education. The mobile screenshot of the AIDCARE application is shown in Figure 1. With its user-friendly mobile interface and integrated notification system, the AIDCARE application simplifies the daily health management process for patients with diabetes. Through a dedicated web-based management panel for doctors, medical staff can instantly monitor patients' glucose measurements, blood pressure values, insulin dosages, and other health data. This enhances doctor-patient interactions, accelerates checkups, and improves the efficiency of healthcare services.

One of the most distinctive features of the system is its ability to not only collect data but also analyze it to provide meaningful health recommendations. With its rule-based and AI-driven analytical infrastructure, AIDCARE offers diabetes risk prediction, blood glucose trend forecasting, insulin dosage recommendations, and personalized health advice. In the AIDCARE application, when the user clicks the “Calculate HbA1c Value” button under “Risk module”, the system displays the predicted HbA1c value. The application then determines the corresponding risk level and informs the patient.

The HbA1c prediction using CatBoost model is implemented as a modular, three-layer architecture, illustrated in Figure 4. Users interact via desktop or mobile clients, while the FastAPI application runs inside a Docker container. The API exposes a /Results endpoint for retrieving current or historical predictions. Real-time HbA1c predictions are automatically triggered when new glucose measurements arrive in the PostgreSQL database. An asynchronous background process continuously polls the database every 10 seconds, retrieves new data, extracts statistical and temporal features, and applies the same preprocessing pipeline used during model training (PowerTransformer and StandardScaler).

The preprocessed features are input to the pre-trained CatBoost model to generate log-transformed HbA1c predictions, which are converted back to the original scale. Predicted values are classified into risk categories—Normal, Prediabetes, and Diabetes—and stored in the database. Users can access their predictions and associated risk information. This design ensures reproducibility, real-time performance, and scalable deployment suitable for clinical or research applications.

The HbA1c prediction using CatBoost model is implemented as a modular, three-layer architecture, illustrated in Figure 4. Users interact via desktop or mobile clients, while the FastAPI application runs inside a Docker container. The API exposes a /Results endpoint for retrieving current or historical predictions. Real-time HbA1c predictions are automatically triggered when new glucose measurements arrive in the PostgreSQL database. An asynchronous background process

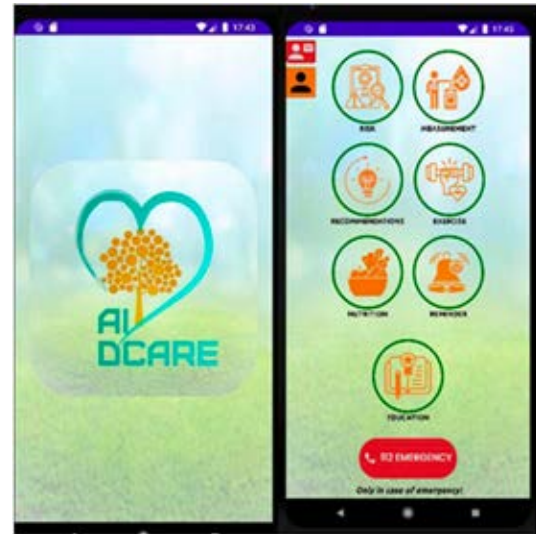


Fig.1 AIDCARE Mobile Application for Diabetes Management



Fig.2 Screenshot of HbA1c Prediction Implemented on AIDCARE Application.

continuously polls the database every 10 seconds, retrieves new data, extracts statistical and temporal features, and applies the same preprocessing pipeline used during model training (PowerTransformer and StandardScaler). The preprocessed features are input to the pre-trained CatBoost model to generate log-transformed HbA1c predictions, which are converted back to the original scale. Predicted values are classified into risk categories—Normal, Prediabetes, and Diabetes—and stored in the database. Users can access their predictions and associated risk information. This design ensures reproducibility, real-time performance, and scalable deployment suitable for clinical or research applications.

V. CONCLUSION

This study presented a real-time machine learning framework, grounded in gradient boosting models, for predicting HbA1c levels from Continuous Glucose Monitoring (CGM) data. The incorporation of data augmentation techniques, including regression-based SMOTE and jittering, enhanced the overall learning performance. Among the evaluated models, CatBoost achieved superior results, demonstrating higher accuracy, robustness, and reliability in HbA1c prediction.

The HbA1c prediction system was further deployed within a modular, containerized architecture utilizing FastAPI and PostgreSQL, with compatibility across both web and mobile platforms. Predictions were automatically initiated upon the insertion of new records into the database, thereby enabling structured risk assessments to support real-time diabetes self-monitoring.

The proposed approach highlights the potential for non-invasive, continuous HbA1c monitoring, offering a scalable and reproducible solution suitable for integration into personalized diabetes management platforms.

ACKNOWLEDGMENT

The study presented in this paper is conducted within the AIDCARE project funded by Health Institutes of Türkiye (TUSEB), Turkish Ministry of Health under grant TUSEB 33987.

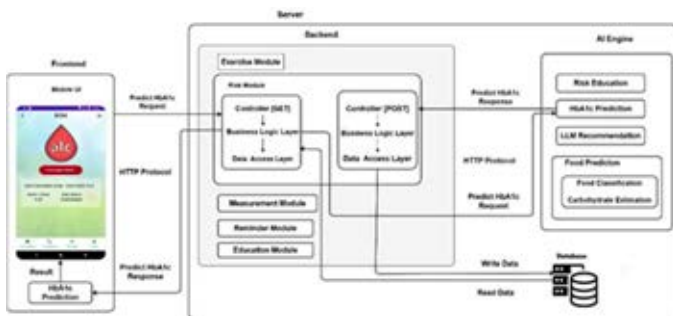


Fig.3 Implementation Methodology of HbA1C Prediction on AIDCARE Platform

REFERENCES

- [1] H. Chehregosha, M.E. Khamseh, M. Malek, F. Hosseinpanah, and F. Ismail-Beigi, "A view beyond HbA1c: role of continuous glucose monitoring," *Diabetes Therapy*, vol. 10 (3), pp. 853-863, 2019.
- [2] T. Elhadd, R. Mall, M. Bashir, J. Palotti, L. Fernandez-Luque, F. Farooq, D. Al Mohanadi, Z. Dabbous, R.A. Malik, and A.B. Abou-Samra, "Artificial intelligence (AI) based machine learning models predict glucose variability and hypoglycaemia risk in patients with type 2 diabetes on a multiple drug regimen who fast during ramadan (The PROFAST-IT Ramadan study)," *Diabetes Research and Clinical Practice*, vol. 169, 108388, 2020.
- [3] M. Qaraqe, A. Elzein, S. Belhaouari, M.S. Ilam, and G. Petrovski, "A novel few shot learning derived architecture for long-term HbA1c prediction," *Scientific Reports*, vol.14(1), 482, 2024.
- [4] C. Duckworth, M.J. Guy, A. Kumaran, A.A. O'Kane, A. Ayobi, A. Chapman, P. Marshall, and M. Boniface, "Explainable machine learning for real-time hypoglycemia and hyperglycemia prediction and personalized control recommendations," *Journal of Diabetes Science and Technology*, vol. 18(1), pp. 113-123, 2024.
- [5] D. Dave, D.J. DeSalvo, B. Haridas, S. McKay, A. Shenoy, C.J. Koh, M. Lawley, M. Erraguntla, "Feature-based machine learning model for real-time hypoglycemia prediction," *Journal of Diabetes Science and Technology*, vol. 15(4), pp. 842-855, 2021.
- [6] Y. Deng, L. Lu, L. Aponte, A.M. Angelidi, V. Novak, G.E. Karniadakis, and C.S. Mantzoros, "Deep transfer learning and data augmentation improve glucose levels prediction in type 2 diabetes patients.," *NPJ Digital Medicine*, vol. 4(1), 109, 2021.
- [7] R. Gutierrez-Osuna, D. Kerr, B. Mortazavi and A. Das, "CGMacros: A scientific dataset for personalized nutrition and diet monitoring (version 1.0.0)," *PhysioNet*, 2025, <https://doi.org/10.13026/3z8q-x658>
- [8] A. Géron, (2022). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow*. O'Reilly Media Inc., 2022.
- [9] Ö.F. Üçer, A. Kavak, Y. Demirhan, M. Uzun, B. Savas, A.B. Inner, Ö. Alkan, B. Çetinarslan, Z. Cantürk, A. Selek, E. Gezer, U. Yiğit, A. Aktaş, A.T. Fırat, M.A. Demirtaş, G. Okandan, K. Ünlü, S. Ünersoy, and Ö. Çakır, "Implementation of a decision support mechanism on a mobile platform using clinical evidence based dynamic insulin dosage adjustment for artificial intelligence enabled diabetes care (AIDCARE) system," *MDPI Engineering Proceedings*, vol 104(27), 2025.