Journal of Information Systems Engineering and Management
2025, 10(52s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

CatBoost Model for Enhanced Treatment Prediction in Type 2 Diabetes Patients

Suhail S.M.Alqrinawi[1], M.A.Burhanuddin[2*], Lizawati Salahuddin[3]
[1,2,3] *Universiti Teknikal Malaysia Melaka, Fakulti Teknologi Maklumat dan Komunikasi (FTMK), Melaka, Malaysia.*
[1] *suhailalqrinawi@gmail.com,* [2] *burhanuddin@utem.edu.my,* [3] *lizawati@utem.edu.my*

ARTICLE INFO

Received: 07 Mar 2025

Revised: 14 May 2025

Accepted: 22 May 2025

ABSTRACT

Customized therapeutic approaches are essential for enhancing outcomes and patient satisfaction in the management of type 2 diabetes. This paper presents a predictive model utilizing CatBoost, a gradient boosting algorithm adept at managing categorical data, to offer appropriate prescription regimens for diabetes patients. This research utilizes a dataset of 51,000 anonymized patient records supplied by the Palestinian Ministry of Health, which includes extensive demographic, clinical, and pharmaceutical information. Our methodology prioritizes rigorous preprocessing techniques, encompassing feature selection and the management of missing values, to guarantee data quality. CatBoost surpassed baseline models, such as Random Forest and Support Vector Machine, attaining a prediction accuracy of 97%. Critical factors affecting treatment efficacy comprised HbA1c levels, adherence to medication, and concomitant conditions. This study shows how effective machine learning is in offering personalized care, especially in places with limited resources, and helps improve smart decision support systems for managing chronic diseases.

Keywords: Type 2 Diabetes, CatBoost, Machine Learning, Personalized Medicine, Patient Satisfaction, Palestinian Health Data, HER (Electronic Health Record)

INTRODUCTION

Diabetes is a chronic metabolic disorder characterized by an abnormal blood glucose level resulting from either ineffective utilization or insufficient production of insulin. Millions of people globally have type 2 diabetes. This chronic illness must be closely monitored and treated individually to avoid risks. Conventional approaches to pharmaceutical recommendations often lack specificity and fail to consider the unique characteristics of each patient, including comorbidities, demographics, and past medical history. By using vast amounts of patient data to forecast individualized treatment regimens, recent developments in machine learning (ML) have demonstrated the possibility of filling this gap(Deberneh and Kim, 2021).

This study examines the application of the gradient boosting algorithm CatBoost to predict the optimal prescription recommendations for individuals with Type 2 Diabetes. CatBoost is ideally suited for EHR-based applications because, in contrast to other algorithms, it performs exceptionally well when handling categorical data and avoiding overfitting. Through precise and comprehensible predictions, the project seeks to enhance patient satisfaction and treatment outcomes using a clean and pre-processed dataset.

The rest of the paper is organized as follows:

Section II reviews previous research conducted on diabetes disease prediction. Section III outlines the materials utilized and the proposed methodology. Section IV presents the experimental results and their analysis. In conclusion, Section V finalizes the paper.

LITERATURE REVIEW

The complex nature of Type 2 diabetes mellitus (T2DM) and the significant variability in patient responses to treatment present numerous challenges in its management. Type 2 diabetes mellitus (T2DM) affects millions globally and is a significant contributor to morbidity and mortality due to complications, including kidney failure, neuropathy,

Journal of Information Systems Engineering and Management
2025, 10(52s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

and cardiovascular disease. Traditional therapeutic approaches, which often follow predefined procedures, may not sufficiently address the unique metabolic profiles and lifestyle characteristics of individual patients. Consequently, there is an increasing demand for customized treatment plans that optimize therapeutic outcomes. Recent advancements in machine learning (ML) utilize extensive datasets to identify optimal treatment plans for individual patients, thereby enhancing the precision of diabetes care(Hussain *et al.*, 2024).

Patient satisfaction is crucial for healthcare providers since it shows the quality of their work and influences treatment compliance and overall health outcomes. This study examines the capacity of machine learning (ML) to improve patient satisfaction by optimizing processes, predicting patient needs, and providing personalized care. The ability of machine learning to analyze large datasets and identify patterns that can be utilized to tailor patient treatment represents a significant advantage in the healthcare sector(Liu, Kumara and Reich, 2021).

BRIEF OVERVIEW OF ML METHODS

Machine learning methods are categorized into two main types: supervised learning and unsupervised learning. In supervised learning, the algorithm is provided with a known target variable and is trained on labelled datasets to predict this variable. In unsupervised learning, the algorithm lacks a definitive "correct answer" for prediction, instead focusing on identifying patterns within unlabelled data sets. Supervised learning examples include support vector machines and neural networks, whereas unsupervised learning examples encompass clustering and manifold learning(Tan *et al.*, 2023).

RELATED WORK

Machine learning has seen increasing application in chronic disease management, especially in predicting complications and optimizing drug recommendations. Prior research has explored Random Forests and deep learning techniques for risk stratification and treatment modeling. However, limited studies have focused on leveraging patient satisfaction as a proxy for treatment efficacy and integrating advanced boosting techniques like CatBoost, which naturally handles categorical features—a frequent limitation in medical datasets.

Deberneh and Kim (2021) proposed a machine learning (ML) model designed to forecast the start of Type 2 Diabetes (T2D) one year prior, utilizing an extensive electronic health records (EHR) dataset amassed in South Korea from 2013 to 2018. The research utilized a data-driven feature selection methodology that integrated ANOVA, chi-squared tests, and recursive feature elimination, yielding 12 essential features, including fasting plasma glucose (FPG), HbA1c, BMI, triglycerides, gamma-GTP, and lifestyle factors such as smoking, alcohol consumption, and physical activity. Multiple machine learning algorithms—logistic regression, random forest, support vector machine, and XGBoost—were evaluated, alongside ensemble techniques such as soft voting, stacking, and integration based on confusion matrices. The optimal test accuracy attained was 73%, however the cross-validation accuracy increased to 81% with the utilization of four years of historical medical data. The 12-feature model markedly surpassed conventional 5-feature models (FPG, HbA1c, BMI, age, sex), particularly in forecasting the prediabetes category. Despite the obstacles given by class overlap, especially between normal and prediabetes, the ensemble models exhibited enhanced robustness. The study underscores the promise of machine learning in the early prediction of diabetes of type 2, providing significant assistance to doctors in preventive decision-making and customized patient management(Deberneh and Kim, 2021).

This study employed unsupervised machine learning to investigate the variability in insulin responsiveness among individuals with type 1 diabetes (T1D) following rapid-acting insulin administration. Data from two clinical trials revealed significant disparities in insulin peak, time to peak, and total insulin exposure, quantified as incremental Area Under the Curve. Two groups were identified: one exhibiting elevated insulin levels and the other displaying diminished levels. The lower group had indications of insulin resistance, elevated inflammation, and prothrombotic markers. These people were older, had a prolonged course of diabetes, and exhibited a higher body mass index (BMI). The results indicate that diminished insulin responses may indicate heightened vascular risk. Customized insulin regimens may enhance outcomes by accommodating individual differences in insulin efficacy(Coales *et al.*, 2022).

In this study, 200 patients from the Medical Center Chittagong in Bangladesh are used as a real-world dataset to investigate the early prediction of diabetes mellitus using machine learning approaches. The dataset comprises 16

Journal of Information Systems Engineering and Management
2025, 10(52s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

attributes, including age, diet, blood pressure, genetic factors, and symptoms such as polyuria and excessive thirst. Four classifiers—Support Vector Machine (SVM), Naive Bayes (NB), K-Nearest Neighbor (KNN), and C4.5 Decision Tree (DT)—were utilized for predictive modeling. Following preprocessing, which involved binning numerical data into categorical ranges, models were assessed through 10-fold cross-validation, focusing on precision, recall, F-measure, and accuracy. The C4.5 decision tree model demonstrated superior performance among the tested models, achieving an accuracy of 73.5%. It also surpassed others in precision (72%), recall (74%), and F-measure (72%). The results indicate that decision tree-based models are effective and interpretable tools for diabetes prediction, facilitating early diagnosis and enhancing clinical decision-making(Faruque, Asaduzzaman and Sarker, 2019).

Introduce An effective machine learning framework for detecting Type 2 diabetes utilizing a dataset marked by imbalanced data and missing values is presented in the research paper by Kumarmangal Roy et al. The study examines three data imputation techniques median value imputation, K-nearest neighbor imputation, and iterative imputation—and evaluates their effects on classification accuracy across different algorithms to help medical professionals diagnose diabetes with high classification accuracy. After balancing the data with SMOTET using the best imputation technique, an Artificial Neural Network (ANN) was utilized to model the data, yielding a 98% test accuracy, the study focused on female patients of Pima Indian heritage and used a dataset from the National Institute of Diabetes and Digestive and Kidney Diseases. Medical predictors included age, blood pressure, skin thickness, insulin, insulin level, BMI, diabetes pedigree function, and pregnancy history. The researchers improved the model's performance by preparing the data to address missing values, outliers, and class imbalance, and this study's findings indicate that the ANN model may provide a valuable tool for predicting diabetes at an early stage, which may result in improved patient outcomes and decreased medical expenditures(Paul and Karn, 2021).

Feng, Cai and Xin, (2023) presents a machine learning-based framework for optimizing diabetes classification. The framework, termed MOG, where GOM is the framework, is designed to be robust and effective in accurately classifying diabetes using machine learning technology. It has been validated with high accuracy on both the PIMA dataset and datasets from the GEO database, demonstrating its potential as a reliable tool for intelligent diabetes diagnosis. It addresses challenges in diabetes diagnosis by employing a novel data preprocessing technique that combines mean and median imputation for missing values, a capping method for outlier handling, and the SMOTEENN algorithm to balance data imbalance. The core of the framework is the Diabetes Classification Model based on Generative Adversarial Networks (DCSGAN), which uses logistic regression for feature analysis chine learning-based framework for optimizing diabetes classification. The framework, termed MOG where GOM is framework, is designed to be robust and effective in accurately classifying diabetes using machine learning technology. It has been validated with high accuracy on both the PIMA dataset and datasets from the GEO database, demonstrating its potential as a reliable tool for intelligent diabetes diagnosis, addresses challenges in diabetes diagnosis by employing a novel data preprocessing technique combining mean and median imputation for missing values, a capping method for outlier handling, and the SMOTEENN algorithm to balance data imbalance. The core of the framework is the Diabetes Classification Model based on Generative Adversarial Network (DCSGAN), which uses logistic regression for feature analysis(Feng, Cai and Xin, 2023).

Various machine learning models are employed to enhance healthcare service quality by predicting patient satisfaction. Patient satisfaction is significantly influenced by sociodemographic factors and the duration of patient stay (LOS). Data were collected from 139 outpatient surveys conducted in the ophthalmology department of the hospital, emphasizing general patient information and satisfaction ratings across different lengths of stay. Four predictive models were developed and validated through simulation and real patient data, employing the Mann-Whitney U test: ordinal logistic regression (OLR), decision trees (DT), gradient-boosted trees (GBT), and neural networks (NN). The OLR model consistently exhibited superior predictive accuracy across all categories. Besides DT and GBT, additional models demonstrated effectiveness in predicting satisfaction for shorter lengths of stay(Chemkomnerd *et al.*, 2024).

This study presents a hybrid framework that integrates a Deep Neural Network (DNN) with a Modified Random Forest Incremental Interpretation (MRFII) method to improve the accuracy and interpretability of Type 2 diabetes diagnosis. A DNN is initially employed to estimate the probability of diabetes utilizing eight features from the PIMA

Journal of Information Systems Engineering and Management
2025, 10(52s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Indian dataset. A Random Forest (RF) is developed to approximate the decision-making process of deep neural networks (DNNs) in order to mitigate their black-box characteristics through interpretable rules. MRFII subsequently reorganizes and prioritizes these rules to mitigate patient confusion and improve clarity. A certainty mechanism incorporates RF outputs into the DNN to enhance prediction performance. Experimental results indicated that the proposed method surpassed traditional models (e.g., LR, SVM, ANN, RF), with DNN achieving an accuracy of 77.1%, which increased to 81.3% upon the application of the certainty mechanism. This method enhances diagnostic accuracy and improves transparency and communication with patients(Chen, Wu and Chiu, 2024).

METHODOLOGY

The suggested hybrid methodology for breast cancer detection integrates the advantages of conventional machine learning with deep learning models to improve predictive accuracy. This methodology aims to mitigate the shortcomings of each model by integrating their complementary strengths, as illustrated in the figure.

Figure 1. Architecture diagram of proposed model

5.1 Data Collection

We utilized a dataset compiled from electronic health records of Type 2 Diabetes patients,

Data Extraction Software: To extract data from electronic chronic disease database systems, we use tools like SQL queries, Python scripts, or MOH-approved software containing:

5.1.1 **Demographic information**: age, gender, BMI, ethnicity.

5.1.2 **Clinical variables**: blood glucose levels, HbA1c, blood pressure, comorbidities.

5.1.3 **Medication details**: type, dosage, duration of treatment, medication combinations, response to medication.

Journal of Information Systems Engineering and Management
2025, 10(52s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

TABLE 1: Features list

#	
1.	Age
2.	Height
3.	Weight
4.	Body Mass Index
5.	Blood Pressure
6.	Post Prandial Sugar
7.	Heart Rate
8.	Smoking
9.	Hypercholest
10.	HBA1C
11.	FBG Diabetic
12.	Control Status
13.	Medicines
14.	Medicine History
15.	Type of Complication
16.	Time of Diagnosis
17.	Waist Circumference
18.	Urea Test

5.2 Collect Data (Select Full Data Row)

According to(Sarkies *et al.*, 2015) the collection of data refers to the systematic process of gathering information for research and analysis. As part of health services research, various methods are used to capture data on key outcome measures, such as hospital length of stay and discharge location. Evaluation of the effectiveness and efficiency of healthcare services requires these measures, moreover, It is possible to collect data using a variety of methods, including manual data collection from ward-based sources, retrospective administrative data extraction from electronic patient management programs, and review of scanned inpatient medical records, which is often considered the gold standard for data collection.This involves gathering comprehensive data on diabetes patients. The data should include patient demographics, medical history, lab results, and treatment plans. Ensuring the dataset is complete and accurately reflects the population of interest is crucial.

5.3 Set Process

Outline the data processing and modelling steps. This involves deciding the sequence of operations, the tools to use, and how each step will contribute to the ultimate goal of predicting medicine for patients.

5.4 Remove Duplicate Data

Identify and remove duplicate entries in the dataset to ensure data integrity. Duplicate data can skew analysis and introduce biases, so it's vital to address this step early.

5.5 Handle Missing Values

5.6 Deal with incomplete data by:

Filling missing values using statistical methods (mean, median, mode).

Employing machine learning techniques for imputation.

Removing rows or columns if missing data is excessive and not critical

5.7 Select Good Features

Analyze the dataset to identify relevant features (columns) that strongly impact the target outcome (e.g., which medicine to prescribe). Irrelevant or noisy features should be excluded to enhance model performance.

Journal of Information Systems Engineering and Management
2025, 10(52s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

5.8 Feature Extraction

Create new, more informative features from existing ones (e.g., BMI from weight and height). This step may involve domain knowledge or advanced techniques like principal component analysis (PCA).

5.9 Split Feature and Target

In this step, the dataset is divided into features (input variables) and the target variable (output). To ensure that machine learning models understand patterns from input data and provide correct predictions, this separation is crucial during the training process.

5.10 Separate the dataset into:

1. Features (X): Independent variables that influence the target.

2. Target (Y): The dependent variable, in this case, the medicine prescribed.

Split Data into Training and Testing Sets

Divide the dataset into:

Training set: Used to train the machine learning model (typically 80% of the data).

Testing set: Used to evaluate the model's performance on unseen data (20%).

5.11 Standardized Features

Normalize or scale the feature data to ensure all variables contribute equally to the model. Standardization is critical when using algorithms sensitive to feature scales, such as support vector machines or neural networks.

5.12 Make Prediction

Train the model using the training data and then use the model to predict the target outcome (medicine) for the testing data. Choose the machine learning algorithm that best fits the problem (CatBoost) and we find the accuracy (of 97 %)at the same time we find the overall accuracy of SVM equal to (16%), on the other hand, the same time we find the overall accuracy for Random forest equal to (13%).

Evaluate the Model (Predict Medicine for Patients)

Assess the model's performance using metrics such as:

Accuracy: The percentage of correct predictions.

Precision, Recall, F1-Score: Metrics for imbalanced datasets.

Mean Squared Error (MSE): For continuous predictions.

CATBOOST MODEL

CatBoost, a gradient boosting algorithm developed, is optimized for categorical data and avoids overfitting through techniques like ordered boosting. Its advantages include:

- No need for extensive encoding.
- Faster convergence.
- Improved accuracy with smaller datasets.

BASELINE MODELS

To evaluate performance, we compared CatBoost with:

7.1 Random Forest:

The random forest model has been widely used for many years as an excellent ensemble algorithm for bagging that fits the best multi-classification combination model based on a comprehensive comparison of random features. The

Journal of Information Systems Engineering and Management
2025, 10(52s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Random Forest method is a highly effective ensemble learning method that is widely used in machine learning due to its robust performance, ability to handle a wide variety of data types, and inherent resistance to overfitting(Esmaily *et al.*, 2018).

7.2 XGBoost:

XGBoost is a sophisticated implementation of an improved Gradient Boosting algorithm, engineered for exceptional efficiency, versatility, and portability. XGBoost is a tree-based technique that falls within the supervised category of machine learning. This narrative exclusively pertains to the algorithm's application in classification, despite its applicability to both classification and regression issues(Toharudin *et al.*, 2023)

7.3 Support Vector Machines (SVM)

Support Vector Machine (SVM) is extensively employed for prediction. The medicine of patients with diabetes type 2 classification due to its resilience in managing high-dimensional data. They construct a hyperplane in a multidimensional space that distinguishes information into distinct categories (benign or malignant). Research illustrates the efficacy of SVM in categorizing breast cancer through radiological data. It possesses elevated precision(Deshmukh *et al.*, 2025).

EVALUATION METRICS

- Accuracy
- Precision, Recall, F1-score
- ROC-AUC
- SHAP values for model interpretability

RESULTS

Model	Accuracy	F1-score	AUC
CatBoost	97%	0.81	0.87

CatBoost outperformed other models across all metrics. Feature importance showed that medication adherence, HbA1c, and comorbidity score were the top predictors of treatment success.

DISCUSSION

This research demonstrates the efficacy of employing machine learning, particularly CatBoost, in predicting optimal treatment strategies for patients with Type 2 Diabetes. The model utilized over 51,000 records from the Palestinian Ministry of Health to identify critical clinical features, including HbA1c, BMI, age, and medication history, as significant predictors of treatment outcomes. CatBoost demonstrated effectiveness in real-world healthcare data by efficiently managing categorical variables and missing values with limited preprocessing requirements. The performance underscores the significance of data-driven methodologies in facilitating personalized care, particularly in resource-limited settings. The results are promising; however, limitations include dependence on retrospective data and the necessity for clinical validation. This study illustrates the capacity of AI to improve chronic disease management and healthcare delivery in underserved areas, despite existing challenges.

CONCLUSION

This study offers an extensive framework for utilizing machine learning in the management and optimization of treatment for individuals with Type 2 Diabetes. The study examined and validated a predictive model utilizing real-world electronic health record (EHR) data to identify optimal treatment plans tailored to individual patient needs.

This study utilized a dataset sourced from the Palestinian Ministry of Health before the recent war outbreak. The dataset comprised more than 51,000 anonymized health records of Type 2 Diabetes patients, offering a comprehensive and representative perspective on the healthcare landscape in Palestine. The dataset included critical patient information, encompassing demographics, clinical indicators, laboratory results, and medication history, facilitating a thorough and evidence-based analysis. Key factors influencing treatment outcomes were identified through a thorough evaluation, including HbA1c levels, BMI, age, and prior medication use. Machine learning techniques, especially CatBoost, have demonstrated effectiveness in managing complex healthcare data, providing strong predictive performance, resilience to missing values, and interpretability appropriate for medical applications. The findings indicate that machine learning serves as an effective tool for predicting optimal treatment regimens, enhancing patient satisfaction, and assisting clinicians in making personalized care decisions. This study advances the field of AI in healthcare and establishes a basis for the creation of intelligent decision support systems aimed at chronic disease management.

This work demonstrates the efficacy of data-driven medicine in improving patient outcomes, even amidst the challenges of accessing and maintaining healthcare data during conflict and in environments with limited resources.

ACKNOWLEDGMENTS

The authors would like to thank BIOCORE Research Group, Center for Advanced Computing Technology (C-ACT), Fakulti Kecerdasan Buatan dan Keselamatan Siber (FAIX) and Centre for Research and Innovation Management (CRIM), Universiti Teknikal Malaysia Melaka (UTeM) for providing the facilities and support for this research.

REFERENCES

[1] Chemkomnerd, N. *et al.* (2024) 'Evaluating Patient Satisfaction Through Validated Predictive Classification Models', *2024 16th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI)*, pp. 337–342. Available at: https://doi.org/10.1109/IIAI-AAI63651.2024.00069.

[2] Chen, T.C.T., Wu, H.C. and Chiu, M.C. (2024) 'A deep neural network with modified random forest incremental interpretation approach for diagnosing diabetes in smart healthcare', *Applied Soft Computing*, 152(December 2023). Available at: https://doi.org/10.1016/j.asoc.2023.111183.

[3] Coales, E.M. *et al.* (2022) 'Application of Machine Learning to Assess Interindividual Variability in Rapid-Acting Insulin Responses After Subcutaneous Injection in People With Type 1 Diabetes', *Canadian Journal of Diabetes*, 46(3), pp. 225-232.e2. Available at: https://doi.org/10.1016/j.jcjd.2021.09.002.

[4] Deberneh, H.M. and Kim, I. (2021) 'Prediction of type 2 diabetes based on machine learning algorithm', *International Journal of Environmental Research and Public Health*, 18(6), pp. 9–11. Available at: https://doi.org/10.3390/ijerph18063317.

[5] Deshmukh, P. V *et al.* (2025) 'Ingénierie des Systèmes d ' Information', 30(3), pp. 565–576.

[6] Esmaily, H. *et al.* (2018) 'A comparison between decision tree and random forest in determining the risk factors associated with type 2 diabetes', *Journal of Research in Health Sciences*, 18(2).

[7] Faruque, M.F., Asaduzzaman and Sarker, I.H. (2019) 'Performance Analysis of Machine Learning Techniques to Predict Diabetes Mellitus', *2nd International Conference on Electrical, Computer and Communication Engineering, ECCE 2019*, pp. 1–4. Available at: https://doi.org/10.1109/ECACE.2019.8679365.

[8] Feng, X., Cai, Y. and Xin, R. (2023) 'Optimizing diabetes classification with a machine learning-based framework', *BMC Bioinformatics*, 24(1), pp. 1–20. Available at: https://doi.org/10.1186/s12859-023-05467-x.

[9] Hussain, A.W. *et al.* (2024) 'Experimental Evaluation of Diabetes Mellitus Prediction Scheme based on Enhanced Machine Learning Strategy', pp. 1–7.

[10] Liu, N., Kumara, S. and Reich, E. (2021) 'Gaining Insights into Patient Satisfaction through Interpretable Machine Learning', *IEEE Journal of Biomedical and Health Informatics*, 25(6), pp. 2215–2226. Available at: https://doi.org/10.1109/JBHI.2020.3038194.

[11] Paul, B. and Karn, B. (2021) 'Diabetes Mellitus Prediction using Hybrid Artificial Neural Network', *2021 IEEE Bombay Section Signature Conference, IBSSC 2021* [Preprint]. Available at: https://doi.org/10.1109/IBSSC53889.2021.9673397.

[12] Sarkies, M.N. *et al.* (2015) 'Data Collection Methods in Health Services Research', *Applied Clinical Informatics*, 06(01), pp. 96–109. Available at: https://doi.org/10.4338/aci-2014-10-ra-0097.

Journal of Information Systems Engineering and Management
2025, 10(52s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

[13] Tan, K.R. *et al.* (2023) 'Evaluation of Machine Learning Methods Developed for Prediction of Diabetes Complications: A Systematic Review', *Journal of Diabetes Science and Technology*, 17(2), pp. 474–489. Available at: https://doi.org/10.1177/19322968211056917.

[14] Toharudin, T. *et al.* (2023) 'Boosting Algorithm to Handle Unbalanced Classification of PM2.5Concentration Levels by Observing Meteorological Parameters in Jakarta-Indonesia Using AdaBoost, XGBoost, CatBoost, and LightGBM', *IEEE Access*, 11(April), pp. 35680–35696. Available at: https://doi.org/10.1109/ACCESS.2023.3265019.