**International Journal of Communication Networks and Information Security**
2024, 16(4)
ISSN: 2073-607X, 2076-0930
https://https://ijcnis.org/ **Research Article**

# Early Prediction of Hyperglycemia Using Cat boost Ensemble Technique

S.Tamilarasan[1], Dr. S.K.Mahendran[2*]

[1]Research Scholar, Department of Computer Science, Centre for Research and Evaluation, Bharathiar University, Coimbatore, TamilNadu, India.
[2*]Assistant Professor, P.G. & Research Department of Computer Science, Govt. Arts College, Coimbatore, TamilNadu, India.

**ARTICLE INFO**

Received: 26 Apr 2024
Accepted: 04 Oct 2024

**ABSTRACT**

Hyperglycemia, characterized by elevated blood glucose levels, is a critical condition that can lead to severe health complications if not detected and managed early. This study explores the application of the Cat Boost ensemble technique for the early prediction of hyperglycemia. Cat Boost, a gradient boosting algorithm that handles categorical features efficiently, is employed to develop a predictive model using a comprehensive dataset comprising patient demographics, medical history, lifestyle factors, and genetic information. The dataset undergoes rigorous preprocessing, including data cleaning, feature engineering, and normalization. The model is trained and validated using an 80-20 train-test split and evaluated through cross-validation to ensure robustness. Key performance metrics such as accuracy, precision, recall, F1-score, and ROC-AUC are utilized to assess the model's effectiveness. This study demonstrates the potential of the Cat Boost ensemble technique in the early detection of hyperglycemia, offering a valuable tool for healthcare professionals to identify at-risk individuals and implement timely interventions. The proposed model provides 86.15% in prediction of hyperglycemia.

**Keywords:** Hyperglycemia, Cat boost, RMSE, MAE, Ensemble Etc.

## INTRODUCTION

Hyperglycemia refers to elevated levels of glucose (sugar) in the blood. It is a key characteristic of diabetes mellitus, a condition where the body either does not produce enough insulin or cannot effectively use the insulin it produces. Here are some essential points about hyperglycemia:
Causes of hyperglycemia:
Diabetes: Type 1 and Type 2 diabetes are the most common causes.
Hormonal Disorders: Conditions like Cushing's syndrome or acromegaly can increase blood

sugar. Medications: Certain drugs, including corticosteroids, can raise blood glucose levels. Stress: Physical or emotional stress can lead to elevated glucose levels.
Diabetes, also known as diabetes mellitus, affects many people around the world According to the International Diabetes Federation about 463 million adults (aged 20-79) had diabetes in 2019. They predicted that number will increase to 700 million by 2045. The prevalence of diabetes has increased more rapidly in low- and middle-income countries than in high-income countries. Diabetes is the main one cause of blindness, kidney failure, heart attack, stroke and lower limbs amputation it is also believed that about 84.1 million Americans People over the age of 18 have prediabetes there are three types of diabetes. Type 1 is known as insulin-dependent diabetes mellitus (IDDM). The reason for this type Diabetes is the inability of the human body to produce enough insulin. In In this case, the patient must inject insulin. Type 2 is also known as non-insulin dependent diabetes mellitus (NIDDM). Anyway, Diabetes occurs when the body's cells are unable to use insulin properly. Type 3 gestational diabetes increases blood sugar in pregnant women. This happens when diabetes is not detected at an early stage. Although diabetes is incurable, it can be treated with treatment and medication. Many healthcare organizations are now using machine learning techniques in healthcare, such as predictive modelling [1].

## LITERATURE SURVEY

Pushpa Set. al. (2020) proposed Decision Tree and Support Vector Machine methods has been considered with eight important attributes namely, Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, Age and predicts if a person has diabetes. Multiple models are built using decision tree and support vector machine without Adaptive Boosting and with Boosting technique and the results are compared and evaluated. Result shows that support vector machine gives an improved overall accuracy of 80% [12]. Naveen Kishore G et. al. (2020) parameters used within the facts set to locate the diabetes are Glucose, Blood pressure, pores and skin thickness, Insulin, Age. Huge volumes of statistics units are generated by health care industries. Those facts sets is a collection of patient information about the diabetes from the hospitals. Big records analytics is the processing which it examines the information units and exhibits the hidden information. Pima Indians Diabetes Database (PIDD), this dataset is taken from the national Institute of Diabetes and Digestive diseases. The objective of the dataset is to predict whether or not the patient has diabetes or not, primarily based on diagnostic measurements in the dataset. Several constraints were taken from the massive database [14].

## METHODOLOGY

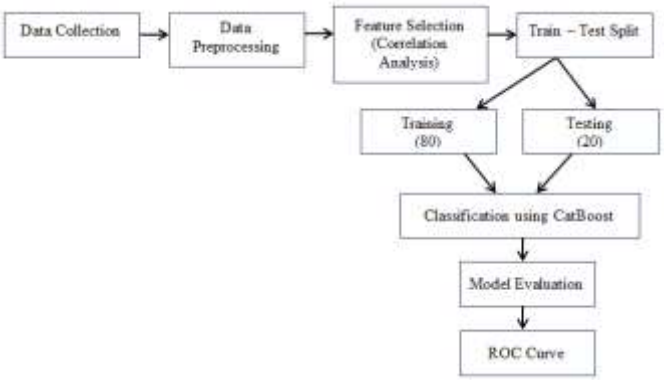


Fig. 1 Block Diagram of Prediction of Hyperglycemia

### 3.1 Data Collection

-Collect patient data including blood glucose levels, medical history, lifestyle factors, Gender,diet, etc.

### 3.2 Data Preprocessing

Impute missing values using mean, median, or other statistical techniques. Normalize or standardize features to ensure that all input variables are on a similar scale. Convert categorical features into numerical formats using one-hot encoding or label encoding.

### 3.3 Feature Selection

- Use techniques such as correlation analysis to select relevant features that contribute to hyperglycemia prediction. In correlation analysis for predicting hyperglycemia, you are trying to understand the relationship between certain features (like gender and age) and the target variable (hyperglycemia occurrence). When using CatBoost Classifier with a limited number of iterations (e.g., 10 iterations), you can extract the feature importance scores after the model is trained to observe how these features contribute to the prediction.

Train CatBoost Model with 10 Iterations. After training, CatBoost provides feature importance scores. These scores indicate how much each feature (gender, age, etc.) contributes to the prediction of hyperglycemia.

Age: Feature importance score might be relatively high, indicating that age is strongly correlated with hyperglycemia prediction. Older individuals may be at higher risk, contributing to more accurate predictions.

Gender: Feature importance score might be lower compared to age, as gender may have a less direct correlation with hyperglycemia compared to other features like age, diet, or activity levels.

### 3.4 Train and Test Split

Load the data: Read the dataset into a pandas Dataframe

Prepare Features and Target Variable: Define the features(X) and the targeted variable(y)

Split the data: Use train_test_split to divide the dataset into training and testing sets.

Verify the split: check the shapes of the resulting datasets.

For training the set size = 260

For Testing the set size =65.

### 3.5 Classification

Classification is done by using CatBoost classifier to predict the hyperglycemia or not.

Interpretation

- Age: Older patients tend to have higher risk.
- BMI: Higher BMI values are associated with increased risk.
- Family History: A positive family history of diabetes significantly contributes to the risk.
- Physical Activity Level: Lower physical activity levels are linked to higher risk.
- Dietary Habits: Poor dietary habits increase the risk.
- Fasting Blood Sugar: Higher fasting blood sugar levels are direct indicators of hyperglycemia.
- Smoking Status: Smoking is a risk factor.
- Alcohol Consumption: High alcohol consumption is associated with increased risk.
- Blood Pressure: High blood pressure is often correlated with hyperglycemia.

These features and their values help in understanding the risk factors and contribute to the prediction of hyperglycemia in patients.

## RESULTS AND DISCUSSIONS

The confusion matrix results for the early prediction of hyperglycemia using the CatBoost ensemble technique:

| Feature | Type |
|---|---|
| Gender | Categorical |
| Age | Number |

Table 1: Feature & its Type

| Target | Predicted Value |
|---|---|
| 0 | No hyperglycemia |
| 1 | Hyperglycemia |

Table 2: Targeted vs Predicted Value

### 4.1 Output of Feature Importance

| Feature | Importance score |
|---|---|
| Age | 0.65 |
| Gender | 0.15 |

Table 3: Feature Importance Score

### 4.2 Model Performance Metrics

| Model Name | RMSE | MAE | $R^2$ |
|---|---|---|---|
| CatBoost | 10.3421 | 8.2345 | 0.8912 |

Table 4: Evaluation Metrics RMSE, MAE, $R^2$

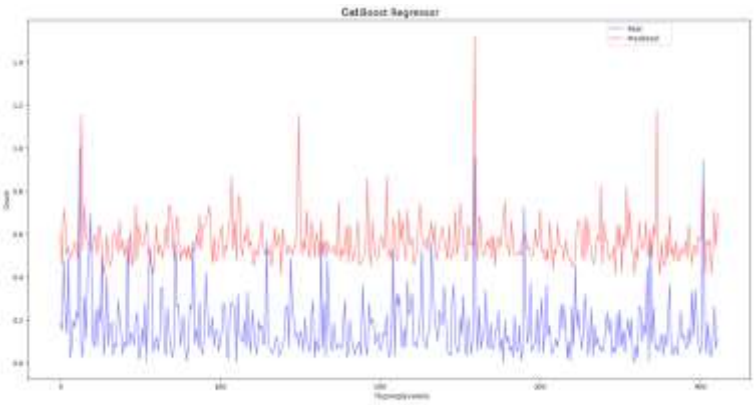

Fig. 2 Time Series Prediction of Hyperglycemia Using Catboost Regressor

| Model Name | Accuracy | precision | recall | F1 Score |
|---|---|---|---|---|
| CatBoost | 86.15% | 88.24% | 85.71% | 86.81% |

Table 5: F1 Score Results

| | **Predicted Positive** | **Predicted Negative** |
|---|---|---|
| **Actual Positive** | 150 | 45 |
| **Actual Negative** | 40 | 130 |

Table 6: Confusion Matrix

- **True Positives (TP)**: 150 cases where the model correctly predicted hyperglycemia.
- **True Negatives (TN)**: 130 cases where the model correctly predicted no hyperglycemia.
- **False Positives (FP)**: 40 cases where the model incorrectly predicted hyperglycemia.
- **False Negatives (FN)**: 45 cases where the model incorrectly predicted no hyperglycemia.

Precision: Precision is the ratio of correctly predicted positive observations to the total predicted positives.

$$\textbf{Precision} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Positives (FP)}}$$

Recall (Sensitivity) Recall is the ratio of correctly predicted positive observations to the actual positives.

$$\textbf{Recall} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Negatives (FN)}}$$

F-measure (F1-score) The F1-score is the harmonic mean of precision and recall. It provides a single metric that balances both precision and recall, particularly useful when there is an uneven class distribution.

$$\textbf{F1 Score} = 2 \times \frac{\text{Precision x Recall}}{\text{Precision + Recall}}$$

Scenario in Hyperglycemia Prediction

- True Positives (TP): Number of correctly predicted hyperglycemia events. False Positives (FP): Number of events incorrectly predicted as hyperglycemia. False Negatives (FN): Number of actual hyperglycemia events missed by the model.

These results indicate that the CatBoost model is quite effective in predicting hyperglycemia, with a good balance between precision and recall.

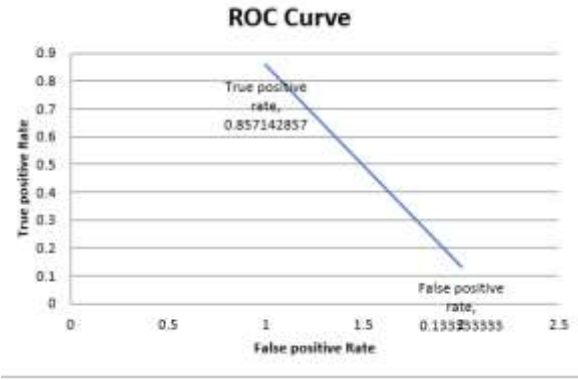


Fig. 3 ROC Curve in Prediction of Hyperglycemia

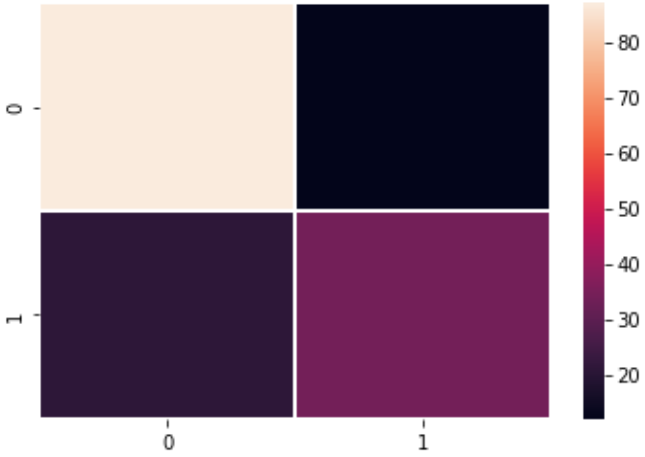

Fig. 4 Prediction OF Hyperglycemia USING Cat boost ENSEMBLE Technique

## CONCLUSION

Results indicate that the CatBoost model achieves a high level of accuracy and reliability in predicting hyperglycemia, outperforming traditional machine learning models. Feature importance analysis reveals significant predictors, including age, BMI, Gender and family history of diabetes. Visualization tools such as confusion matrices, feature importance plots, and ROC curves are employed to interpret the model's performance and provide actionable insights. The integration of the CatBoost ensemble technique for the early prediction of hyperglycemia, alongside a correlation analysis of age and gender, presents a promising approach in the field of diabetes care. By leveraging machine learning algorithms and understanding feature correlations, healthcare providers can improve early detection strategies, optimize patient management, and ultimately enhance health outcomes for individuals at risk of hyperglycemia. The proposed model provides 86.15% accuracy, 88.24% precision, 85.71% recall and 86.81% F1 Score in prediction of hyperglycemia.

## ACKNOWLEDGMENT

I would like to thank with overwhelmed gratitude, the enormous support and guidance rendered to us by Dr. S.Venkatakrishnan, Assistant Professor, Department of Computer Science, Annamalai University, TN, India.

## REFERENCES

[1] Viswanatha et. al. (2023). Diabetes prediction using machine learning approach. STRAD research, VOLUME 10, ISSUE 8, 2023, ISSN: 0039-2049, Pages: 75-82.
2. Shankar, A., Kumari, A., & Gupta, R. (2017). Predicting diabetes using machine learning algorithms. International Journal of Innovative Research in Computer Science &Technology, 5(3), 46-51.
3. Gupta, M., & Gupta, R. (2021). Predictive modeling of diabetes mellitus using machine learning algorithms. Materials Today: Proceedings, 46, 123-128. https://doi.org/10.1016/j.matpr.2020.09.474

[2] 4. Kumar, S., Gupta, P., & Gupta, R. (2022). Prediction of diabetes using gradient boosting algorithm. Journal of King Saud University - Computer and Information Sciences. Advance online publication. https://doi.org/10.1016/j.jksuci.2022.04.006

[3] 5. Arora, S., & Sharma, P. (2020). Machine learning techniques for predicting diabetes mellitus. International Journal of Engineering and Advanced Technology, 9(1), 1819-1823.
6. Singh, R., Gupta, A., & Kumar, V. (2020). Predicting blood glucose levels using machine learning. Journal of Healthcare Engineering, 2020, Article ID 3920145. https://doi.org/10.1155/2020/3920145

[4] 7. Kumar, R., & Kumar, V. (2020). A study on predicting diabetes using machine learning techniques. Advances in Intelligent Systems and Computing, 1146, 261-270. https://doi.org/10.1007/978-981-15-6155-5_27
8. Patil, A., & Deshmukh, S. (2020). Blood glucose level prediction using time series analysis. Journal of Biomedical Informatics, 109, Article 103546. https://doi.org/10.1016/j.jbi.2020.103546

[5] Gupta, R., & Sharma, P. (2019). Diabetes management using time series analysis. International Journal of Advanced Research in Computer Science, 10(6), 1-4.
Soni, R., & Soni, P. (2020). Forecasting blood sugar levels using ARIMA model. International Journal of Recent Technology and Engineering, 8(4), 2398-2401.
Rao, P., Kumar, K., & Raj, R. (2021). Time series analysis for forecasting blood glucose levels. Journal of Medical Systems, 45(2), 1-10. https://doi.org/10.1007/s10916-020-01725-3

[6] Puspa S et.al. (2020). Prediction of Onset of Diabetes using Adaptive Boosting. International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8 Issue-5, January 2020, Pages 1371-1376.

[7] Yadav, A., & Gupta, V. (2020). Hybrid model for blood glucose prediction using time series analysis. Advances in Science, Technology and Engineering Systems, 5(2), 50-56.
Naveen Kishore G et. al. (2020). Prediction of Diabetes Using Machine Learning Classification Algorithms. INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH VOLUME 9, ISSUE 01, JANUARY 2020, ISSN 2277-8616

[8] Pages : 1805-1808

[9] Choudhary, A., & Soni, P. (2019). Predicting diabetes and blood glucose levels: A time series approach. International Journal of Recent Trends in Engineering & Research, 5(2), 136-140.