dHealth 2025
M. Baumgartner et al. (Eds.)
© 2025 The Authors.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/SHTI250166

Classifying Type 2 Diabetes Using N-Glycan Profiling and Machine Learning Algorithms

Veronika GOMBAS[a], Rebeka TOROK[b], Marta VITAI[c], Laszlo KORANYI[c], Gabor JARVAS[b], Andras GUTTMAN[b,d] and Agnes VATHY-FOGARASSY[a,1]
[a] *Department of Computer Science and Systems Technology, University of Pannonia, H-8200 Veszprem, Hungary*
[b] *Research Institute of Biomolecular and Chemical Engineering, University of Pannonia, H-8200 Veszprem, Hungary*
[c] *DRC Drug Research Center Ltd., H-8230 Balatonfured, Hungary*
[d] *Horváth Csaba Memorial Laboratory of Bioseparation Sciences, Research Center for Molecular Medicine, Doctoral School of Molecular Medicine, Faculty of Medicine, University of Debrecen, H-4032 Debrecen, Hungary*

Abstract. *Background:* Type 2 diabetes (T2D) continues to present a global public health challenge due to its increasing prevalence. Early diagnosis is critical for preventing complications, but current screening methods often fail to detect early diabetic conditions. *Objectives:* This study aimed to classify T2D patients from healthy individuals using high-resolution N-glycan profiling. *Methods:* Glycan profiling was performed on serum samples from 161 individuals using capillary electrophoresis with laser-induced fluorescence detection. Different classification methods were fine-tuned using hyperparameter optimization and feature selection techniques, and their performance was comprehensively evaluated based on quality metrics. *Results:* The Extra Trees Classifier outperformed the other models with the highest median AUC, demonstrating robust accuracy (0.8982), sensitivity (0.8966), and specificity (0.9000). *Conclusion:* N-glycan profiling combined with machine learning provides a promising approach for early T2D detection. The Extra Trees Classifier showed exceptional predictive performance, warranting further investigation with larger datasets to validate its clinical applicability.

Keywords. Classification, Type 2 Diabetes, N-glycan, Machine learning, Hyperparameter optimization

1. Introduction

Type 2 diabetes (T2D) is a form of diabetes mellitus and as of 2025, it is still a global public health burden since its incidences are continuously rising. The current screening protocol for T2D is simple, patients should be tested for blood sugar levels at three-year intervals, especially people who are obese. Screening of apparently healthy individuals may lead to early detection and the associated treatment that could prevent or delay the development of related complications [1]. Next to the overnight/two-hour fasting blood

[1] Corresponding Author: Agnes Vathy-Fogarassy, Department of Computer Science and Systems Technology, University of Pannonia, Veszprem, Hungary, E-Mail: vathy.agnes@mik.uni-pannon.hu

glucose level check and oral glucose tolerance test, glycated hemoglobin A1c level is an additional test that has been frequently used to screen and diagnose T2D. The performance of current screening methods may significantly differ among populations [2, 3], and not suitable for early (i.e., well before) detection of diabetic and pre-diabetic conditions.

N-linked glycosylation is a post-translational modification of proteins and its alteration is reportedly a reliable biomarker for various diseases including cancer metastasis, chronic inflammatory diseases, and viral pathogenesis [4, 5]. Previous studies also demonstrated that N-glycosylation changes in T2D can be used to distinguish from healthy controls, based on the N-glycome profile [6-8]. Several liquid phase separation-based analytical methods are available for structural elucidation of carbohydrates, among which capillary electrophoreses (CE) had one of the highest resolutions. Current CE-supported glycan structure identification utilizes manual data interpretation and the reported workflows have limited support from automated computational platforms. GUcal is a pioneering tool for database-assisted, high-precision glycan structure identification [9], but it has some limitations, i.e., not featuring direct data mining options or artificial intelligence (AI) supported data evaluation. However, the ultimate complexity of glycans [10] motivates both academic and industrial research communities to involve AI-based data analysis in comprehensive glycomics workflows.

This study adapted and tested numerous classifier methods for differentiating diabetic and healthy patients using high-resolution N-glycomics data obtained by capillary electrophoresis coupled with laser-induced fluorescence detection. The classification methods included in the study had different principles of operation, providing us with a wide range of analyses. Five classification models were developed and applied: Logistic Regression, Support Vector Machine, Random Forest, Extra Trees Classifier, and Light Gradient Boosting Logistic Regression [11] predicts binary class labels by modeling the relationship between the attributes and the probability of the outcomes using a linear combination of the attributes, transformed by the logistic function. The Support Vector Machine SVM [12] classifier aims to find a hyperplane in a high-dimensional space that best separates data points into different classes. The optimal hyperplane provides a robust decision boundary, making SVM suitable for various classification tasks, especially when dealing with complex relationships in the data. The Random Forest (RF) [13], Extra Trees Classifier [14], and Light Gradient Boosting (LightGBM) [15] classifiers are tree-based ensemble techniques in which classification is determined by majority voting or, in the case of LightGBM, by an optimized boosting mechanism. The choice of ensemble models was based on the fact that ensemble methods reduce the overfitting of simple decision trees and provide high predictive accuracy. While the operation of Extra Trees Classifiers is entirely stochastic, the trees are built in a completely random manner; the Random Forest algorithm uses the bootstrap aggregating technique. That means it creates different training sets using replacement sampling, and the algorithm builds diverse decision tree models on slightly different subsets of the data. Both classifiers build trees independently; in contrast, the LightGBM classifier uses a boosting technique, and each created tree aims to minimize the error of the loss function of the previously built tree.

2. Methods

2.1. Dataset Creation

The sample preparation started with collecting samples from diabetic and healthy patients. The samples underwent pretreatment following a comprehensive protocol described in [6], including denaturation, glycan release, fluorophore labeling, and magnetic bead-mediated cleanup. The labeled samples were analyzed by capillary electrophoresis with laser-induced fluorescence detection (CE-LIF) using parameters similar to those in the publication [6]. The relative peak profiles from each serum sample, applied in triplicates, were averaged before the data analysis.

The available sample size for the analysis was limited, consisting of 161 samples from 85 diabetic and 76 healthy patients. The preprocessed dataset included the anonymized patient identification code, the relative peak intensities of samples (35 attributes), and information indicating the presence or absence of diabetes. The binary classification task was defined based on the information about health status. The class labels of diabetic patients were marked 1, and the class labels of healthy people were marked 0. The distribution of the class labels was more or less balanced in the available dataset; approximately 52.7% of them represented diabetes patients, and around 47.2% represented healthy patients. The dataset used for classification was split into a training set and a test set in the ratio of 80%-20%.

2.2. Preprocessing of Data and Classification Models

The effectiveness of the classifications is affected by information stored in the attributes of the dataset. In some cases, the attributes contain important information for the classifications, while in other cases, the attributes present non-relevant information or noise from the perspective of the problem to be solved. The permutation importance [16] method was executed iteratively and separately for each classifier to identify the most relevant N-glycan structures on the dataset. The permutation importance method measures the impact of attributes on the predictive performance of the model by permutating their values and re-evaluating the model on the altered dataset to assess the effect of the changes. The permutation importance method was applied 50 times to each classifier using random train-test splits, and the resulting importance scores were averaged. Each classification model was run on the attributes having high importance scores and their union. As more accurate classification models can be built by removing attributes that are not relevant to the classification task, a reduced dataset was determined based on the quality metrics of the resulting models, containing only the relevant attributes (N-glycans) for classification.

Hyperparameter optimization was performed on the reduced dataset for each classifier. First, the search space of the hyperparameters to be tuned was determined, and then all classifiers were tuned using the GridSearch method [17] for each classifier. Considering the limited amount of data, all optimization was executed ten times on the randomly split reduced dataset, with disjoint training and validation sets created each time and with 5-fold cross-validation. Finally, each fine-tuned classifier was run 1000 times on the reduced dataset, and the performances of the resulting models were calculated as the average results of one thousand runs based on random train-test splits in a ratio of 80%-20%. To evaluate the performance of the classification models, below quality metrics were applied: accuracy, sensitivity, specificity, F1-score, and AUC value.

3. Results

The result of the feature selection process is presented in Figure 1. It can be observed that various N-glycan structures are relevant to different classifiers; however, some similarities are also noticeable within the types of classifiers. Logistic Regression and SVC are marked blue. The average values of permutation importance for these models are the highest for the same eight N-glycan: g33, g19, g13, g16, g29, g32, g20, and g2, but for Logistic Regression additional five attributes appeared important. In the case of ensemble models (marked green), similar results can be observed. The most important feature of the ensemble model is the N-glycan g10, but the relevance of the other attributes has identical trends in each case. Based on the quality metrics of the classifier models, only one new, reduced dataset was generated containing all relevant N-glycan structures (g1, g2, g3, g6, g10, g13, g14, g15, g16, g18, g19, g20, g21, g29, g32, g33).

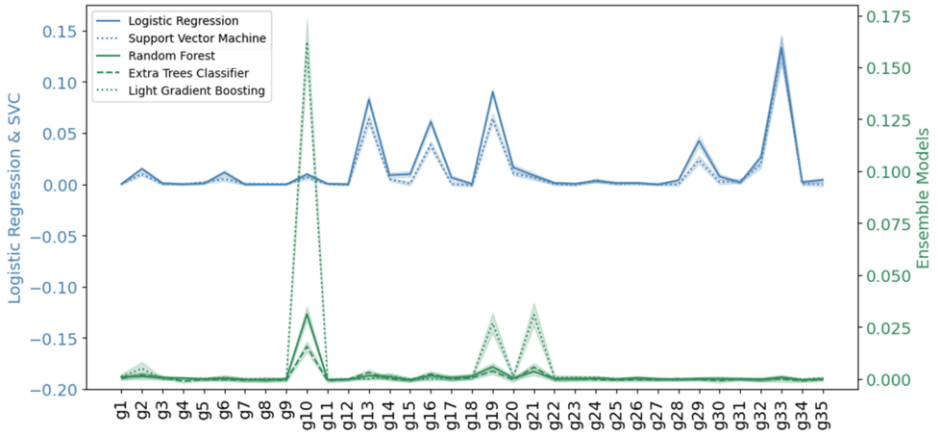

Figure 1. Multi-axis line chart showing the average permutation importance results by classification models.

Following this, the hyperparameter optimization was performed on the reduced dataset. The best parameter and hyperparameter values determined from the fine-tuning procedures are summarized in Table 1.

Table 1. The most relevant parameter and hyperparameter values of the fine-tuned classification models.

Classifier	Parameters / Hyperparameters
Logistic Regression	C=100, max_iter=500, penalty='l1', solver='saga'
Support Vector Machine	C=10, degree=3,gamma='scale', kernel='rbf'
Random Forest	bootstrap=True, max_depth=10, min_samples_leaf=1, min_samples_split=2, n_estimators=100
Extra Trees Classifier	max_depth=50, min_samples_leaf=1, min_samples_split=2, n_estimators=100
Light Gradient Boosting	boosting_type='dart', learning_rate=0.1, max_depth=-1, n_estimators=50, num_leaves=31

A comprehensive overview of the results of all fine-tuned classifiers is presented in Table 2 and Figure 2. As shown in Table 2, each classifier demonstrated high values on all metrics across all classes; the lowest value is 0.8480. Logistic Regression as a linear model provides the weakest results; however, the average metrics are around 0.85. The Extra Trees Classifier has the highest quality values in diabetes prediction.

Table 2. The results of the fine-tuned classifiers on reduced datasets.

Classifier	Accuracy	AUC	F1	Sensitivity	Specificity
Logistic Regression	0.8589	0.8592	0.8582	0.8480	0.8705
Support Vector Machine	0.8851	0.8856	0.8847	0.8685	0.9028
Random Forest	0.8799	0.8797	0.8794	0.8848	0.8746
Extra Trees Classifier	0.8982	0.8983	0.8979	0.8966	0.9000
Light Gradient Boosting	0.8626	0.8622	0.8619	0.8753	0.8491

Boxplots in Figure 2 compare the AUC values of the five classifiers across 1000 runs. Each boxplot presents the distribution of AUC scores for one model, providing insight into their performance consistency and overall effectiveness. The Extra Trees classifier achieved the highest median of AUC scores. However, Random Forest and Support Vector Machine with radial basis function kernel also indicate reliable performance across different random train-test splits. All models exhibit relatively compact interquartile ranges, suggesting lower variability in their results.

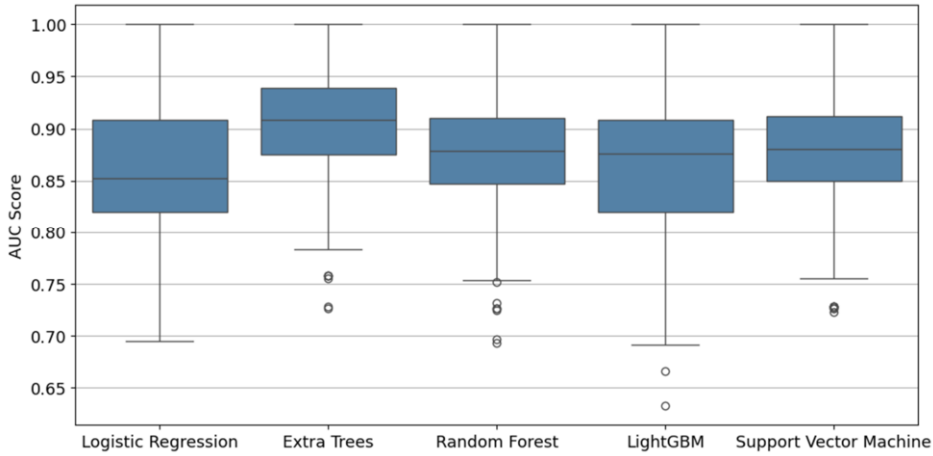

Figure 2. Boxplots presenting the distribution of AUC scores of the five classifiers across 1000 runs.

4. Discussion

Our research focused the applicability of different machine learning algorithms for early identification of type 2 diabetes based on N-glycan profiles of patients. N-glycan profiles were created by capillary electrophoresis with laser-induced fluorescence detection. Feature selection was applied to extract relevant information from the N-glycan profiles, and the effectiveness of five different classifiers (Logistic Regression, Support Vector Machine, Random Forest, Extra Trees Classifier, Light Gradient Boosting) were evaluated on the reduced datasets.

It was observed that the Extra Trees Classifier identified type 2 diabetes based on N-glycan profiles with the highest accuracy. All metrics for the Extra Trees Classifier were close to exceeding 0.9, indicating its potential for further research opportunities. It is important to note that the examined dataset was small from a data analysis perspective, and further samples and investigations are necessary to establish a stable and reliable method and results. Our future goal is to continue to refine the results based on additional blood samples and analyze the role of the different N-glycan structures in the disease.

Acknowledgement

This work was supported by the 2024-2.1.1-EKÖP-2024-00025 University Research Fellowship Program and by the 2020-1.1.2-PIACI-KFI-2020-00045 ('Collaborative Hospital Information Platform,' Development of a process-oriented medical system to support inpatient specialized care) project with the support provided by the Ministry of Culture and Innovation of Hungary from the National Research, Development and Innovation Fund. The support of the University of Debrecen Program for Scientific Publications is also acknowledged. This is contribution #220 of the Horváth Csaba Memorial Laboratory of Bioseparation Sciences. Authors also gratefully acknowledge the support from the following sources: ATBG Korea V4 joint project of the National Research, Development and Innovation Office of Hungary #2023-1.2.1-ERA_NET-2023-00015, the Andras Koranyi Foundation and the Cooperative Doctoral Program of the Ministry of Culture and Innovation. This work was also supported by the #150780 grant from National Research, Development and Innovation Office of Hungary.

Institutional Review Board Statement

The study was conducted according to the guidelines of the Declaration of Helsinki, and approved by the Ethics Committee of Medical Research Council of Ministry of Human Resources (protocol code 22051-3/2013/EKU (278/2013) approved on 27 August 2013).

References

[1] Peer, N., Y. Balakrishna, and S. Durao, Screening for type 2 diabetes mellitus, *Cochrane Database Syst Rev*, **5**(5) (2020), CD005266.
[2] Şahin, M., et al., Comparison of the effectiveness of screening methods for the diagnosis of gestational diabetes mellitus in pregnant women: A cross-sectional study, *International Journal of Clinical Practice*, **75**(11) (2021), e14857.
[3] Ishikawa, Y., et al., Comparison of diagnostic screening methods for diabetes in patients with heart failure, *Diabetes Epidemiology and Management*, **9 (**2023), 100109.
[4] Reily, C., et al., Glycosylation in health and disease. *Nature Reviews Nephrology*, **15**(6) (2019) 346-366.
[5] Esmail, S. and M.F. Manolson, Advances in understanding N-glycosylation structure, function, and regulation in health and disease, *Eur J Cell Biol*, **100**(7-8) (2021), 151186.
[6] Torok, R., et al. N-Glycosylation Profiling of Human Blood in Type 2 Diabetes by Capillary Electrophoresis: A Preliminary Study, *Molecules*, **26**(21) **(**2021), 6399.
[7] Štambuk, T. and O. Gornik, Protein Glycosylation in Diabetes, *Adv Exp Med Biol*, **1325 (**2021), 285-305.
[8] Keser, T., et al., Correction to: Increased plasma N-glycome complexity is associated with higher risk of type 2 diabetes, *Diabetologia*, **61**(2) (2018), 506.
[9] Jarvas, G., et al., Expanding the capillary electrophoresis-based glucose unit database of the GUcal app, *Glycobiology*, **30**(6) (2020), 362-364.
[10] Varki, A., Biological roles of glycans, *Glycobiology*, **27**(1) (2017), 3-49.
[11] David W. Hosmer Jr., Stanley Lemeshow, Rodney X. Sturdivant: *Applied Logistic Regression*, John Wiley & Sons, Hoboken New Jersey, *2013.*
[12] Hearst, M.A., et al., Support vector machines. *Intelligent Systems and their Applications,* IEEE, **13**(4) (1998), 18–28.
[13] Breiman L., Random Forests, *Machine learning* **45** (2001), 5–32.
[14] P. Geurts, D. Ernst., L. Wehenkel, Extremely randomized trees, *Machine Learning*, **63**(1) (2006), 3-42.
[15] Guolin Ke, et al., LightGBM: A Highly Efficient Gradient Boosting Decision Tree, A*dvances in Neural Information Processing Systems 30, (2017)* , 3149-3157.
[16] Christoph Molnar, Interpretable Machine Learning: A Guide for Making Black Box Models Explainable, https://christophm.github.io/interpretable-ml-book/ , last access: 15.01.*2025.*
[17] Hutter, Frank, Lars Kotthoff, and Joaquin Vanschoren. *Automated machine learning: methods, systems, challenges*. Springer Nature, Barcelona, Spain, 2019