Application of UV-VIS Spectroscopy and Machine Learning Methods in Glucosuria Diagnostics: A Phantom Study

Maria Babińska,* Adam Władziński

Department of Metrology and Optoelectronics, Faculty of Electronics, Telecommunications and Informatics, Gdańsk University of Technology, Gdańsk, Poland

Received February 13; accepted March 26, 2025; published March 31, 2025

Abstract— In this study, UV-VIS spectroscopy was used as a tool for detecting low glucose concentrations in urine. Measurements were performed on artificial urine samples and solutions with 0.1% and 0.2% glucose, covering both normal and pathological thresholds. Among the evaluated models, Random Forest reached 0.887 for the 0.1% glucose sample, while Logistic Regression achieved 0.7796 for the 0.2% glucose sample, demonstrating high effectiveness in distinguishing glucose levels. The results confirm that the integration of UV-VIS spectroscopy and machine learning has the potential to serve as a fast and non-invasive screening tool for the early detection of metabolic disorders.

Early detection of glucose in urine, known as glucosuria, is essential for diagnosing diabetes and other metabolic disorders[1]. Traditional methods for detecting glucose in biological samples often require complex instrumentation and expensive reagents. UV-VIS spectroscopy offers a rapid and non-invasive alternative for determining glucose concentrations [2–5]. In this exploratory study, we investigate the feasibility of using UV-VIS spectroscopy combined with machine learning techniques to detect and classify low glucose concentrations in artificial urine. This study presents an analysis of artificial urine samples with different glucose concentrations ranging from trace levels (0.1%, 0.2%) to higher values (up to 10%) and the application of a machine learning model to classify glucose levels based on the obtained spectra [6-8]. Our primary focus is on evaluating the detectability of the lowest concentrations, which are the most relevant for early-stage diagnosis.

In this study, artificial urine was prepared with glucose concentrations of 0%, 0.1% (100 mg/dL), 0.2% (200 mg/dL). The renal threshold for glucose is approximately 180 mg/dL. However, urinary glucose concentrations exceeding 25 mg/dL may already be concerning [1,9]. The solution was made using four easily accessible and safe ingredients: black tea, Phosphate-Buffered Saline (PBS), citric acid, and aspirin. The samples were then analyzed using a spectrophotometer in the ultraviolet and visible light range from 200 to 800 nm. Fig. 1 presents a schematic of the ingredients used for sample preparation as well as the measurement setup.

E-mail: marbabi1@pg.edu.pl

Fig. 1. Schematic of the ingredients used for sample preparation and the measurement setup.

The UV-VIS measurements were performed using the NanoDrop One device manufactured by Thermo Fisher Scientific in the USA. The sample volume was 2 µL, measurements were conducted in transmission mode, each measurement had an integration time of approximately 8 seconds, the wavelength range was 190–850 nm with a wavelength accuracy of ±1 nm, and the pathlength was automatically adjusted between 0.03 mm and 1 mm [10]. Each sample underwent 50 repeated measurements to ensure accuracy and reproducibility. Absorbance spectra were recorded, and characteristic bands associated with glucose presence were analyzed. The artificial urine exhibited spectral properties in the UV-VIS range similar to those of natural urine [2], allowing for its analysis as a viable substitute. Additionally, Fig. 2 compares the UV-VIS spectra of human urine with those of the artificial urine used in this study.

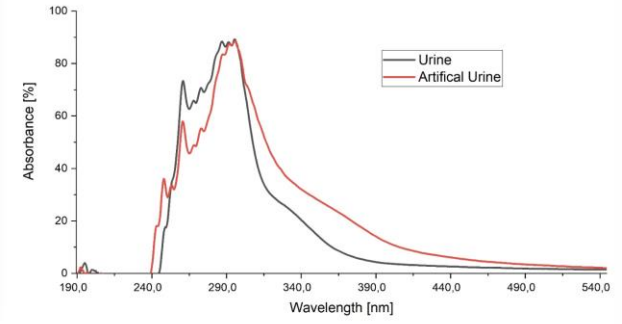

Fig. 2. Comparison of the UV-VIS spectra of human urine with those of the artificial urine used in this study.

The UV-VIS spectra data were subjected to preprocessing, including parsing a tab-separated values (TSV) file containing metadata on measurements and experimental parameters. Each sample was assigned a unique identifier, which was then mapped to the corresponding glucose concentration class label. Subsequently, validation and statistical analysis were conducted to assess data quality and integrity. The study involved visualizing representative absorption spectra for different glucose concentration classes, calculating basic descriptive statistics, and generating box plots to evaluate absorbance distribution. Principal Component Analysis (PCA) was performed to determine whether dimensionality reduction could preserve key diagnostic information [11].

The identification of biomarker thresholds in small datasets is a well-established feature exploration technique, particularly when utilizing Support Vector Machines (SVM) and Random Forest algorithms [12–14].

The dataset was preprocessed by standardizing spectral values and encoding glucose concentrations as categorical labels. Principal Component Analysis (PCA) was employed to assess whether dimensionality reduction could retain key diagnostic information. The optimal number of principal components (PCs) required to explain 95% of the variance was eight, indicating that spectral features could be effectively reduced while preserving essential classification information. A diagram of the management and data processing of the best-performing Random Forest algorithm is shown in Fig. 3.

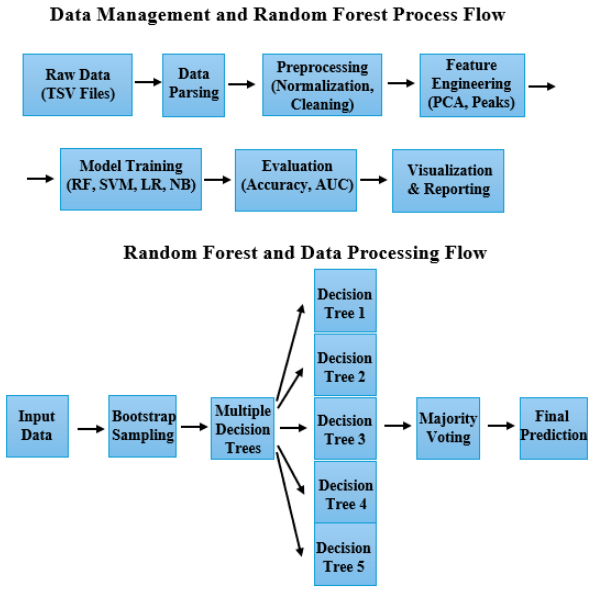

Fig. 3. A diagram of the management and data processing of the Random Forest algorithm.

Machine learning models were trained to classify glucose concentrations in artificial urine samples. The evaluated models included Support Vector Machine (SVM), Logistic Regression (LR), Naïve Bayes (NB), and Random Forest (RF). The classification accuracy for each model is summarized in Table 1.

Table 1. Classification Accuracy of Different Machine Learning Models.

Model	Accuracy
Support Vector Machine	0.4638
Logistic Regression	0.4203
Naive Bayes	0.4493
Random Forest	0.4928

Among the evaluated models, Random Forest achieved the highest classification accuracy of 49.28%, followed by Naïve Bayes. When exploring the dataset with multiple algorithms, earlier conclusions for disease prediction were confirmed [15]. The confusion matrix for the best-performing model (Random Forest) is shown in Fig. 4, illustrating its capability in classifying glucose concentrations. The average number of repetitions for the algorithms was 1000 trials.

The ability to differentiate between standard (0%), borderline (0.1%), and pathological (0.2%) glucose concentrations is critical for early detection of metabolic disorders. Glucosuria is defined as a glucose concentration of 0.2% or higher, while 0.1% glucose in urine is considered an alarming sign requiring further medical evaluation [1].

Area Under Curve (ACU) scores were computed for 0.1% and 0.2% glucose compared to pure synthetic urine to assess the detectability of these critical glucose concentrations. The results are summarized in Table 2.

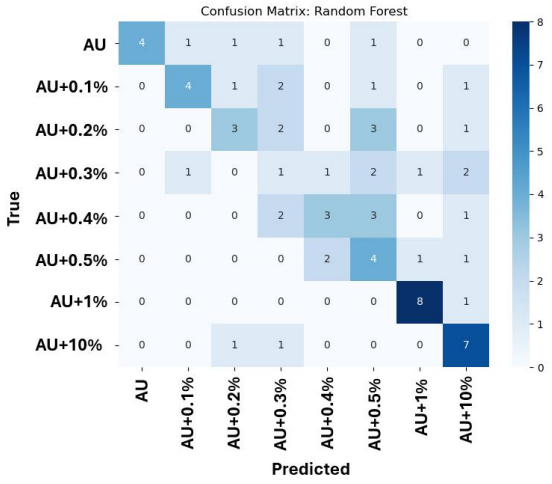

Fig. 4. Confusion matrix for Random Forest classification of glucose concentrations (0.1%–10%) in artificial urine (AU). The extended concentration range was used for exploratory evaluation of classification performance, with a focus on early detection. AU: artificial urine; AU+X%: artificial urine with X% glucose.

Table 2. Detection AUC Scores for Low Glucose Concentrations.

	Support Vector Machine	Logistic Regression	Naive Bayes	Random Forest
0.1% Glucose	0.7593	0.8278	0.7481	0.8870
0.2% Glucose	0.7407	0.7796	0.7500	0.7648

These results indicate that Random Forest provides the highest detection capability for 0.1% glucose (AUC = 0.8870), while Logistic Regression performs slightly better for 0.2% glucose (AUC = 0.7796). This demonstrates that the methodology can be utilized for early identification of individuals at risk of glycosuria by detecting glucose levels exceeding 0.1%.

The ability to distinguish between pure urine and samples containing 0.1% and 0.2% glucose concentrations may offer a non-invasive, rapid screening tool for detecting metabolic imbalances. This approach could be relevant for diabetes screening programs, where early intervention is crucial to prevent severe complications. Glucosuria, defined as the presence of glucose in urine, typically occurs when blood glucose levels exceed the renal threshold of approximately 180 mg/dL. Detecting glucose concentrations above 0.25 mg/mL in urine may indicate an underlying health issue. The current study was designed as an initial exploratory assessment of UV-VIS spectroscopy combined with machine learning techniques for low-concentration glucose detection in artificial urine samples. While integrating these methods shows promise for non-invasive glucose monitoring, the approach remains in the developmental phase. Therefore, further optimization of both the spectral acquisition process and model training is necessary to enhance its diagnostic accuracy and reliability for clinical applications.

This research was supported by the Ministry of Education and Science under project NdS-II/SP/0438/2024/01 at Gdansk Medical University, by the DS programs of the Faculty of Electronics, Telecommunications and Informatics at Gdansk University of Technology, by grant 14/1/2024/IDUB/III.4.1/Tc and 7/1/2024/IDUB/III.4c/Tc under the TECHNETIUM Talent Management Grants and by COST Action [CA21159].

References

[1] S.L. Cowart, M.E. Stachura, *Clinical Methods: The History, Physical, and Laboratory Examinations*, 3rd edition, chapter 139 (1990).
[2] P. Sokołowski, P. Wityk, K. Cierpiak, M. Babińska, W. Graczyk, B. Krawczyk, M. Markuszewski, M. Szczerska, Optical Sensing and Detection VIII, SPIE, 449 (2024).
[3] T.-T. Wang, K. Guo, X.-M .Hu, J. Liang, X.-D. Li, Z.-F. Zhang, J. Xie, Chemosensors **8**(1), 10 (2020).
[4] P. Sokołowski, K. Cierpiak, M. Szczerska, M. Wróbel, A. Łuczkiewicz, S. Fudala-Księżek, P. Wityk, J. Biophotonics, e202300523 (2024).
[5] M. Jędrzejewska-Szczerska, M. Gnyba, M. Sobaszek, E. Krystian, Sensors Actuators A: Physical **202**, 8 (2013).
[6] P. Wityk, P. Sokołowksi, M. Szczerska, K. Cierpiak, B. Krawczyk, M.J. Markuszewski, J. Biophotonics **16**(9), p. e202300095 (2013).
[7] R.T. Yunardi, R. Apsari, M. Yasin, J. Electronics Electromedical Engineering and Medical Informatics **2**(2), 33 (2020).
[8] A. Althnian, D. AlSaeed, H. Al-Baity, A. Samha, A.B. Dris, N. Alzakari, A. Abou Elwafa, H. Kurdi, App. Sciences **11**(2), 2, (2021).
[9] S. Walford, M. McB Page, S. P. Allison, Diabetes Care **3**(9), 672 (1980).
[10] Thermo Fisher Scientific, *NanoDrop One Microvolume UV-Vis Spectrophotometers Product Specifications*.
[11] A.M.C. Davies, T. Fearn, *Back to basics: the principles of principal component analysis*, Spectroscopy Europe/World, Dec. 2004.
[12] N. Sancar, S.S. Tabrizi, BMC Medical Informatics and Decision Making **23**: 219, 1 (2023).
[13] T. Pranckevičius, V. Marcinkevičius, Baltic J. Modern Computing 5(2), 221 (2017).
[14] M. Marzejon, M. Kosowska, D. Majchrowicz, B. Buło-Piontecka, M. Wąsowicz, M. Jędrzejewska-Szczerska, J. Biophotonics **12**(4), e201800273 (2018).
[15] S. Uddin, A. Khan, M. Hossain, M. Ali Moni , BMC Medical Informatics and Decision Making **19**, 281 (2019).