

УДК 004.89:616.379-008.64

DOI [https://doi.org/10.24144/2616-7700.2025.46\(1\).226-233](https://doi.org/10.24144/2616-7700.2025.46(1).226-233)**М. М. Мальяр¹, Н. Е. Кондрук², Є. Б. Кондрук³, В. А. Нерода⁴**

¹ ДВНЗ «Ужгородський національний університет»,
професор кафедри кібернетики і прикладної математики,
доктор технічних наук
mykola.malyar@uzhnu.edu.ua
ORCID: <https://orcid.org/0000-0002-2544-1959>

² ДВНЗ «Ужгородський національний університет»,
доцентка кафедри кібернетики і прикладної математики,
кандидат технічних наук
natalia.kondruk@uzhnu.edu.ua
ORCID: <https://orcid.org/0000-0002-9277-5131>

³ ДВНЗ «Ужгородський національний університет»,
аспірант кафедри кібернетики і прикладної математики
yevhen.kondruk@uzhnu.edu.ua
ORCID: <https://orcid.org/0009-0002-8555-3351>

⁴ ДВНЗ «Ужгородський національний університет»,
аспірант кафедри кібернетики і прикладної математики
vladyslav.neroda@uzhnu.edu.ua
ORCID: <https://orcid.org/0009-0009-2037-6631>

РОЗРОБКА ІНТЕЛЕКТУАЛЬНОЇ СИСТЕМИ ПРИЙНЯТТЯ РІШЕНЬ ДЛЯ ДІАГНОСТУВАННЯ ДІАБЕТУ

У статті представлено результати розробки інтелектуальної системи підтримки прийняття рішень для діагностики цукрового діабету з використанням методів машинного навчання. На основі відкритого медичного датасету було побудовано та проаналізовано низку класифікаційних моделей, зокрема логістичну регресію, Random Forest та XGBoost. З метою підвищення точності було застосовано метод стратифікації при виділенні валідаційних даних, масштабування ознак, крос-валідацію та оптимізацію гіперпараметрів моделей. Особливу увагу приділено аналізу матриць невідповідності та оцінці впливу ключових ознак. Найкраща кросвалідаційна точність досягнута моделлю Random Forest (94.43%), що свідчить про її здатність добре узагальнювати закономірності у незбалансованих даних. Найнижча кількість помилок другого роду ($FN = 16$) спостерігається в моделі XGBoost. Запропонований підхід може бути ефективно використаний для попереднього діагностування діабету.

Ключові слова: система прийняття рішень, класифікація, діагностування діабету, інтелектуальна система, машинне навчання.

1. Вступ. Проблема захворювання на діабет є однією з найбільш актуальних у сучасних умовах, оскільки вона має не тільки медичні, але й суттєві соціо-економічні наслідки. Зростання кількості випадків діабету призводить до збільшення витрат на медичне обслуговування, зниження продуктивності праці та погіршення якості життя населення.

Раннє виявлення ризику розвитку діабету є важливим завданням для системи охорони здоров'я, оскільки дозволяє своєчасно запроваджувати профілактичні заходи і ефективно розподіляти ресурси.

Одним із перспективних підходів до вирішення цієї задачі є розробка систем підтримки прийняття рішень (СППР), які базуються на методах прогнозування

та класифікації. Використання таких систем дозволяє автоматизувати процес діагностики, підвищити його точність та обґрунтованість.

Мета даної роботи полягає у розробці системи прийняття рішень для діагностування діабету з використанням методів класифікації. У рамках дослідження проведено аналіз ефективності різних класифікаторів, що дозволить обґрунтувати вибір найбільш оптимального підходу для практичного використання в соціо-економічних дослідженнях та медичних інформаційних системах.

2. Огляд літератури. Проблема розробки систем прийняття рішень для діагностики діабету набуває значної уваги в сучасних дослідженнях, зокрема завдяки застосуванню методів машинного навчання та штучного інтелекту. В останні роки багато авторів досліджували різні підходи до підвищення точності прогнозування та автоматизації аналізу медичних даних.

Одним із перспективних напрямків є застосування глибоких нейронних мереж для аналізу медичних записів та автоматичного діагностування діабету. У дослідженні [1] автори продемонстрували ефективність використання глибоких нейронних мереж у моделюванні та класифікації медичних даних. Вони порівняли різні архітектури мереж та показали, що комбіновані моделі можуть значно покращити точність діагностики.

Машинне навчання широко застосовується для покращення точності прогнозування діабету. Автори у праці [2] розглядають ключові алгоритми, що використовуються в сучасних клінічних системах підтримки прийняття рішень. Автори аналізують різні моделі машинного навчання, їх переваги та недоліки, а також шляхи покращення продуктивності за рахунок вдосконалення методів обробки даних.

У дослідженні [3] розглядається застосування методів машинного навчання для прогнозування діабету, аналізуються різні алгоритми, включаючи моделі прихованих Марковських процесів (Hidden Markov Models, HMM), які використовуються для виявлення закономірностей у клінічних даних.

Порівняння квантового машинного навчання (QML) та глибокого навчання (DL) для прогнозування діабету на основі набору даних «PIMA Indian Diabetes» розглянуто у [4]. Автори аналізують їх продуктивність. Дослідження також підкреслює значення нормалізації даних і усунення аномалій для покращення діагностичних результатів.

3. Постановка проблеми. Для досягнення поставленої мети необхідно вирішити наступні завдання:

1. Підібрати набір даних про захворювання діабету та провести попередню обробку даних.
2. Побудувати та налаштувати моделі машинного навчання (логістична регресія, Random Forest, XGBoost) із налаштуванням гіперпараметрів.
3. Здійснити оцінку ефективності моделей за допомогою крос-валідації та побудови матриць невідповідності.
4. Дослідити вплив моделей на співвідношення помилок першого та другого роду.
5. Визначити найбільш інформативні ознаки для діагностики діабету на основі аналізу важливості у моделях Random Forest та XGBoost.

4. Методи дослідження. Серед найефективніших алгоритмів для вирішення задач класифікацій із незбалансованими класами є логістична регресія,

метод випадкового лісу та XGBoost.

Логістична регресія є одним із класичних статистичних методів, який добре працює у випадках, коли залежність між ознаками та цільовою змінною має лінійний характер. Вона використовується для моделювання ймовірності належності до певного класу та часто застосовується у медичних прогнозах. Дослідження останніх років показують, що дана модель залишається ефективною при аналізі великих медичних наборів даних, хоча її точність може бути нижчою порівняно з більш складними моделями машинного навчання. Зокрема, у дослідженні [5-7] розглядаються основні аспекти застосування логістичної регресії у прогнозуванні медичних станів та оцінку ефективності моделі у діагностиці захворювань.

Random Forest Classifier є ансамблевим методом, який використовує сукупність дерев рішень для підвищення точності класифікації та регресії. Його основна ідея полягає у побудові ряду дерев, кожне з яких навчається на випадковій підмножині даних, а потім агрегує їхні результати для отримання остаточного прогнозу. Завдяки цьому підходу алгоритм демонструє високу точність, стійкість до перенавчання та здатність ефективно працювати з великими наборами даних. Дослідження останніх років підтверджують ефективність даної моделі у медичних прогнозах. У [8] аналізуються різні підходи до вибору ознак у моделі Random Forest, що є важливим аспектом для підвищення точності прогнозування. Стаття [9] аналізує застосування Random Forest Classifier для прогнозування діабету.

XGBoost є одним із найпотужніших алгоритмів градієнтного бустингу, який широко використовується у задачах прогнозування. Його основна ідея полягає у побудові ансамблю дерев рішень, які поступово коригують помилки попередніх моделей, що дозволяє досягти високої точності прогнозування. Завдяки ефективному використанню пам'яті та оптимізованим обчисленням XGBoost демонструє високу продуктивність навіть на великих наборах даних. У [10] аналізуються різні варіанти градієнтного бустингу, а в [11] досліджується його застосування для класифікації незбалансованих даних. Автори аналізують проблему нерівномірного розподілу класів у наборах даних, що часто зустрічається у медичних діагностичних задачах, інформаційній безпеці та текстовому аналізі.

5. Експерименти. У межах експерименту було реалізовано інформаційно-аналітичну систему для діагностики цукрового діабету з використанням методів машинного навчання. Робота починалась із попереднього аналізу та обробки медичного датасету [12], який було зібрано відповідно до показників та стандартів ВООЗ, та містив 17 ознак, включно з демографічними, клінічними, біохімічними та поведінковими характеристиками 4303 пацієнтів. Цільова ознака відображає наявність або відсутність діабету. Аналіз частот показав наявність дисбалансу класів, що вимагало спеціальних налаштувань алгоритмів.

Масштабування числових ознак на основі стандартного нормалізатора (StandardScaler) дозволило вирівняти шкали показників і уникнути переважного впливу ознак з великими значеннями.

Після масштабування дані було розділено на тренувальну та тестову вибірки у співвідношенні 80/20 методом стратифікації. Це дозволило забезпечити об'єктивну оцінку якості моделей на прихованих даних. Було розглянуто три ал-

горитми класифікаторів: логістична регресія, Random Forest та XGBoost який розглядався на стиснених методом головних компонент та нестиснених даних. Для кожної з моделей було проведено тюнінг гіперпараметрів з використанням перехресної перевірки (GridSearchCV з 5-фолд крос-валідацією), а також досліджено вплив таких параметрів як регуляризація, кількість дерев, глибина дерева та ваги класів.

Для оцінки якості моделей використовувалась середня точність крос-валідації, а також матриці невідповідності, побудовані на тестовій вибірці. Особливу увагу було приділено аналізу помилок другого роду (FN), оскільки в контексті медичної діагностики вони можуть призвести до невчасного виявлення захворювання. Для логістичної регресії окремо було проаналізовано зміну класифікаційного порогу з 0.5 до 0.4, що дало змогу зменшити кількість хибно негативних рішень за рахунок прийняттого зростання хибно позитивних.

Окремо було реалізовано підхід зі зниженням розмірності за допомогою методу головних компонент (PCA), який описаний у [13], що могло б покращити ефективність моделі. Визначено кількість компонент, що зберігають не менше 95% дисперсії, після чого модель XGBoost було навчено на новому зменшеному просторі ознак.

6. Результати дослідження та обговорення. У процесі дослідження було реалізовано та протестовано моделі для задачі діагностики цукрового діабету, результати їх ефективності наведено у таблиці 1. Основну увагу приділено оцінці точності моделей за крос-валідацією та аналізу матриць невідповідності для тестових даних.

Таблиця 1.

Порівняльна таблиця результатів

Модель	Точність крос-валідації	TN	Помилка I роду, FP	Помилка II роду, FN	TP
Логістична регресія (поріг = 0.5)	0.9314	256	5	22	239
Логістична (поріг = 0.4)	0.9313	254	7	17	244
Random Forest	0.9443	253	8	17	244
XGBoost	0.9429	252	9	16	245
XGBoost + PCA	0.9319	256	5	24	237

У таблиці 1 позначено TP — істинно позитивні прогнози, TN — істинно негативні прогнози, FP — хибно позитивні прогнози (помилка I роду), FN — хибно негативні прогнози (помилка II роду).

Найкраща загальна точність досягнута моделлю Random Forest (94.43%), що свідчить про її здатність добре узагальнювати закономірності у незбалансованих даних. Найнижча кількість помилок другого роду (FN = 16) спостерігається в моделі XGBoost, що є суттєвим плюсом, оскільки помилка другого роду

в медичній сфері (невиявлення хворого пацієнта) має серйозні наслідки. Водночас ця модель має найвищу кількість FP (9). Логістична регресія із зниженим порогом (0.4) демонструє зменшення FN з 22 до 17 порівняно з класичним підходом (пори́г = 0.5), що дозволяє краще виявляти хворих пацієнтів, але ціною збільшення FP з 5 до 7. Такий компроміс часто є прийнятним у клінічних задачах, де важливіша чутливість (sensitivity), ніж специфічність (specificity).

Модель XGBoost з PCA показує зниження точності (93.19%) і зростання FN (24), що свідчить про те, що зниження розмірності через PCA у цьому випадку не покращує, а навіть трохи погіршує діагностичну здатність.

Також для моделей Random Forest та XGBoost було проаналізовано важливість ознак, що показано на рисунку 1. Визначено, що найбільший вплив на класифікаційні рішення мають такі ознаки, як рівень HDL (ліпопротеїни високої щільності), FPG (рівень глюкози в плазмі крові натщесерце), паління, вживання алкоголю, вік та рівень LDL (ліпопротеїни низької щільності). Це дозволяє не лише підвищити ефективність діагностики, а й сформулювати рекомендації щодо зменшення ризиків.

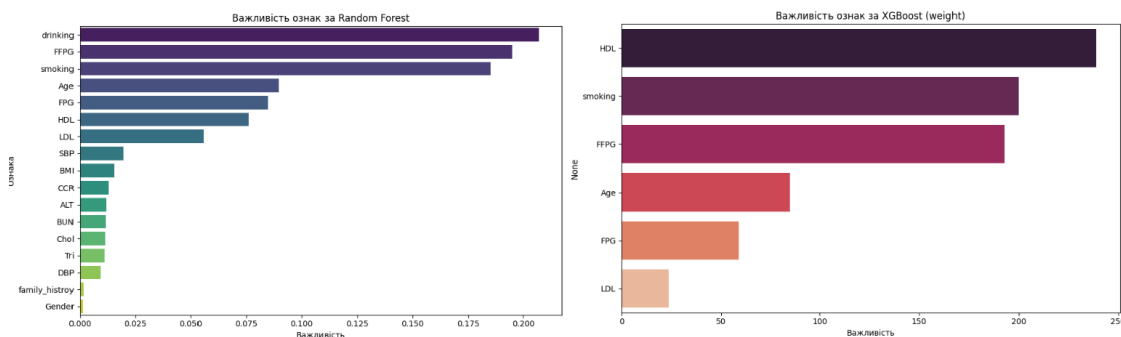


Рис. 1. Важливість ознак Random Forest та XGBoost.

У контексті медичної діагностики, мінімізація помилок другого роду (FN) є критичною. У цьому аспекті XGBoost та логістична регресія з порогом 0.4 виглядають найбільш релевантними. Якщо важливо забезпечити високий загальний баланс точності при обмеженні FP та FN — найкращим вибором є Random Forest. Логістична регресія із порогом 0.4 показала добрий компроміс і може бути рекомендована для впровадження як проста та інтерпретована модель. Метод PCA, хоча і популярний для зменшення розмірності, в даному випадку не дав позитивного ефекту, що свідчить про високу інформативність початкових ознак.

Аналіз важливості ознак виявив ключові фактори ризику розвитку діабету, зокрема рівні HDL, FFPG, паління та вживання алкоголю. Це узгоджується з медичними знаннями і демонструє, що моделі не лише здатні до класифікації, але й мають потенціал для інтерпретації і формування рекомендацій. Отримані результати підтверджують доцільність комплексного підходу, що включає збалансування класів, тюнінг моделей, оптимізацію порогу та інтерпретацію важливості ознак.

7. Висновки. У результаті проведеного дослідження було реалізовано підхід до побудови інтелектуальної системи прийняття рішень для діагностики

цукрового діабету на основі методів машинного навчання.

Найвищу точність крос-валідації (94.43%) продемонструвала модель Random Forest, однак за мінімальним рівнем помилок другого роду на тестових даних (що є важливим показником у медичній діагностиці) найкращих результатів досягнуто за допомогою XGBoost (FN = 16). З огляду на це, зазначені моделі є найбільш придатними для інтеграції в СППР.

Таким чином, результати дослідження засвідчують практичну ефективність застосування ансамблевих методів машинного навчання у задачах медичної діагностики, а також важливість гнучкого налаштування моделей з урахуванням специфіки предметної галузі. Розроблена система може слугувати ефективним допоміжним інструментом для медичних фахівців у процесі прийняття клінічних рішень щодо діагностики діабету.

Список використаної літератури

1. Rabie O., Alghazzawi D., Asghar J., Saddozai F. K., Asghar M. Z. A decision support system for diagnosing diabetes using deep neural network. *Frontiers in Public Health*. 2022. Vol. 10. DOI: <https://doi.org/10.3389/fpubh.2022.861062>
2. Tuppad A., Patil S. D. Machine learning for diabetes clinical decision support: a review. *Advances in Computational Intelligence*. 2022. Vol. 2, No. 2. P. 22. DOI: <https://doi.org/10.1007/s43674-022-00034-y>
3. Omana J., Moorthi M. Predictive analysis and prognostic approach of diabetes prediction with machine learning techniques. *Wireless Personal Communications*. 2022. Vol. 127, No. 1. P. 465–478. DOI: <https://doi.org/10.1007/s11277-021-08274-w>
4. Gupta H., Varshney H., Sharma T. K., Pachauri N., Verma O. P. Comparative performance analysis of quantum machine learning with deep learning for diabetes prediction. *Complex & Intelligent Systems*. 2022. Vol. 8, No. 4. P. 3073–3087. DOI: <https://doi.org/10.1007/s40747-021-00398-7>
5. Olowe K. J., Edoh N. L., Zouo S. J. C., Olamijuwon J. Comprehensive review of logistic regression techniques in predicting health outcomes and trends. *World Journal of Advanced Pharmaceutical and Life Sciences*. 2024. Vol. 7, No. 2. P. 16–26. DOI: <https://doi.org/10.53346/wjapls.2024.7.2.0039>
6. Badawy M., Ramadan N., Hefny H. A. Healthcare predictive analytics using machine learning and deep learning techniques: a survey. *Journal of Electrical Systems and Information Technology*. 2023. Vol. 10, No. 1. Article No. 40. DOI: <https://doi.org/10.21203/rs.3.rs-1885746/v2>
7. Sievering A. W. and others. Comparison of machine learning methods with logistic regression analysis in creating predictive models for risk of critical in-hospital events in COVID-19 patients on hospital admission. *BMC Medical Informatics and Decision Making*. 2022. Vol. 22, No. 1. Article No. 309. DOI: <https://doi.org/10.1186/s12911-022-02057-4>
8. Iranzad R., Liu X. A review of random forest-based feature selection methods for data science education and applications. *International Journal of Data Science and Analytics*. 2024. P. 1–15. DOI: <https://doi.org/10.1007/s41060-024-00509-w>
9. Palimkar P., Shaw R. N., Ghosh A. Machine learning technique to prognosis diabetes disease: Random forest classifier approach. *Advanced computing and intelligent technologies : proceedings of ICACIT 2021*. Singapore : Springer Singapore, 2021. P. 219–244. DOI: https://doi.org/10.1007/978-981-16-2164-2_19
10. Bentéjac C., Csörgő A., Martínez-Muñoz G. A comparative analysis of gradient boosting algorithms. *Artificial Intelligence Review*. 2021. Vol. 54. P. 1937–1967. DOI: <https://doi.org/10.1007/s10462-020-09896-5>
11. Zhang P., Jia Y., Shang Y. Research and application of XGBoost in imbalanced data. *International Journal of Distributed Sensor Networks*. 2022. Vol. 18, No. 6. DOI: <https://doi.org/10.1177/1550132922110693>
12. Darabi P. K. Diabetes dataset with 18 features. 2023. URL: <https://www.kaggle.com/datasets/pkdarabi/diabetes-dataset-with-18-features> (date of

access: 15.03.2025).

13. Кондрук Н. Е. Аналіз технік зменшення розмірності в машинному навчанні. *Науковий вісник Ужгородського університету. Серія «Математика і інформатика»*. 2023. Вип. 42, № 1. С. 181–187. DOI: [https://doi.org/10.24144/2616-7700.2023.42\(1\).181-187](https://doi.org/10.24144/2616-7700.2023.42(1).181-187)

Malyar M. M., Kondruk N. E., Kondruk E. B., Neroda V. A. Development of an Intelligent Decision Support System for Diabetes Diagnosis.

The article presents the results of the development of an intelligent decision support system for the diagnosis of diabetes mellitus using machine learning methods. Based on an open medical dataset, a number of classification models were built and analyzed, including logistic regression, Random Forest, and XGBoost. In order to improve the accuracy, the stratification method was used to select validation data, feature scaling, cross-validation, and optimization of model hyperparameters. Particular attention was paid to analyzing the discrepancy matrices and assessing the impact of key features. The best cross-validation accuracy was achieved by the Random Forest model (94.43%), which indicates its ability to generalize patterns in unbalanced data well. The lowest number of errors of the second kind (FN = 16) is observed in the XGBoost model. The proposed approach can be effectively used for the preliminary diagnosis of diabetes.

Keywords: decision-making system, classification, diabetes diagnosis, intelligent system, machine learning.

References

1. Rabie, O., Alghazzawi, D., Asghar, J., Saddozai, F. K., & Asghar, M. Z. (2022). A decision support system for diagnosing diabetes using deep neural network. *Frontiers in Public Health*, 10, 861062. <https://doi.org/10.3389/fpubh.2022.861062>
2. Tuppad, A., & Patil, S. D. (2022). Machine learning for diabetes clinical decision support: A review. *Advances in Computational Intelligence*, 2(2), 22. <https://doi.org/10.1007/s43674-022-00034-y>
3. Omana, J., & Moorthi, M. (2022). Predictive analysis and prognostic approach of diabetes prediction with machine learning techniques. *Wireless Personal Communications*, 127(1), 465–478. <https://doi.org/10.1007/s11277-021-08274-w>
4. Gupta, H., Varshney, H., Sharma, T. K., Pachauri, N., & Verma, O. P. (2022). Comparative performance analysis of quantum machine learning with deep learning for diabetes prediction. *Complex & Intelligent Systems*, 8(4), 3073–3087. <https://doi.org/10.1007/s40747-021-00398-7>
5. Olowe, K. J., Edoh, N. L., Zouo, S. J. C., & Olamijuwon, J. (2024). Comprehensive review of logistic regression techniques in predicting health outcomes and trends. *World Journal of Advanced Pharmaceutical and Life Sciences*, 7(2), 16–26. <https://doi.org/10.53346/wjapls.2024.7.2.0039>
6. Badawy, M., Ramadan, N., & Hefny, H. A. (2023). Healthcare predictive analytics using machine learning and deep learning techniques: A survey. *Journal of Electrical Systems and Information Technology*, 10(1), 40. <https://doi.org/10.21203/rs.3.rs-1885746/v2>
7. Sievering, A. W., & others. (2022). Comparison of machine learning methods with logistic regression analysis in creating predictive models for risk of critical in-hospital events in COVID-19 patients on hospital admission. *BMC Medical Informatics and Decision Making*, 22(1), 309. <https://doi.org/10.1186/s12911-022-02057-4>
8. Iranzad, R., & Liu, X. (2024). A review of random forest-based feature selection methods for data science education and applications. *International Journal of Data Science and Analytics*. Advance online publication. <https://doi.org/10.1007/s41060-024-00509-w>
9. Palimkar, P., Shaw, R. N., & Ghosh, A. (2021). Machine learning technique to prognosis diabetes disease: Random forest classifier approach. In *Advanced computing and intelligent technologies: Proceedings of ICACIT 2021*. Springer: Singapore. https://doi.org/10.1007/978-981-16-2164-2_19
10. Bentéjac, C., Csörgő, A., & Martínez-Muñoz, G. (2021). A comparative analy-

- sis of gradient boosting algorithms. *Artificial Intelligence Review*, 54(3), 1937–1967. <https://doi.org/10.1007/s10462-020-09896-5>
11. Zhang, P., Jia, Y., & Shang, Y. (2022). Research and application of XGBoost in imbalanced data. *International Journal of Distributed Sensor Networks*, 18(6). <https://doi.org/10.1177/1550132922110693>
 12. Darabi, P. K. (2023). *Diabetes dataset with 18 features*. [Data set]. Kaggle. Retrieved from <https://www.kaggle.com/datasets/pkdarabi/diabetes-dataset-with-18-features>
 13. Kondruk, N. E. (2023). Analysis of dimensionality reduction techniques in machine learning. *Scientific Bulletin of Uzhhorod University, Series "Mathematics and Informatics"*, 42(1), 181–187. [https://doi.org/10.24144/2616-7700.2023.42\(1\).181-187](https://doi.org/10.24144/2616-7700.2023.42(1).181-187)

Одержано 25.03.2025