

УДК 004.85

ВИКОРИСТАННЯ МОДЕЛЕЙ ШТУЧНОГО ІНТЕЛЕКТУ ДЛЯ КЛАСИФІКАЦІЇ ЧИСЛОВИХ МЕДИЧНИХ ДАНИХ

Волобуєва Д.С., Чала Л. Е.

e-mail: daryna.volobuieva@nure.ua, larysa.chala@nure.ua

Харківський національний університет радіоелектроніки, каф. ШІ
м. Харків, Україна

This paper explores the use of artificial intelligence models for the classification of numerical medical data. Machine learning methods are widely used in medical diagnostics to analyze clinical indicators and support decision-making. In this study, several models are compared, including Logistic Regression, Random Forest, Gradient Boosting, and Support Vector Machine. The models are evaluated using Accuracy, F1-score, ROC-AUC, and MCC metrics. The experimental results show that ensemble methods provide high classification performance. The Random Forest model demonstrates the best results among the considered algorithms. Based on the obtained results, an approach to improving classification using ensemble models and feature importance analysis is proposed.

У сучасному світі штучний інтелект став одним із ключових напрямів розвитку цифрових технологій, що суттєво трансформують різні сфери суспільства, зокрема й медицину. Алгоритми машинного навчання та інтелектуальні системи дедалі активніше застосовуються для аналізу великих масивів даних, автоматизації прийняття рішень і підвищення точності прогнозування.

Особливої актуальності набуває використання моделей штучного інтелекту для обробки та класифікації структурованих числових даних, які становлять значну частину медичної інформації. На відміну від традиційних статистичних підходів, моделі штучного інтелекту здатні виявляти складні нелінійні залежності між ознаками та адаптуватися до особливостей конкретних вибірок.

Важливою частиною проблеми класифікації поточної ситуації є вибір правил прийняття діагностичних рішень, насамперед за результатами аналізу клінічних даних [1]. Саме тому застосування методів машинного навчання для автоматизованого аналізу числових медичних показників дозволяє підвищити обґрунтованість діагностичних висновків і підтримати процес прийняття медичних рішень.

У роботі використовується синтетичний датасет ALLER-Search (n=1200), сформований відповідно до клінічних діапазонів показників алергологічних обстежень. Датасет містить 18 числових ознак: загальний та специфічний IgE (пилок, кліщі, харчові алергени), абсолютна та відносна кількість еозинофілів, результати шкірних прик-тестів (мм), рівні цитокінів IL-4, IL-5, IL-13, показники ВАТ, загальний аналіз крові

(гемоглобін, лейкоцити, тромбоцити), вік, спадковість, тривалість симптомів, кількість сенсibiliзацій. Цільова змінна – бінарна: 1 – алергія підтверджена (55.6%), 0 – відсутня (44.4%). Для реалістичності задачі до набору даних введено 5% випадкових пропущених значень (заповнення медіаною по ознаці) та 8% перевернутих міток, що імітує «граничні» клінічні випадки.

Для класифікації обрано чотири алгоритми: Logistic Regression (LR) – як базова лінійна модель; Random Forest (RF) – ансамблевий метод на основі дерев рішень; Gradient Boosting (GB) – послідовний ансамблевий метод; Support Vector Machine з RBF-ядром (SVM).

Передобробка даних здійснювалась за допомогою Min-Max Scaling, що дозволило привести всі ознаки до одного інтервалу. Оцінка моделі здійснювалась через 5-кратну стратифіковану крос-валідацію (StratifiedKFold) для збереження пропорцій класів.

Для порівняння результатів застосовувалися кілька метрик: Accuracy – для загальної частки правильних передбачень, F1-score – для балансу між точністю та повнотою при дисбалансі класів, ROC-AUC – для оцінки здатності моделі розрізняти класи при різних порогах, та Matthews Correlation Coefficient (MCC) – для комплексного врахування всіх типів прогнозів при дисбалансі класів.

Отримані результати свідчать про ефективність застосування моделей машинного навчання для класифікації числових медичних даних (рис. 1).

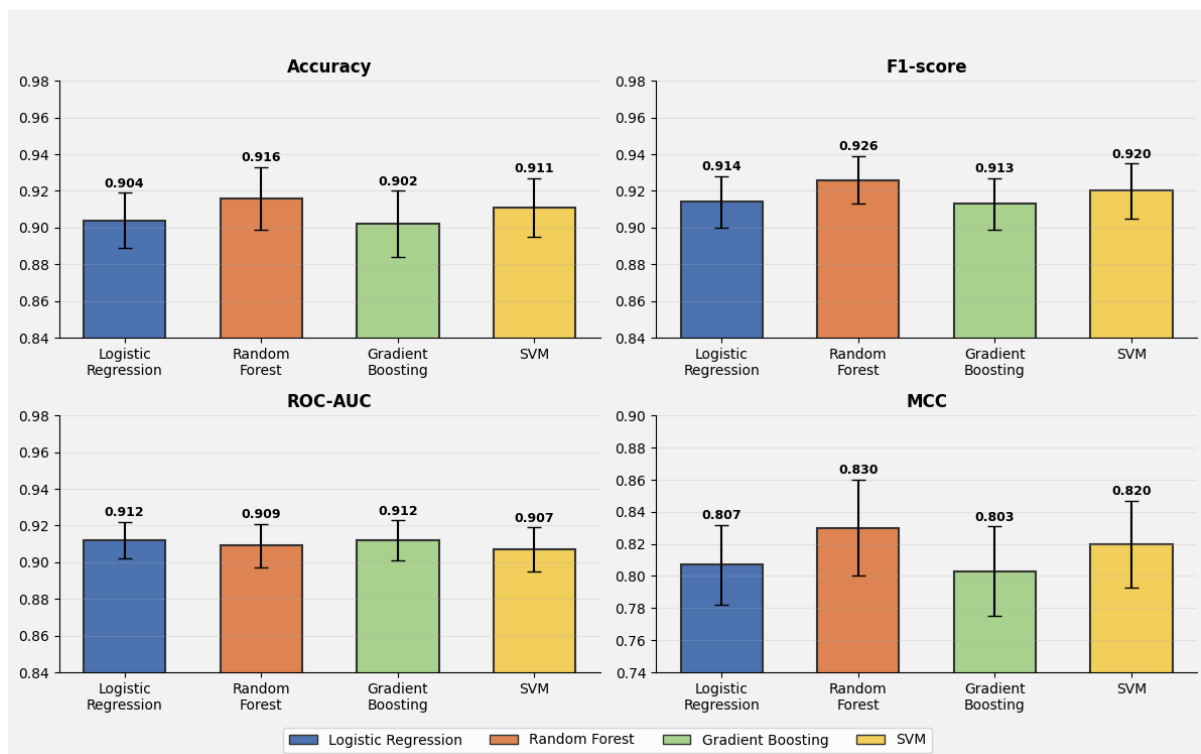


Рисунок 1 – Порівняльна гістограма метрик класифікації моделей МН

Найкращі показники продемонструвала модель Random Forest, яка досягла найвищих значень Accuracy (0.916), F1-score (0.926) та MCC (0.830). Моделі Logistic Regression та Gradient Boosting показали найвищі значення ROC-AUC (0.912), що свідчить про їхню високу здатність розрізняти класи [2, 3]. SVM також продемонструвала стабільні результати за більшістю метрик.

На основі отриманих результатів у роботі пропонується підхід до вдосконалення класифікації числових медичних даних шляхом використання адаптивного ансамблевого комбінування моделей машинного навчання. Запропонований підхід передбачає інтеграцію результатів кількох базових алгоритмів (Logistic Regression, Random Forest, Gradient Boosting, SVM) із використанням вагових коефіцієнтів, що визначаються відповідно до ефективності моделей за ключовими метриками якості.

Крім того, перспективним напрямом слід вважати врахування важливості окремих медичних ознак під час побудови класифікаційної моделі. Використання методів оцінювання значущості ознак може дозволити виділити найбільш інформативні клінічні показники та підвищити точність прогнозування. Це особливо важливо для медичних задач, де різні показники можуть мати різний вплив на результат діагностики.

Таким чином, подальший розвиток ансамблевих методів та інтеграція інформації про важливість ознак можуть сприяти підвищенню ефективності аналізу числових медичних даних. Запропонований підхід може бути використаний для створення інтелектуальних систем підтримки прийняття медичних рішень, що дозволить автоматизувати процес аналізу клінічних показників та підвищити точність діагностики алергічних захворювань.

Список використаних джерел:

1. Удовенко, С., Чала, Л., & Шергін, В. (2018). Байєсівська класифікація ситуацій в інформаційних системах медичної діагностики. «Інформаційні системи та технології» ICT-2018, 261 <http://openarchive.nure.ua/handle/document/7778> (дата звернення: 03.03.2026).
2. Кравець, П., Пасічник, В., Проданюк, М., & Кісь, Я. (2024). Комп'ютерне моделювання логістичної регресії для бінарної класифікації // Вісник Національного університету "Львівська політехніка". Серія: Інформаційні системи та мережі, Вип. 16, с. 167–190. <https://doi.org/10.23939/sisn2024.16.167> (дата звернення: 04.03.2026).
3. Zhang, Z., Zhao, Y., Canes, A., Steinberg, D., Lyashevskaya, O.; written on behalf of AME Big-Data Clinical Trial Collaborative Group (2019). Predictive analytics with gradient boosting in clinical medicine. *Annals of Translational Medicine*, 7(7), 152. <https://doi.org/10.21037/atm.2019.03.29> (дата звернення: 04.03.2026).