

Використання алгоритмів NAIVE BAYES, RANDOM FOREST та SVM для класифікації захворювань

Толстих М.Ю., науковий керівник Семко В.В.

Автоматизоване виявлення захворювань за допомогою штучного інтелекту набуває особливого значення в умовах цифровізації медицини. Алгоритми машинного навчання здатні ефективно класифікувати пацієнтів за симптомами, покращуючи точність діагностики, зменшуючи навантаження на лікарів і підвищуючи швидкість реагування. У цьому дослідженні було розглянуто та порівняно ефективність трьох популярних підходів до задачі класифікації хвороб: Naive Bayes, Random Forest та Support Vector Machine (SVM).

Naive Bayes — це статистичний метод, який базується на теоремі Байєса з припущенням умовної незалежності ознак [1]. Його основними перевагами є швидкість роботи, низькі обчислювальні витрати та здатність працювати навіть на невеликих наборах даних. Алгоритм широко використовується у медичній діагностиці, особливо для первинного аналізу симптомів, де необхідна швидка оцінка ймовірності наявності захворювання.

У дослідженнях, що стосуються серцево-судинних захворювань, модель Naive Bayes досягала точності до 75%, демонструючи хороший баланс між простотою та ефективністю [2]. Проте головним недоліком цього методу є те, що він погано справляється із сильно корельованими ознаками, що часто трапляється в медичних даних [3].

Random Forest — це ансамблевий метод навчання, що створює велику кількість дерев рішень і об'єднує їхні передбачення для досягнення стабільної та надійної класифікації [4]. Однією з ключових переваг є його стійкість до перенавчання, здатність обробляти великі обсяги даних і працювати з неструктурованими вхідними ознаками. У медичній практиці RF активно використовується для діагностики гепатиту, панкреатиту, хвороби Альцгеймера та онкологічних захворювань. У дослідженнях нейровізуалізації точність класифікації за допомогою RF досягала понад 85%, що свідчить про здатність моделі ефективно обробляти складні медичні ознаки. Також RF дозволяє оцінити важливість кожного симптому, що є корисним для лікарів у прийнятті клінічних рішень.

Support Vector Machine — це потужний алгоритм, що створює оптимальну гіперплощину, яка розділяє дані різних класів із максимально можливим зазором. Метод добре працює з нелінійно роздільними даними завдяки використанню ядерних функцій. Особливо ефективним SVM є у задачах із високою розмірністю ознак, таких як геномні або зображувальні дані.

У класифікації раку молочної залози, діабету та при роботі з магнітно-резонансними зображеннями точність SVM перевищує 90% . Проте метод має високі обчислювальні витрати та вимагає ретельного налаштування гіперпараметрів, що може ускладнити його використання у клінічних системах реального часу .

Порівняння результатів.

Проведене порівняння свідчить про те, що вибір моделі залежить від структури даних, обсягу вибірки, наявності корельованих ознак і обчислювальних ресурсів.

Naive Bayes демонструє найвищу швидкість навчання, але поступається в точності іншим методам у випадках залежних ознак [1,3].

Random Forest забезпечує високу точність, стабільність та інтерпретованість результатів, і є хорошим вибором для задач зі складними залежностями .

SVM забезпечує найвищу точність у складних задачах з високорозмірними даними, але потребує значних обчислювальних ресурсів .

У роботах, присвячених комплексному порівнянню цих підходів, було встановлено, що гібридні моделі або ансамблі кількох алгоритмів здатні покращити результати класифікації до рівня понад 95% .

Кожен із розглянутих методів має свої переваги та обмеження.

Naïve Bayes доцільно застосовувати на етапі первинної діагностики або коли доступні невеликі вибірки.

Random Forest — універсальний інструмент для медичної класифікації завдяки високій точності та гнучкості.

SVM доцільно використовувати в задачах із великим обсягом даних та складними ознаками, особливо в генетиці та медичній візуалізації. Найперспективнішим напрямом розвитку є побудова ансамблевих та гібридних систем, які поєднують переваги кількох алгоритмів і дозволяють отримати точнішу та надійнішу діагностику.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Patil, S., & Waghmare, L. (2023). *Medical Diagnosis using Naive Bayes Algorithm*. International Journal of Advanced Computer Science.
2. Nandi, S., et al. (2023). *Limitations of Naive Bayes in correlated medical data*. Expert Systems with Applications.
3. Breiman, L. (2001). *Random Forests*. Machine Learning, 45(1), 5–32.
4. Wang, L. et al. (2023). *Diagnosis of Hepatitis using Random Forests*. Computers in Biology and Medicine.

**НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ БІОРЕСУРСІВ І
ПРИРОДОКОРИСТУВАННЯ УКРАЇНИ
ФАКУЛЬТЕТ ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ**



**ТЕОРЕТИЧНІ ТА ПРИКЛАДНІ
АСПЕКТИ РОЗРОБКИ
КОМП'ЮТЕРНИХ СИСТЕМ**

**ЗБІРНИК НАУКОВИХ ПРАЦЬ ЗА МАТЕРІАЛАМИ
VII ВСЕУКРАЇНСЬКОЇ НАУКОВО-ПРАКТИЧНОЇ КОНФЕРЕНЦІЇ
СТУДЕНТІВ І АСПІРАНТІВ
*24 квітня 2025 року***

Київ 2025

УДК 004

Відповідальна за випуск: М.І. Лендел

Збірник наукових праць за матеріалами VII Всеукраїнської науково-практичної конференції студентів і аспірантів «ТЕОРЕТИЧНІ ТА ПРИКЛАДНІ АСПЕКТИ РОЗРОБКИ КОМП'ЮТЕРНИХ СИСТЕМ '2025», 24 квітня 2025 року, НУБіП України, Київ. – 452 с. (електронне видання)

Відповідальність за зміст публікацій несуть автори.

Передрук матеріалів, а також використання їх будь-якій формі допускається лише з дозволу авторів