УДК 004.8 DOi 10.30837/ bi.2025.2(103).06

Н. Б. Гулієв[1], О. С. Назаров[2]

[1]ХНУРЕ, м. Харків, Україна, nural.huliiev@nure.ua, ORCID iD: 0000-0003-2123-0377

[2]ХНУРЕ, м. Харків, Україна, oleksii.nazarov1@nure.ua, ORCID iD: 0000-0001-8682-5000

ДОСЛІДЖЕННЯ МЕТОДІВ НАЛАШТУВАНЬ ГІПЕРПАРАМЕТРІВ ДЛЯ РЕАЛІЗАЦІЇ АЛГОРИТМУ ВИПАДКОВИЙ ЛІС НА ОСНОВІ МЕДИЧНИХ ТА ПСИХОЛОГІЧНИХ ДАНИХ

Випадковий ліс є одним із найпоширеніших алгоритмів машинного навчання, що належить до методів ансамблевого навчання. Його застосовують у медицині, фінансах, соціальних науках, екології, IT та багатьох інших сферах. Сутність алгоритму полягає у створенні великої кількості дерев рішень і подальшому об'єднанні їхніх результатів для отримання точного та стабільного прогнозу. Попри численні переваги, випадковий ліс має й недоліки, зокрема низьку стійкість до різнорідності даних, що часто трапляється в медицині. У дослідженні алгоритм застосовується для аналізу медичних даних із психологічними показниками. Медичні дані мають порогові значення, які можуть давати неочікувані результати, тому оптимізація випадкового лісу залишається актуальною. Для аналізу альтернативних варіантів удосконалення обрано метод лінійної адитивної згортки. Він дозволяє обчислювати зважену суму нормалізованих показників, щоб порівнювати різні рішення. Цей метод є універсальним, простим у реалізації та придатним для задач із багатьма різнорідними критеріями. Досліджено способи оптимізації алгоритму випадковий ліс через налаштування гіперпараметрів. Розглянуто рандомізований пошук, пошук за сіткою та байєсівську оптимізацію. Проаналізовано їхні реалізації, особливості та можливі комбінації. На основі оцінки ефективності визначено, що для медичних і психологічних даних найкращим підходом є байєсівська оптимізація. Вона забезпечує більш точні та стабільні результати. Зрештою обрано найбільш оптимальний спосіб удосконалення алгоритму.

БАЙЄСІВСЬКА ОПТИМІЗАЦІЯ, ВИПАДКОВИЙ ЛІС, ГІПОТИРЕОЗ, ГІПЕРТИРЕОЗ, ДЕРЕВА РІШЕНЬ, ОПТИМІЗАЦІЯ, ПОШУК ЗА СІТКОЮ, ПСИХОЛОГІЧНІ РОЗЛАДИ, РАНДОМІЗОВАНИЙ ПОШУК.

N. B. Huliiev, O. S. Nazarov. Study of hyperparameter tuning methods for implementing the Random Forest algorithm based on medical and psychological data. Random Forest is one of the most widely used machine learning algorithms and belongs to ensemble learning methods. It is applied in medicine, finance, social sciences, ecology, IT, and many other fields. The essence of the algorithm lies in creating a large number of decision trees and then combining their results to obtain a more accurate and stable prediction. Despite its numerous advantages, Random Forest also has drawbacks, including low robustness to heterogeneous data, which is common in medical datasets. In this study, the algorithm is used to analyze medical data with psychological indicators. Medical data contains threshold values that may produce unexpected results, which makes the optimization of Random Forest still relevant. To analyze alternative improvement options, the linear additive convolution method was chosen. It allows calculating a weighted sum of normalized indicators to compare different solutions. This method is universal, easy to implement, and suitable for problems involving many heterogeneous criteria. Various approaches to optimizing the Random Forest algorithm through hyperparameter tuning were examined. Random search, grid search, and Bayesian optimization were considered. Their implementations, characteristics, and possible combinations were analyzed. Based on the evaluation of effectiveness, Bayesian optimization was identified as the best approach for medical and psychological data. It provides more accurate and stable results. Ultimately, the most optimal method for improving the algorithm was selected.

BAYESIAN OPTIMIZATION, RANDOM FOREST, HYPOTHYROIDISM, HYPERTHYROIDISM, DECISION TREES, OPTIMIZATION, GRID SEARCH, PSYCHOLOGICAL DISORDERS, RANDOMIZED SEARCH.

Вступ

Медична сфера налічує чимало випадків ускладнень через наявні хвороби, що також стосується гіпо- та гіпертиреозу, адже вони спричиняють чимало інших хвороб, що може заважати лікувати першочергову недугу. Одним із таких чинників є психологічні розлади.

Об'єкт даного дослідження є саме процес прогнозування психологічних розладів у пацієнтів із гіпо- та гіпертиреозом на основі медичних і психологічних показників.

Наразі автори вже робили аналіз даної проблеми та дійшли до того, що застосовуватимуть алгоритм випадковий ліс задля оцінки стану пацієнтів для завчасного визначення потенційної можливості погіршення психологічного стану хворих.

Предмет дослідження – методи оптимізації гіперпараметрів алгоритму Random Forest, зокрема жадібний підхід, рандомізований пошук по сітці, байєсівська оптимізація та комбінований підхід.

1. Дослідження оптимізації алгоритму випадковий ліс для аналізу даних пацієнтів

Випадковий ліс – це метод машинного навчання, який застосовується в прогнозуванні та класифікації. Алгоритми навчання моделей в машинному навчанні та штучному інтелекті вимагають великий обсяг даних задля якомога точного та якісного результату. Інформація про продуктивність розробленої системи дозволяє вдосконалити аналогічні алгоритми, підвищити ефективність апаратного та програмного забезпечення,

процеси прийняття рішень, прогнозів, вирішення проблем, що загалом покращить показники точності побудованої моделі. Вирішуючи кожну проблему, необхідно інтегрувати різні способи збору та обробки вхідних даних, що також допомагає підвищувати рівень точності. Дослідницький процес міждисциплінарних спостережень може містити різні види методологій, що застосовується в прийнятті рішень, розпізнаванні образів, вирішенні проблем та прогнозах, а також допомагає досягати інноваційності.

Випадковий ліс вважається потужним механізмом у сфері машинного навчання. Особливо популярний в прогнозуванні, супервізорному навчанні та категоризації. Випадковий ліс вирізняється чималою кількістю переваг, тому і є популярним алгоритмом в машинному навчанні та серед способів передбачення.

Застосування даного методу полягає в наступному:

– високий показник точності, що підтверджує ефективність та надійність класифікації та прогнозування, проведених за його допомогою;

– зменшення проблеми перенавчання та узагальнення результатів за рахунок будови дерев рішень на різних підмножинах даних;

– стійкість до відсутніх та незбалансованих даних;

– визначають важливість ознаки задля будування причинно-наслідкових зв'язків їх впливу на остаточний результат;

– техніка ансамблевого навчання допомагає підвищити стабільність прогнозів у порівнянні з іншими існуючими методами класифікації;

– прості в інтеграції та реалізації.

Цей метод наразі популярний серед інновацій, бо застосовує методи дерев рішень, розробляючи колекцію дерев та надаючи результат на основі усіх них. Кожне дерево будується та навчається за рахунок випадково обраної підмножини даних. Остаточне рішення надається об'єднанням усіх побудованих моделей.

Не дивлячись на те, що випадкові ліси – популярний та потужний інструмент в прогнозуванні, він потребує оптимізації у застосуванні з наборами вхідних даних, отриманих у ході біомедичних досліджень, які мають рідкісні результати та коваріати [1].

Метою роботи є підвищення точності та ефективності моделі Random Forest для раннього виявлення ризику психологічних розладів шляхом визначення найоптимальнішого методу налаштування її гіперпараметрів.

Незважаючи на багато способів удосконалення алгоритму випадковий ліс, вирішальну позицію в цьому займає саме оптимізація гіперпараметрів, а саме глибина дерева, кількість дерев рішень, мінімальний розмір вибірки. Для налаштування даних показників зазвичай застосовують рандомізований пошук, пошук по сітці та байєсівську оптимізацію. Кожен має свої переваги та недоліки, тому слід розглянути кожний окремо, щоб визначити якомога кращий, тому предметом дослідження є аналіз різних видів оптимізації налаштування гіперпараметрів за допомогою їх реалізації на мові Python та лінійної адитивної згортки.

Завдання дослідження:

– проаналізувати медичні та психологічні показники пацієнтів з Kaggle для формування набору даних;

– провести попередню обробку даних та підготовку вибірки для навчання моделі Random Forest;

– реалізувати та порівняти різні методи оптимізації гіперпараметрів:

– жадібний метод;

– рандомізований пошук по сітці;

– байєсівську оптимізацію;

– комбінований підхід (Bayesian Optimization + Random Search);

– оцінити якість налаштованих моделей за метриками Accuracy, Precision, Recall, F1-score, а також за часом оптимізації;

– визначити метод гіперпараметричної оптимізації, який забезпечує найкраще співвідношення точності та обчислювальних витрат.

Розглянемо методи та засоби спостереження більш детально – згортку, яка використовується в задачах прийняття рішень, що і є задачею експерименту.

2. Матеріали і методи досліджень

Розв'язуючи багатокритеріальні задачі, результатом завжди є кращі альтернативи, які відповідають поставленим вимогам. Найчастіше тут використовуються методи двох видів: перша полягає у виключенні кількості критеріїв оцінки, а друга зменшує кількість варіантів аналізу на його початку. Для нашого дослідження найбільш підходящим є саме метод із першої групи. Такими способами є метод граничних та головного критеріїв, відстані та згорток.

Методи згорток поділяються на лінійні адитивні, мультиплікативні та максимінні. Метою застосування згорток є узагальнення усіх критеріїв аналізу.

Адитивна розраховується за наступною формулою:

$$K(x)=\sum_{j=1}^{n} a_j K_j(x) \qquad (1),$$

де $K(x)$ – загальний критерій для альтернативи $x \in X$, $(K_1(x), ..., K_j(x), ..., K_n(x))$ – набір вихідних критеріїв; n – число вихідних критеріїв; $a_j(x)$ – нормуючий множник, який вказує на вагу альтернативи.

Найкращий із усіх можливих альтернатив задачі обчислюється за допомогою наступної формули:

$$x^n = \arg\max_{x \in X} K(x) \qquad (2),$$

Тобто результатом є найбільше значення, отримане методом згортки.

Мультиплікативна згортка розраховується за допомогою такої формули:

$$K(x)=\prod_{j=1}^{n}K_j^{a_j}(x) \quad (3)$$

Максимінна згортка знаходиться за формулою:

$$K(x)=\max_i \min_j a_{ij}K_j(x) \quad (4).$$

Найкращі результати за мультиплікативною та максимінною згортками обчислюється за формулою (2).

Метод граничних критеріїв застосовується в задачах проектування і планування, в яких порогові значення критеріїв набувають значень $k_j(x)\geq k_{jo}$; $j=1$, ..., n. Формула обчислення цього способу наступна:

$$K(x)=\min_j\left(\frac{K_j(x)}{K_{jo}(x)}\right) \quad (5)$$

Найкращий результат обирається формулою 2.

Метод відстані використовує відстань, яка є додатковою метрикою. Наприклад, для вибору ідеального рішення цілком достатньо інформації. Обчислимо відстань до значення максимуму d(x) для кожної альтернативи. Тоді найкраща альтернатива буде відомою із застосуванням формули:

$$x^*=\arg\min_{x\in X} d(x) \quad (6)$$

Застосування методу з першої множини іноді вимагає один із способи із другої, наприклад – принцип Парето: альтернативи, які за всіма критеріями програють іншому або іншим варіантам, видаляються до початку дослідження.

Також бувають випадки, в яких параметри, які неконтрольовані через різні причини, ускладнюють будову моделі для подальшого аналізу. Тут у нагоді може стати метод гарантованого результату, мета якого полягає у визначенні найгіршої реакції та гарантованого значення.

Для спостереження варто використати згортку, адже важко визначити порогові значення критеріїв аналізу, а саме лінійну адитивну, яка є найпоширенішою та найпростішою, та метод із другої множини способів – принцип Парето, якщо одна із альтернатив прозоро гірша за інші.

Спершу необхідно обрати критерії, за якими проводитиметься дослідження: альтернативи порівнюватимуться за допомогою цих ознак.

Значення кожної з них може мати як кількісне, так і якісне походження. Згортка оперує із першими, тому у випадку других, необхідно конвертувати їх у кількісні та побудувати нову таблицю вхідних даних варіантів аналізу.

На третьому кроці виключатимемо альтернативи за допомогою принципу Парето, якщо усі її показники за усіма її ознаками менші з-поміж інших можливих значень інших варіантів експерименту. Варто зазначити, якщо показники альтернатив в різних проміжках або мірах вимірювання, необхідно нормалізувати дані максимізацією або мінімізацією даних, щоб точність та коректність результатів відповідала дійсності.

Четвертим етапом є ранжування показників – обчислення вагомих коефіцієнтів. Існують різні способи, в даному спостереженні можна застосувати один із найбільш популярних методів: для кожного критерію один ділитимемо на суму усіх її значень.

Останнім етапом залишається обчислення значення згортки для кожної із альтернатив: розрахунок суми добутків кожної пари значень вагових коефіцієнтів та критеріїв [5].

Задачею дослідження якраз є вибір найкращого або найкращих методів удосконалення способу налаштування гіперпараметрів для оптимізації алгоритму випадковий ліс.

Проведемо експеримент та розв'яжемо багатокритеріальну задачу вибору способу налаштування гіперпараметрів для реалізації алгоритму випадковий ліс для застосування в задачі аналізу психічних розладів серед хворих гіпотиреозом та гіпертиреозом. Розпочнемо із застосування існуючих методів.

3. Аналіз літературних джерел

Наразі можна розрізнити пошукові та підтверджувальні експерименти, а тому розуміти, коли їхнє застосування доцільне. У цій статті описано використання саме алгоритму випадкових лісів, які спроможні надати кращі прогнози, ніж регресія, та визначати нелінійні ефекти. Даний алгоритм застосовується вже в багатьох сферах: банківські справи, фармацевтика, біржа, охорона здоров'я тощо. Однак, автори мають думку, що в психології – набагато рідше, тому вирішили розглянути його використання в контексті психологічних досліджень. Особливу увагу вони приділили обмеженням, які можуть виникнути в цьому випадку, та шляхам їх уникнення та вирішення, розглянувши існуючі дослідження детальніше далі.

Програмне забезпечення може зазнавати збоїв під час роботи. Задля мінімізації даних проблем необхідно ефективно прогнозувати можливі помилки. Існує дослідження, мета якого була якраз передбачення несправностей в роботі програмного забезпечення за допомогою вдосконаленого алгоритму випадкового лісу на основі даних NASA JM1, які налічували 21 програмну метрику. Спочатку метод усував дисбаланс класів способом надмірної вибірки синтетичної меншини (SMOTE). Суттю нового підходу було налаштування класифікатора випадкових лісів з увагою на оптимізацію гіперпараметрів. Під час порівняння модифікованого методу із стандартними у машинному навчанні він мав кращі показники точності та F1. Підкреслено, що важливо пам'ятати про можливі дефекти програм та шляхи їх передбачення задля підвищення продуктивності програмного забезпечення.

Спочатку проходить обробка вхідних даних: видаляються нульові значення та інші проблемні дані. Потім для усування дисбалансу використовується метод

SMOTE. На наступному кроці відбираються ознаки випадкового лісу, які безпосередньо впливатимуть на процес будування та структуру дерев рішень. Для більшої ефективності роботи алгоритму було скомбіновано два методи оптимізації алгоритму: усунення дисбалансу класів та налаштування гіперпараметрів рандомізованим пошуком по сітці. [2]

Ефективність комунікаційних та радіолокаційних систем залежить від інверсії атмосферних каналів. А на продуктивність та якість прогнозування моделі машинного навчання впливають параметри, які безпосередньо беруть участь в її реалізації. Тому в одному із експериментів розроблено модель випадкового лісу, вдосконалену за допомогою методу байєсівської оптимізації, для прогнозування атмосферних каналів. Оптимізацію застосовано задля пошуку певних гіперпараметрів під час навчання. Додатково використано метод K-кратної перехресної перевірки для визначення кращого способу поділу моделі та уникнення проблеми її перенавчання. Для перевірки реалізованого алгоритму його результати порівнювались із результатами, розрахованими за допомогою інших популярних методів прогнозування: класичний алгоритм випадкового лісу, метод k-найближчого сусіда з та без байєсівської оптимізації та метод градієнтного підсилення з та без байєсівської оптимізації. Зрештою, показники нового методу коефіцієнту детермінації R2, MAE та MSE були більшими, а результати прогнозування більш точними. Також визначено, що результати кращі навіть у випадку наявності шуму в даних.[3]

Енергетична безпека забезпечується відсутністю вторгнень в енергетичні промислові системи, тому метод їх виявлення є конче важливим в даній галузі. Існують два способи, але вони мають недоліки: вони добре працюють з гіперпараметрами, але їхня оптимізація може суттєво підвищити показники точності моделі виявлення вторгнень, а також вони взагалі застосовуються для контролю безпеки інформаційних систем, а не окремо для моніторингу атак на управління енергетичних систем. Тому в одному із експериментів було запропоновано модель випадкового лісу для виявлення вторгнень, де було використано метод improved grid search algorithm в якості оптимізації гіперпараметрів для покращення показників ефективності майбутньої моделі. Новий метод аналізувався на основі даних управління державної енергетичної системи. Точність досягла 98%.

У статті описано модифіковану модель виявлення вторгнень у промислові енергетичні системи задля вирішення наявних недоліків існуючої реалізації. Було запропоновано оптимізувати гіперпараметри покращеним методом сіткового пошуку: параметри налаштовувались у порядку важливості для збільшення продуктивності моделі. Швидкість нового способу була в 165 разів вищою, аніж показники швидкості роботи звичайного сіткового пошуку. Тому даний метод може застосовуватися в даній галузі, однак він ще не є ідеальним підходом: можна застосувати більше алгоритмів машинного навчання, наприклад, глибоке навчання або спробувати оптимізувати гіперпараметри за допомогою алгоритмів метаевристичного пошуку. Тим паче наразі дана модель недостатньо інтерпретована. [4]

4. Експериментальні дослідження

Експерименти проводилися на двох наборах даних: FF++ та DFDC [12]. FF ++ — це великий набір даних по маніпуляціях з обличчям, створений з використанням state-of-the- art методів редагування відеозаписів. Цей набір даних містить два класичних підходи маніпуляції обличчями,а саме Face2Face і FaceSwap, разом з двома стратегіями, ґрунтованими на навчанні (DeepFake і NeuralTextures). Кожен метод застосовувався до 1000 високоякісних відеозаписів, завантажених з YouTube, щоб показувати зображення без перешкод і зайвих об'єктів. Усі послідовності містили не менше 280 кадрів. Для імітації реалістичних налаштувань відеозаписи було стиснено з використанням кодека H.264. Відеозаписи високої та низької якості генерувалися з використанням параметра квантування з постійною швидкістю, рівною 23 і 40 відповідно.

Проведемо дослідження та оберемо найпідходящий спосіб налаштування гіперпараметрів для алгоритму випадковий ліс, написаного задля аналізу медичних та психологічних показників.

У дослідженні альтернативами виступатимуть глобальний пошук за сіткою, випадковий пошук, байєсівська оптимізація.

Для цього поділимо дослідження на дві складові: теоретичну та практичну. Застосуємо згортку для них.

4.1. Теоретична складова

Критеріями розгляду, за якими будуть будуватись три моделі, будуть такі атрибути, як: age – вік, sex – стать, on_thyroxine – чи приймає тироксин, query_on_thyroxine – запит на тироксин, on_antithyroid_meds – чи приймає антитиреоїдні ліки, sick – чи хворий, pregnant – вагітність, thyroid_surgery – чи робилась операція на щитоподібній, I131_treatment – лікування радіоактивним йодом, query_hypothyroid – підозра на гіпотиреоз, query_hyperthyroid – підозра на гіпертиреоз, lithium – чи приймає літій, goitre – зоб, tumor – пухлина, hypopituitary – гіпопітуїтаризм, psych – чи є психічні розлади, TSH_measured – чи вимірювався TSH, TSH – значення TSH, T3_measured – чи вимірювався T3, T3 – значення T3, TT4_measured – чи вимірювався TT4, TT4 – значення TT4, T4U_measured – чи вимірювався T4U, T4U – значення T4U, FTI_measured – чи вимірювався FTI, FTI – значення FTI, TBG_measured – чи вимірювався TBG, TBG – значення TBG, referral_source – джерело направлення, target – цільовий клас, patient_id – ID пацієнта.

Побудуємо спочатку таблицю теоретичних відомостей трьох моделей (див. табл. 1).

Таблиця 1

Характеристики способів налаштування гіперпараметрів

Метод	Витрати часу	Обчислювальна складність	Гарантія оптимуму	Гнучкість	Простота реалізації	Використання ресурсів
Grid Search	Високі	Експоненційна	Так, якщо оптимальні параметри є в сітці	Низька	Висока	Високі
Random Search	Середні	Лінійна або нижча, ніж у Grid Search	Ні, але може знайти хороший набір параметрів	Висока	Висока (легко реалізується)	Менші, ніж у Grid Search
Bayesian Optimization	Низькі (порівняно з Grid/Random)	Середня	Висока ймовірність знаходження оптимуму	Висока	Середня	Оптимізоване

Наступним кроком експерименту є конвертування якісних показників у кількісні.

Чим менше часу необхідно для роботи алгоритму, тим краще. Якщо витрати часу високі, то оцінкою буде 1 бал, якщо середні – 2 бали, а коли низькі в порівнянні із глобальним пошуком за сіткою та випадковим способом – 1 бал.

Розглянемо, наскільки важко проводити розрахунки. У випадку, коли обчислювальна складність експоненційна, тобто залежить від кількості параметрів, то дана характеристика описується як 1 бал, якщо ж середня – 2 бали, а коли – лінійна – 3 бали.

Якщо гарантія оптимуму достовірна, то це – 2 бали, якщо ні – 0, а у випадку, коли є висока ймовірність, – 1 бал.

Гнучкість може бути низькою (1 балів) та високою (2 бали).

Чим простіше, тим реалізація менш складна. Якщо простота реалізації висока, то це – 1 бал, а коли середня – 2 бали.

Якщо ж застосування ресурсів високе, то це 1 бал, коли воно краще, ніж «високе» – 2 бали, коли оптимізоване – 3 бали.

Заповнимо нову таблицю з кількісними даними способів налаштувань гіперпараметрів (див. табл. 2).

Таблиця 2

Кількісні показники способів налаштування гіперпараметрів

Метод	Витрати часу	Обчислювальна складність	Гарантія оптимуму	Гнучкість	Простота реалізації	Використання ресурсів
Grid Search	1	1	3	1	2	1
Random Search	2	3	1	2	2	2
Bayesian Optimization	3	2	2	2	1	3

Обчислимо значення згортки для кожного способу та визначимо найкращий із них (див. табл. 3).

Таблиця 3

Результати

Метод	Результати згортки
Grid Search	1,6
Random Search	2,13333333
Bayesian Optimization	2,26666667

4.2. Практична складова

Додамо четвертий спосіб, який теж не рідше використовується, а саме – комбінацію рандомізованого способу та байєсівської оптимізації.

Написаний код на python показав, що чотири алгоритми мають такі показники та одразу обчислимо значення лінійної адитивної згортки (див. табл. 4):

Таблиця 4

Числові характеристики алгоритмів

Метод	Час	Accuracy	Precision	Recall	F1	Згортка
Grid Search	6,85	0,9427	0,6106	0,5611	0,5757	1,115054794
Random Search	11,9	0,9427	0,6106	0,5611	0,5757	1,196717149
Bayesian Optimization	21,69	0,9427	0,6139	0,5491	0,5654	1,346458789
Random + Bayes	21,4	0,9427	0,6139	0,5491	0,5654	1,341769268

Результати лінійної адитивної згортки показують, найкращим способом налаштування гіперпараметрів є байєсівська оптимізація. Даний метод бере до уваги попередні показники для того, щоб будувати моделі функції втрат та ефективніше обирає наступні параметри. Порівняно з алгоритмами глобального пошуку за ставкою та рандомізованого пошуку, його витрати часу низькі. Звичайно, його обчислювальна складність та простота реалізації середні, адже залежать від обраної моделі аналізу та необхідні спеціальні бібліотеки для реалізації, але гарантія оптимуму та гнучкість високі, адже байєсівська оптимізація здатна пристосуватися до параметрів. Результати показують, що подальший розвиток оптимізацію алгоритму випадковий ліс на основі медичних та психологічних даних слід розпочинати з налаштування гіперпараметрів за допомогою байєсівської оптимізації.

Наукова новизна отриманих результатів полягає у встановленні ефективності байєсівської оптимізації та її комбінації з рандомізованим пошуком у задачі передбачення психологічних розладів, пов'язаних із порушеннями функції щитоподібної залози, що не була предметом спеціальних порівняльних досліджень у відкритих джерелах. Отримано нові результати щодо співвідношення точності та часу обчислень для різних стратегій оптимізації гіперпараметрів Random Forest у медичних задачах.

Практична значущість результатів показує, що полягає у можливості застосування оптимізованої моделі Random Forest як інструмента раннього виявлення ризику психологічних порушень у пацієнтів із гіпо- та гіпертиреозом. Запропонована методика налаштування гіперпараметрів може бути впроваджена у медичні інформаційні системи, скринінгові програми та системи підтримки клінічних рішень для підвищення точності діагностики та зменшення навантаження на медичних фахівців.

Байєсівська оптимізація вважається ще новим інструментом налаштування гіперпараметрів та загальної оптимізації функцій типу «чорного ящику». Вже багато досліджень, присвячених цьому методу, відображають все більше його застосувань [6].

Гіперпараметри відіграють важливу роль в ефективності роботи моделей машинного навчання. Сьогодення налічує чимало відомих алгоритмів, які застосовуються майже в будь-якій галузі, що потребує професійності та відповідного досвіду, тому задля успішної результативності вибір гіперпараметрів вкрай важливий. Від їх значень залежать остаточні показники, надані моделлю аналізу. Тому їхня оптимізація конче потрібна в реалізації будь-яких алгоритмів машинного навчання. Модель випадкового лісу має визначати прогнози якомога коректніше, адже вони впливатимуть на подальше спостереження за здоров'ям пацієнтів, хворих на гіпотиреоз та гіпертиреоз. Тому проблема правильного налаштування гіперпараметрів є оптимізаційною багатокритеріальною задачею вибору, розв'язком якого став метод байєсівської оптимізації.

Даний спосіб заснований на теоремі Байєса, що виходить з її назви. Вона визначає значення апостеріору над функцією оптимізації та збирає дані із попередніх підмножин даних, щоб оновити показник апостеріор. Функція корисності обирає наступну точку в даних для того, щоб максимізувати значення функції оптимізації. [7-8]

Отже, вдосконаленням алгоритму випадковий ліс буде новий комбінований спосіб із таких методів, як балансування класів, зменшення кореляції між деревами та рішень та налаштування гіперпараметрів байєсівською оптимізацією.

Висновки

Випадковий ліс – один із найпотужніших інструментів машинного навчання, який широко застосовується в різних сферах як механізм класифікації та прогнозування. Не дивлячись на чималу кількість переваг даної моделі, вона має передумови для свого удосконалення [9].

Метою дослідження було визначити способи налаштування гіперпараметрів алгоритму випадковий ліс задля одного із способів оптимізації точних результатів прогнозування розвитку психологічних розладів серед людей, хворих на гіпотиреоз та гіпертиреоз. Переглянуто спостереження, які теж були спрямовані на це.

Налаштування гіперпараметрів має декілька варіантів, тому було проведено окреме дослідження: розв'язувалась багатокритеріальна задача вибору методу налаштування гіперпараметрів за допомогою лінійної адитивної згортки.

Альтернативами були рандомізований пошук, пошук за сіткою, байєсівська оптимізація комбінований спосіб із рандомізованим пошуком та байєсівською оптимізацією, а критеріями вибору витрати часу, обчислювальна складність, гарантія оптимуму, гнучкість, простота реалізації, використання ресурсів, accuracy, precision, recall та f1. Після розрахунків згортки визначено, що найоптимальнішим варіантом є саме байєсівська оптимізація [10].

Тому найкращим методом реалізації випадкового лісу є алгоритм випадковий ліс із налаштуванням гіперпараметрів за допомогою байєсівської оптимізації.

Список літератури:

[1] Salman, H.A., Kalakech, A. i Steiti, A. (2024). Random Forest Algorithm Overview. Babylonian Journal of Machine Learning, 2024, с. 69–79. DOI: https://doi.org/10.58496/BJML/2024/007.

[2] Thomas, N.S. та Kaliraj, S. (2024). An Improved and Optimized Random Forest Based Approach to Predict the Software Faults. SN Computer Science, 5(5), с. 530. DOI: 10.1007/s42979-024-02764-x.

[3] Yang, C., Wang, Y., Zhang, A., Fan, H. та Guo, L. (2023). A Random Forest Algorithm Combined with Bayesian Optimization for Atmospheric Duct Estimation. Remote Sensing, 15(17), с. 4296. DOI: https://doi.org/10.3390/rs15174296.

[4] Zhu, N., Zhu, C., Zhou, L., Zhu, Y., & Zhang, X. (2022). Optimization of the Random Forest Hyperparameters for Power Industrial Control Systems Intrusion Detection Using an Improved Grid Search Algorithm. Applied Sciences, 12(20), 10456. https://doi.org/10.3390/app122010456

[5] Huliiiev, N. (2025). Study of methods for constructing decision trees for the implementation of the random forest algorithm in the medical field. Measuring and Computing Devices in Technological Processes, (1), 36–43. DOI: https://doi.org/10.31891/2219-9365-2025-81-5.

[6] V. Nguyen, "Bayesian Optimization for Accelerating Hyper-Parameter Tuning," 2019 IEEE Second International Conference on Artificial Intelligence and Knowledge Engineering (AIKE), Sardinia, Italy, 2019, pp. 302-305, doi: 10.1109/AIKE.2019.00060.

[7] Wu, J., Chen, X.-Y., Zhang, H., Xiong, L.-D., Lei, H. та Deng, S.-H. (2019). Hyperparameter optimization for machine learning models based on Bayesian optimization. Journal of Electronic Science and Technology, 17(1), с. 26–40. DOI: https://doi.org/10.11989/JEST.1674-862X.80904120.

[8] Lujan-Moreno, G.A., Howard, P.R., Rojas, O.G. та Montgomery, D.C. (2018). Design of experiments and response surface methodology to tune machine learning hyperparameters, with a random forest case-study. Expert Systems With Applications, 109, с. 195–205. DOI: https://doi.org/10.1016/j.eswa.2018.05.024.

[9] Siji George C G and B.Sumathi. “Grid Search Tuning of Hyperparameters in Random Forest Classifier for Customer Feedback Sentiment Prediction”. International Journal of Advanced Computer Science and Applications (IJACSA) 11.9 (2020). http://dx.doi.org/10.14569/IJACSA.2020.0110920

[10] Xilu Wang, Yaochu Jin, Sebastian Schmitt, and Markus Olhofer. 2022. Recent Advances in Bayesian Optimization. In Proceedings of ACM Conference (Conference’17). ACM, New York, NY, USA, 25 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

Надійшла до редколегії 04.12.2025