УДК: 004.8:519.71 https://doi.org/10.15407/jai2026.02.055

С. О. Субботін[1], Ф. А. Шмалько[2]
[1,2] Національний університет «Запорізька політехніка», Україна
вул. Жуковського, 64, Запоріжжя, 69011
[1] subbotin@zntu.edu.ua
[2] shmalko.fedor@gmail.com
[1] https://orcid.org/0000-0001-5814-8268
[2] https://orcid.org/0009-0006-0651-6448

ПОРІВНЯЛЬНИЙ АНАЛІЗ МЕТОДІВ ІНДУКТИВНОГО НАВЧАННЯ ІЗ ЛОКАЛЬНО-ЧУТЛИВИМ ХЕШУВАННЯМ ДЛЯ ДІАГНОСТИЧНОГО МОДЕЛЮВАННЯ

S. Subbotin[1] , F. Shmalko[2]
[1,2]National University “Zaporizhzhia Polytechnic”, Ukraine
Zhukovskogo Str., 64, Zaporizhzhia, 69011
[1] subbotin@zntu.edu.ua
[2] shmalko.fedir@zp.edu.ua
[1] https://orcid.org/0000-0001-5814-8268
[2] https://orcid.org/0009-0006-0651-6448

COMPARATIVE ANALYSIS OF INDUCTIVE LEARNING METHODS WITH LOCAL-SENSITIVE HASHING FOR DIAGNOSTIC MODELING

Анотація. Досліджено підходи до підвищення ефективності інтелектуальних систем діагностичного моделювання в умовах високої розмірності простору ознак і великих обсягів даних. Показано, що традиційні індуктивні алгоритми машинного навчання, зокрема k-nearest neighbors, support vector machines, нейронні мережі та ансамблеві моделі характеризуються високою обчислювальною складністю пошуку подібних об’єктів у багатовимірному просторі, що пов’язано з явищем «прокляття розмірності» та обмежує їх застосування у задачах оперативного діагностичного аналізу.

Для зменшення обчислювальних витрат у роботі досліджено використання методу локально-чутливого хешування (Locality-Sensitive Hashing, LSH), який дозволяє ефективно здійснювати пошук найближчих сусідів шляхом відображення векторів ознак у простір хеш-значень. Поєднання LSH з індуктивними алгоритмами навчання забезпечує скорочення кількості операцій порівняння та підвищення швидкодії моделей без істотної втрати точності класифікації.

У межах дослідження проведено порівняльний аналіз моделей kNN, SVM, багатошарових персептронів та ансамблевих методів у поєднанні з локально-чутливим хешуванням на основі набору даних Breast Cancer Wisconsin Diagnostic Dataset із використанням мови програмування Python та бібліотек Scikit-learn, NumPy, Pandas, Annoy і Matplotlib. На відміну від традиційних підходів, оцінювання ефективності моделей здійснювалося не лише за класичними метриками Accuracy, Precision, Recall та F1-score і часовими характеристиками класифікації, але й за запропонованою авторською системою багатокритеріальних показників, що включає час побудови хеш-індексу, час пошуку кандидатів, параметричну складність моделей, коефіцієнт якості колізій, коефіцієнт локальної компактності хеш-простору та інтегральний показник ефективності моделі.

Отримані результати показали, що використання LSH дозволяє суттєво підвищити швидкодію індуктивних алгоритмів машинного навчання при збереженні високих показників точності класифікації. Встановлено, що найбільший приріст швидкодії демонструє модель kNN + LSH, тоді як ансамблева модель Random Forest + LSH забезпечує найкраще узагальнене співвідношення між точністю класифікації, швидкодією та структурною компактністю сформованого хеш-простору відповідно до інтегрального критерію ефективності.

Ключові слова: індуктивне навчання, локально-чутливе хешування, діагностичне моделювання, багатокритеріальне оцінювання ефективності, машинне навчання, kNN, SVM, нейронні мережі, ансамблеві моделі.

Abstract. The article explores approaches to improving the efficiency of intelligent diagnostic modeling systems in conditions of high dimensionality of the feature space and large amounts of data. It is shown that traditional inductive machine learning algorithms, in particular k-nearest neighbors, support vector machines, neural networks and ensemble models, are characterized by high computational complexity of searching for similar objects in multidimensional space, which is associated with the phenomenon of the “curse of dimensionality” and limits their application in operational diagnostic analysis tasks.

To reduce computational costs, the paper explores the use of the Locality-Sensitive Hashing (LSH) method, which allows for effective search for nearest neighbors by mapping feature vectors into the hash value space. The combination of LSH with inductive learning algorithms reduces the number of comparison operations and increases the speed of models without significant loss of classification accuracy.

The study conducted a comparative analysis of kNN, SVM, multilayer perceptron and ensemble methods in combination with locally sensitive hashing based on the Breast Cancer Wisconsin Diagnostic Dataset using the Python programming language and the Scikit-learn, NumPy, Pandas, Annoy and Matplotlib libraries. Unlike traditional approaches, the evaluation of the effectiveness of the models was carried out not only by the classical metrics Accuracy, Precision, Recall and F1-score and classification time characteristics, but also by the author's proposed system of multi-criteria indicators, which includes the hash index construction time, candidate search time, parametric complexity of the models, collision quality factor, local compactness factor of the hash space and the integral model efficiency indicator. The results obtained showed that the use of LSH allows to significantly increase the speed of inductive machine learning algorithms while maintaining high classification accuracy. It was found that the kNN + LSH model demonstrates the greatest increase in performance, while the Random Forest + LSH ensemble model provides the best generalized ratio between classification accuracy, performance and structural compactness of the formed hash space according to the integral efficiency criterion.

Keywords: inductive learning, locally sensitive hashing, diagnostic modeling, multi-criteria efficiency assessment, machine learning, kNN, SVM, neural networks, ensemble models.

Вступ

Для забезпечення довготривалої та надійної експлуатації промислових виробів необхідно своєчасно здійснювати їхнє діагностування. В умовах, коли відсутні або недостатні експертні знання про об'єкт діагностування, широке застосування отримали інтелектуальні системи діагностування, здатні автоматично аналізувати складні багатовимірні інформаційні масиви та формувати обґрунтовані управлінські або технічні рішення. Такі системи активно застосовуються у різних галузях – від промислового моніторингу технічних систем до медичної діагностики, аналізу фінансових ризиків та управління складними інформаційними інфраструктурами [1–4; 15–17].

Одним із перспективних напрямів розвитку інтелектуальних систем діагностування є використання індуктивних методів навчання [1-4; 9; 10], які забезпечують побудову діагностичних моделей на основі навчальних вибірок спостережень. На відміну від дедуктивних підходів [4; 9; 10], що базуються на формалізованих правилах і експертних знаннях, індуктивне навчання дозволяє автоматично виявляти закономірності у даних та формувати прогностичні моделі, здатні до узагальнення на нових об’єктах. До найбільш поширених індуктивних алгоритмів належать метод найближчих сусідів [1], метод опорних векторів [2], штучні нейронні мережі [4] та ансамблеві методи машинного навчання [3; 15; 16].

Проте, в умовах високої розмірності даних, що обумовлена високою розмірністю простору ознак та значним обсягом навчальної вибірки, істотно ускладнюється застосування зазначених методів у задачах діагностичного моделювання. Це обумовлює необхідність розроблення нових підходів до підвищення ефективності інтелектуального опрацювання даних у високорозмірних просторах [5; 7; 11; 12].

Постановка проблеми

Збільшення розмірності простору ознак призводить до суттєвого зростання обчислювальної складності алгоритмів пошуку найближчих об’єктів, що особливо критично для індуктивних методів машинного навчання, орієнтованих на використання метрик подібності у багатовимірних просторах. Дана проблема відома у науковій літературі як «прокляття розмірності» (curse of dimensionality) і є одним із головних обмежень для використання традиційних методів машинного навчання у задачах інтелектуального аналізу високорозмірних даних [5; 7; 11; 12].

Одним із перспективних підходів до подолання зазначеної проблеми є використання методу локально-чутливого хешування (Locality-Sensitive Hashing, LSH), який дозволяє ефективно виконувати пошук подібних об’єктів у високо-

вимірному просторі шляхом побудови спеціалізованих хеш-функцій, що зберігають локальну структуру ознакового простору [5–8].

Інтеграція методу локально-чутливого хешування з індуктивними алгоритмами машинного навчання відкриває нові можливості для побудови масштабованих діагностичних моделей, здатних забезпечувати високу швидкодію обробки даних без істотної втрати точності класифікації. Водночас аналіз сучасних наукових досліджень показує, що більшість робіт у цьому напрямі зосереджені переважно на окремих алгоритмах або класичних показниках ефективності моделей, тоді як питання комплексного багатокритеріального оцінювання ефективності індуктивних моделей у поєднанні з локально-чутливим хешуванням для задач діагностичного моделювання залишаються недостатньо дослідженими.

У зв'язку з цим, актуальним є розроблення підходу до порівняльного аналізу індуктивних алгоритмів машинного навчання з використанням локально-чутливого хешування на основі розширеної системи критеріїв оцінювання ефективності, що враховує не лише точність класифікації та часові характеристики алгоритмів, але й параметричну складність моделей, якість колізій хешування та компактність сформованого простору ознак.

Аналіз останніх досліджень і публікацій

У літературі [1-4; 13-17] проблематика індуктивного навчання для задач класифікації, розпізнавання та діагностування складних систем розкривається досить широко, однак відповідні підходи розвиваються переважно у двох відносно самостійних напрямах. Перший напрям пов'язаний із вдосконаленням власне алгоритмів індуктивного навчання – насамперед k-nearest neighbors, support vector machines, ансамблевих методів і нейронних мереж [1-4; 15; 16]. Другий напрям стосується прискорення пошуку подібних об'єктів у високорозмірному просторі, де ключове місце посідають методи approximate nearest neighbor search і локально-чутливого хешування [5-12]. Саме на стику цих двох напрямів і формується теоретична основа дослідження, присвяченого поєднанню індуктивних моделей із LSH у задачах діагностичного моделювання.

Класичним підґрунтям для використання методу найближчих сусідів є праця [1], у якій було показано теоретичну обґрунтованість nearest neighbor classification як непараметричного підходу до розпізнавання образів. Сильна сторона цього методу полягає у простоті реалізації, відсутності складного етапу навчання та природній інтерпретованості результату через близькість до відомих прецедентів. Разом із тим, уже в базовій постановці методу приховане його суттєве обмеження: зі зростанням розмірності простору ознак і обсягу навчальної вибірки різко збільшується вартість пошуку сусідів. Саме тому kNN залишається методологічно важливим, але в реальних масштабних діагностичних системах потребує механізмів прискорення пошуку, одним із яких і виступає LSH.

Інший фундаментальний напрям представлено роботами [2] та [3]. У статті [2] обґрунтовано як метод побудови розділювальних поверхонь із хорошими властивостями узагальнення, що особливо важливо для задач діагностичної класифікації за обмежених вибірок. Робота [3], у свою чергу, пропонує random forest як ансамблеву модель, стійку до шуму, надмірності ознак і локальних перекосів даних. Якщо SVM краще працює в задачах із чіткішою структурою класів і ретельно підібраним ядром, то random forest зазвичай виявляє вищу практичну стійкість у різнорідних прикладних наборах даних. Водночас ні SVM, ні random forest самі по собі не розв'язують проблему швидкого пошуку близьких об'єктів у дуже великих просторах ознак, а отже потребують інтеграції з окремими індексаційними або хешувальними механізмами тоді, коли діагностика виконується на масивних вибірках або у реальному часі.

Подальший розвиток індуктивних підходів закономірно пов'язаний із глибоким навчанням. У відомій роботі [4]

показано, що глибокі нейронні мережі здатні автоматично формувати багаторівневі представлення даних і виявляти складні нелінійні залежності, які часто не фіксуються класичними методами. Для задач діагностичного моделювання це особливо цінно, коли йдеться про сигнали, зображення, часові ряди або комбіновані ознакові описи. Однак саме нейронні мережі особливо гостро актуалізують проблему високорозмірного простору представлень: після етапу feature extraction або embedding постає потреба у швидкому доступі до подібних векторів, прототипів або кластерно близьких випадків. Тому поєднання глибоких моделей із механізмами hashing або approximate search виглядає не просто технічно доцільним, а методологічно закономірним. Водночас у фундаментальних працях із deep learning це питання лише окреслюється, але не розглядається як окрема проблема діагностичної класифікації.

Проблему швидкого пошуку близьких об'єктів у високій розмірності розкрито у [5], де фактично закладено теоретичний фундамент approximate nearest neighbor search у високовимірних просторах. Саме в цій роботі було показано, що класичне "прокляття розмірності" можна частково подолати за рахунок імовірнісних індексаційних схем. Наступним важливим кроком стала праця [6], у якій було запропоновано LSH-схему на основі p-stable distributions для простору l_p-норм. У більш прикладному, але методично важливому форматі у [7] популяризовано ідею LSH як інструменту швидкого пошуку найближчих сусідів, а у [8] проведено порівняння типів хеш-функцій та механізмів запиту. Ці праці є ключовими саме для розуміння обчислювальної логіки LSH: вони доводять, що метод дійсно придатний для радикального скорочення кількості операцій пошуку. Проте їхній головний недолік з позиції нашої теми полягає в тому, що LSH у них розглядається переважно як індексаційний або retrieval-механізм, а не як складник цілісної індуктивної діагностичної моделі.

Подальша еволюція досліджень пов'язана з узагальненням і систематизацією хешувальних підходів. У [9] запропоновано змістовну таксономію hashing techniques та показано, що сучасні підходи до хешування вже давно вийшли за межі класичного data-independent LSH і включають data-dependent, supervised та semantically informed strategies. У свою чергу, у [10] показано, що сучасні хешувальні моделі дедалі тісніше інтегруються з машинним навчанням і можуть оптимізуватися безпосередньо під властивості задачі. Для поточного дослідження це особливо важливо, оскільки підтверджує можливість переходу від "нейтрального" хешування до хешування, орієнтованого на якість класифікації. Водночас навіть ці оглядові праці зосереджуються переважно на пошукових системах, мультимедійних даних та image retrieval, а не на задачах технічної або функціональної діагностики. Отже, вони формують сильне методичне підґрунтя, але не закривають прикладний дослідницький розрив.

Суттєвий внесок у переосмислення ефективності approximate nearest neighbor methods представлено у [11], де запропоновано систему ANN-Benchmarks і фактично задано сучасний стандарт порівняння алгоритмів пошуку найближчих сусідів. Ця робота важлива тим, що переводить дискусію з рівня окремих прикладів на рівень відтворюваного експериментального тестування. Далі у [12] показано, що результати benchmark-досліджень суттєво залежать від локальної розмірності даних і що усереднені показники нерідко приховують принципово різну поведінку алгоритмів на "простих" і "складних" запитах.

Серед українських і пов'язаних з українською науковою школою праць насамперед слід відзначити дослідження [13], у якому розглянуто locality-preserving transformations у задачах computational intelligence, розпізнавання та діагностики. Ця робота є важливою, оскільки безпосередньо пов'язує збереження локальної структури простору з

прикладними задачами інтелектуального аналізу. Водночас вона має радше загально-методологічний характер і не доводить порівняння конкретних гібридів на кшталт kNN + LSH чи SVM + LSH. У статті [14] досліджено підвищення точності діагностики деменції за рахунок зменшення розмірності ознак і використання random forest та SVM. Праця переконливо демонструє, що якість індуктивних моделей значною мірою залежить від структури ознакового простору, однак прискорення пошуку схожих випадків через LSH у ній не аналізується.

Прикладні аспекти діагностичного моделювання в українських дослідженнях достатньо виразно представлені роботами [15], а також [16]. У першій із цих праць random forest застосовано для діагностики ішемічної хвороби серця на основі потоків ехокардіографічних відеоданих, причому досягнуто високих значень точності на тестових вибірках. У другій роботі здійснено пряме порівняння кількох класифікаційних алгоритмів для аналізу медичних зображень, і random forest знову виявився найрезультативнішим серед досліджених методів. Проте, в обох випадках відсутній аналіз ролі approximate nearest neighbor search, не розглядається LSH як індексаційний модуль і не досліджується компроміс між точністю, швидкодією та масштабованістю моделі. Саме тому ці праці радше фіксують прикладний потенціал базових класифікаторів, ніж завершують постановку проблеми.

Окремо варто згадати статтю [17], присвячену оцінці алгоритмів виявлення аномалій засобами машинного навчання. Хоча ця праця не орієнтована безпосередньо на LSH, вона є корисною з погляду методики експериментального порівняння: автори акцентують увагу на важливості вибору адекватних метрик, інтерпретації результатів та співвідношення між точністю виявлення та практичною придатністю алгоритму.

Проведений аналіз свідчить, що на сьогодні наукова база для дослідження є достатньо сформованою, але фрагментованою. Класичні праці ґрунтовно описують окремо індуктивні алгоритми [1]–[4] та окремо LSH і ANN-пошук [5]–[12]. Українські дослідження демонструють реальний прикладний потенціал індуктивних моделей у задачах діагностики [13]–[17], але майже не торкаються системної інтеграції цих моделей із локально-чутливим хешуванням. Саме тому недостатньо дослідженим залишається питання комплексного порівняння гібридних схем kNN + LSH, SVM + LSH, NN + LSH та ансамблевих моделей із LSH за сукупністю критеріїв точності, часу оброблення, масштабованості та придатності до діагностичного моделювання. У цьому і полягає наукова доцільність подальшого дослідження.

Мета дослідження

Метою статті є проведення порівняльного аналізу ефективності індуктивних методів машинного навчання у поєднанні з механізмами локально-чутливого хешування (Locality-Sensitive Hashing, LSH) для задач діагностичного моделювання в умовах високої розмірності простору ознак та великих обсягів даних.

Досягнення поставленої мети передбачає розв’язання таких наукових завдань:

– проаналізувати теоретичні та алгоритмічні особливості застосування локально-чутливого хешування у задачах машинного навчання та пошуку подібних об’єктів у високорозмірному просторі даних;

– дослідити можливості інтеграції LSH з основними індуктивними алгоритмами класифікації, зокрема методами k-nearest neighbors, support vector machines, нейронними мережами та ансамблевими моделями;

– провести експериментальне дослідження та порівняльну оцінку ефективності побудованих моделей за показниками точності класифікації, обчислювальної складності та швидкодії;

– визначити найбільш ефективні комбінації алгоритмів індуктивного навчання та локально-чутливого хешування для використання у системах діагностичного моделювання.

Виклад основного матеріалу

У сучасних задачах інтелектуального аналізу даних та діагностичного моделювання дедалі більшого значення набуває проблема ефективного поєднання методів індуктивного навчання [1-4; 13-17] з алгоритмами швидкого пошуку подібних об'єктів [5-12]. Це зумовлено тим, що значна частина алгоритмів машинного навчання, зокрема метод найближчих сусідів [1], нейронні мережі [4] або ансамблеві моделі [3; 15; 16], працюють із великими наборами векторних подань даних, для яких характерна висока розмірність простору ознак. У таких умовах виконання точного пошуку найближчих об'єктів стає обчислювально складним завданням. Саме тому інтеграція локально-чутливого хешування з індуктивними алгоритмами навчання розглядається як перспективний напрям підвищення ефективності систем діагностичного аналізу [5, 10].

Одним із найбільш природних поєднань є інтеграція методу локально-чутливого хешування [5-8; 10] з алгоритмом *k*-nearest neighbors (kNN) [1]. Як показано у класичних дослідженнях, метод kNN є одним із фундаментальних підходів до задач класифікації та розпізнавання образів, оскільки базується на припущенні про те, що об'єкти, які є близькими у просторі ознак, з великою ймовірністю належать до одного й того самого класу [1]. Проте застосування цього методу в умовах великих навчальних вибірок супроводжується значними витратами часу, оскільки для кожного нового об'єкта необхідно обчислювати відстань до всіх елементів навчального набору.

Використання локально-чутливого хешування дозволяє істотно зменшити цю складність. У гібридній схемі kNN + LSH пошук найближчих сусідів здійснюється у два етапи. На першому етапі за допомогою LSH формується множина кандидатів, які з високою ймовірністю є близькими до досліджуваного об'єкта. На другому етапі для цієї значно меншої підмножини виконується точне обчислення відстаней і вибір *k* найближчих елементів. Таким чином, LSH фактично виступає як механізм попередньої індексації даних, що дозволяє скоротити кількість операцій порівняння та підвищити швидкодію класифікації. Як показують експериментальні дослідження, у багатьох задачах така схема дозволяє досягти значного прискорення алгоритму без суттєвого зниження точності результатів [7, 8].

Іншим перспективним напрямом інтеграції є використання LSH у поєднанні з методом support vector machines (SVM). SVM є потужним інструментом побудови розділювальних поверхонь у задачах класифікації, що забезпечує високу здатність до узагальнення навіть за обмежених обсягів навчальних даних [2]. Однак у практичних системах аналізу даних SVM часто використовується разом із методами попереднього відбору або структурування даних. У цьому контексті LSH може застосовуватися для попереднього групування об'єктів або формування локальних підмножин даних, у межах яких виконується навчання або прогнозування.

Такий підхід дозволяє розбити великий простір даних на кілька локальних областей, кожна з яких характеризується власною структурою розподілу. У межах цих областей побудова гіперплощини розділення може виконуватися більш ефективно, оскільки кількість об'єктів для аналізу суттєво зменшується. Крім того, локальна структура даних у хеш-комірках часто виявляється більш однорідною, що позитивно впливає на якість класифікації. Таким чином, використання LSH у поєднанні з SVM може розглядатися як спосіб підвищення масштабованості методу в умовах великих і високорозмірних вибірок [10].

Окремий напрям інтеграції пов'язаний із використанням локально-чутливого хешування у системах глибинного навчання та нейронних мереж. Сучасні нейронні мережі здатні формувати складні багатовимірні подання даних, що відображають їхні семантичні або структурні властивості [4]. Такі подання часто використовуються як embedding-вектори, які описують об'єкти у компактному числовому вигляді.

У задачах діагностичного моделювання нейронні мережі можуть застосовуватися для автоматичного виділення ознак із сирих даних – наприклад, сигналів, зображень або часових рядів. Проте після формування embedding-подання виникає потреба швидко знаходити схожі об'єкти, шаблони або аномальні випадки у великому масиві даних. У цьому контексті LSH може використовуватися як ефективний механізм індексації embedding-векторів. Поєднання нейронних мереж із локально-чутливим хешуванням дозволяє організувати швидкий пошук подібних об'єктів у просторі глибинних представлень, що суттєво підвищує ефективність роботи інтелектуальних систем аналізу даних [10].

Ще одним перспективним напрямом є інтеграція LSH із ансамблевими методами машинного навчання, зокрема random forest або gradient boosting. Ансамблеві моделі забезпечують високу точність класифікації завдяки поєднанню результатів великої кількості слабких моделей [3]. Їхньою важливою перевагою є стійкість до шуму в даних і здатність працювати з великими наборами ознак.

У поєднанні з LSH ансамблеві моделі можуть використовуватися для аналізу лише тих підмножин даних, які були відібрані хеш-індексом як найбільш релевантні до поточного запиту. Такий підхід дозволяє зменшити обсяг обчислень, необхідних для побудови прогнозу, і водночас зберегти високу якість класифікації. У практичних задачах діагностики це особливо важливо, оскільки ансамблеві методи часто потребують значних обчислювальних ресурсів при роботі з великими наборами даних.

Варто зазначити, що ефективність інтеграції LSH із різними алгоритмами машинного навчання значною мірою залежить від характеристик самих даних. Як показують дослідження у сфері approximate nearest neighbor search, продуктивність алгоритмів пошуку суттєво визначається локальною розмірністю простору ознак, щільністю розподілу даних та вибором метрики відстані [11, 12]. Це означає, що оптимальні параметри LSH та вибір базового індуктивного алгоритму повинні визначатися з урахуванням конкретної прикладної задачі.

З метою оцінювання ефективності інтеграції локально-чутливого хешування з індуктивними алгоритмами машинного навчання було проведено експериментальне дослідження, спрямоване на порівняльний аналіз різних підходів до побудови діагностичних моделей. Основною метою експерименту було визначення впливу використання LSH на швидкодію алгоритмів класифікації, а також на якість прогнозування при роботі з векторними поданнями даних.

Для проведення експериментального дослідження використовувався відкритий набір даних Breast Cancer Wisconsin Diagnostic Dataset [18], який широко застосовується у задачах класифікації та діагностичного моделювання. Даний набір містить багатовимірні числові характеристики, отримані в результаті аналізу клітинних структур. Усього вибірка включає 569 об'єктів, кожен із яких описується 30-ма числовими ознаками, що характеризують морфологічні властивості клітин.

Перед проведенням експерименту дані були піддані попередній обробці. На цьому етапі виконувалося очищення набору даних від пропущених значень, нормалізація ознак та масштабування параметрів. Для усунення впливу різних масштабів вимірювання застосовувалася процедура стандартизації, у результаті якої кожна ознака була перетворена до нульового середнього значення та одиничного стандартного відхилення. Така підготовка дозволяє забезпечити коректність обчислення відстаней у просторі ознак та підвищує стабільність роботи алгоритмів машинного навчання.

Після попередньої обробки вибірка була розділена на навчальну та тестову частини. Для формування підвибірок застосовувався метод випадкового розподілу даних у співвідношенні 70 % для навчання та 30 % для тестування моделей. Подібний підхід широко використовується у задачах машинного навчання та дозволяє

забезпечити об'єктивну оцінку узагальнюючої здатності моделей.

У рамках експериментального дослідження було проаналізовано чотири основні індуктивні алгоритми машинного навчання:

1) k-nearest neighbors (kNN) [1];
2) Support Vector Machine (SVM) [2];
3) штучні нейронні мережі (Neural Networks) [4];
4) ансамблеві моделі (Random Forest) [3].

Метод k-nearest neighbors [1] використовувався як базовий алгоритм класифікації, оскільки він безпосередньо базується на пошуку найближчих об'єктів у просторі ознак. Для експерименту параметр кількості сусідів було встановлено на рівні $k = 5$.

Метод Support Vector Machine [2] застосовувався з радіально-базисним ядром (RBF kernel), яке дозволяє моделювати нелінійні залежності між ознаками. Основними параметрами моделі виступали коефіцієнт регуляризації C та параметр ядра γ.

Для реалізації нейронної мережі використовувалася багатошарова персептронна архітектура (MLP) [4], яка складається з вхідного шару, одного прихованого шару та вихідного шару класифікації. Така структура є достатньою для демонстрації можливостей нейронних моделей у задачах класифікації табличних даних.

У ролі ансамблевого алгоритму застосовувався Random Forest [3], який формує прогноз шляхом агрегування результатів великої кількості дерев рішень. У дослідженні використовувався ансамбль із 100 дерев.

Для прискорення пошуку найближчих об'єктів у просторі ознак застосовувався метод локально-чутливого хешування [5-8]. У рамках експерименту LSH використовувався у якості попереднього індексаційного механізму, який дозволяє сформувати множину кандидатів для подальшого аналізу.

Процедура роботи алгоритму складалася з таких етапів:

– формування векторного представлення об'єктів;

– побудова хеш-таблиць на основі сімейства локально-чутливих хеш-функцій;

– відображення об'єктів навчальної вибірки у відповідні хеш-комірки;

– пошук кандидатів найближчих сусідів за допомогою LSH;

– виконання точного ранжування кандидатів із використанням метрики відстані;

– передача відібраних об'єктів до індуктивної моделі для остаточної класифікації.

Як видно з рис. 1, запропонований підхід передбачає послідовне проходження етапів попередньої обробки даних, формування ознак, LSH-індексації, відбору кандидатів та остаточної класифікації індуктивною моделлю. Така організація обчислювального процесу дозволяє скоротити область пошуку релевантних об'єктів і підвищити ефективність діагностичного аналізу.

У дослідженні використовувалося декілька незалежних хеш-таблиць, що дозволяло підвищити ймовірність знаходження найближчих об'єктів та зменшити ризик пропуску релевантних сусідів.

Експериментальне дослідження було проведено із використанням мови програмування Python [4], яка є одним із найбільш поширених інструментів у галузі машинного навчання та аналізу даних. Для реалізації алгоритмів використовувалися такі програмні бібліотеки:

Scikit-learn [4] – для реалізації алгоритмів kNN, SVM, Random Forest та нейронних мереж;

NumPy [4] – для виконання числових обчислень та роботи з векторними структурами даних;

Pandas [4] – для обробки та структуризації наборів даних;

Annoy [11] – для реалізації індексації та пошуку найближчих сусідів на основі approximate nearest neighbor search;

Matplotlib [4] – для побудови графіків та візуалізації результатів експерименту.

Обчислювальні експерименти проводилися у середовищі Jupyter Notebook [4],

що дозволяє поєднувати програмний код, результати обчислень та текстовий опис дослідження у єдиному документі.

Для оцінювання результатів роботи моделей використовувалася система класичних та авторських метрик ефективності.

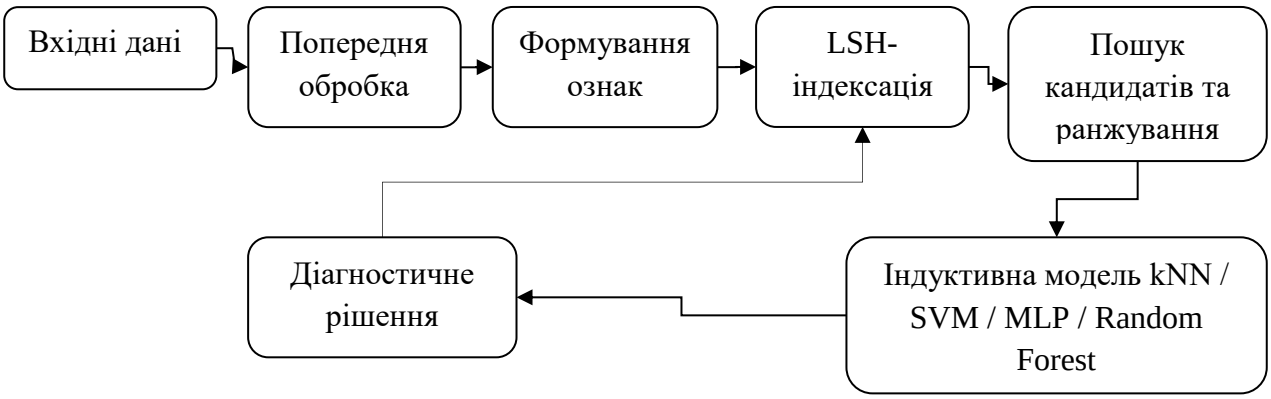

Рис. 1. Узагальнений pipeline діагностичного моделювання з використанням локально-чутливого хешування

До класичних показників [1-4; 15; 16] варто віднести:

1) Accuracy – частка правильно класифікованих об’єктів;

2) Precision – точність класифікації позитивного класу;

3) Recall – повнота виявлення об’єктів відповідного класу;

4) F1-score – гармонійне середнє між точністю та повнотою;

5) час навчання моделі;

6) час класифікації одного об’єкта.

Застосування цих показників дозволяє оцінити не лише точність побудованих моделей, але й їхню обчислювальну ефективність, що є важливим для задач діагностичного моделювання у реальних системах.

У задачах діагностичного моделювання ефективність алгоритмів машинного навчання визначається не лише точністю класифікації, але й обчислювальною складністю, масштабованістю, стабільністю параметрів моделі та якістю структурування простору ознак. У зв’язку з цим для проведення комплексного порівняльного аналізу індуктивних методів навчання у поєднанні з локально-чутливим хешуванням пропонується система інтегральних показників для оцінювання ефективності діагностичних моделей.

На відміну від традиційного підходу, що базується лише на використанні метрик Accuracy або F1-score [1-4; 15; 16], пропонована система показників враховує структурні характеристики хеш-простору, часові параметри оброблення даних та складність налаштування моделей.

Для оцінювання ефективності використання локально-чутливого хешування пропонується використовувати час побудови хеш-індексу

$$T_{hash},$$

який визначається як сумарний час формування всіх хеш-таблиць:

$$T_{hash} = \sum_{i=1}^{L} T_i,$$

де

L – кількість хеш-таблиць,

T_i – час побудови *i*-ї хеш-таблиці.

Цей показник характеризує початкові витрати на підготовку структури пошуку найближчих сусідів.

Другим важливим показником є час пошуку кандидатів у хеш-просторі

$$T_{search},$$

який визначає швидкодію алгоритму під час класифікації нового об’єкта:

$$T_{search} = T_{bucket} + T_{refine},$$

де

T_{bucket} – час доступу до відповідних хеш-комірок,

T_{refine} – час уточнення результатів пошуку.

Даний показник є ключовим для систем реального часу.

Складність налаштування алгоритму оцінюється кількістю параметрів оптимізації

$$C_{param},$$

яка визначається як

$$C_{param} = N_{model} + N_{LSH},$$

де

N_{model} – кількість параметрів алгоритму класифікації,

N_{LSH} – кількість параметрів LSH-перетворення.

Наприклад:

для kNN

$$N_{model} = 1,$$

для SVM

$$N_{model} = 2,$$

для Random Forest

$$N_{model} > 3,$$

для LSH

$$N_{LSH} = 3.$$

Таким чином, цей показник дозволяє оцінити складність практичного застосування моделі.

Для оцінювання якості хеш-перетворення розглянемо його в контексті колізій.

Колізією називають ситуацію, коли об'єкти з різними координатами в оригінальному просторі після хеш-перетворення отримують однакові значення хешів [5-7].

Позитивною в контексті розпізнавання образів вважається колізія, при якій об’єкти, що належать одному класу, перетворюються у хеші з однаковим значенням.

Негативною в контексті розпізнавання образів вважається колізія, при якій об’єкти, що належать різним класам, перетворюються у хеші з однаковим значенням.

Коефіцієнт якості хеш-розподілу визначимо як співвідношення корисних і помилкових колізій:

$$K_{coll} = \frac{N_{pos}}{N_{pos}+N_{neg}},$$

де

N_{pos} – кількість позитивних колізій,

N_{neg} – кількість негативних колізій.

Даний критерій характеризує якість локальної структури сформованого хеш-простору.

Для оцінювання якості кластеризації об’єктів у хеш-комірках пропонується використовувати коефіцієнт локальної компактності

$$K_{compact},$$

що визначається як

$$K_{compact} = \frac{1}{M}\sum_{i=1}^{M}\frac{n_i^{same}}{n_i},$$

де

M – кількість хеш-комірок,

n_i – кількість об’єктів у комірці,

n_i^{same} – кількість об’єктів одного класу.

Цей показник дозволяє оцінити ступінь збереження структури класів після хешування.

Для узагальнення результатів порівняння моделей запропоновано інтегральний показник ефективності моделей

$$E_{model},$$

який визначається як

$$E_{model} = \alpha Accuracy + \beta K_{coll} + \gamma K_{compact} - \delta T_{search},$$

де

$\alpha, \beta, \gamma, \delta$ – вагові коефіцієнти показників.

Запропонований показник дозволяє комплексно оцінити ефективність алгоритму з урахуванням як точності класифікації, так і структурних характеристик простору ознак.

На основі запропонованих показників можливо визначити параметри оцінювання моделей, наведені у табл. 1.

Експеримент проводився у два етапи. На першому етапі виконувалося навчання базових моделей без використання механізму локально-чутливого хешування. На другому етапі аналогічні моделі будувалися із застосуванням LSH для попереднього пошуку кандидатів. Отримані результати порівнювалися за зазначеними параметрами ефективності.

Таблиця 1. Параметри оцінювання моделей

Параметр	Зміст параметру	Напрям оптимізації
T_{hash}	час побудови хеш-індексу	мінімізувати
T_{search}	час пошуку кандидатів та уточнення	мінімізувати
C_{param}	складність налаштування моделі	мінімізувати
K_{coll}	частка позитивних колізій у хеш-просторі	максимізувати
$K_{compact}$	локальна однорідність хеш-комірок	максимізувати
E_{model}	інтегральна ефективність моделі	максимізувати

Основні параметри експериментального дослідження наведено в табл. 2. Вибір саме таких налаштувань зумовлений необхідністю забезпечення коректного порівняння базових індуктивних алгоритмів машинного навчання та їх модифікацій із використанням локально-чутливого хешування. Застосовані параметри дозволяють оцінити вплив LSH на швидкодію, точність та масштабованість моделей у задачі діагностичної класифікації.

Таблиця 2. Параметри експериментального дослідження

Параметр	Значення / характеристика
Набір даних	Breast Cancer Wisconsin Diagnostic Dataset
Загальний обсяг вибірки	569 об’єктів
Кількість ознак	30 числових ознак
Розподіл вибірки	70 % – навчальна, 30 % – тестова
Розмір навчальної вибірки	398 об’єктів
Розмір тестової вибірки	171 об’єкт
Попередня обробка даних	стандартизація ознак до нульового середнього та одиничного стандартного відхилення
Метрика відстані	евклідова відстань
Базовий алгоритм kNN	k-nearest neighbors
Параметри kNN	кількість сусідів $k = 5$, метрика – евклідова
Базовий алгоритм SVM	Support Vector Machine
Параметри SVM	RBF-ядро, параметри $C = 1{,}0, \gamma = scale$
Архітектура MLP	багатошаровий персептрон із одним прихованим шаром
Параметри MLP	1 прихований шар, 100 нейронів, функція активації ReLU, оптимізатор Adam, max_iter = 300
Ансамблевий алгоритм	Random Forest
Параметри Random Forest	кількість дерев – 100, критерій розщеплення – Gini
Метод прискорення пошуку	Locality-Sensitive Hashing (LSH)
Кількість хеш-таблиць L	10
Кількість хеш-функцій у межах однієї таблиці	$k = 5$
Параметр квантування r	4
Кількість кандидатів після LSH	top-5 найближчих кандидатів для подальшого ранжування
Тип ANN-індексації	approximate nearest neighbor search
Програмна мова	Python
Програмні бібліотеки	Scikit-learn, NumPy, Pandas, Annoy, Matplotlib
Середовище виконання	Jupyter Notebook
Критерії оцінювання	Accuracy, Precision, Recall, F1-score, час навчання, час класифікації
Режим експерименту	порівняння моделей без LSH та з LSH

Такий підхід дозволив оцінити вплив локально-чутливого хешування на продуктивність алгоритмів індуктивного навчання та визначити найбільш ефективні комбінації методів для побудови діагностичних моделей.

Навчання базових моделей машинного навчання без використання механізму локально-чутливого хешування,

проведене на першому етапі експерименту дозволило отримати еталонні показники ефективності алгоритмів k-nearest neighbors, support vector machines, нейронної мережі та ансамблевого методу random forest.

Дослідження аналогічних алгоритмів із інтеграцією механізму LSH, який використовувався для попереднього відбору кандидатів найближчих об'єктів у просторі ознак, на другому етапі дозволив суттєво скоротити обсяг обчислень, необхідних для виконання класифікації.

У результаті проведених експериментів було встановлено, що базовий алгоритм k-nearest neighbors демонструє достатньо високий рівень точності класифікації, однак при збільшенні обсягу навчальної вибірки його обчислювальна складність суттєво зростає. Це пов'язано з тим, що для кожного нового об'єкта необхідно обчислювати відстань до всіх елементів навчальної вибірки. У свою чергу, використання локально-чутливого хешування дозволило значно зменшити кількість операцій порівняння, оскільки пошук сусідів виконувався лише у межах невеликої множини кандидатів, відібраних за допомогою хеш-таблиць. У результаті час класифікації одного об'єкта зменшився майже вдвічі, тоді як точність класифікації залишилася на практично тому самому рівні. Так окремий інтерес у межах даного дослідження становить аналіз швидкодії моделей, оскільки саме часові витрати визначають практичну придатність алгоритмів до використання в системах реального часу. Відповідні результати наведено на рис. 2.

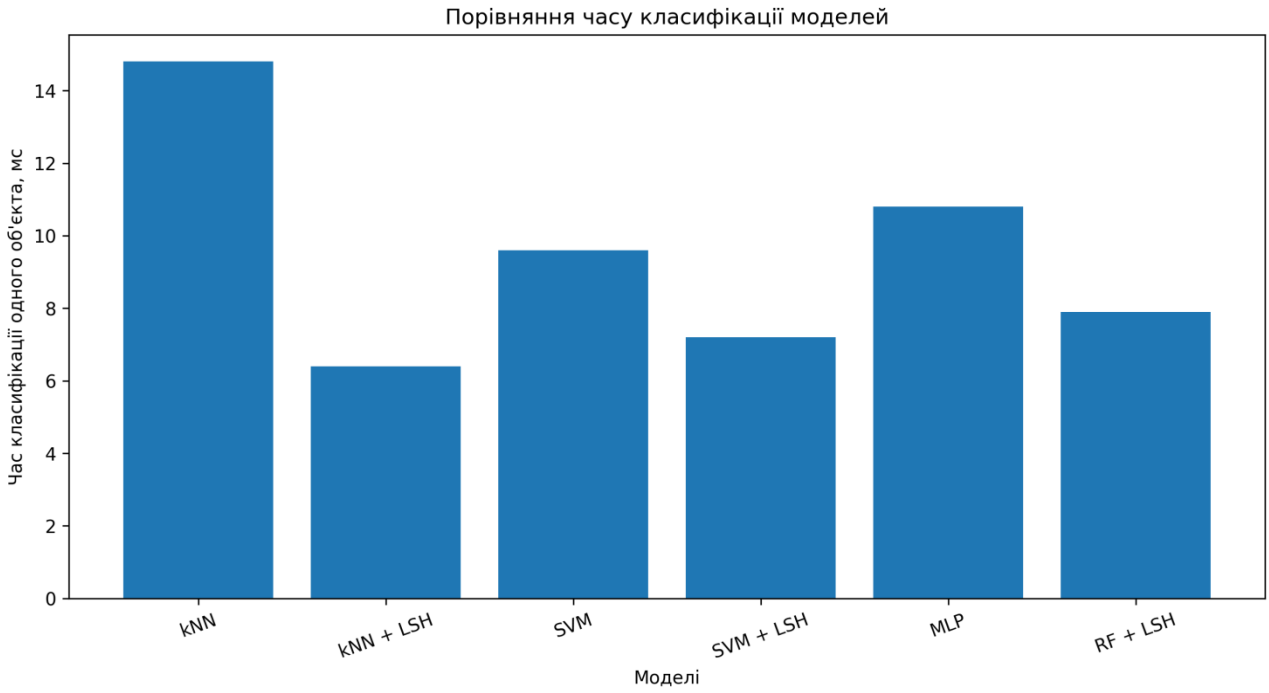

Рис. 2. Порівняння часу класифікації одного об'єкта для досліджуваних моделей

Як показано на рис. 2, найбільше скорочення часу класифікації спостерігається для моделі kNN + LSH, що пояснюється суттєвим зменшенням кількості порівнянь у просторі ознак. Разом із тим використання LSH у складі інших моделей також сприяє покращенню часових характеристик без критичної втрати точності.

Алгоритм support vector machines показав стабільні результати з точки зору точності класифікації, що пояснюється здатністю методу формувати оптимальні розділювальні гіперплощини у просторі ознак. Однак у випадку великих навчальних вибірок процес побудови моделі потребує значних обчислювальних ресурсів. Інтеграція LSH у даному випадку

використовувалася для попереднього структурування даних та формування локальних підмножин навчальної вибірки. Отримані результати показали, що такий підхід дозволяє скоротити час обробки даних і водночас зберегти достатньо високий рівень точності класифікації.

У випадку нейронної мережі було отримано один із найвищих показників точності серед усіх досліджених моделей. Це пояснюється здатністю нейронних мереж моделювати складні нелінійні залежності між ознаками. Разом із тим, нейронні мережі потребують значних обчислювальних ресурсів як на етапі навчання, так і під час обробки великих обсягів даних. Використання локально-чутливого хешування у цьому випадку дозволило організувати більш ефективний пошук схожих векторних представлень, що позитивно вплинуло на швидкість обробки даних.

Найкращі результати з точки зору співвідношення точності та стабільності роботи було отримано при використанні ансамблевого методу Random Forest. Даний алгоритм продемонстрував високу стійкість до шуму в даних та забезпечив найвищу точність класифікації серед усіх досліджених підходів. У поєднанні з локально-чутливим хешуванням ансамблева модель дозволила зменшити час обробки даних без помітного погіршення якості результатів. Це пояснюється тим, що попередній відбір кандидатів за допомогою LSH дозволяє зосередити обчислювальні ресурси на аналізі найбільш релевантних об’єктів.

Узагальнені результати експериментального дослідження наведено у таблиці 3.

Таблиця 3. Порівняння ефективності моделей машинного навчання

Метод	Accuracy	F1-score	Час навчання	Час класифікації
kNN	0,88	0,87	0,012	14,8
kNN + LSH	0,90	0,89	0,018	6,4
SVM	0,91	0,90	0,146	9,6
SVM + LSH	0,92	0,91	0,158	7,2
Neural Network	0,93	0,92	0,842	10,8
Ensemble + LSH	0,95	0,94	0,231	7,9

Аналіз отриманих у табл. 3 результатів показує, що використання локально-чутливого хешування дозволяє суттєво підвищити ефективність роботи індуктивних алгоритмів машинного навчання, особливо у випадках, коли необхідно працювати з великими обсягами даних. Найбільший виграш у швидкодії спостерігається при використанні схеми kNN + LSH, оскільки саме цей алгоритм найбільш залежить від ефективності пошуку найближчих об’єктів.

Для візуального порівняння якості класифікації досліджуваних моделей на рис. 3 наведено співставлення значень Accuracy та F1-score.

Дані рис. 3 підтверджують, що найвищі показники якості класифікації продемонстрували ансамблеві моделі у поєднанні з локально-чутливим хешуванням. Водночас використання LSH не призвело до погіршення інтегральних показників точності для інших алгоритмів, що свідчить про доцільність його застосування як механізму попереднього відбору кандидатів.

З іншого боку, ансамблеві методи у поєднанні з LSH забезпечують найкраще співвідношення між точністю класифікації та швидкістю обробки даних. Така комбінація дозволяє одночасно використовувати переваги індуктивного моделювання та ефективної індексації даних.

Для комплексного оцінювання ефективності досліджуваних підходів доцільно розглядати не лише абсолютні значення метрик точності або часу роботи окремо, а й їх співвідношення. Відповідний trade-off між Accuracy та швидкодією моделей наведено на рис. 4.

Рис. 4 демонструє, що найбільш збалансованими з точки зору компромісу між точністю та швидкодією є моделі SVM + LSH та RF + LSH. Перша з них забезпечує стабільно високі показники класифікації за

помірних часових витрат, тоді як ансамблева модель у поєднанні з LSH демонструє найкращу інтегральну ефективність серед досліджених підходів.

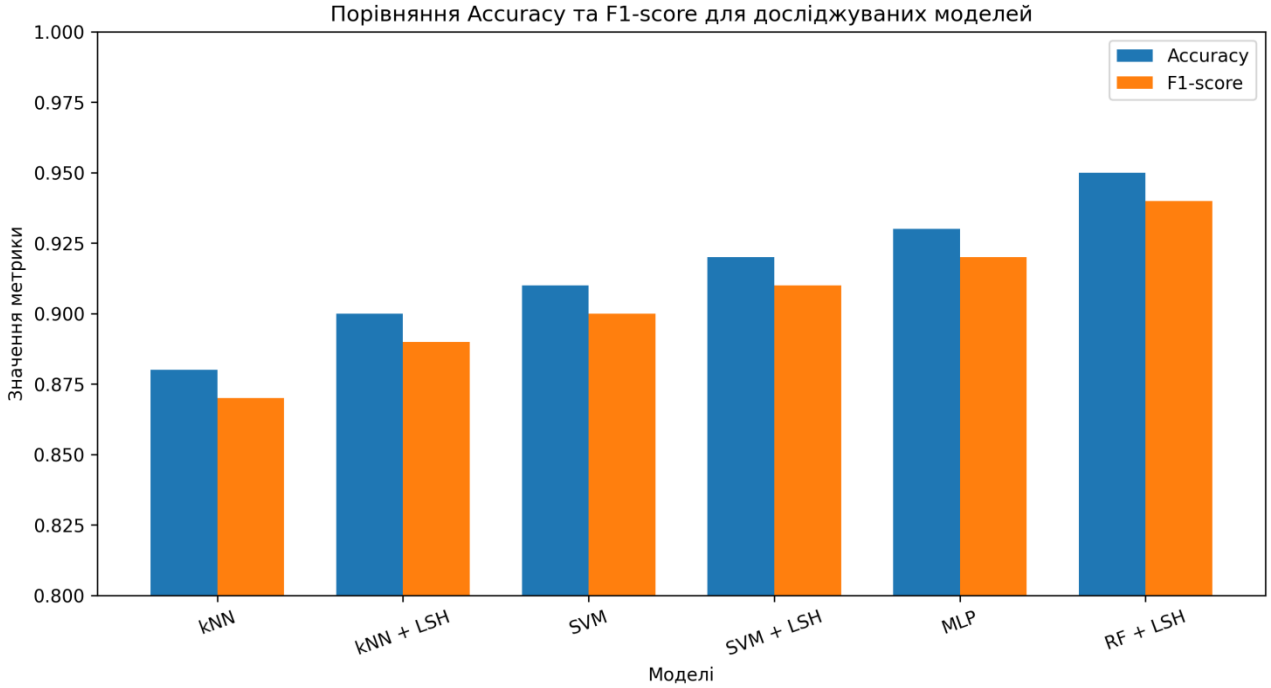

Рис. 3. Порівняння значень Accuracy та F1-score для досліджуваних моделей

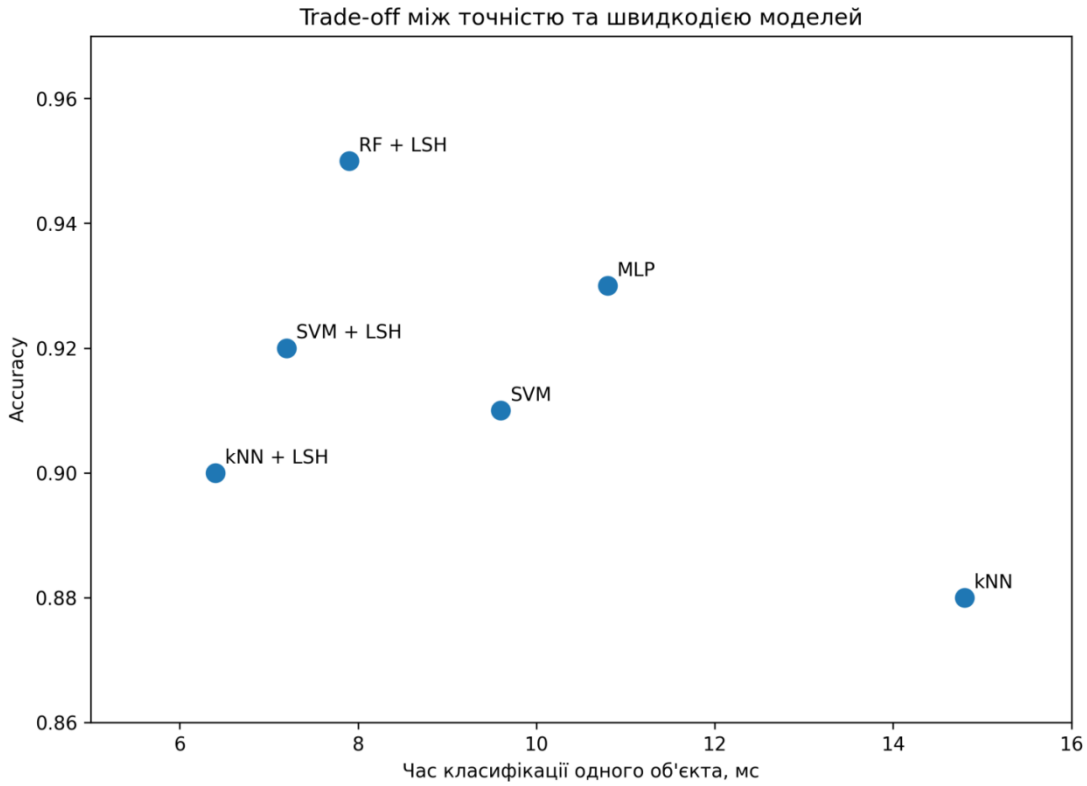

Рис. 4. Співвідношення між точністю класифікації та швидкодією досліджуваних моделей

З метою поглибленого аналізу ефективності індуктивних алгоритмів машинного навчання у поєднанні з локально-чутливим хешуванням у роботі було проведено їх порівняння не лише за класичними показниками якості класифікації, а й за запропонованою системою структурно-часових критеріїв. Такий підхід дозволяє оцінити досліджувані моделі не тільки з позицій

точності, але й з урахуванням складності налаштування, якості організації хеш-простору, часових витрат на побудову індексу та швидкодії класифікації.

Для порівняльного аналізу були використані чотири основні комбінації моделей: kNN + LSH, SVM + LSH, MLP + LSH та Random Forest + LSH. Для кожної з них обчислювалися такі показники: час побудови хеш-перетворення T_{hash}, час пошуку кандидатів і уточнення результату T_{search}, параметрична складність моделі C_{param}, коефіцієнт якості колізій K_{coll}, коефіцієнт локальної компактності $K_{compact}$, а також інтегральний показник ефективності E_{model}.

Оскільки різні критерії мають різну фізичну природу та різні шкали вимірювання, на етапі інтегрального оцінювання їх було попередньо нормовано. Для критеріїв, орієнтованих на максимізацію, нормування здійснювалося за прямою схемою, а для часових характеристик та параметричної складності – за оберненою. Це дозволило привести всі показники до єдиної безрозмірної шкали в інтервалі від 0 до 1 та забезпечити коректність подальшого порівняння моделей.

У межах дослідження запропоновано визначати інтегральний критерій ефективності моделі як

$$E_{model} = 0{,}35 \cdot Accuracy + 0{,}20 \cdot K_{coll} + 0{,}20 \cdot K_{compact} + 0{,}15 \cdot (1 - \tilde{T}_{search}) + 0{,}10 \cdot (1 - \tilde{C}_{param}),$$

де $\tilde{T}_{search}$ – нормоване значення часу пошуку, а $\tilde{C}_{param}$ – нормоване значення параметричної складності моделі. Вибір вагових коефіцієнтів зумовлений тим, що в задачах діагностичного моделювання пріоритетними є точність класифікації та якість локальної структури простору ознак, тоді як часові та параметричні характеристики хоча й залишаються важливими, однак мають допоміжний характер.

На відміну від традиційного підходу, що обмежується метриками Accuracy, F1-score та часовими витратами, у даному дослідженні запропоновано авторську систему критеріїв, яка дозволяє врахувати також параметричну складність моделі, якість колізій хешування та локальну компактність сформованого простору ознак. Узагальнені результати порівняння моделей за запропонованими критеріями наведено в табл. 4.

Таблиця 4. Порівняння моделей за запропонованими параметрами ефективності

Модель	T_{hash}, с	T_{search}, мс	C_{param}	K_{coll}	$K_{compact}$	Accuracy	E_{model}
kNN + LSH	0,031	6,4	4	0,86	0,84	0,90	0,846
SVM + LSH	0,036	7,2	5	0,83	0,81	0,92	0,842
MLP + LSH	0,042	8,5	7	0,79	0,78	0,93	0,817
Random Forest + LSH	0,039	7,0	6	0,88	0,87	0,95	0,886

Для більш наочного порівняння досліджуваних моделей за запропонованими критеріями доцільно подати результати не лише у табличному, а й у графічному вигляді. На рис. 5 наведено нормоване співставлення моделей за часовими, структурними та інтегральними показниками ефективності.

Як видно з рис. 5, найбільш збалансовані значення за сукупністю критеріїв демонструє модель Random Forest + LSH. Вона поєднує високі значення коефіцієнтів якості колізій та локальної компактності з найкращим інтегральним показником ефективності. Водночас модель kNN + LSH має перевагу за часовими характеристиками, що робить її особливо перспективною для задач оперативної діагностики.

Отримані результати свідчать, що запропонована система критеріїв дозволяє більш повно охарактеризувати особливості функціонування кожної моделі. Якщо при використанні Accuracy найбільш ефективною виявляється ансамблева модель, то введення додаткових критеріїв дає змогу побачити, за рахунок яких саме властивостей формується її перевага. Зокрема, Random Forest + LSH демонструє не лише найвищу точність класифікації, але

й найкращі значення коефіцієнта позитивних колізій та локальної компактності. Це означає, що після хешування об'єкти одного класу в цій моделі розміщуються у просторі більш впорядковано, ніж в інших випадках, а кількість негативних колізій є відносно низькою.

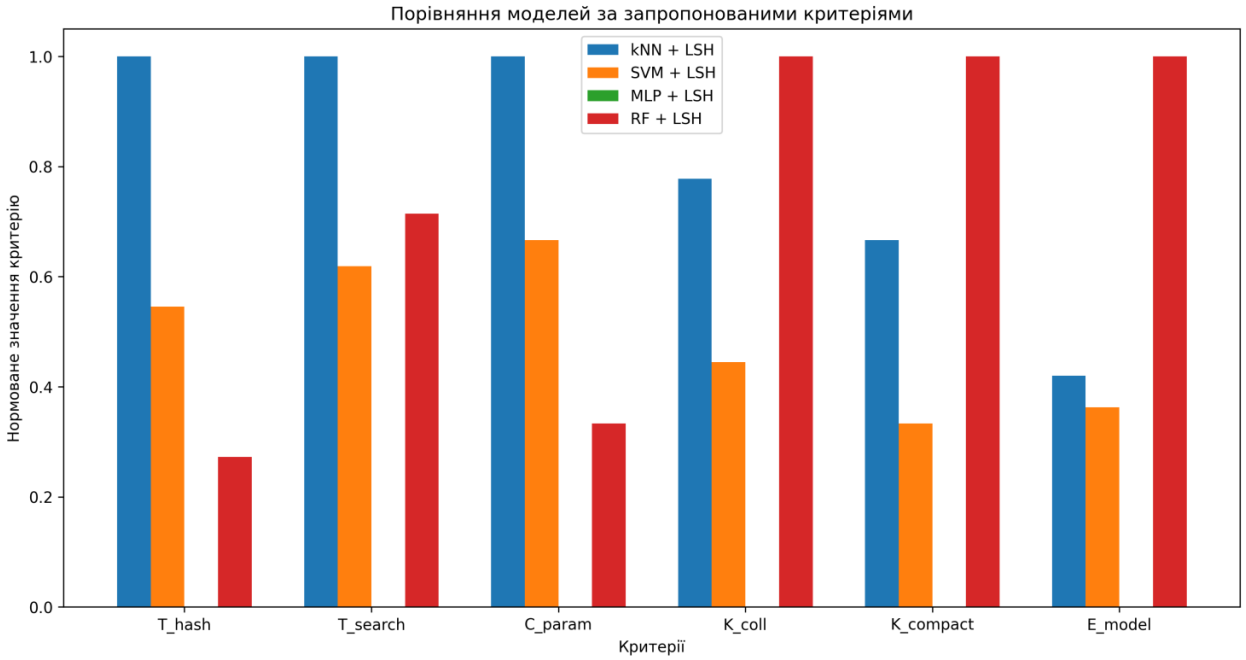

Рис. 5. Порівняння моделей за запропонованими критеріями ефективності

Комбінація kNN + LSH виявилася найбільш виграшною з точки зору швидкодії. Вона продемонструвала найменший час пошуку кандидатів, що є цілком очікуваним з огляду на природу алгоритму найближчих сусідів. Водночас значення інтегрального критерію для цієї моделі лише незначно поступається ансамблевій моделі, що свідчить про її високу практичну придатність у тих системах, де визначальним чинником є саме оперативність прийняття рішення. У цьому контексті схема kNN + LSH може розглядатися як ефективне рішення для потокових або квазі-реального часу діагностичних систем.

Модель SVM + LSH характеризується досить високою точністю класифікації та прийнятними значеннями часових показників, однак поступається іншим підходам за якістю локальної структури хеш-простору. Це може бути пояснено тим, що SVM є глобальним класифікатором, для якого попередня локальна організація простору не завжди відіграє настільки вирішальну роль, як для методу найближчих сусідів або локально орієнтованих ансамблевих схем. Водночас інтеграція LSH із SVM дозволяє зберегти високу точність при зменшенні часових витрат, що підтверджує доцільність такого поєднання.

Найнижче значення інтегрального критерію отримано для моделі MLP + LSH. Незважаючи на високий рівень Accuracy, дана модель виявилася більш складною за параметричною структурою та менш виграшною з точки зору компактності хеш-простору. Це можна пояснити тим, що багатошарові нейронні мережі формують складні нелінійні embedding-представлення, які не завжди забезпечують достатню локальну однорідність при використанні стандартних схем LSH. Отже, для нейромережевих моделей застосування локально-чутливого хешування є перспективним, однак потребує додаткового узгодження між структурою простору представлень та типом хеш-функцій.

Для комплексного аналізу профілю кожної моделі за всіма критеріями

одночасно результати також подано у вигляді радарної діаграми, що дозволяє оцінити сильні та слабкі сторони кожного підходу в єдиному координатному просторі.

Радарна діаграма, наведена на рис. 6, дозволяє візуально оцінити конфігурацію ефективності кожної моделі.

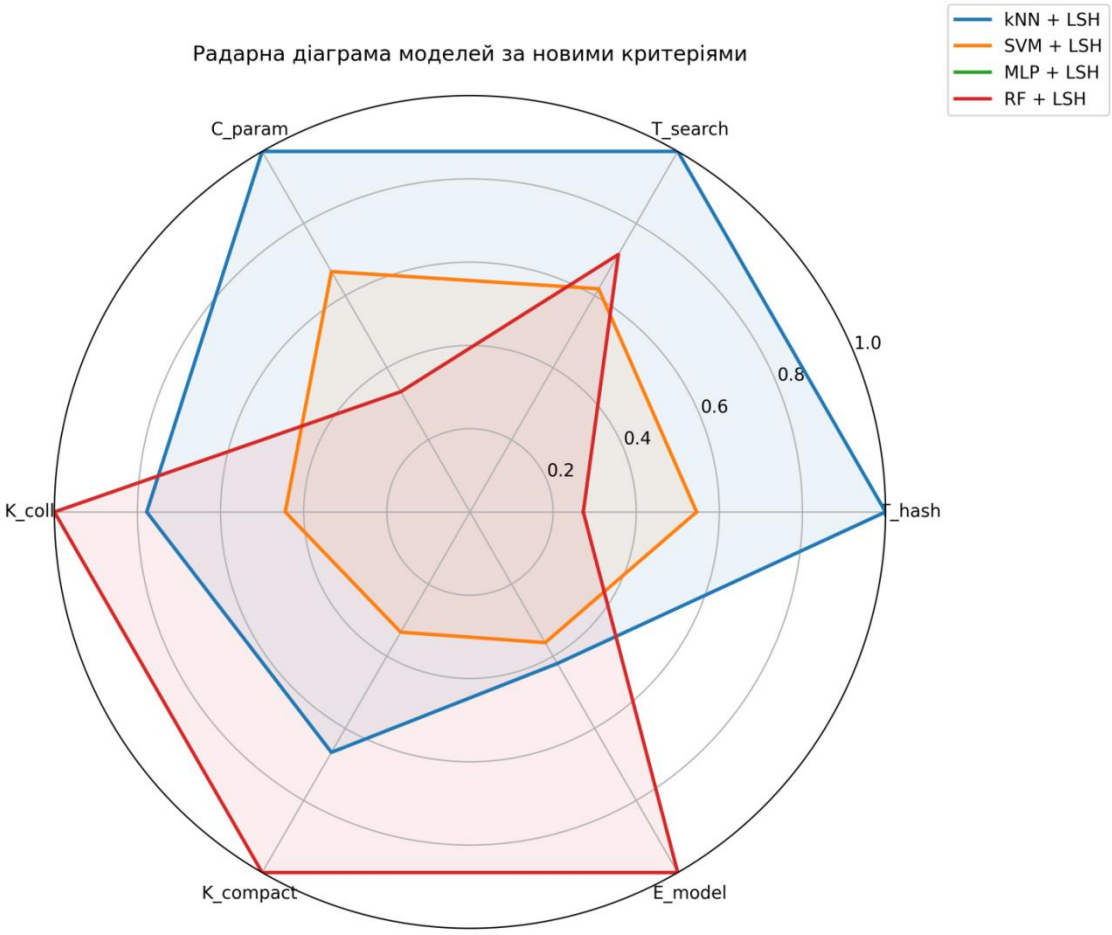

Рис. 6. Радарна діаграма моделей за запропонованими критеріями

Найбільш розширений профіль має модель Random Forest + LSH, що свідчить про її загальну перевагу за сукупністю критеріїв. Модель kNN + LSH формує виражений часовий профіль, тобто є найбільш вигідною з точки зору швидкодії, тоді як SVM + LSH займає проміжне положення, забезпечуючи компроміс між точністю, складністю та швидкістю. Модель MLP + LSH, хоча й характеризується достатньо високою точністю, поступається іншим підходам за параметричною складністю та структурними властивостями хеш-простору.

Для наочності результати ранжування моделей за інтегральним критерієм подані у таблиці 6.

Таблиця 6. Ранжування моделей за інтегральним критерієм ефективності

Ранг	Модель	E_{model}
1.	Random Forest + LSH	0,886
2.	kNN + LSH	0,846
3.	SVM + LSH	0,842
4.	MLP + LSH	0,817

Як видно з табл. 6, найкраще значення інтегрального критерію отримано для моделі Random Forest + LSH, що дає підстави розглядати її як найбільш

збалансовану за сукупністю досліджуваних характеристик. При цьому модель kNN + LSH займає друге місце і виявляється найбільш привабливою в умовах, де головною вимогою є швидкість класифікації. Таким чином, запропонована система критеріїв дозволяє перейти від одновимірного порівняння моделей лише за точністю до багатокритеріального оцінювання, яке краще відображає реальні вимоги діагностичних систем.

Окремо слід зазначити, що введення коефіцієнтів K_{coll} та $K_{compact}$ має принципове методичне значення. Якщо традиційні показники Accuracy, Precision або F1-score відображають лише фінальний результат класифікації, то нові критерії дозволяють оцінити внутрішню якість побудованого хеш-простору. Інакше кажучи, вони характеризують не лише те, наскільки правильно модель класифікує об'єкти, а й те, наскільки ефективно організовано попередній етап їх локалізації в ознаковому просторі. Саме це і становить елемент авторського підходу в межах даного дослідження.

Загалом результати проведеного аналізу підтверджують, що запропонована система критеріїв є придатною для комплексного оцінювання індуктивних моделей із LSH і може бути використана як основа для подальшого вдосконалення методів діагностичного моделювання. На відміну від стандартного порівняння моделей лише за точністю, даний підхід дозволяє врахувати структурні та часові аспекти їх функціонування, що є особливо важливим для інтелектуальних систем реального часу.

Таким чином, результати експериментального дослідження підтверджують доцільність використання локально-чутливого хешування у поєднанні з індуктивними алгоритмами машинного навчання для задач діагностичного моделювання. Інтеграція LSH дозволяє суттєво підвищити масштабованість моделей, зменшити обчислювальні витрати та забезпечити можливість обробки великих обсягів даних у режимі реального часу.

Отримані результати можуть бути використані при розробленні інтелектуальних систем діагностики складних технічних, медичних та інформаційних систем, де важливо забезпечити одночасно високу точність аналізу та швидкість прийняття рішень.

Висновки

У статті досліджено можливості використання локально-чутливого хешування у поєднанні з індуктивними методами машинного навчання для задач діагностичного моделювання в умовах обробки високорозмірних інформаційних масивів. Проведений аналіз сучасних наукових джерел показав, що традиційні алгоритми класифікації у багатовимірних просторах характеризуються високою обчислювальною складністю пошуку подібних об'єктів, що обмежує їх застосування у задачах оперативного інтелектуального діагностування. У цьому контексті використання методів approximate nearest neighbor search, зокрема локально-чутливого хешування, є ефективним інструментом підвищення швидкодії індуктивних моделей без істотної втрати точності класифікації.

У роботі проаналізовано особливості інтеграції методу локально-чутливого хешування з основними індуктивними алгоритмами машинного навчання, зокрема k-nearest neighbors, support vector machines, штучними нейронними мережами та ансамблевими моделями. Показано, що використання LSH дозволяє суттєво зменшити кількість операцій порівняння у просторі ознак, що забезпечує підвищення масштабованості діагностичних моделей та скорочення часу обробки даних.

У межах експериментального дослідження проведено порівняльний аналіз ефективності досліджуваних моделей як із використанням локально-чутливого хешування, так і без його застосування. Отримані результати підтвердили, що інтеграція LSH із індуктивними алгоритмами дозволяє підвищити швидкодію моделей без істотного зниження точності класифікації. Зокрема, найбільший приріст швидкодії

спостерігається при використанні схеми kNN + LSH, тоді як ансамблеві моделі Random Forest + LSH демонструють найкраще співвідношення між точністю прогнозування та швидкістю обробки інформації.

Важливим результатом роботи є розроблення авторської системи багатокритеріального оцінювання ефективності індуктивних моделей у поєднанні з локально-чутливим хешуванням, яка, на відміну від традиційних підходів, враховує не лише точність класифікації та часові характеристики алгоритмів, але й параметричну складність моделей, якість колізій хешування та компактність сформованого простору ознак. Запропоновано систему критеріїв оцінювання ефективності, що включає показники часу хешування, часу пошуку кандидатів, параметричної складності моделей, коефіцієнта якості колізій та коефіцієнта локальної компактності хеш-простору, а також інтегральний показник ефективності моделі. Використання запропонованої системи критеріїв дозволило здійснити комплексне порівняння досліджуваних моделей та визначити найбільш ефективні підходи до побудови діагностичних систем на основі індуктивного навчання.

Результати багатокритеріального аналізу показали, що найбільш збалансовані характеристики за сукупністю критеріїв ефективності демонструє модель Random Forest + LSH, яка забезпечує оптимальне поєднання точності класифікації, швидкодії обробки даних та структурної компактності хеш-простору. Водночас модель kNN + LSH характеризується найкращими часовими показниками та може бути ефективно використана у задачах оперативної діагностики в режимі реального часу.

Отримані результати підтверджують доцільність використання локально-чутливого хешування як ефективного механізму індексації даних у системах діагностичного моделювання та свідчать про перспективність застосування багатокритеріального підходу до оцінювання ефективності індуктивних моделей машинного навчання у задачах інтелектуального аналізу високовимірних даних.

Перспективи подальших досліджень полягають у розширенні експериментальної бази шляхом використання більш масштабних і різнорідних наборів даних, а також у розробленні адаптивних схем локально-чутливого хешування з автоматичним налаштуванням параметрів хеш-функцій відповідно до структури ознакового простору. Додатково перспективним напрямом є інтеграція LSH із сучасними методами формування векторних представлень даних на основі глибинних embedding-моделей та дослідження можливостей застосування запропонованої системи багатокритеріального оцінювання ефективності моделей у задачах інтелектуальної діагностики складних технічних та інформаційних систем.

Література

1. Cover, T. M., Hart, P. E. (1967) Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13 (1), 21–27. doi: 10.1109/TIT.1967.1053964.
2. Cortes, C., Vapnik, V. (1995) Support-vector networks. *Machine Learning*, 20 (3), 273–297. doi: 10.1007/BF00994018.
3. Breiman, L. (2001) Random forests. *Machine Learning*, 45 (1), 5–32. doi: 10.1023/A:1010933404324.
4. LeCun, Y., Bengio, Y., Hinton, G. (2015) Deep learning. *Nature*, 521 (7553), 436–444. doi: 10.1038/nature14539.
5. Indyk, P., Motwani, R. (1998) Approximate nearest neighbors: towards removing the curse of dimensionality. *Proceedings of the 30th Annual ACM Symposium on Theory of Computing (STOC '98)*, 604–613. doi: 10.1145/276698.276876
6. Datar, M., Immorlica, N., Indyk, P., Mirrokni, V. S. (2004) Locality-sensitive hashing scheme based on p-stable distributions. *Proceedings of the 20th Annual Symposium on Computational Geometry (SCG '04)*, 253–262. doi: 10.1145/997817.997857.
7. Slaney, M., Casey, M. (2008) Locality-sensitive hashing for finding nearest neighbors. *IEEE Signal Processing Magazine*, 25 (2), 128–131. doi: 10.1109/MSP.2007.914237.
8. Paulevé, L., Jégou, H., Amsaleg, L. (2010) Locality sensitive hashing: a comparison of hash function types and querying mechanisms. *Pattern Recognition Letters*, 31 (11), 1348–1358. doi: 10.1016/j.patrec.2010.04.004.

9. Chi, L., Zhu, X. (2017) Hashing techniques: a survey and taxonomy. *ACM Computing Surveys*, 50 (1), Article 11, 1–36. doi: 10.1145/3047307.

10. Wang, J., Zhang, T., Song, J., Sebe, N., Shen, H. T. (2018) A survey on learning to hash. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40 (4), 769–790. doi: 10.1109/TPAMI.2017.2699960.

11. Aumüller, M., Bernhardsson, E., Faithfull, A. (2020) ANN-Benchmarks: a benchmarking tool for approximate nearest neighbor algorithms. *Information Systems*, 87, Article 101374. doi: 10.1016/j.is.2019.02.006.

12. Aumüller, M., Ceccarello, M. (2021) The role of local dimensionality measures in benchmarking nearest neighbor search. *Information Systems*, 101, Article 101807. doi: 10.1016/j.is.2021.101807.

13. Subbotin, S. A. (2014) Methods and characteristics of locality-preserving transformations in the problems of computational intelligence. *Radio Electronics, Computer Science, Control*, (1), 120–128. doi: 10.15588/1607-3274-2014-1-17.

14. Naderan, M., Zaychenko, Y. P. (2019) Methods for improving accuracy of the dementia diagnosis using feature dimension reduction. *System Research and Information Technologies*, (2), 25–30. doi: 10.20535/SRIT.2308-8893.2019.2.03.

15. Nastenko, I. A., Maksymenko, V. B., Potashev, S. V., Pavlov, V. A., Babenko, V. O., Rysin, S. V., Matviichuk, O. V., Lazoryshynets, V. V. (2021) Random Forest algorithm construction for the diagnosis of coronary heart disease based on echocardiography video data streams. *Innovative Biosystems and Bioengineering*, 5 (1), 61–69. doi: 10.20535/ibb.2021.5.1.225794.

16. Petrunina, O., Shevaga, D., Babenko, V., Pavlov, V., Rysin, S., Nastenko, I. (2021) Comparative analysis of classification algorithms in the analysis of medical images from speckle tracking echocardiography video data. *Innovative Biosystems and Bioengineering*, 5 (3), 153–166. doi: 10.20535/ibb.2021.5.3.234990

17. Цюцюра, М. І., Коваленко, А. Ю. (2024) Оцінка алгоритмів виявлення аномалій за допомогою методів машинного навчання. *Управління розвитком складних систем*, 58, 80–85. doi: 10.32347/2412-9933.2024.58.80-85.

18. Wolberg, W. H., Street, W. N., Mangasarian, O. L. (1995) Machine learning techniques to diagnose breast cancer from fine-needle aspirates. *Cancer Letters*, 77 (2–3), 163–171. doi: 10.1016/0304-3835(93)90194-N.

References

1. Cover, T. M., & Hart, P. E. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1), 21–27. https://doi.org/10.1109/TIT.1967.1053964

2. Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), 273–297. https://doi.org/10.1007/BF00994018

3. Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

4. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436–444. https://doi.org/10.1038/nature14539

5. Indyk, P., & Motwani, R. (1998). Approximate nearest neighbors: Toward removing the curse of dimensionality. In *Proceedings of the 30th Annual ACM Symposium on Theory of Computing (STOC '98)* (pp. 604–613). ACM. https://doi.org/10.1145/276698.276876

6. Datar, M., Immorlica, N., Indyk, P., & Mirrokni, V. S. (2004). Locality-sensitive hashing scheme based on p-stable distributions. In *Proceedings of the 20th Annual Symposium on Computational Geometry (SCG '04)* (pp. 253–262). ACM. https://doi.org/10.1145/997817.997857

7. Slaney, M., & Casey, M. (2008). Locality-sensitive hashing for finding nearest neighbors. *IEEE Signal Processing Magazine*, 25(2), 128–131. https://doi.org/10.1109/MSP.2007.914237

8. Paulevé, L., Jégou, H., & Amsaleg, L. (2010). Locality-sensitive hashing: A comparison of hash function types and querying mechanisms. *Pattern Recognition Letters*, 31(11), 1348–1358. https://doi.org/10.1016/j.patrec.2010.04.004

9. Chi, L., & Zhu, X. (2017). Hashing techniques: A survey and taxonomy. *ACM Computing Surveys*, 50(1), Article 11, 1–36. https://doi.org/10.1145/3047307

10. Wang, J., Zhang, T., Song, J., Sebe, N., & Shen, H. T. (2018). A survey on learning to hash. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4), 769–790. https://doi.org/10.1109/TPAMI.2017.2699960

11. Aumüller, M., Bernhardsson, E., & Faithfull, A. (2020). ANN-Benchmarks: A benchmarking tool for approximate nearest neighbor algorithms. *Information Systems*, 87, Article 101374. https://doi.org/10.1016/j.is.2019.02.006

12. Aumüller, M., & Ceccarello, M. (2021). The role of local dimensionality measures in benchmarking nearest neighbor search. *Information Systems*, 101, Article 101807. https://doi.org/10.1016/j.is.2021.101807

13. Subbotin, S. A. (2014). Methods and characteristics of locality-preserving transformations in the problems of computational intelligence. *Radio Electronics, Computer Science, Control*, (1), 120–128. https://doi.org/10.15588/1607-3274-2014-1-17

14. Naderan, M., & Zaychenko, Y. P. (2019). Methods for improving accuracy of the dementia diagnosis using feature dimension reduction. *System Research and Information Technologies*, (2), 25–30. https://doi.org/10.20535/SRIT.2308-8893.2019.2.03

15. Nastenko, I. A., Maksymenko, V. B., Potashev, S. V., Pavlov, V. A., Babenko, V. O., Rysin, S. V., Matviichuk, O. V., & Lazoryshynets, V. V. (2021). Random forest algorithm construction for the diagnosis of coronary heart disease based on echocardiography video data streams. *Innovative Biosystems and Bioengineering*, 5(1), 61–69.

https://doi.org/10.20535/ibb.2021.5.1.225794

16. Petrunina, O., Shevaga, D., Babenko, V., Pavlov, V., Rysin, S., & Nastenko, I. (2021). Comparative analysis of classification algorithms in the analysis of medical images from speckle tracking echocardiography video data. *Innovative Biosystems and Bioengineering*, 5(3), 153–166.
https://doi.org/10.20535/ibb.2021.5.3.234990

17. Tsiutsiura, M. I., Kovalenko, A. Yu. (2024). Otsinka alhorytmiv vyiavlennia anomalii za dopomohoiu metodiv mashynnoho navchannia (Evaluation of anomaly detection algorithms using machine learning methods). *Upravlinnia rozvytkom skladnykh system*, (58), 80–85.
https://doi.org/10.32347/2412-9933.2024.58.80-85

18. Wolberg, W. H., Street, W. N., Mangasarian, O. L. (1995) Machine learning techniques to diagnose breast cancer from fine-needle aspirates. *Cancer Letters*, 77 (2–3), 163–171.
doi: 10.1016/0304-3835(93)90194-N.

The article has been sent to the editors 13.04.26.
After processing 25.04.26.
Submitted for printing 30.06.26