

КВАЛІФІКАЦІЙНА РОБОТА

на здобуття освітнього ступеня

магістр

(назва освітнього ступеня)

на тему: Аналіз ансамблевих методів машинного
навчання та їх інтеграція з алгоритмами глибинного навчання

Виконав(ла): студент(ка) 6 курсу, групи СНмн-61
спеціальності 122 Комп'ютерні науки

(шифр і назва спеціальності)

	<u>Боб О.О.</u>
(підпис)	(прізвище та ініціали)
Керівник	<u>Марценко С.В.</u>
(підпис)	(прізвище та ініціали)
Нормоконтроль	<u>Дуда О.М.</u>
(підпис)	(прізвище та ініціали)
Завідувач кафедри	<u>Боднарчук І.О.</u>
(підпис)	(прізвище та ініціали)
Рецензент	<u></u>
(підпис)	(прізвище та ініціали)

Міністерство освіти і науки України
Тернопільський національний технічний університет імені Івана Пулюя

Факультет комп'ютерно-інформаційних систем і програмної інженерії
(повна назва факультету)

Кафедра комп'ютерних наук
(повна назва кафедри)

ЗАТВЕРДЖУЮ
Завідувач кафедри
Боднарчук І.О.
(підпис) (прізвище та ініціали)

"13" квітня 2026 р.

ЗАВДАННЯ НА КВАЛІФІКАЦІЙНУ РОБОТУ

на здобуття освітнього ступеня Магістр
(назва освітнього ступеня)

за спеціальністю 122 Комп'ютерні науки
(шифр і назва спеціальності)

студенту Боб Олександр Олександрович
(прізвище, ім'я, по батькові)

1. Тема роботи Аналіз ансамблевих методів машинного навчання та їх інтеграція з алгоритмами глибинного навчання.

Керівник роботи к.т.н., доц. Марценко С.В.
(прізвище, ім'я, по батькові, науковий ступінь, вчене звання)

Затверджені наказом ректора від "10" березня 2026 року № 4/9-150

2. Термін подання студентом завершеної роботи 25 травня 2026 р.

3. Вихідні дані до роботи Літературні джерела з тематики роботи

4. Зміст роботи (перелік питань, які потрібно розробити)
ВСТУП; 1 ОГЛЯД АНСАМБЛЕВИХ МЕТОДІВ НАВЧАННЯ; 1.1 Загальний опис ансамблевих методів; 1.2 Загальний опис бегінгу; 1.3 Загальний опис бустингу; 1.4 Загальний огляд стекінгу; 2 АНАЛІЗ МЕТОДІВ ІНТЕГРАЦІЇ МОДЕЛІЙ МАШИННОГО І ГИБОКОГО НАВЧАННЯ; 2.1 Тенденції ансамблевого навчання; 2.2 Основи ансамблевого навчання; 2.3 Вибірка даних; 2.4 Класифікатори базової лінії навчання; 2.5 Метод злиття; 3 ОПИС ІНТЕГРАЦІЇ МЕТОДІВ МАШИННОГО НАВЧАННЯ; 3.1 Ансамблеві методи без інтеграції з глибоким навчанням; 3.2 Ансамблеві методи з глибоким навчанням; 3.3 Оцінювання ансамблевих методів; 3.4 Застосування ансамблю глибокого навчання; 4 ОХОРОНА ПРАЦІ ТА БЕЗПЕКА В НАДЗВИЧАЙНИХ СИТУАЦІЯХ; 4.1 Питання щодо охорони праці; 4.2 Питання щодо безпеки в надзвичайних ситуаціях; ВИСНОВКИ; СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ; ДОДАТКИ

5. Перелік графічного матеріалу (з точним зазначенням обов'язкових креслень, слайдів)

6. Консультанти розділів роботи

Розділ	Прізвище, ініціали та посада консультанта	Підпис, дата	
		завдання видав	завдання прийняв
Охорона праці	Сенчишин В.С., к.т.н., доц. каф. МТ		
Безпека в надзвичайних ситуаціях	Теслюк В.М., проректор з адміністративно-господарської роботи та будівництва		

7. Дата видачі завдання 13 квітня 2026 р.

КАЛЕНДАРНИЙ ПЛАН

№ з/п	Назва етапів роботи	Термін виконання етапів роботи	Примітка
1.	Ознайомлення з завданням до кваліфікаційної роботи	13.04.26-18.04.26	Виконано
2.	Підбір наукових джерел за темою роботи	19.04.26-21.04.26	Виконано
3.	Переклад та опрацювання наукових джерел за темою кваліфікаційної роботи	20.04.26-23.04.26	Виконано
4.	Виконання дослідження щодо теми кваліфікаційної роботи	24.04.26-10.05.26	Виконано
5.	Оформлення першого розділу	04.05.26-05.05.26	Виконано
6.	Оформлення другого розділу	05.05.26-13.05.26	Виконано
7.	Оформлення третього розділу	13.05.26-14.05.26	Виконано
8.	Виконання завдання до підрозділу "Охорона праці"	08.05.26-09.05.26	Виконано
9.	Виконання завдання до підрозділу "Безпека в надзвичайних ситуаціях"	10.05.26-05.05.26	Виконано
10.	Оформлення кваліфікаційної роботи	05.05.26-13.05.26	Виконано
11.	Нормоконтроль	14.05.26-15.05.26	Виконано
12.	Перевірка на плагіат	16.05.2026	Виконано
13.	Попередній захист кваліфікаційної роботи	20.05.2026	Виконано
14.	Захист кваліфікаційної роботи	26.05.2026	

Студент

_____ (підпис)

Боб О.О.

_____ (прізвище та ініціали)

Керівник роботи

_____ (підпис)

Марценко С.В.

_____ (прізвище та ініціали)

АНОТАЦІЯ

"Аналіз ансамблевих методів машинного навчання та їх інтеграція з алгоритмами глибинного навчання" // Кваліфікаційна робота освітнього рівня "Магістр" // Боб Олександр Олександрович // Тернопільський національний технічний університет ім. І. Пулюя, факультет комп'ютерно-інформаційних систем і програмної інженерії, кафедра комп'ютерних наук, група СНм-61 // Тернопіль, 2026 // с. – 67, рис. – 16, табл. – 4, джерел – 39.

Ключові слова: машинне навчання, штучний інтелект, ансамблеві методи, нейронна мережа

У галузі машинного навчання два підходи демонструють суттєві переваги над традиційними алгоритмами – ансамблеве та глибоке навчання. Ансамблеве навчання охоплює методи, що передбачають інтеграцію кількох базових моделей у межах єдиного фреймворку з метою формування більш ефективної моделі, здатної перевершувати окремі компоненти ансамблю. Ефективність ансамблевих методів залежить від низки чинників, зокрема способу навчання базових моделей та механізмів їх комбінування. У науковій літературі описано загальні підходи до побудови ансамблевих моделей, які успішно застосовуються в різних предметних областях.

Своєю чергою, моделі глибокого навчання забезпечили суттєве підвищення точності прогнозування у широкому спектрі задач машинного навчання. Незважаючи на різноманіття архітектур глибокого навчання, їх здатність розв'язувати складні задачі та автоматично виділяти ознаки, суттєвою проблемою залишається необхідність підбору оптимальних гіперпараметрів, що потребує значного практичного досвіду та спеціалізованих знань. Це ускладнює процес налаштування моделей і робить його трудомістким.

У зв'язку з цим значна кількість досліджень спрямована на подолання зазначених обмежень шляхом поєднання підходів ансамблевого та глибокого

навчання. Більшість наявних досліджень зосереджені на застосуванні простих ансамблевих методів, які характеризуються певними обмеженнями. Відтак дана робота присвячена аналізу різних стратегій ансамблевого навчання, зокрема в контексті їх інтеграції з методами глибокого навчання. Окрему увагу приділено чинникам та особливостям, що визначають ефективність ансамблевих методів.

ANNOTATION

“Study of Ensemble Machine Learning Methods and Their Integration with Deep Learning Algorithms” // Master’s degree qualification paper // Bob Oleksandr // Ternopil Ivan Puluj National Technical University, Faculty of Computer Information Systems and Software Engineering, Computer Science Department, group CHHM-61 // Ternopil, 2026 // p. – 67, fig. – 16, tables – 4, references – 39.

Key words: machine learning, artificial intelligence, ensemble methods, neural network

In the field of machine learning, two approaches demonstrate significant advantages over traditional algorithms: ensemble learning and deep learning. Ensemble learning encompasses methods that involve the integration of multiple base models within a unified framework in order to construct a more effective model capable of outperforming its individual components. The effectiveness of ensemble methods depends on several factors, including the training procedures of the base models and the mechanisms used for their combination. The scientific literature describes general approaches to the construction of ensemble models that have been successfully applied across various domains.

In turn, deep learning models have substantially improved predictive accuracy in a wide range of machine learning tasks. Despite the diversity of deep learning architectures, their capability to solve complex problems, and their ability to automatically extract features, a major challenge remains the selection of optimal hyperparameters, which requires considerable practical experience and specialized knowledge. This significantly complicates the model configuration process and makes it time-consuming.

In this regard, a considerable amount of research has been devoted to overcoming these limitations through the integration of ensemble learning and deep learning approaches. Most existing studies focus on the application of relatively simple

ensemble methods, which are characterized by certain limitations. Therefore, this review paper is devoted to the analysis of various ensemble learning strategies, particularly in the context of their integration with deep learning methods. Special attention is given to the factors and characteristics that determine the effectiveness of ensemble methods.

ЗМІСТ

ВСТУП.....	8
1 ОГЛЯД АНСАМБЛЕВИХ МЕТОДІВ НАВЧАННЯ	12
1.1 Загальний опис ансамблевих методів	12
1.2 Загальний опис бегінгу	16
1.3 Загальний опис бустингу	22
1.4 Загальний огляд стекінгу	28
2 АНАЛІЗ МЕТОДІВ ІНТЕГРАЦІЇ МОДЕЛІ МАШИННОГО І ГИБОКОГО НАВЧАННЯ.....	32
2.1 Тенденції ансамблевого навчання	32
2.2 Основи ансамблевого навчання.....	33
2.3 Вибірка даних	36
2.4 Класифікатори базової лінії навчання	36
2.5 Метод злиття.....	38
3 ОПИС ІНТЕГРАЦІЇ МЕТОДІВ МАШИННОГО НАВЧАННЯ	41
3.1 Ансамблеві методи без інтеграції з глибоким навчанням.....	41
3.2 Ансамблеві методи з глибоким навчанням	42
3.3 Оцінювання ансамблевих методів.....	45
3.4 Застосування ансамблю глибокого навчання.....	48
4 ОХОРОНА ПРАЦІ ТА БЕЗПЕКА В НАДЗВИЧАЙНИХ СИТУАЦІЯХ	53
4.1 Питання щодо охорони праці	53
4.1.1 Ергономіка	53
4.1.2 Освітлення	55
4.1.3 Параметри мікроклімату	56
4.1.4 Емоційна психогігієна	57
4.2 Питання щодо безпеки в надзвичайних ситуаціях	59
ВИСНОВКИ.....	61
СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ.....	63
ДОДАТКИ.....	67

ВСТУП

Актуальність задачі.

У світі, повному різноманітних та різноманітних джерел даних, машинне навчання стало однією з найважливіших та домінуючих галузей методів штучного інтелекту, яка застосовується в багатьох галузях. Існує багато різних алгоритмів та методів навчання. Підводні камені та недоліки кожного методу вимірюються з точки зору кількох факторів, включаючи продуктивність та масштабованість. Грунтуючись на численних дослідженнях у галузі машинного навчання, два методи домінують серед алгоритмів навчання, а саме: глибоке навчання [1] та ансамблеве навчання [2]. Методи глибокого навчання можуть масштабувати та обробляти складні проблеми, а також пропонують автоматичне вилучення ознак з неструктурованих даних. Також методи глибокого навчання містять кілька типів мережевих архітектур для різних завдань, таких як нейронні мережі прямого поширення, згорткові нейронні мережі, рекурентні нейронні мережі. Однак процес навчання моделей глибокого навчання вимагає величезних зусиль, а налаштування оптимальних гіперпараметрів вимагає досвіду та ретельних випробувань, що є виснажливим та трудомістким завданням. Крім того, навчання більш складної глибокої нейронної мережі збільшує ймовірність перенавчання.

Ансамблеве навчання, з іншого боку, стосується методології навчання, яка поєднує кілька базових моделей для створення більшої, але потужнішої моделі, ніж її складові. Крім того, ансамблеве навчання може зменшити ризик перенавчання завдяки різноманітності базових моделей. Ансамблеве навчання було успішно застосовано в різних галузях і областях і перевершує окремі моделі. Існує кілька ансамблевих методів, що відрізняються тим, як навчаються та поєднуються різні базові моделі. Найбільш широко використовувані ансамблеві методи включають усереднення, пакетування, випадковий ліс, стекування та підвищення. У літературі є багато оглядів методів та технік ансамблевого навчання [3]. Традиційне ансамблеве навчання базується на

інтеграції традиційних моделей машинного навчання та їх застосуванні в різних галузях. Однак ці зусилля обмежувалися простими одиничними моделями. В останні роки було зроблено численні спроби наблизити ансамблеве навчання до глибокого навчання [4]. Однак більшість цих спроб сформульовано з використанням методу усереднення голосування базових моделей глибокого навчання. Однак ансамблевий процес, що використовує методи усереднення голосування, упереджений до слабких базових учнів і не є розумною стратегією для їх об'єднання. Незважаючи на кілька стратегій поєднання учнів базового рівня, які можна застосувати до ансамблевого глибокого навчання, ці стратегії мають деякі обмеження з точки зору узагальнення, труднощів у навчанні та інших проблем [5]. У літературі деякі оглядові роботи ввели концепцію глибокого ансамблевого навчання. Однак ці зусилля обмежуються застосуванням ансамблю в окремих областях з оглядами традиційних ансамблевих підходів.

Мета роботи.

Таким чином метою роботи є всебічно розглянути різні стратегії застосування ансамблевого глибокого навчання. У ній також представлено кілька аспектів, що впливають на успіх ансамблевих методів, таких як тип використовуваних моделей базового навчання, методи вибірки даних, що використовуються в навчанні, різноманітність використання різних базових класифікаторів та методи об'єднання базових глибоких моделей. Крім того, обговорюються переваги та недоліки кожної стратегії.

Для досягнення цієї мети потрібно вирішити ряд завдань.

По-перше, ми отримати кількісну аналітику ансамблевого навчання.

По-друге, представити фундаментальні концепції та загальну архітектуру ансамблевого навчання, стратегії створення різноманітності серед базових класифікаторів та фактори, що впливають на будь-який ансамблевий метод.

По-третє, представити структуру кожного з кількох ансамблевих методів, а також переваги, недоліки та загальні класифікації для кожного методу.

По-четверте, проаналізувати різні стратегії моделей ансамблевого глибокого навчання.

Таким чином буде розглянуто численні дослідницькі зусилля, які використовували ансамблеве навчання в різних застосуваннях.

Об'єкт дослідження: процеси інтеграції алгоритмів ансамблевих моделей машинного навчання з алгоритмами глибокого навчання.

Предмет дослідження: методи і засоби сумісного використання алгоритмів ансамблевих моделей машинного навчання з алгоритмами глибокого навчання.

Наукова новизна отриманих результатів.

Наукова новизна проведеного дослідження полягає у розробці та теоретико-методологічному обґрунтуванні комплексного підходу до інтеграції алгоритмів ансамблевого навчання з архітектурами глибокого навчання, що дозволяє системно підвищити точність та стабільність класифікації у складних прикладних задачах.

У межах дослідження вперше або з новим ступенем деталізації отримано такі наукові результати.

1. Удосконалено методологію оцінювання ансамблевих моделей шляхом формування кількісної аналітики їхнього функціонування. Це дозволяє перейти від емпіричного підбору конфігурацій до математично обґрунтованого прогнозування ефективності роботи розгалужених систем класифікації.

2. Систематизовано та поглиблено концептуальний базис ансамблевого навчання. Сформульовано узагальнену архітектуру побудови таких систем, а також класифіковано ключові стратегії штучного формування різноманітності (diversity) серед базових класифікаторів. Визначено та формалізовано внутрішні й зовнішні фактори, що критично впливають на стійкість та узагальнюючу здатність будь-якого ансамблевого методу.

3. Розроблено порівняльно-структурний аналіз базових та гібридних ансамблевих методів. Для кожного з них чітко розмежовано архітектурні переваги, обчислювальні недоліки та специфіку застосування, що дало змогу

створити багатокритеріальну класифікацію методів для оптимізації вибору моделей під конкретні типи даних.

4. Дістали подальшого розвитку підходи до інтеграції глибокого навчання (Deep Learning) та ансамблевих стратегій. Проаналізовано та типізовано різнорідні стратегії об'єднання глибоких нейромереж в єдині ансамблеві комплекси, що забезпечує синергетичний ефект: поєднання високої здатності глибоких мереж до вилучення ознак із надійністю та низькою дисперсією ансамблевих методів.

Практичне значення отриманих результатів.

Практична цінність та новизна підтверджуються наскрізним аналізом існуючих дослідницьких зусиль, що дозволило адаптувати розроблені методи і засоби сумісного використання алгоритмів для ефективного вирішення різнопланових прикладних завдань у сфері інтелектуального аналізу даних. В наукових роботах ТНТУ ім. Івана Пулюя є немало, присвячених тематиці використання машинного навчання. Зокрема, в [12] описано алгоритми опрацювання медичних даних, в [13] здійснено порівняльний огляд алгоритмів МН для фінтеку, робота [14] містить матеріал про порівнюють архітектури нейронних мереж, роботи [15] і [16] використовують нейронні мережі для області матеріалознавства. Проте в жодній з них не аналізується проблема інтеграції алгоритмів МН з нейронними мережами. Тому дана робота має наукову та практичну цінність.

Апробація результатів та особистий внесок здобувача.

Основні положення роботи доповідались, розглядались та обговорювались на науковій конференції Тернопільського національного технічного університету імені Івана Пулюя у тезах студентської науково-технічної конференції "Природничі та гуманітарні науки. Актуальні питання – 2026", яка проходила у ТНТУ.

1 ОГЛЯД АНСАМБЛЕВИХ МЕТОДІВ НАВЧАННЯ

Ансамблеве навчання є однією з найрезультативніших парадигм сучасного машинного навчання, що неодноразово засвідчувала свою ефективність у численних прикладних задачах з публічним доступом їх розв'язків, зокрема на платформі Kaggle [6]. В основі цього підходу лежить гіпотеза, згідно з якою агрегування скінченної множини окремих моделей дозволяє побудувати єдину модель із суттєво вищою прогностичною спроможністю порівняно з кожною зі складових окремо.

Метою роботи, як зазначалось вище, є систематичний виклад ключових понять ансамблевого навчання та формування теоретичного підґрунтя, необхідного для коректного застосування відповідних методів і розроблення адаптованих рішень для конкретних задач. У роботі розглянуто базові концепції галузі, а саме: бутстреп, бегінг, випадковий ліс, бустинг і стекінг. З метою забезпечення цілісного розуміння взаємозв'язків між зазначеними методами їх подано в межах єдиного логічно впорядкованого концептуального каркаса.

1.1 Загальний опис ансамблевих методів

Ансамблеве навчання визначається як парадигма машинного навчання, у межах якої сукупність моделей (так званих слабких учнів) навчається для розв'язання однієї й тієї самої задачі з подальшим їх поєднанням задля підвищення якості прогнозування. Засадничою є гіпотеза про те, що за умови коректного поєднання слабких моделей може бути отримано точніші та/або стійкіші моделі.

У машинному навчанні, незалежно від типу розв'язуваної задачі (класифікація чи регресія), вибір моделі має визначальне значення для досягнення прийнятної якості результатів. Цей вибір зумовлюється низкою характеристик задачі, серед яких обсяг наявних даних, розмірність простору ознак та апріорні припущення щодо розподілу даних.

Низьке зміщення та низька дисперсія, які здебільшого змінюються в протилежних напрямках, є двома фундаментальними властивостями, очікуваними від моделі. Для адекватного розв'язання задачі модель повинна мати достатню кількість ступенів свободи, аби відтворити складність досліджуваних даних, проте їх надлишок призводить до зростання дисперсії та зниження стійкості. Зазначене протиріччя відоме в літературі як компроміс зміщення-дисперсія (рис. 1.1).

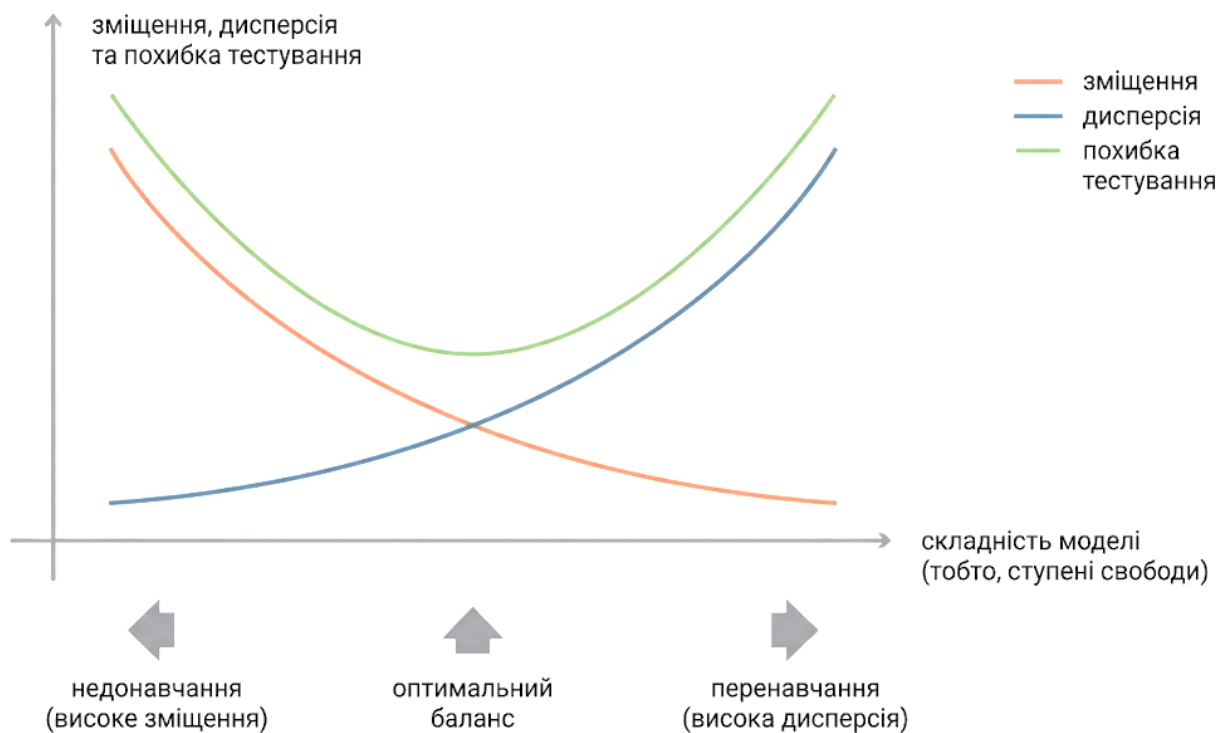


Рисунок 1.1 – Ілюстрація компромісу зміщення-дисперсія

У теорії ансамблевого навчання слабкими учнями (або базовими моделями) називають моделі, що використовуються як структурні елементи для побудови складніших моделей шляхом поєднання кількох із них. Як правило, такі базові моделі окремо демонструють невисоку якість: або внаслідок високого зміщення (моделі з малою кількістю ступенів свободи), або внаслідок надмірної дисперсії, що знижує їхню стійкість (моделі з великою кількістю ступенів свободи). Відтак сутність ансамблевих методів полягає у зменшенні зміщення

та/або дисперсії слабких учнів через їх поєднання, унаслідок чого формується сильний учень (ансамблева модель) із вищими показниками якості.

Поєднання слабких учнів

Реалізація методу ансамблевого навчання передбачає насамперед вибір базових моделей, що підлягають агрегуванню. У більшості випадків (зокрема в методах бегінгу та бустингу) застосовується єдиний базовий алгоритм навчання, унаслідок чого формуються однорідні слабкі учні, навчені в різний спосіб; відповідну ансамблеву модель називають однорідною. Водночас існують методи, що послуговуються різними типами базових алгоритмів; у такому разі різнорідні слабкі учні поєднуються в різнорідну ансамблеву модель.

Принципово важливим є узгодження вибору слабких учнів зі способом їх агрегування. У разі застосування базових моделей із низьким зміщенням та високою дисперсією доцільним є метод агрегування, спрямований на зменшення дисперсії; натомість для базових моделей із низькою дисперсією та високим зміщенням слід обирати метод агрегування, що зменшує зміщення.

Окреслене вище зумовлює питання про способи поєднання моделей. Можна виокремити три основні типи метаалгоритмів, призначених для поєднання слабких учнів:

- бегінг, що зазвичай оперує однорідними слабкими учнями, навчає їх незалежно та паралельно й поєднує за допомогою детермінованого процесу усереднення;
- бустинг, що зазвичай оперує однорідними слабкими учнями, навчає їх послідовно та адаптивно (кожна базова модель залежить від попередніх) і поєднує за детермінованою стратегією;
- стекінг, що зазвичай оперує різнорідними слабкими учнями, навчає їх паралельно й поєднує за допомогою метамоделі, яка формує прогноз на основі прогнозів окремих слабких моделей.

Узагальнено можна стверджувати, що бегінг спрямований передусім на формування ансамблевої моделі з меншою дисперсією порівняно з її складовими, тоді як бустинг і стекінг орієнтовані переважно на побудову

сильних моделей із меншим зміщенням, ніж у їхніх складових, хоча дисперсія при цьому також може зменшуватися.

У подальших розділах детально розглянуто бегінг і бустинг, що мають дещо ширше застосування, ніж стекінг, і дозволяють розкрити низку ключових понять ансамблевого навчання; після цього наведено стислий огляд стекінгу. (див. рис. 1.2).

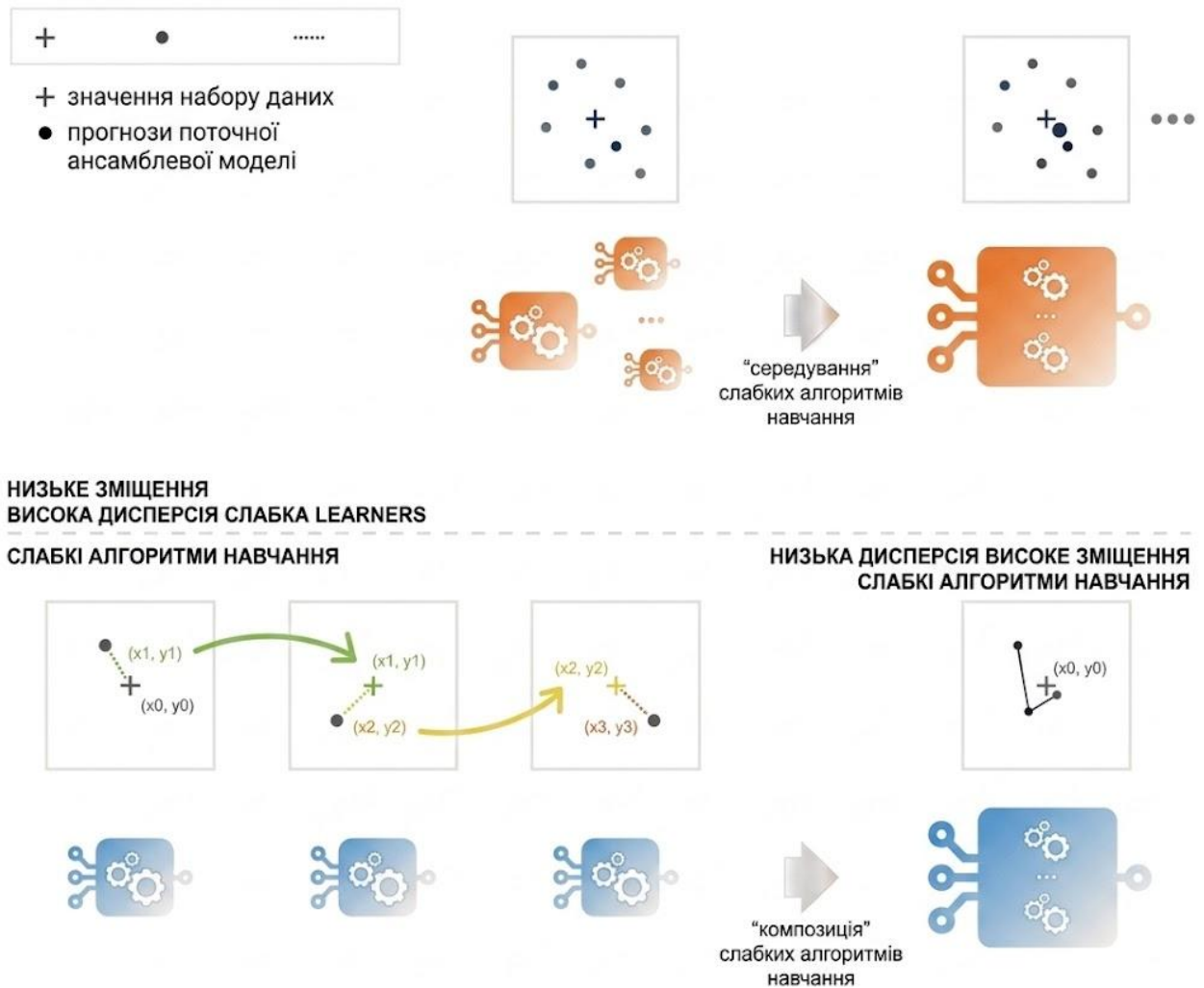


Рисунок 1.2 – Стекінг моделей машинного навчання

Слабких учнів можна поєднати, щоб отримати модель із кращими показниками. Спосіб поєднання базових моделей слід адаптувати до їхніх типів. Слабкі моделі з низьким зміщенням і високою дисперсією слід поєднувати так, щоб зробити сильну модель стійкішою, тоді як базові моделі з низькою

дисперсією та високим зміщенням краще поєднувати так, щоб зробити ансамблеву модель менш зміщеною.

1.2 Загальний опис бегінгу

У паралельних методах слабкі учні навчаються незалежно один від одного, що уможлиблює їх одночасне навчання. Найвідомішим із таких підходів є бегінг (англ. bootstrap aggregating – бутстреп-агрегування), метою якого є побудова ансамблевої моделі, стійкішої за окремі моделі, що її утворюють.

Розгляд доцільно розпочати з означення бутстрепа. Ця статистична техніка полягає у формуванні вибірок розміру B (бутстреп-вибірок) із початкового набору даних розміру N шляхом випадкового відбору B спостережень із поверненням [6] (див. рис. 1.3).

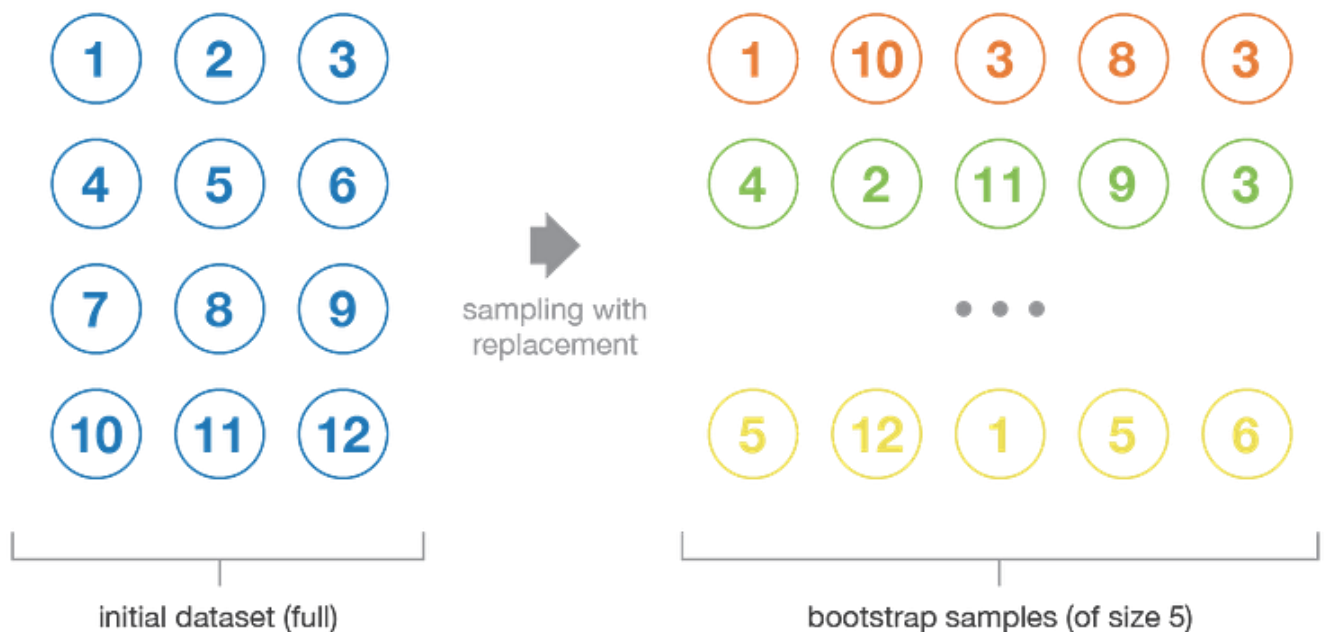


Рисунок 1.3 – Ілюстрація процесу бутстрепа

За певних припущень такі вибірки мають задовільні статистичні властивості: у першому наближенні їх можна розглядати як відібрані безпосередньо з істинного (зазвичай невідомого) розподілу даних і незалежно одна від одної, а отже, вважати репрезентативними та незалежними вибірками з

істинного розподілу (наближено незалежними та однаково розподіленими, н.о.р.). Чинність цього наближення зумовлюється виконанням двох умов. По-перше, розмір N початкового набору даних має бути достатньо великим, щоб охопити більшу частину складності розподілу, унаслідок чого відбір із набору даних слугує адекватним наближенням відбору з реального розподілу (умова репрезентативності). По-друге, розмір N має бути достатньо великим порівняно з розміром B бутстреп-вбірок, що забезпечує їх слабку корельованість (умова незалежності). Слід наголосити, що зазначені властивості мають наближений характер.



Рисунок 1.4 – Бутстреп часто використовують для оцінювання дисперсії або довірчого інтервалу деяких статистичних оцінювачів

Бутстреп-вбірки широко застосовуються, зокрема, для оцінювання дисперсії або довірчих інтервалів статистичних оцінювачів. За означенням статистичний оцінювач є функцією спостережень, а отже, випадковою величиною, дисперсія якої зумовлена цими спостереженнями. Оцінювання дисперсії такого оцінювача потребує його обчислення на кількох незалежних вибірках із досліджуваного розподілу. Оскільки в більшості випадків формування справді незалежних вибірок вимагало б обсягу даних, що перевищує реально доступний, застосовують бутстреп для генерування кількох наближено репрезентативних і наближено незалежних (майже н.о.р.) вибірок. Обчислення

значення оцінювача на кожній із них дозволяє наблизити його дисперсію (див. рис. 1.4).

Унаслідок навчання моделі, незалежно від типу задачі (класифікація чи регресія), отримують функцію, яка ставить у відповідність вхідним даним певний результат і визначена відносно навчального набору даних. З огляду на теоретичну дисперсію навчального набору (що є спостереженою вибіркою з істинного невідомого розподілу) підібрана модель також характеризується мінливістю: за умови спостереження іншого набору даних було б отримано іншу модель.

Сутність бегінгу полягає в підборі кількох незалежних моделей та усередненні їхніх прогнозів задля отримання моделі з нижчою дисперсією [2]. Оскільки на практиці підбір цілком незалежних моделей є неможливим через обмеженість обсягу даних, використовуються наближені властивості бутстреп-вбірок (репрезентативність і незалежність), що дозволяють підібрати наближено незалежні моделі.

На першому етапі формується сукупність бутстреп-вбірок, кожна з яких слугує наближено незалежним набором даних, відібраним з істинного розподілу. Для кожної вибірки підбирається слабкий учень, після чого отримані моделі агрегуються шляхом усереднення результатів, що забезпечує побудову ансамблевої моделі з меншою дисперсією порівняно з її складовими. Оскільки бутстреп-вбірки є наближено незалежними та однаково розподіленими (н.о.р.), такими ж властивостями характеризуються й навчені базові моделі. Відтак усереднення результатів слабких учнів зберігає математичне сподівання відповіді, проте зменшує її дисперсію, аналогічно до усереднення н.о.р. випадкових величин.

Тож, припускаючи, що ми маємо L бутстреп-вбірок (наближень L незалежних наборів даних) розміру B , позначених в (1.1).

$$\{z_1^1, z_2^1, \dots, z_B^1\}, \{z_1^2, z_2^2, \dots, z_B^2\}, \dots, \{z_1^L, z_2^L, \dots, z_B^L\} \quad (1.1)$$

Тут $z_b^l \equiv b$ -та реалізація l -го зразка.

Таким чином можна підібрати L майже незалежних слабких учнів (по одному на кожен набір даних) $w_1(\cdot), w_2(\cdot), \dots, w_L(\cdot)$, а потім агрегувати їх за допомогою певного процесу усереднення, щоб отримати ансамблеву модель із нижчою дисперсією. Наприклад, ми можемо визначити нашу сильну модель так, щоби вирішити задачу регресії (1.2) чи класифікації (1.3).

$$s_L(\cdot) = \frac{1}{L} \sum_{l=1}^L w_l(\cdot), \quad (1.2)$$

$$s_L(\cdot) = \frac{1}{L} \sum_{l=1}^L w_l(\cdot)$$

$$s_L(\cdot) = \arg \max_k [\text{card}(l|w_l(\cdot) = k)]. \quad (1.3)$$

Агрегування паралельно підібраних моделей може здійснюватися кількома способами. У задачах регресії результат ансамблевої моделі отримують безпосереднім усередненням результатів окремих моделей (1.2). У задачах класифікації клас, що його повертає кожна модель, інтерпретується як голос, а ансамблева модель повертає клас, який набрав більшість голосів (так зване жорстке голосування) (1.3). Альтернативно усереднюють надані всіма моделями ймовірності належності до класів і обирають клас із найвищою усередненою ймовірністю (м'яке голосування). Як усереднення, так і голосування можуть бути простими або зваженими за наявності відповідних вагових коефіцієнтів.

Істотною перевагою бегінгу є можливість його розпаралелювання. З огляду на незалежність підбору окремих моделей за потреби можуть застосовуватися інтенсивні методи паралельних обчислень (див. рис. 1.5).

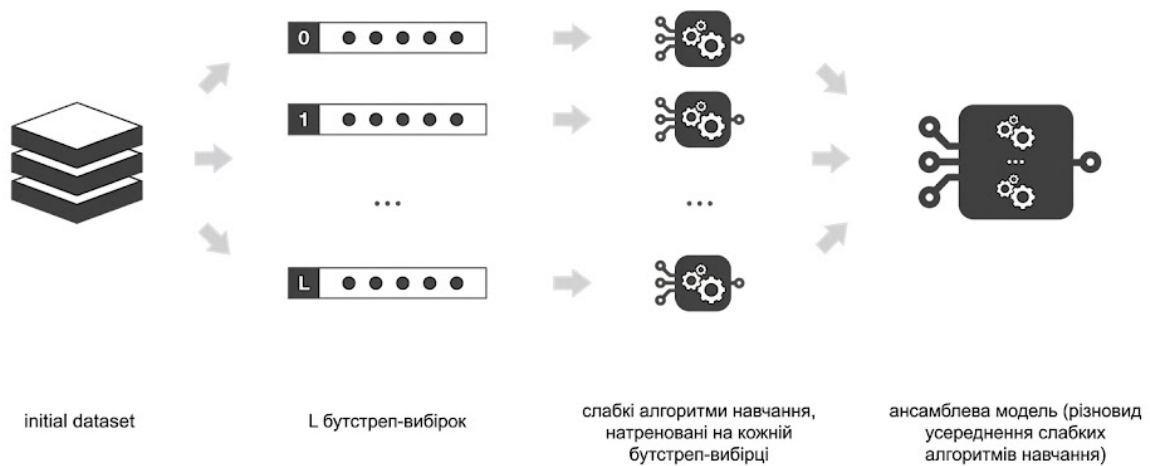


Рисунок 1.5 – Бегінг полягає в підборі кількох базових моделей на різних бутстреп-вибірках і побудові ансамблевої моделі, яка “усереднює” результати цих слабких учнів

Дерева рішень належать до найпоширеніших базових моделей в ансамблевих методах. Сильних учнів, утворених із сукупності дерев, прийнято називати лісами. Дерева, що формують ліс, можуть бути як неглибокими (з малою глибиною), так і глибокими (зі значною глибиною за умови неповного вирощування). Неглибокі дерева характеризуються меншою дисперсією, проте вищим зміщенням, що робить їх доцільнішими для послідовних методів, розглянутих далі. Натомість глибокі дерева мають низьке зміщення та високу дисперсію, а отже, відповідають методу бегінгу, орієнтованому передусім на зменшення дисперсії.

Підхід випадкового лісу – це метод бегінгу, у якому глибокі дерева, підбрані на бутстреп-вибірках, поєднуються для отримання результату з нижчою дисперсією. Проте випадкові ліси використовують ще один прийом, щоб зробити кілька підібраних дерев трохи менш корельованими між собою: під час вирощування кожного дерева, замість того щоб робити відбір лише з спостережень у наборі даних для формування бутстреп-вибірки, ми також робимо відбір з ознак і залишаємо лише випадкову підмножину з них для побудови дерева.

Відбір за ознаками зумовлює використання різними деревами неоднакової інформації під час ухвалення рішень, що зменшує кореляцію між їхніми результатами. Додатковою перевагою такого відбору є підвищення стійкості процесу ухвалення рішень до пропущених даних: спостереження з пропусками можуть бути піддані регресії або класифікації на основі дерев, що враховують лише ознаки без пропущених значень. Отже, алгоритм випадкового лісу інтегрує принципи бегінгу та випадкового вибору підпростору ознак, забезпечуючи побудову стійкіших моделей [7] (див. рис. 1.6).

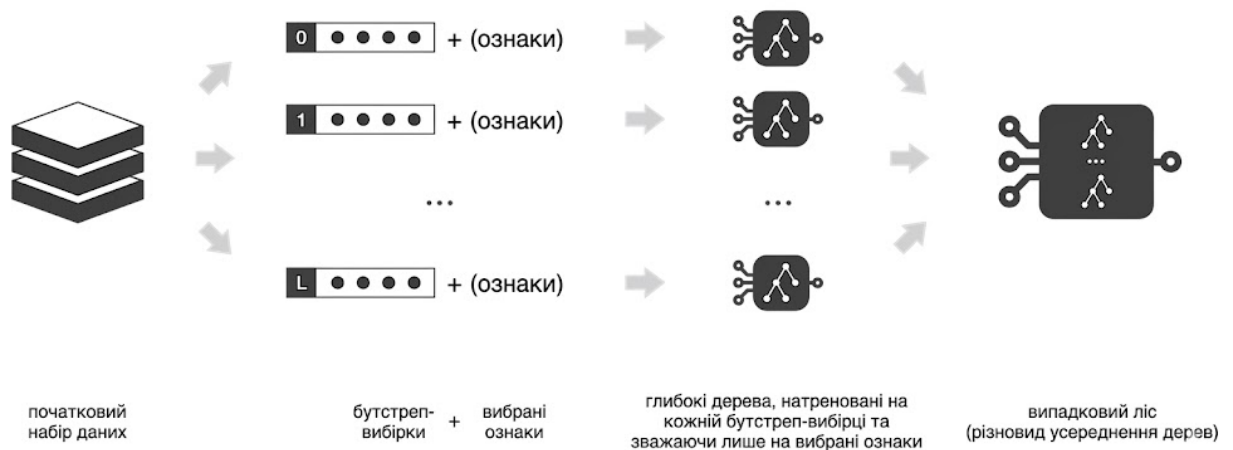


Рисунок 1.6 – Метод випадкового лісу – це метод бегінгу з деревами як слабкими учнями. Кожне дерево підбирається на бутстреп-вибірці, враховуючи лише випадково обрану підмножину змінних

У послідовних методах поєднувані слабкі моделі підбираються не незалежно, а ітеративно, причому навчання моделі на кожному кроці залежить від моделей, підібраних на попередніх кроках. Найвідомішим із таких підходів є бустинг, що формує ансамблеву модель із меншим зміщенням порівняно зі слабкими учнями, які її утворюють.

1.3 Загальний опис бустингу

Методи бустингу ґрунтуються на тих самих засадах, що й методи бегінгу: будується сімейство моделей, які агрегуються в сильного учня з вищою якістю. Проте, на відміну від бегінгу, орієнтованого передусім на зменшення дисперсії, бустинг передбачає послідовний адаптивний підбір слабких учнів, за якого кожна наступна модель надає більшої ваги тим спостереженням, що були незадовільно опрацьовані попередніми моделями. Унаслідок такого зосередження на найскладніших для підбору спостереженнях наприкінці процесу формується сильний учень із нижчим зміщенням (при цьому можливе й зменшення дисперсії). Подібно до бегінгу, бустинг застосовний як до задач регресії, так і до задач класифікації.

Оскільки бустинг орієнтований переважно на зменшення зміщення, як базові для нього зазвичай обирають моделі з низькою дисперсією та високим зміщенням. Зокрема, у разі використання дерев перевагу надають неглибоким деревам рішень із малою глибиною. Додатковим аргументом на користь застосування таких моделей є їхня менша обчислювальна витратність під час підбору, зумовлена малою кількістю ступенів свободи. Це є істотним з огляду на те, що, на відміну від бегінгу, обчислення в бустингу не піддаються розпаралелюванню, унаслідок чого послідовний підбір складних моделей виявився б надмірно витратним.

Після вибору слабких учнів необхідно визначити порядок їх послідовного підбору (яка інформація з попередніх моделей враховується під час підбору поточної) та спосіб їх агрегування (як поточна модель поєднується з попередніми). Зазначені питання розглянуто у двох наступних підрозділах на прикладі двох важливих алгоритмів бустингу – adaboost і градієнтного бустингу.

Зазначені метаалгоритми відрізняються способом створення та агрегування слабких учнів у межах послідовного процесу. Адаптивний бустинг оновлює вагові коефіцієнти, поставлені у відповідність кожному спостереженню навчального набору, тоді як градієнтний бустинг оновлює значення цих

спостережень. Така відмінність зумовлена різними підходами обох методів до розв’язання задачі оптимізації пошуку найкращої моделі, що подається у вигляді зваженої суми слабких учнів (див. рис. 1.7).

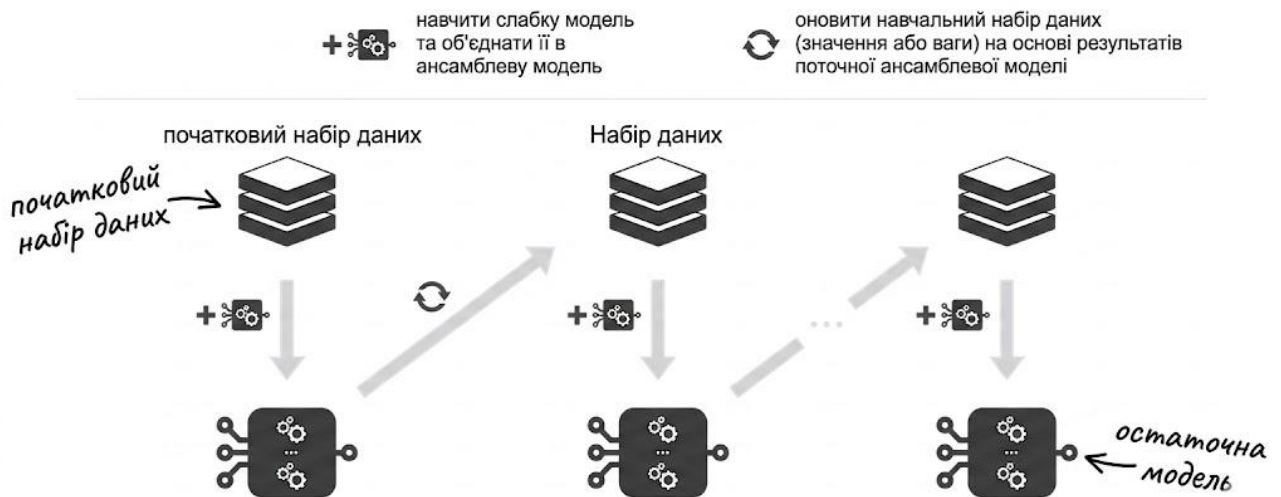


Рисунок 1.7 – Схема роботи методу бустингу

Бустинг полягає в ітеративному підборі слабкого учня, його агрегуванні до ансамблевої моделі та “оновленні” навчального набору даних, щоб краще врахувати сильні та слабкі сторони поточної ансамблевої моделі під час підбору наступної базової моделі.

В адаптивному бустингу (adaboost) ансамблевую модель означають як зважену суму L слабких учнів (1.4).

$$s_L(\cdot) = \sum_{l=1}^L c_l \times w_l(\cdot). \quad (1.4)$$

Тут c – вагові коефіцієнти, а w – слабкі учні.

Пошук найкращої ансамблевої моделі зазначеного вигляду є складною задачею оптимізації. Замість її одночасного розв’язання (визначення всіх коефіцієнтів і слабких учнів, що формують оптимальну адитивну модель) застосовується ітеративний процес оптимізації, значно зручніший для реалізації,

хоча й здатний приводити до субоптимального розв'язку. Слабкі учні додаються послідовно, причому на кожній ітерації відшукується найкраща пара (коефіцієнт, слабкий учень) для долучення до поточної ансамблевої моделі. Формально послідовність (s_l) визначається рекурентно у вигляді $s_l(\cdot) = s_{l-1}(\cdot) + c_l \times w_l(\cdot)$, де (c_l) і (w_l) обираються так, щоб модель (s_l) найкраще узгоджувалася з навчальними даними й відповідала найбільшому можливому покращенню порівняно зі (s_{l-1}) . Відтак можна записати:

$$(c_l, w_l(\cdot)) = \arg \min_{c, w(\cdot)} E(s_{l-1}(\cdot) + c \times w(\cdot)) = \arg \min_{c, w(\cdot)} \sum_{n=1}^N e(y_n, s_{l-1}(x_n) + c \times w(x_n)),$$

де $E(\cdot)$ – помилка підгонки заданої моделі, а $e(\cdot, \cdot)$ – функція втрат. Таким чином, замість глобальної оптимізації за всіма L моделями суми оптимум наближають локально – шляхом послідовної побудови та долучення слабких учнів до сильної моделі.

Зокрема, для задачі бінарної класифікації можна показати, що алгоритм adaboost подається у вигляді процесу з такими складовими. По-перше, оновлюються вагові коефіцієнти спостережень, після чого навчається новий слабкий учень з акцентуванням на спостереженнях, неправильно класифікованих поточною ансамблевою моделлю. По-друге, цей слабкий учень долучається до зваженої суми відповідно до коефіцієнта оновлення, що відображає якість моделі: чим вища якість слабого учня, тим більшим є його внесок у сильного учня.

Розгляньмо задачу бінарної класифікації з N спостереженнями за умови застосування алгоритму adaboost із заданим сімейством слабких моделей. На початковому етапі (перша модель послідовності) усім спостереженням приписують однакові ваги $1/N$. Далі L разів (для L учнів послідовності) виконують такі кроки:

- підбирають найкращу слабку модель за поточних ваг спостережень;
- обчислюють коефіцієнт оновлення – скалярну метрику якості слабого учня, що визначає міру його врахування в ансамблевій моделі;

- оновлюють сильного учня шляхом долучення нового слабкого учня, помноженого на його коефіцієнт оновлення;
- обчислюють нові ваги спостережень, які визначають пріоритетність спостережень на наступній ітерації (ваги неправильно спрогнозованих агрегованою моделлю спостережень зростають, а правильно спрогнозованих – зменшуються).

Унаслідок повторення наведених кроків послідовно будують L моделей та агрегують їх у лінійну комбінацію, зважену коефіцієнтами, що відображають якість кожного учня. Слід зазначити, що існують модифікації базового алгоритму adaboost, зокрема LogitBoost (для класифікації) та L2Boost (для регресії), які різняться переважно вибором функції втрат [9] (див. рис. 1.8).

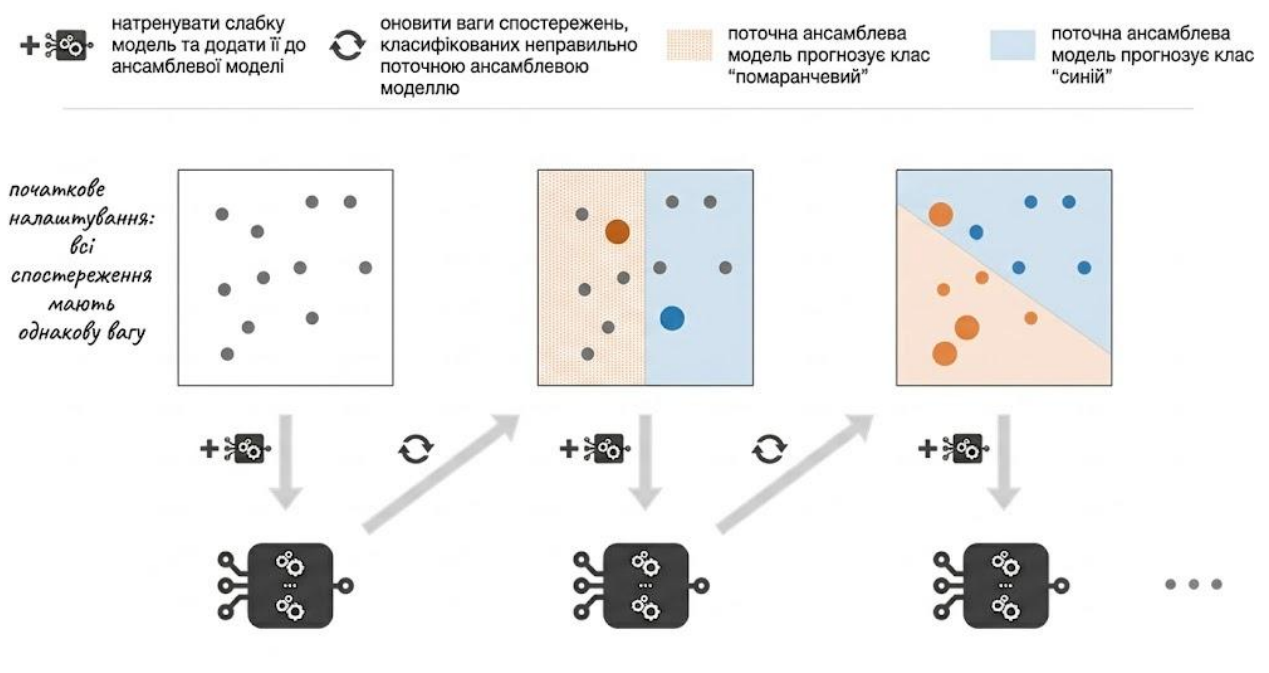


Рисунок 1.8 – AdaBoost оновлює ваги спостережень на кожній ітерації

Ваги правильно класифікованих спостережень зменшуються відносно ваг неправильно класифікованих спостережень. Моделі, що працюють краще, мають вищі ваги в підсумковій ансамблевій моделі.

У градієнтному бустингу шукана ансамблева модель також подається як зважена сума слабких учнів (див. вираз 1.4 вище).

Як ми вже згадували для adaboost, пошук оптимальної моделі такого вигляду надто складний, і потрібен ітеративний підхід. Головна відмінність від адаптивного бустингу полягає у визначенні послідовного процесу оптимізації. Справді, градієнтний бустинг зводить задачу до задачі градієнтного спуску: на кожній ітерації ми підбираємо слабкого учня до протилежного значення градієнта поточної помилки підгонки відносно поточної ансамблевої моделі. Спробуймо прояснити цей останній момент. По-перше, теоретичний процес градієнтного спуску за ансамблевою моделлю можна записати:

$$s_l(\cdot) = s_{l-1}(\cdot) - c_l \times \nabla_{s_{l-1}} E(s_{l-1})(\cdot). \quad (1.5)$$

Тут $E(\cdot)$ – помилка підгонки заданої моделі, c_l – коефіцієнт, що відповідає розміру кроку, а $-\nabla_{s_{l-1}} E(s_{l-1})(\cdot)$ – це протилежне значення градієнта помилки підгонки відносно ансамблевої моделі на кроці $l-1$. Це (доволі абстрактне) протилежне значення градієнта є функцією, яку на практиці можна обчислити лише для спостережень у навчальному наборі даних (для яких ми знаємо входи й виходи): ці обчислення називають псевдозалишками, прикріпленими до кожного спостереження. Ба більше, навіть якщо ми знаємо для спостережень значення цих псевдозалишків, ми не хочемо додавати до нашої ансамблевої моделі будь-яку функцію: ми хочемо додати лише новий екземпляр слабкої моделі. Тож природним кроком є те, щоб підібрати слабкого учня до псевдозалишків, обчислених для кожного спостереження. Нарешті, коефіцієнт c_l обчислюється за допомогою одновимірного процесу оптимізації (лінійний пошук для отримання найкращого розміру кроку c_l).

Розгляньмо застосування техніки градієнтного бустингу із заданим сімейством слабких моделей. На початковому етапі (перша модель послідовності) псевдозалишки покладають рівними значенням спостережень. Далі L разів (для L моделей послідовності) виконують такі кроки:

- підбирають найкращого слабкого учня до псевдозалишків (наближають протилежне значення градієнта відносно поточного сильного учня);

- обчислюють оптимальний розмір кроку, що визначає міру оновлення ансамблевої моделі в напрямі нового слабкого учня;
- оновлюють ансамблеву модель шляхом долучення нового слабкого учня, помноженого на розмір кроку (виконують крок градієнтного спуску);
- обчислюють нові псевдозалишки, що визначають для кожного спостереження напрям подальшого оновлення прогнозів ансамблевої моделі.

Повторюючи ці кроки, ми послідовно будемо наші L моделей та агрегуємо їх за підходом градієнтного спуску. Зауважимо, що тоді як адаптивний бустинг намагається на кожній ітерації розв'язати точно “локальну” задачу оптимізації (знайти найкращого слабкого учня та його коефіцієнт для додавання до сильної моделі), градієнтний бустинг натомість використовує підхід градієнтного спуску й може легше адаптуватися до великої кількості функцій втрат. Таким чином, градієнтний бустинг можна розглядати як узагальнення *adaboost* на довільні диференційовні функції втрат [10] (див. рис. 1.9).

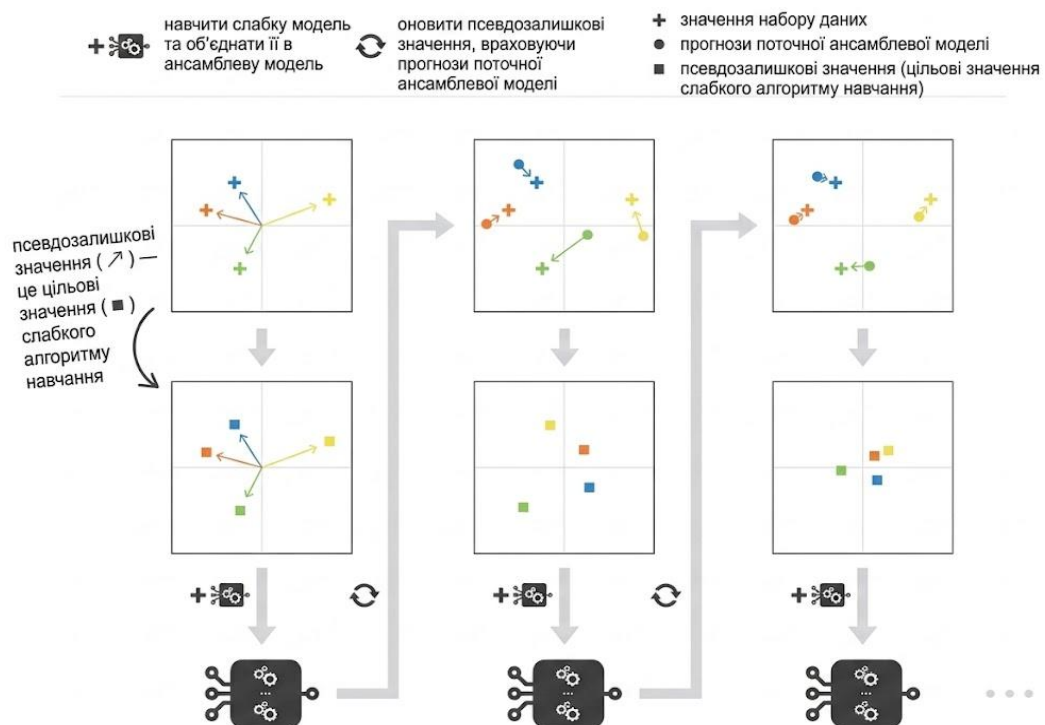


Рисунок 1.9 – Градієнтний бустинг оновлює значення спостережень на кожній ітерації

Слабких учнів навчають підбиратися до псевдозалишків, що вказують, у якому напрямку коригувати прогнози поточної ансамблевої моделі, щоб зменшити помилку.

1.4 Загальний огляд стекінгу

Стекінг відрізняється від бегінгу та бустингу за двома основними ознаками. По-перше, він зазвичай оперує різнорідними слабкими учнями, поєднуючи різні алгоритми навчання, тоді як бегінг і бустинг послуговуються переважно однорідними слабкими учнями. По-друге, стекінг навчається поєднувати базові моделі за допомогою метамоделі, тоді як бегінг і бустинг агрегують слабких учнів за детермінованими правилами.

Як ми вже згадували, ідея стекінгу полягає в тому, щоб навчити кілька різних слабких учнів і поєднати їх, навчивши метамодель видавати прогнози на основі кількох прогнозів, повернутих цими слабкими моделями. Тож нам потрібно визначити дві речі, щоб побудувати нашу модель стекінгу: L учнів, яких ми хочемо підібрати, і метамодель, що їх поєднує.

Наприклад, для задачі класифікації як слабких учнів можна обрати класифікатор KNN, логістичну регресію та метод опорних векторів (SVM), а як метамодель – нейронну мережу. У такому разі нейронна мережа отримує на вхід результати трьох слабких учнів і навчається формувати на їх основі остаточні прогнози.

Розгляньмо побудову ансамблю стекінгу, утвореного з L слабких учнів. Для цього необхідно виконати такі кроки:

- розділяють навчальні дані на дві частини (folds);
- обирають L слабких учнів і підбирають їх до даних першої частини;
- для кожного з L слабких учнів формують прогнози для спостережень другої частини;
- підбирають метамодель на другій частині, використовуючи прогнози слабких учнів як вхідні дані.

Розділення набору даних на дві частини зумовлене тим, що прогнози на даних, які використовувалися для навчання слабких учнів, є нерелевантними для навчання метамоделі. Очевидним недоліком такого розділення є те, що для навчання базових моделей і для навчання метамоделі використовується лише половина наявних даних. Для подолання цього обмеження може застосовуватися підхід k-кратного перехресного навчання, аналогічний до k-кратної перехресної перевірки, за якого для навчання метамоделі залучаються всі спостереження: прогноз для кожного спостереження формується екземплярами слабких учнів, навченими на k-1 частинах, що не містять цього спостереження. Інакше кажучи, навчання на k-1 частинах із подальшим прогнозуванням на залишковій частині виконується ітеративно для отримання прогнозів за всіма частинами, що дає змогу сформувати релевантні прогнози для кожного спостереження й навчити на них метамодель (див. рис. 1.10).

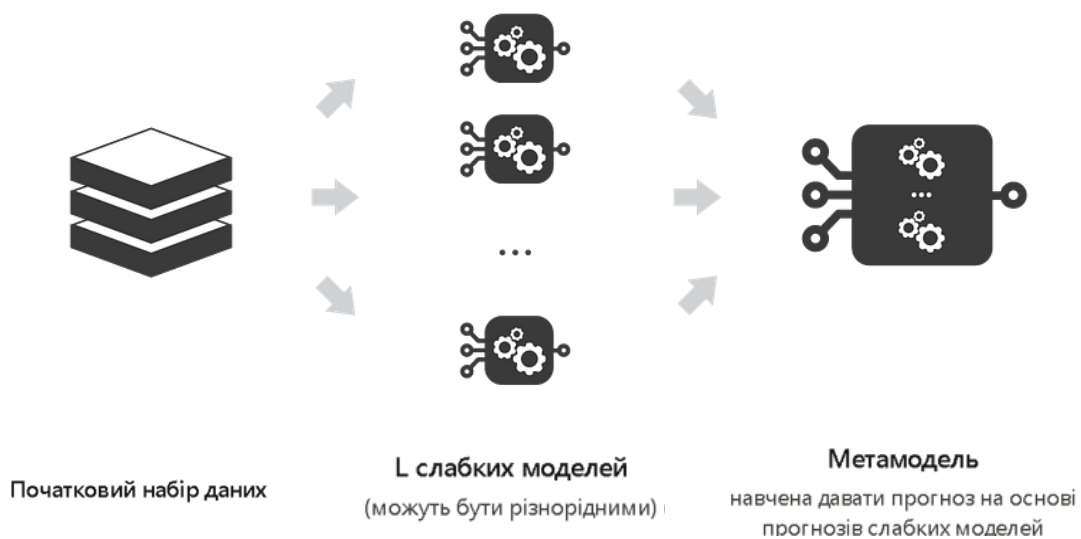


Рисунок 1.10 – Схематичне зображення стекінгу

Стекінг полягає в навчанні метамоделі видавати результати на основі результатів, повернутих слабкими учнями нижнього шару.

Можливим розширенням стекінгу є багаторівневий стекінг, що передбачає виконання стекінгу з кількома шарами. Як приклад розгляньмо трирівневий стекінг. На першому рівні підбирають L обраних слабких учнів. На другому рівні

замість єдиної метамоделі, що навчається на прогнозах слабких моделей, підбирають M таких метамodelей. На третьому рівні підбирають підсумкову метамодель, яка отримує на вхід прогнози M метамodelей попереднього рівня.

З практичного погляду для кожної метамоделі різних рівнів багаторівневої ансамблевої моделі стекінгу алгоритм навчання може обиратися довільно, зокрема й серед алгоритмів, застосованих на нижчих рівнях. Слід зауважити, що додавання рівнів є витратним або щодо обсягу даних (за відсутності техніки на кшталт k -кратного поділу, що потребує більшого обсягу даних), або щодо часу (за її використання, що потребує підбору значної кількості моделей) (див. рис. 1.11).

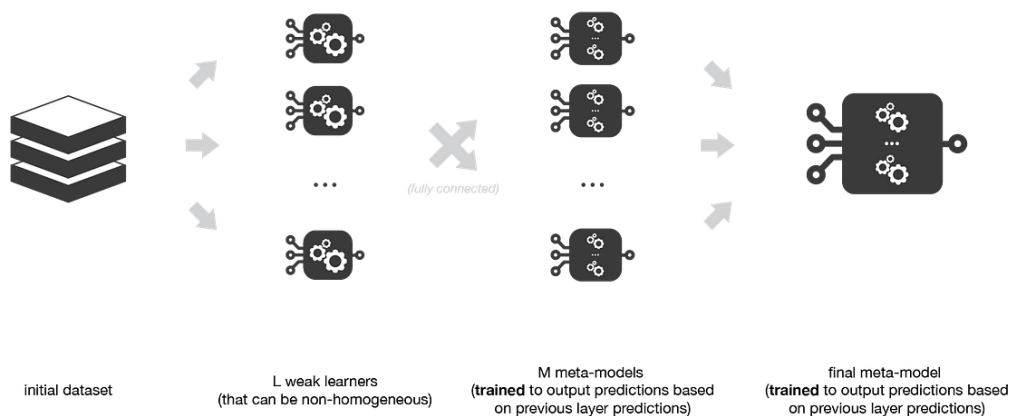


Рисунок 1.11 – Схематичне зображення багаторівневого стекінгу

Багаторівневий стекінг розглядає кілька шарів стекінгу: деякі метамоделі навчаються на результатах, повернутих метамоделями нижнього шару, і так далі. Тут ми зобразили 3-шарову модель стекінгу.

Основні результати розділу можуть бути узагальнені в таких положеннях:

- ансамблеве навчання є парадигмою машинного навчання, у межах якої сукупність моделей (слабких учнів, або базових моделей) навчається для розв'язання спільної задачі та поєднується задля підвищення якості прогнозування;

- засадничою є гіпотеза про те, що коректне поєднання слабких учнів забезпечує отримання точніших та/або стійкіших моделей;
- у методах бегінгу сукупність екземплярів однієї базової моделі навчається паралельно та незалежно на різних бутстреп-вибірках із подальшим агрегуванням за допомогою процесу усереднення;
- усереднення наближено н.о.р. підібраних моделей у бегінгу забезпечує побудову ансамблевої моделі з нижчою дисперсією порівняно з її складовими, чим зумовлюється доцільність застосування базових моделей із низьким зміщенням і високою дисперсією;
- у методах бустингу сукупність екземплярів однієї базової моделі навчається послідовно, причому навчання поточного слабкого учня на кожній ітерації залежить від попередніх учнів і, зокрема, від їхньої якості на даних;
- ітеративна стратегія навчання в бустингу, що адаптується до недоліків попередніх моделей, забезпечує побудову ансамблевої моделі з нижчим зміщенням порівняно з її складовими, чим зумовлюється доцільність застосування слабких учнів із низькою дисперсією та високим зміщенням;
- у методах стекінгу різнорідні слабкі учні підбираються незалежно, після чого над ними навчається метамодель, яка формує прогнози на основі результатів базових моделей.

Ансамблеве навчання полягає в поєднанні базових моделей задля отримання ансамблевої моделі з вищими показниками якості. Попри те що бегінг, бустинг і стекінг є найпоширенішими ансамблевими методами, можливими є й інші варіанти, які можуть бути розроблені для адаптації до конкретних задач. Це потребує насамперед глибокого розуміння розв'язуваної задачі та творчого підходу до її вирішення.

2 АНАЛІЗ МЕТОДІВ ІНТЕГРАЦІЇ МОДЕЛІ МАШИННОГО І ГИБОКОГО НАВЧАННЯ

2.1 Тенденції ансамблевого навчання

Завдяки силі та ефективності системи ансамблевого навчання для покращення прогностичної продуктивності моделей, ансамблеве навчання стало важливою дослідницькою тенденцією останніх років, що призвело до збільшення кількості досліджень, що використовуються для ансамблевого навчання в кількох сферах застосування.

Отже, в цьому розділі представлена ця важлива тенденція в одній з найпотужніших баз даних, «Scopus». Щоб показати, наскільки щороку збільшувалася кількість індексованих опублікованих статей для ансамблевого навчання, а також різні прикладні галузі ансамблевого навчання з 2018 по 2024 рік, пошуковий запит у цій базі даних – «Ансамблеве навчання» та «Ансамблеве глибоке навчання». Пошук здійснювався за назвами статей, анотацією та ключовими словами.

Кількість статей, знайдених за цим терміном, оцінюється в понад 25000, що свідчить про зростання тенденції ансамблевого навчання протягом кількох років. Можна зазначити також, що галузь комп'ютерних наук має найбільшу оціночну кількість згаданих статей, яка оцінюється в понад 16000 документів.

Згідно з наведеною вище статистичною інформацією, очевидно, що дослідження в галузі ансамблевого навчання та ансамблевого глибокого навчання щороку зростають швидше завдяки їхній здатності покращувати ефективність прогнозування. За оцінками, найбільша кількість статей з використанням «ансамблевого навчання» та «ансамблевого глибокого навчання» у 2024 році становила понад 7000 та понад 2300 документів відповідно. Крім того, ансамблеве навчання та глибоке ансамблеве навчання застосовувалися в кількох галузях, особливо в інформатиці, з найвищим рівнем використання

ансамблевого навчання та глибокого ансамблевого навчання 30% та 35,1% відповідно.

2.2 Основи ансамблевого навчання

Загальна основа будь-якої ансамблевої системи навчання полягає у використанні агрегуючої функції G для об'єднання набору h базових класифікаторів, c_1, c_2, \dots, c_h , з метою отримання єдиного вихідного результату. Нехай маємо набір даних розміром n та ознак розмірністю $m, D = \{(x_i, y_i)\}, 1 \leq i \leq n, x_i \in R^m$. Прогнозування вихідного значення на основі цього ансамблевого методу задається рівнянням $y_i = \phi(x_i) = G(c_1, c_2, \dots, c_k)$.

На рис. 2.1 ілюструється загальна абстрактна структура ансамблю навчання.

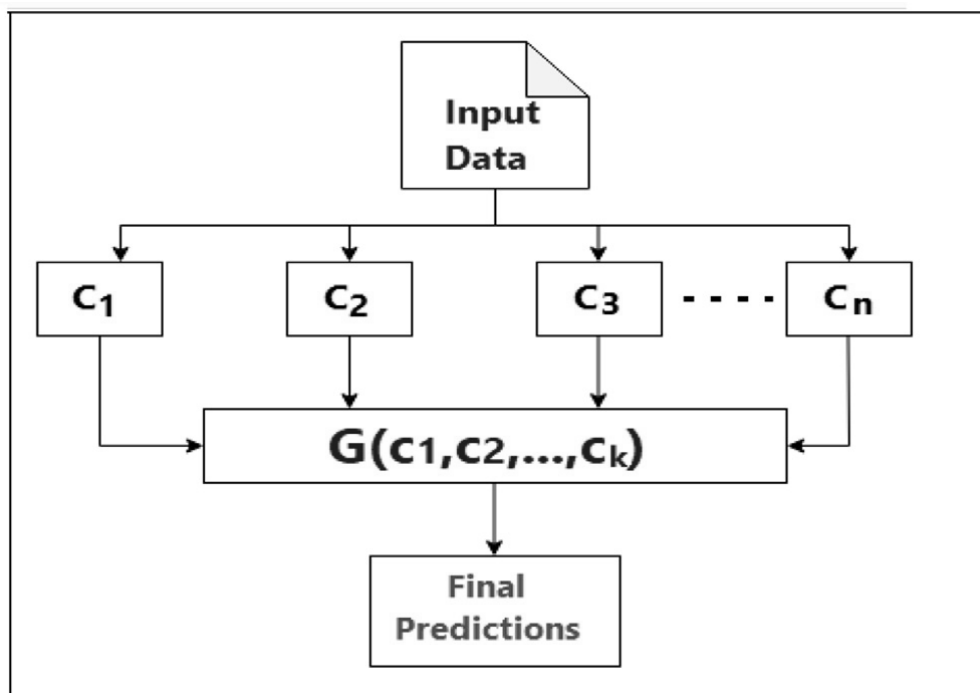


Рисунок 2.1 – Загальний фреймворк ансамблевого навчання

Усі ансамблі складаються з колекції базових класифікаторів (ансамблю класифікаторів), які були навчені на вхідних даних, що створюють прогнози, що об'єднуються для створення сукупної прогнозованої моделі. Ансамблеві стратегії відрізняються способом вибору базових класифікаторів, що

навчаються. Дві стратегії генерують різноманітність серед базових класифікаторів на основі їхньої природи, або однорідні, або неоднорідні ансамблі, як показано на рис. 2.2.

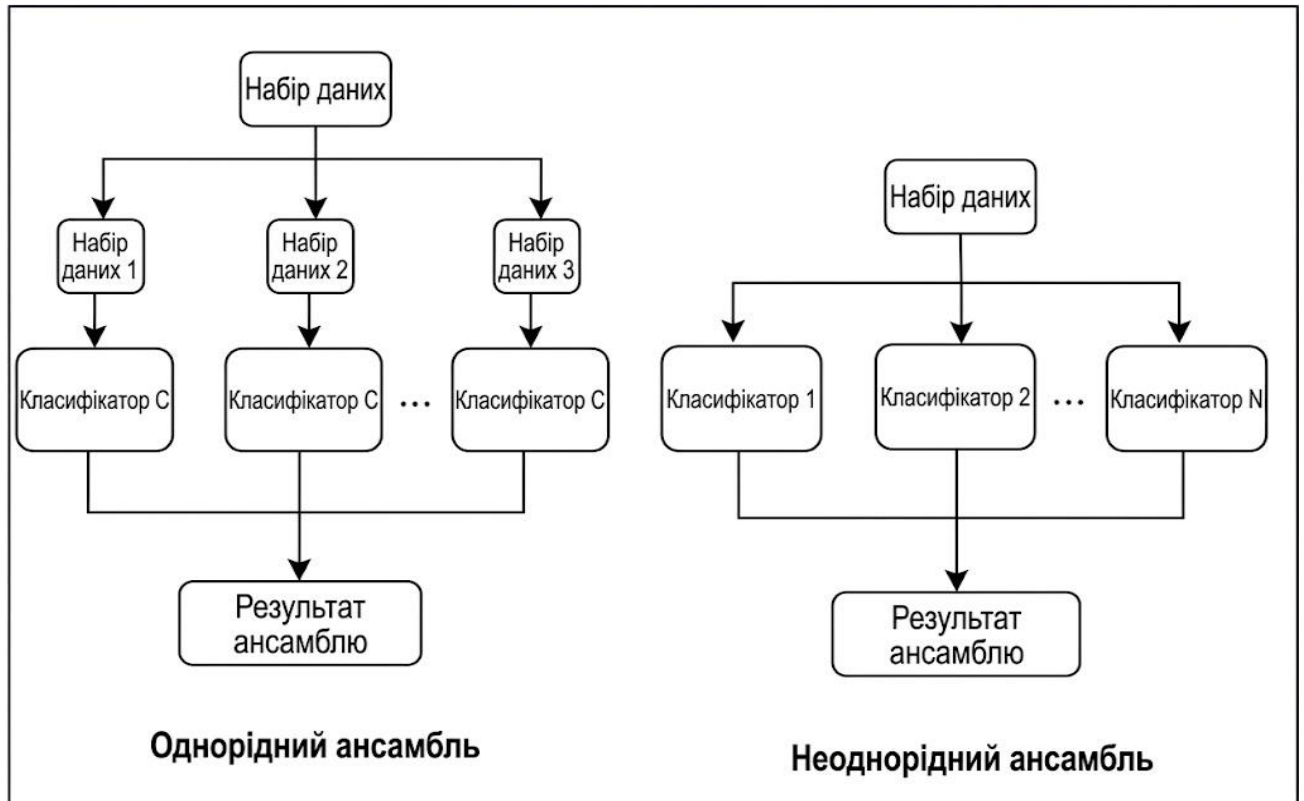


Рисунок 2.2 – Загальне подання гомогенного та гетерогенного ансамблів

Однорідний (гомогенний) ансамбль складається з базових класифікаторів одного типу, причому кожен класифікатор базується на різних даних. Метод вибору ознак у цій стратегії однаковий для різних навчальних даних. Основна складність в однорідній формі полягає в генерації різноманітності з одного й того ж алгоритму навчання. Тоді як гетерогенні ансамблі складаються з різної кількості базових класифікаторів, оскільки кожен класифікатор базується на одних і тих самих даних. У гетерогенних класифікаторах метод вибору ознак відрізняється для одних і тих самих навчальних даних. Зрештою, методи гомогенних ансамблів є більш привабливими для дослідників, оскільки їх легше зрозуміти та застосувати. Крім того, побудова гомогенних ансамблів є менш витратною, ніж гетерогенних.

Загалом, будь-який ансамблевий фреймворк можна розглядати та визначати за допомогою трьох характеристик, які впливають на його продуктивність. Перша з них – це залежність від навчених базових моделей, незалежно від того, чи є вони послідовними чи паралельними. Друга характеристика – це методи об'єднання, які передбачають вибір відповідного процесу для об'єднання виходів базових класифікаторів з використанням методу голосування за вагою або метанавчання. Третя характеристика – це гетерогенність задіяних базових класифікаторів, незалежно від того, чи є вони однорідними, чи гетерогенними. У таблиці 2.1 підсумовано характеристики популярних ансамблевих методів.

Таблиця 2.1 – Категорії ансамблевих методів

Метод	Залежність (побудова)	Метод об'єднання	Гетерогенність (однорідність)
Bagging (Бегінг)	Паралельна	Зважене голосування	Однорідна (Гомогенна)
Random Forest (Випадковий ліс)	Паралельна	Зважене голосування	Однорідна (Гомогенна)
Boosting (Бустинг)	Послідовна	Зважене голосування	Однорідна (Гомогенна)
AdaBoost	Послідовна	Зважене голосування	Однорідна (Гомогенна)
Gradient Boosting (Градiєнтний бустинг)	Послідовна	Зважене голосування	Однорідна (Гомогенна)
Extreme Gradient Boosting (XGBoost)	Послідовна	Зважене голосування	Однорідна (Гомогенна)
Stacking (Стеккінг)	Паралельна	Метанавчання	Неоднорідна (Гетерогенна)
Hybrid Ensemble (Гібридний ансамбль)	Обидва варіанти	Обидва варіанти	Неоднорідна / Однорідна
Hybrid Ensemble (Гібридний ансамбль)	Обидва варіанти	Обидва варіанти	Неоднорідна / Однорідна

Далі ці характеристики будуть детально обговорені.

2.3 Вибірка даних

Вибір методу вибірки даних є одним з найважливіших факторів, що впливають на продуктивність ансамблевої системи. В ансамблевій системі нам потрібна різноманітність у рішеннях щодо вибірки даних базових класифікаторів. Існує дві стратегії методів вибірки з навчального набору даних в ансамблевій системі: стратегія незалежних наборів даних та стратегія залежних наборів даних.

У стратегії незалежних наборів даних підмножини – це ті, що не залежать одна від одної. На відміну від цього, стратегія залежних наборів даних – це підмножини, залежні одна від одної. Основною перевагою використання стратегії незалежних наборів даних є те, що на його піднабір даних не впливає продуктивність інших піднаборів даних, на відміну від використання стратегії залежних наборів даних, де на його піднабір даних впливають результати попереднього піднабору даних. Складність методу вибірки даних в обох стратегіях полягає у визначенні оптимального розміру кожної вибірки та максимальної кількості вибірок. Крім того, визначення відповідної стратегії для вибірок даних відповідно до різних методів ансамблю.

2.4 Класифікатори базової лінії навчання

Різнманітність базових класифікаторів є другим впливовим фактором в ансамблевій системі. В основі будь-якої ансамблевої системи лежать два методи навчання окремих членів ансамблю: метод послідовного ансамблю та метод паралельного ансамблю. У методі послідовного ансамблю різні учні навчаються послідовно через залежність даних. Таким чином, помилки, допущені першою моделлю, послідовно виправляються другою моделлю, як показано на рис. 2.3.

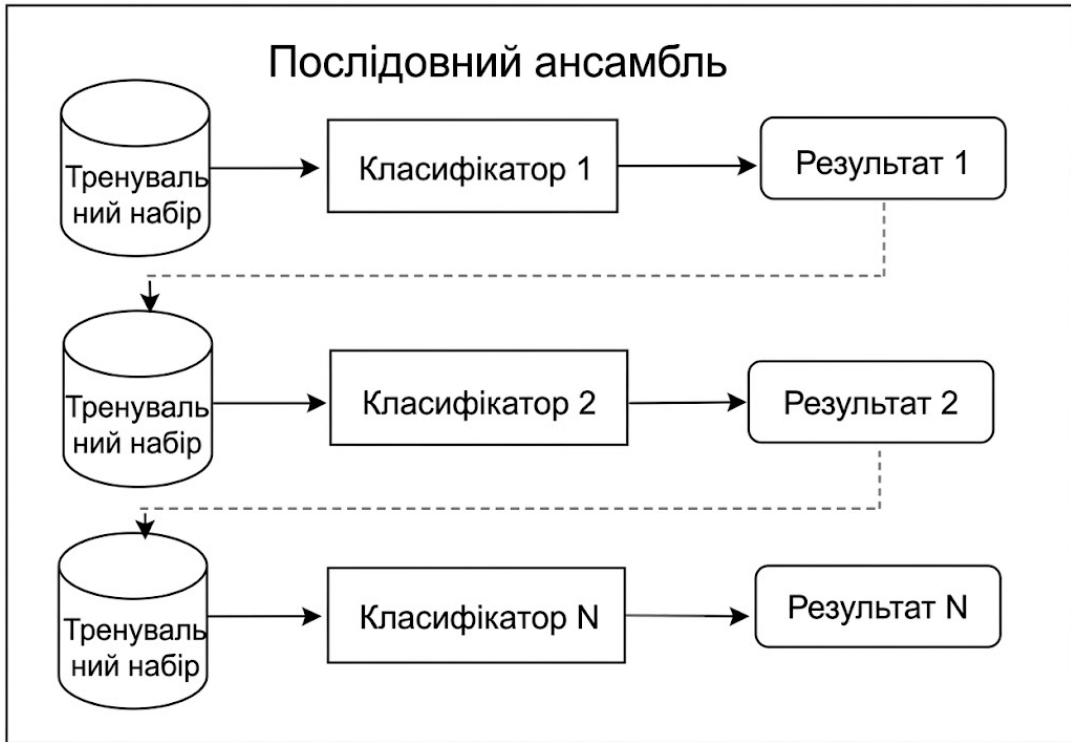


Рисунок 2.3 – Концепція послідовного ансамблевого навчання

Отже, головною перевагою послідовних методів є використання залежності між базовими учнями. Тоді як у методі паралельного ансамблю базові учні генеруються одночасно, оскільки немає залежності від даних. Отже, кожен дани в базовому учні генеруються незалежно, як показано на рис. 2.4.

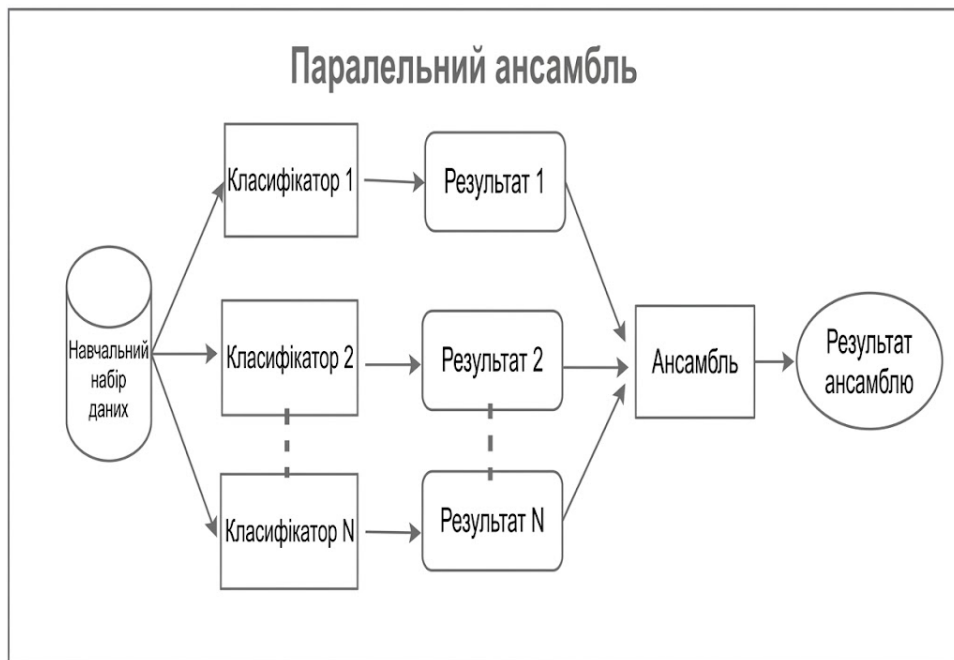


Рисунок 2.4 – Концепція паралельного ансамблевого навчання

Основна перевага цього методу полягає у використанні незалежності між базовими учнями. Таким чином, помилки, допущені однією моделлю, відрізняються від помилок, виявлених в іншій незалежній моделі, що дозволяє ансамблевій моделі обчислити середнє значення помилок.

2.5 Метод злиття

Об'єднання результатів означає інтеграцію результатів базових класифікаторів в один результат. Існує два методи об'єднання: метод голосування та метод метанавчання. У кожному методі описано, як реалізувати інтегрування вихідних даних базових класифікаторів, їхні переваги та складність їх застосування, а також вибрати відповідний метод об'єднання для кожного з ансамблевих методів. Методи об'єднання можна використовувати з незалежними або залежними вибірками даних, а також з паралельними або послідовними базовими класифікаторами.

Методи голосування зазвичай використовуються в задачах класифікації або регресії для покращення прогнозової ефективності. Крім того, методи голосування є відповідним методом інтегрування для методів розподілу та підвищення. Перший метод об'єднання - це ансамбль голосування, який включає три методи: максимальне голосування, усереднення та середньозважене голосування. Ми обговоримо в кожному методі голосування природу його реалізації, а також переваги та недоліки.

Перший і найпопулярніший метод голосування – це максимальне голосування, яке часто називають мажоритарним голосуванням жорстким голосуванням. Ідея максимального голосування передбачає збір прогнозів для кожної мітки класу та прогнозування мітки класу з найбільшою кількістю голосів, як показано у виразі (1.2) . Наприклад, припускаючи, що ми об'єднуємо три класифікатори, C_1 , C_2 та C_3 , які призначають такі класифікації навчальній вибірці: $[0,0,1]$ виводить результат $y = \text{mode} [0,0, 1] = 0$. Ми б класифікували вибірку як "клас 0". Максимальне голосування часто використовується в методі

розподілу. Іншим типом максимального голосування є м'яке голосування. М'яке голосування передбачає збір прогнозованих ймовірностей для кожної мітки класу та прогнозування мітки класу з найбільшою ймовірністю, як показано у виразі (1.3). Максимальне голосування відрізняється від м'якого голосування тим, що як тільки ми знаємо прогноз для будь-якого з базових класифікаторів, нам не потрібно зберігати жодної іншої інформації про розподіл ймовірностей прогнозів.

З іншого боку, м'яке голосування потребує зберігання та використання всіх значень розподілу, що робить його більш обчислювально складним та дорогим для зберігання. Однак, при м'якому голосуванні ми можемо використовувати різні методи для розрахунку прогнозу, такі як обчислення максимальних або середніх значень ймовірності. Загалом, метод максимального голосування має переваги простоти розуміння та найпростішого методу голосування. Недоліки методу максимального голосування включають обчислювальні витрати на використання кількох базових моделей. Крім того, максимальне голосування марне, коли прогнози базових класифікаторів мають однакові результати та можуть не відповідати всім потребам.

Другий метод голосування – це усереднювальне. Ідея усереднювального голосування полягає в тому, що прогнози витягуються з кількох моделей, а середнє значення прогнозів використовується для формування остаточного прогнозу. Середній прогноз обчислюється за допомогою середнього арифметичного, яке є сумою прогнозів, поділеною на загальну кількість зроблених прогнозів. Наприклад, припустимо, що ансамбль класифікаторів містить три елементи: $C1(x) = [0.9, 0.1]$, $C2(x) = [0.2, 0.8]$ та $C3(x) = [0.6, 0.4]$. Середній прогноз буде таким: для обчислення класу 0 у 0 $[0.9 + 0.2 + 0.6/3] = 0.566$. А для обчислення класу 1 у 1 $[0.1 + 0.8 + 0.4/3] = 0.433$, отримаємо прогноз $y = 0$.

Метод усереднення має перевагу в тому, що він є найсильнішим з точки зору прогностичної сили. Крім того, він є точнішим за продуктивністю, ніж мажоритарне голосування, і зменшує перенавчання. Також усереднення є

природним конкурентом методу максимального голосування за розподіл. До недоліків методу усереднення належить його обчислювальна вартість порівняно з методом максимального голосування, оскільки він вимагає усереднення результатів прогнозування всіх базових моделей. Одним з обмежень методу усереднення полягає в тому, що він припускає, що всі базові моделі в ансамблі однаково ефективні. Однак це не так, оскільки деякі моделі можуть бути кращими за інші.

Третій метод голосування – це середньозважене голосування, яке є дещо модифікованою версією усереднювального голосування. Ідея середньозваженого голосування полягає в різних вагових коефіцієнтах, що надаються базовим учням, що вказує на важливість кожної моделі в прогнозуванні. Помноживши кожен прогноз на вагу класифікаторів для отримання зваженої суми, а потім поділивши результат на суму ваг класифікатора, ці ваги можна використовувати для обчислення середньозваженого значення для кожного класу 0 або класу 1.

Наприклад, припустимо, що ансамбль класифікаторів містить три члени: $C1(x) = [97.2, 2.8]$, $C2(x) = [100.0, 0]$ та $C3(x) = [95.8, 4.2]$. Він має постійні ваги для членів ансамблю $[0.84, 0.87, 0.75]$. Щоб обчислити клас 0, $y_0 = ((97,2 * 0,84) + (100,0 * 0,87) + (95,8 * 0,75)) / (0,84 + 0,87 + 0,75) = 97,763$. А щоб обчислити клас 1, $y_1 = ((2,8 * 0,84) + (0 * 0,87) + (4,2 * 0,75)) / (0,84 + 0,87 + 0,75) = 2,235$, отримаємо прогноз $y = 0$.

Метод голосування за середньозваженим значенням є точнішим, ніж простий метод голосування за середнім значенням. Складність використання ансамблю середньозважених значень полягає у виборі відносної ваги кожного члена. Крім того, обчислення є дорожчим, ніж метод голосування за середнім значенням, оскільки воно вимагає обчислення середньозваженого значення результатів прогнозування всіх базових моделей, що робить його малозастосовним.

3 ОПИС ІНТЕГРАЦІЇ МЕТОДІВ МАШИННОГО НАВЧАННЯ

У цьому розділі представлено два аспекти. Перший аспект включає структуру найпопулярніших методів ансамблевого навчання та окремо перелічує переваги, недоліки та проблеми впровадження кожного методу. Другий аспект представляє ідею глибокого ансамблевого навчання та переваги його застосування порівняно з традиційним ансамблевим навчанням. Також обговорюються проблеми глибокого навчання, які долає ансамблеве глибоке навчання. Крім того, представлені різні стратегії застосування ансамблевого глибокого навчання та переваги кожної стратегії з поясненням факторів, які можуть впливати на її ефективність.

3.1 Ансамблеві методи без інтеграції з глибоким навчанням

Ці методи детально розглянуті у першому розділі, тому викладемо коротко їх опис для майбутнього порівняння з ансамблевими методами з інтеграцією глибокого навчання.

Метод бегінгу (агрегування бутстрепом) базується на паралельному створенні кількох підмножин даних з початкового масиву для одночасного навчання незалежних базових елементів. Кінцевий прогноз формується шляхом мажоритарного голосування. Головною перевагою підходу є ефективне зниження дисперсії та усунення перенавчання, зокрема на багатовимірних даних, яскравим прикладом чого є алгоритм випадкового лісу (RF). Водночас метод характеризується високою обчислювальною вартістю, підвищеним зміщенням і втратою інтерпретованості, а його реалізація потребує точного налаштування кількості моделей та оптимального вибору стратегії об'єднання їхніх виходів.

Метод бустингу є послідовним процесом, де кожна наступна адаптивна модель фокусується на виправленні помилок попередніх, надаючи більшої ваги некоректно класифікованим об'єктам. Цей підхід, що охоплює алгоритми AdaBoost, SGB та XGBoost, дозволяє успішно знижувати як дисперсію, так і

зміщення, зберігаючи легкість інтерпретації. Проте через необхідність послідовного навчання бустинг складно масштабується, є чутливим до перенавчання при збільшенні ітерацій і може потребувати більше часу для навчання порівняно з бегінгом. Об'єднання результатів тут також реалізується через різні модифікації зваженого голосування.

Метод стекінгу (стекового узагальнення) передбачає інтеграцію прогнозів кількох різнорідних базових моделей (рівень 0) за допомогою окремої мета-моделі (рівень 1), яка навчається оптимально комбінувати їхні виходи. Такий паралельний підхід зазвичай демонструє вищу точність порівняно з окремими алгоритмами, але створює серйозні ризики перенавчання через дублювання прогнозів. Багаторівневий стекінг є трудомістким та обчислювально дорогим процесом, який ускладнює інтерпретацію та стикається із проблемою "прокляття розмірності", вимагаючи складного обґрунтування кількості й типу базових моделей для посилення загальної прогностичної сили системи.

3.2 Ансамблеві методи з глибоким навчанням

В останні роки глибоке навчання або глибоке нейронне навчання призвело до низки досягнень у різних завданнях. Архітектури глибокого навчання продемонстрували великий успіх майже у всіх задачах, пов'язаних з машинним навчанням у різних галузях, таких як NLP, комп'ютерний зір, розпізнавання мовлення та машинний переклад.

Моделі глибоких нейронних мереж – це нелінійні методи, які навчаються за допомогою стохастичного алгоритму навчання. Це означає, що вони є дуже гнучкими, здатними вивчати складні взаємозв'язки між змінними та наближено оцінювати будь-яку функцію відображення. Недоліком цієї гнучкості є те, що моделі потребують вищої дисперсії.

Високу дисперсію глибокої моделі можна вирішити за допомогою підходу ансамблевого глибокого навчання шляхом навчання кількох глибоких моделей для вирішення проблеми та об'єднання їхніх прогнозів. Отже, методи

ансамблевого глибокого навчання стосуються навчання кількох базових глибоких моделей та об'єднання деяких правил для створення прогнозів. Ансамблеве глибоке навчання прагне ефективно поєднати основні переваги кількох моделей глибокого навчання з перевагами системи ансамблевого навчання.

Незважаючи на потужність методів ансамблевого глибокого навчання у покращенні ефективності прогнозування, більшість літератури з ансамблевого глибокого навчання зосереджена лише на застосуванні більшості алгоритмів голосування для підвищення продуктивності через його простоту.

Ансамблеве навчання на основі моделей глибокого навчання є складнішим, ніж ансамблеве навчання на основі традиційних класифікаторів, через глибокі нейронні мережі, що містять від мільйонів до мільярдів гіперпараметрів, які потребують багато часу та простору для навчання кількох базових моделей глибокого навчання. Таким чином, гіперпараметри є проблемою у застосуванні методів ансамблевого глибокого навчання.

Стратегії ансамблевого навчання формуються в контексті маніпулювання рівнем даних або рівнем базової моделі. При маніпулюванні на рівні даних, шляхом вибірки даних або перехресної перевірки даних (повторної вибірки) для створення нових навчальних наборів для навчання різних базових моделей. При маніпулюванні на рівні базових моделей, глибоке навчання відрізняється більшою різноманітністю стратегій, ніж традиційне або машинне навчання, що полягає в можливості зменшення кількості гіперпараметрів, які використовуються в ансамблі базових глибоких моделях, шляхом вибору тієї ж моделі та зміни гіперпараметрів.

На рис. 3.1 показано чотири стратегії, за допомогою яких можна проводити глибоке навчання на основі ансамблю, представленого: (А) Застосування багатьох різних базових моделей з використанням тих самих даних. (В) Застосування різних структур тієї ж базової моделі з використанням тих самих даних. (С) Застосування багатьох різних базових моделей з використанням

багатьох різних зразків даних. (D) Застосування різних структур однієї й тієї ж базової моделі з використанням багатьох різних зразків даних.

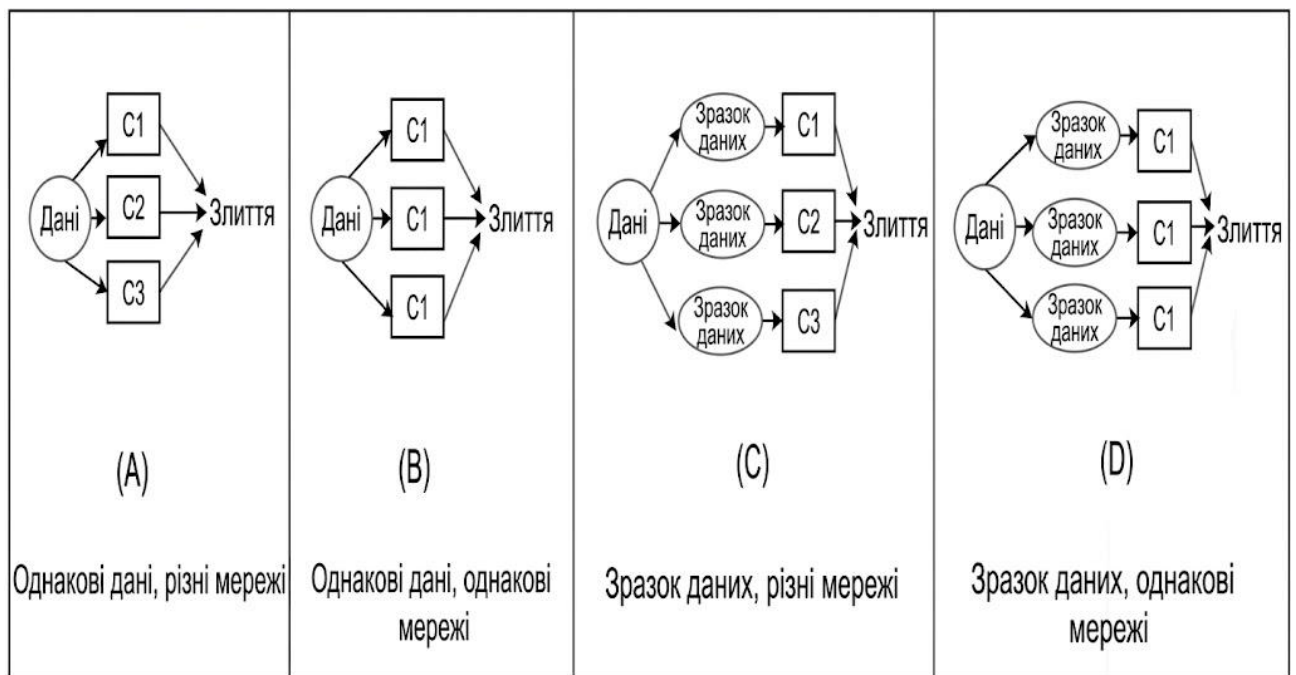


Рисунок 3.1 – Різні випадки ансамблевого глибокого навчання

Порівняння цих стратегій показує, що стратегія А та стратегія С сумісні з моделями глибокого навчання та традиційними методами навчання. Тоді як стратегія В та стратегія D застосовуються лише до моделей глибокого навчання та не можуть бути використані з традиційними методами навчання, що робить ансамблеві стратегії глибокого навчання різноманітними. Крім того, стратегія В та стратегія D дозволяють ансамблевому глибокому навчанню зменшувати гіперпараметри базових глибоких моделей за допомогою різних структур однієї й тієї ж базової моделі, змінюючи деякі значення гіперпараметрів. На додаток до цих стратегій, сила ансамблевої системи глибокого навчання залежить від її дизайну, від визначення найефективніших моделей глибокого навчання для вирішення проблеми та визначення відповідної кількості базових моделей глибокого навчання, таких як три або більше, а також визначення оптимального співвідношення для розділення даних, такого як (80–20 або 70–30 або 60–40). Крім того, ми розглядаємо фактори, які можуть впливати на глибоку ансамблеву систему, такі як визначення характеру генерації даних, навчання глибоких

базових моделей та вибір найбільш підходящого методу об'єднання виходів базових класифікаторів, як згадувалося раніше. Ці три фактори впливають на загальну структуру ансамблевої системи.

3.3 Оцінювання ансамблевих методів

З появою підходів до ансамблевого навчання було проведено багато досліджень для оцінки методів ансамблю. Оцінювання має вирішальне значення для визначення ефективності певного ансамблевого методу. Існує кілька критеріїв для оцінювання ансамблю, включаючи прогностичну ефективність. Інші критерії, такі як обчислювальна складність або зрозумілість згенерованого ансамблю, також можуть бути важливими. Далі ми опишемо різні критерії оцінювання ансамблевого навчання.

Прогнозована продуктивність.

Метрики прогнозованої ефективності завжди були основним критерієм вибору ефективності класифікаторів. Крім того, показники прогнозованої ефективності вважаються об'єктивними та кількісно вимірюваними, тому їх часто використовують для практичного порівняння алгоритмів машинного навчання. Першим кроком до застосування прогнозованої ефективності є використання відповідного набору даних. Метод holdout – це типовий підхід до вимірювання прогнозованої ефективності, де заданий набір даних випадковим чином поділяється на дві підмножини: навчальні та тестові набори. Можуть бути використані й інші версії методу holdout. Зазвичай повторна вибірка даних, що означає їх поділ на навчальні та тестові набори різними способами. Два поширені методи повторної вибірки включають випадкову підвибірку та n-кратну перехресну перевірку.

Існують загальні показники для оцінки ансамблевої моделі. Точність класифікації є одним із популярних і найпростіших показників, який, як визначено в рівнянні (3.1) .

$$\text{Accuracy} = \frac{\text{number of true predictions}}{\text{total number of prediction}}. \quad (3.1)$$

У деяких випадках точність є недостатньою та може бути оманливою при оцінці ансамблевої моделі з незбалансованим розподілом класів. В останньому сценарії як альтернативні міри можна використовувати інші міри, такі як повнота, точність, специфічність та F-міра.

Рекол, також відома як чутливість, вимірює здатність моделі ансамблю ідентифікувати позитивні зразки, що, як визначено в рівнянні (3.2).

$$\text{Recall} = \frac{\text{true positive}}{\text{positive}}, \quad (3.2)$$

де «істинно позитивний» позначає кількість істинно позитивних спостережень, а «позитивне» позначає кількість позитивних спостережень.

Ще одним відомим показником продуктивності є прецизійність (точність позитивних результатів). Він кількісно визначає, скільки випадків, класифікованих як позитивні, насправді є позитивними. Формально рівняння точності визначається як (3.3).

$$\text{Precision} = \frac{\text{true positive}}{\text{true positive} + \text{false positive}}. \quad (3.3)$$

Аналогічно, специфічність вимірює, наскільки добре модель ідентифікує негативні зразки. Рівняння визначається як (3.4).

$$\text{Specificity} = \frac{\text{true negative}}{\text{negative}}, \quad (3.4)$$

де істинно негативне позначає кількість істинно негативних спостережень, а негативне позначає кількість негативних спостережень.

Зазвичай існує компроміс між показниками точності та повноти. Спроба покращити один показник часто призводить до зниження другого. Таким чином, F-міра кількісно визначає цей компроміс, обчислюючи середнє гармонійне як точності, так і повноти. Більш конкретно, цей показник визначається в рівнянні (3.5).

$$F - Measure = \frac{2 \times Precision \times Recall}{Precision + Recall} \cdot \quad (3.5)$$

Обчислювальна складність.

Обчислювальна складність ансамблевого підходу є додатковим важливим аспектом, який слід враховувати. Загалом, обчислювальні витрати стосуються кількості процесорного часу, необхідного для кожної моделі ансамблю. Обчислювальні витрати розподіляються на два показники складності: обчислювальні витрати на навчання та створення моделі ансамблю та обчислювальні витрати на прогнозування нового екземпляра: обчислювальні витрати на прогнозування є відносно невеликими порівняно з обчислювальними витратами на навчання ансамблю. Таким чином, цей показник слід враховувати. Що стосується пам'яті, меншій моделі ансамблю потрібно менше пам'яті для зберігання своїх компонентів. Крім того, менші ансамблі виконують швидше прогнозування.

Інші критерії.

Окрім обчислювальної складності та точності прогнозування, під час вибору найкращого методу ансамблю можна враховувати й інші критерії. Ці критерії включають інтерпретованість, масштабованість, зручність використання та стійкість моделі ансамблю. Інтерпретованість стосується здатності користувача розуміти результати ансамблю. Однак інтерпретованість зазвичай є суб'єктивним показником. Одним із багатьох кількісних показників та показників, які можуть допомогти нам оцінити цей критерій, є показник

компактності. Компактність в ансамблі можна оцінити за допомогою кількості задіяних класифікаторів та складності кожного класифікатора.

З іншого боку, масштабованість стосується здатності ансамблевого підходу побудувати модель класифікації з урахуванням великих обсягів даних. Незалежні ансамблеві методи вважаються більш масштабованими, ніж залежні методи, оскільки класифікатор, що бере участь в ансамблевому підході, може навчатися паралельно. Зручність використання – це ще один показник, який оцінює переваги користувача щодо розуміння того, як налаштовувати ансамблеві моделі, які вони використовують. Загалом кажучи, хороший ансамблевий метод повинен містити комплексний набір параметрів керування, які можна легко налаштувати.

3.4 Застосування ансамблю глибокого навчання

Методи ансамблевого навчання у застосуваннях глибокого навчання перевершують традиційні ансамблеві підходи в багатьох галузях, зокрема у класифікації зображень, обробці природної мови та інших задачах. У таблиці 3.1 наведено узагальнення робіт, присвячених застосуванню методів ансамблевого навчання у поєднанні з глибоким навчанням у різних предметних областях.

Проаналізуємо нашу таблицю 3.1. У сфері класифікації зображень застосовуються методи стекування на основі кількох згорткових нейронних мереж для роботи з набором даних CIFAR-10. Також методи стекування на основі кількох CNN використовуються для усунення розмиття зображень із використанням наборів даних GoPro та Video Deblurring. Запропоновано методи підсилення на основі CNN для пошуку зображень із використанням великих наборів даних.

Крім того, розроблено фреймворки глибокого підсилення шляхом інтеграції згорткових нейронних мереж в алгоритми boosting. Для їх оцінювання використовувалися еталонні набори даних Set12 та BSD68. Також

запропоновано методи голосування на основі CNN для оцінювання віку за зображеннями облич із використанням набору даних IMDB-WIKI.

Таблиця 3.1 – Застосування ансамблевого навчання в підході глибокого навчання [17]

Дослідження	Базові класифікатори	Метод злиття	Домен
[18]	DCN	Бегінг	Класифікація семантичних висловлювань
[32]	DCN	Бегінг	Розуміння розмовної мови
[25]	DNN	Бустинг	Розпізнавання виразів обличчя
[19]	RNN	Бегінг	Розпізнавання мовлення
[31]	RNN, CNN	Бегінг	Розпізнавання мовлення
[28]	CNNM	Бегінг	Класифікація зображень
[20]	DBN	Голосування	Медичне зображення
[22]	CNN, LSTM	Голосування	Аналіз тональності тексту
[26]	BPNN	Бегінг	Прогнозування повеней
[33]	CNN, DRN	Голосування	Медичне зображення
[30]	CNN, RetinaNet, Deep SVDD	Голосування	Медичне зображення
[29]	LSTM, ГРУ, CNN, RCNN, DNN	Голосування, об'єднання	Англійські настрої
[24]	6 моделей	Гібридний ансамбль	Класифікація багатомовних текстів
[21]	CNN	Голосування	Медичне зображення
[34]	LSTM, ГРУ	Голосування	Аналіз тональності тексту
[27]	CNN	Стекінг, бустинг	Аналіз тональності тексту
[Error! Reference source not found.]	CNN, LSTM, GRU, Bi_LSTM	Бегінг	Аналіз тональності тексту

Методи boosting на основі CNN активно застосовуються для пошуку зображень, розпізнавання облич, підрахунку об'єктів на зображеннях та інших задач комп'ютерного зору. Для цього використовуються різні набори даних, зокрема MNIST, CIFAR-10, ImageNet, а також набори даних для підрахунку натовпу, автомобілів і літаків.

Також запропоновано методи стекування на основі спрощення фіксованої нейронної мережі (SNNM) з використанням чотирьох наборів даних зображень облич. У роботі дослідники застосували метод підвищення до набору даних CIFAR-10, що містить 60 000 кольорових зображень, для навчання CNN. Зокрема, в класифікації медичних зображень автори застосували інтелектуальну систему охорони здоров'я для прогнозування серцевих захворювань, використовуючи підходи ансамблевого глибокого навчання та об'єднання ознак. Запропонована система досягла точності 98,5%. Автори запропонували голосувати на основі CNN для завдань візуального розпізнавання (розпізнавання вуха), використовуючи кілька наборів даних вуха.

Автори застосували голосування на основі мереж глибоких переконань, використовуючи великий набір даних Ініціативи нейровізуалізації хвороби Альцгеймера (ADNI). Інші дослідники запропонували голосування на основі мереж залишків (DRN) та CNN для розпізнавання меланоми на дермоскопічних зображеннях. Метод голосування досяг точності 76% завдяки використанню набору даних дермоскопічних зображень, що містить 1279 зображень. Також було застосовано голосування на основі CNN для виявлення туберкульозу за допомогою двох наборів даних рентгенівської рентгенографії туберкульозу. Метод голосування досяг точності 97,5% та 97,69% на наборах даних відповідно.

В іншій роботі запропоновано голосування на основі дев'яти CNN для класифікації особливостей барабанної перетинки. Голосування досягло середньої точності 93,67% завдяки використанню великої бази даних із більше 900000 зображень. Також було запропоновано метод голосування для автоматизованого скринінгу передракових захворювань шийки матки, використовуючи 30 000 зображень з кількох наборів даних. Метод голосування

поєднував оцінку трьох архітектур глибокого навчання: RetinaNet, Deep SVDD та CNN. Середня точність та F-оцінка становили 91,6% та 0,89% відповідно. Крім того, вчені застосували метод голосування на основі CNN для прогнозування захворювань, пов'язаних з нервово-м'язовими розладами, використовуючи два відповідні набори даних.

В області NLP запропоновано новий ансамбль для класифікації багатомовних текстів з використанням шести еталонних наборів даних. Також порівняно ефективність запропонованого та інших ансамблевих методів. Результати доводять, що цей метод перевершує сучасні ансамблеві аналоги. В іншому дослідженні запропоновано метод стекінгу, заснований на проблемах глибокої опуклої мережі (DCN) для розуміння розмовної мови (SLU). Метод стекінгу досяг точності 91,88% завдяки використанню набору даних ATIS, який складається з 5871 речення. Автори в іншій публікації запропонували ансамбль м'якого голосування на основі CNN та LSTM, використовуючи набір даних SemEval 2013. Також було представлено голосування на основі моделі CNN_RNN з використанням великого набору даних документів.

У наукових працях запропоновано метод голосування на основі CNN для прогнозування безпеки ліків з використанням англійських оглядів з форумів охорони здоров'я. Дослідники застосували як голосування, так і стекінг на основі кількох моделей глибокого навчання, а саме CNN, LSTM та GRU, використовуючи сім англійських наборів даних з оглядів фільмів. Також було застосовано голосування на основі Bi_LSTM для виявлення фейкових новин англійською мовою за допомогою NLP4IF 2019. В іншому проекті використали голосування на основі CNN та LSTM, застосувавши п'ять англійських наборів даних з оглядів фільмів. Також було запропоновано CNN на основі баггінгу та стекінгу з використанням кількох англійських наборів даних з оглядів.

Додатково науковці застосували як голосування, так і стекінг на основі кількох моделей глибокого навчання, а саме LSTM, GRU, CNN, RCNN та DNN, використовуючи два набори даних англійських твітів (SemEval та Toxic Comment). Було реалізовано стекінг на основі чотирьох моделей глибокого

навчання, а саме CNN, LSTM, GRU та BiLSTM, використовуючи набір даних англійських оглядів (SemEval). Науковці також запропонували ансамбль стекінгу на основі CNN для класифікації англійських твітів з використанням набору даних SemEval-2016. На противагу цьому, в іншій роботі застосували голосування на основі комбінації моделей CNN та LSTM, використовуючи арабський набір даних (ASTD). Крім того, було застосовано метод голосування на основі LSTM та GRU з використанням п'яти наборів даних з арабських твітів.

У суміжних галузях дослідники запропонували систему, яка спільно вивчає політики захоплення та стекінгу через мережу захоплення для стекінгу (GSNet), що дозволяє роботизованій руці правильно вибирати коробки зі столу та розміщувати їх на платформі. Було запропоновано метод Adaboost на основі DNN для класифікації рівня безпеки. Набір даних являє собою результати оцінки 100 Android-терміналів (включаючи смартфони, смарт-браслети, планшетні ПК) та даних зі шкіл, лікарень, заводів та інших середовищ.

Вчені також застосували посилену мережу глибоких переконань для розпізнавання виразів обличчя або змін форми на основі бази даних СК+, яка містить понад 320 зображень виразів. Автори іншого дослідження застосували метод стекінгу на основі RNN та CNN для розпізнавання мовлення, використовуючи набір даних TIMIT. Крім того, було застосовано стекінг на основі нейронних мереж зворотного поширення (BPNN) для прогнозування повеней. Інші дослідники застосували посилення CNN для розпізнавання одиниць мімічних дій. Також було реалізовано метод стекінгу на основі глибоких опуклих мереж (DCN) для семантичної класифікації висловлювань за набором даних висловлювань користувачів системи розмовного діалогу, а інші автори застосували стекінг RNN для систем розпізнавання мовлення на основі набору даних TIMIT.

4 ОХОРОНА ПРАЦІ ТА БЕЗПЕКА В НАДЗВИЧАЙНИХ СИТУАЦІЯХ

В умовах прискореного розвитку інформаційних технологій, професії у сфері ІТ займають ключове місце в економічній структурі та соціальному устрої. Відповідно, набуває актуальності забезпечення адекватних умов праці та охорони здоров'я для спеціалістів цієї галузі. Попри загальноприйнятту думку про порівняно низький ризик в роботі програмістів порівняно з іншими професіями, ця сфера містить специфічні ризики та особливості, що вимагають індивідуального підходу до питань охорони праці.

Аналіз робочого середовища ІТ-спеціалістів показує, що, незважаючи на зовнішню зручність, існують недоліки з точки зору ергономіки, психологічного навантаження та інших важливих аспектів. У цьому контексті важливим є вивчення основних аспектів охорони праці програмістів, аналіз потенційних ризиків та розробка рекомендацій щодо оптимізації умов праці для фахівців у галузі інформаційних технологій.

4.1 Питання щодо охорони праці

Фактори трудового середовища можуть істотно впливати на здоров'я та працездатність розробника програм. Неправильно організоване робоче місце може викликати проблеми із хребтом, шиєю, зап'ястям та іншими частинами тіла. Довгий час роботи за комп'ютером може призвести до тунельного синдрому зап'ястного каналу.

4.1.1 Ергономіка

У сучасних умовах трудової діяльності значуще місце займає організація робочого місця користувачів персональних комп'ютерів. Законодавство України наголошує на необхідності забезпечення безпечних та комфортних умов праці. Зокрема, відповідно до КЗпП та Закону України "Про охорону праці",

роботодавці мають обов'язки стосовно забезпечення працівників належними умовами.

При організації робочих місць із персональними комп'ютерами важливо забезпечити відстані між боковими частинами столів не менше ніж 1,2 м, а також враховувати, що відстань між задньою частиною одного комп'ютера та екраном іншого має бути 2,5 м. Структура робочого столу повинна бути розроблена відповідно до ергономічних стандартів, дозволяючи ефективно розташовувати необхідне обладнання, таке як дисплей, клавіатура, принтер, а також робочі документи.

Щодо параметрів робочого столу: його висота повинна знаходитися у діапазоні від 680 до 800 мм. Що стосується ширини та глибини, то вони повинні бути такими, щоб працівник міг з легкістю працювати в зоні доступності рук (для цього рекомендовані показники: ширини від 600 до 1400 мм, глибини від 800 до 1000 мм). Також необхідно передбачити комфортний простір для ніг користувача: висота – мінімум 600 мм, ширина – мінімум 500 мм, глибина на рівні колін – 450 мм і на рівні витягнутої ноги – не менше 650 мм.

Стілець на робочому місці повинен мати підйомно-поворотні характеристики, можливість регулювання по висоті, куту нахилу сидушки і спинки, а також відстані від спинки до зовнішнього краю сидіння. Поверхня сидіння має бути рівною, а її зовнішній край має мати округлу форму. Налаштування по кожному параметру повинно бути індивідуальним, інтуїтивним та надійно фіксуватися.

Інтервал регулювання частин стільця: для лінійних розмірів – 15-20 мм, для кутових 2-5°. Зусилля для регулювання не має перевищувати 20 Н. Висота сидіння повинна бути від 400 до 500 мм, а її ширина і глибина – не менше 400 мм. Сидіння має мати можливість нахилу до 15° вперед і до 5° назад. Висота спинки стільця – 300 ± 20 мм, ширина – не менше 380 мм, радіус кривизни горизонтально – 400 мм. Кут нахилу спинки може регулюватися в діапазоні 1-30° від вертикального положення. Відстань від спинки до сидіння – від 260 до 400 мм. Для зменшення напруження в руках рекомендується використовувати

підлокітники довжиною від 250 мм, шириною 50-70 мм, що налаштовуються по висоті від 230 до 260 мм та по відстані між ними від 350 до 500 мм. Матеріал сидіння і спинки має бути напівтвердим, антиковзним, що пропускає повітря і легко миється, а також не збирає статичний заряд.

Робоча зона повинна бути обладнана підніжкою шириною мінімум 300 мм, глибиною не менше 400 мм, з можливістю регулювання по висоті до 150 мм і нахилом до 20°. На підніжці має бути рельєфна поверхня і маленький борт по зовнішньому краю висотою 10 мм. Екран комп'ютера слід розміщувати на оптимальній відстані від користувача, що варіюється в межах 500-700 мм, але не ближче ніж 500 мм, з урахуванням легкості читання тексту і зображень. Екран повинен бути розташований так, щоб забезпечити комфорт при спостереженні, кутом у 30° від вертикалі.

4.1.2 Освітлення

Ефективне та грамотне виробниче світло підвищує якість зорової діяльності, зменшує втомленість, стимулює зростання продуктивності, сприяє комфортному робочому середовищу, додаючи спокій та позитив працівникам, а також підсилює безпеку роботи, зменшуючи ризик травм. Недостатнє або надто яскраве освітлення може спричинити зорове перевтомлення і головний біль.

Недолік світла може призводити до перевантаження очей, зниження уваги та передчасної втоми. Занадто яскраве світло викликає сліпоту, незадоволення та відчуття дискомфорту в очах. Неправильне розташування світла може утворювати відблиски, контрастні тіні та дезорієнтувати працівника. Ці фактори можуть спричинити нещасний випадок або професійні захворювання, тому важливо правильно розраховувати інтенсивність світла.

Розрізняють три типи світла - природне, штучне та комбіноване.

Природне світло – це варіант освітлення, при якому денне світло проникає через вікна або інші прозорі елементи приміщення. Інтенсивність природного світла може сильно коливатися в залежності від доби, сезону та інших факторів.

Штучне світло використовують у вечірній час або коли природне світло недостатнє. Якщо природне світло доповнюється штучним, таке освітлення називають змішаним.

За своїм призначенням штучне світло може бути робочим, евакуаційним, аварійним чи охоронним. Робоче світло поділяється на загальне та місцеве. При загальному освітленні світильники розташовані рівномірно по приміщенню. У системі комбінованого освітлення до загального додається додаткове місцеве світло.

Згідно норм "ДБН В.2.5-28:2018 Природне і штучне освітлення", в комп'ютерних залах слід застосовувати комбінований тип освітлення. Для виконання робіт високої зорової точності (об'єкт розрізнення 0,3 ... 0,5 мм) інтенсивність природного світла має бути не менше 1,5%. Для робіт середньої точності (об'єкт розрізнення 0,5 ... 1,0 мм) - не менше 1,0%. Як джерела штучного світла часто використовують лампи типу ЛБ або ДРЛ, об'єднані в світильниках над робочими столами.

Вимоги до освітленості для робочих місць з комп'ютерами такі: при високій точності зорової роботи – 300 лк (загальне) та 750 лк (комбіноване); при середній точності – 200 та 300 лк відповідно.

Також важливо, щоб усе поле зору було якісно освітлене. Світлова інтенсивність приміщення та яскравість екрану комп'ютера повинні бути близькими, адже надлишкове світло може збільшувати втомленість очей.

4.1.3 Параметри мікроклімату

Для створення комфортних умов праці в приміщеннях з комп'ютерною технікою важливо дотримуватися певних стандартів мікроклімату. В таблиці 4.1 наведені параметри мікроклімату для таких приміщень

Таблиця 4.1 – Параметри мікроклімату для приміщень, де встановлені комп'ютери

Параметри мікроклімату	Літній період	Зимовий період
Температура повітря, °С	20 – 24	22 – 24
Відносна вологість, %	40 – 60	40 – 60
Швидкість руху повітря, м/с	0,1 – 0,2	0,1 – 0,2
Рівень шуму, дБ	до 50	До 50

4.1.4 Емоційна психогігієна

У сфері ІТ охорона праці включає не лише фізичне, але й психічне здоров'я працівників. Фактори, що впливають на психіку працівників у галузі інформаційних технологій, є різноманітними та мають велике значення у створенні здорового робочого середовища.

Таблиця 4.2 – Допустимі значення параметрів неіонізуючого електромагнітного випромінювання

Найменування параметра	Допустимі значення
Напруженість електричної складової електромагнітного поля на відстані 50см від поверхні монітора	10 В / м
Напруженість магнітної складової електромагнітного поля на відстані 50см від поверхні монітора	0,3 А / м
Напруженість електростатичного поля	20кВ / м

До них належать:

– тривалість та інтенсивність робочого дня: довгі години роботи без відпочинку можуть призвести до перевтоми та стресу, впливаючи на психічне здоров'я;

- терміни виконання завдань і робочий тиск: нереалістичні терміни виконання завдань або надмірний робочий тиск можуть викликати стрес та тривогу;
- міжособистісні взаємодії: конфлікти в команді або нездорова робоча атмосфера можуть спричиняти емоційне вигорання;
- робота в умовах ізоляції або віддалена робота: відсутність соціальної взаємодії або підтримки може вплинути на почуття ізоляції та самотності;
- work-life баланс : нездатність знайти баланс між роботою та особистим життям може призвести до стресу та емоційного вигорання;
- значне розумове навантаження: постійна потреба в освоєнні нових технологій та адаптації до змін може бути стресовою;
- відсутність автономії або контролю над роботою: обмежений контроль над робочими процесами та відсутність автономії можуть викликати почуття безпорадності та фрустрації.

Для ефективної профілактики психоемоційних проблем, пов'язаних з роботою в IT-галузі, необхідно розробити комплексний підхід, який враховує різноманітність факторів, що впливають на психічне здоров'я працівників. Значущість цього підходу полягає в його спрямованості на зменшення стресорів у робочому середовищі та підвищення ресурсів для психічного відновлення.

Першочергово, акцентується увага на регулюванні тривалості робочого дня та інтенсивності роботи. Це включає розробку ефективних графіків роботи, що передбачають достатні перерви для відновлення, та уникнення перевантаження завданнями. Крім того, розглядається важливість балансу між роботою та особистим життям, який може бути підтриманий через гнучкі робочі години та можливість дистанційної роботи.

Ще одним важливим аспектом є оптимізація робочого середовища з точки зору ергономіки. Покращення умов роботи на робочому місці, включаючи забезпечення якісного обладнання та комфортних умов, сприятиме зниженню фізичного та психологічного дискомфорту.

Ключову роль відіграє також розвиток корпоративної культури, яка підтримує психологічне благополуччя. Це передбачає створення відкритого, підтримуючого співробітництва та комунікації середовища, заохочення взаємодопомоги між колегами, та реалізацію програм з управління стресом.

Для забезпечення довгострокового психічного благополуччя, розглядається імплементація регулярних тренінгів та навчальних програм, спрямованих на розвиток навичок управління стресом та підвищення особистісної стійкості. Також важливим є застосування систематичного моніторингу психічного стану працівників з використанням психометричних інструментів для раннього виявлення потенційних проблем.

Розуміння та належне врахування цих факторів є ключовим для забезпечення психічного благополуччя працівників у галузі ІТ, що, у свою чергу, позитивно впливає на їх продуктивність та загальне задоволення роботою.

4.2 Питання щодо безпеки в надзвичайних ситуаціях

В умовах швидкого розвитку інформаційних технологій та зростання залежності бізнес-процесів від ІТ-систем, питання безпеки в сфері ІТ набуває особливої актуальності. Розглядаючи безпеку праці в ІТ, особливу увагу необхідно приділити розробці та імплементації процедур реагування на надзвичайні ситуації, що включають технічні збої, кібератаки, витік конфіденційної інформації, фізичні загрози обладнанню та перебої в електропостачанні, а також інші аварійні стани, здатні порушити звичний хід робочих процесів.

Відповідальність за реалізацію ефективної системи безпеки лежить на ІТ-спеціалістах, керівництві компаній та інших зацікавлених сторонах. Необхідно забезпечити встановлення чітких правил та інструкцій, які регламентують дії персоналу у випадку виникнення надзвичайних ситуацій. Це включає визначення швидких комунікаційних каналів, розробку планів відновлення

роботи систем, а також проведення регулярних тренінгів та інструктажів з охорони праці для всіх працівників.

Плани евакуації та реагування мають бути адаптовані до можливих ІТ-ризиків. Вони повинні включати дії для захисту життя та здоров'я працівників, а також процедури збереження даних та обладнання. Плани мають бути детальними та зрозумілими для кожного члена команди, з чіткими інструкціями щодо дій в кризових ситуаціях.

Регулярні тренінги з охорони праці, які охоплюють надзвичайні ситуації в ІТ, є необхідними для підготовки персоналу до ефективного реагування на інциденти. Навчальні програми повинні включати імітації надзвичайних ситуацій, навчання з використанням спеціалізованого обладнання та вивчення найкращих практик у сфері кібербезпеки. Систематичний аналіз ризиків є важливим для ідентифікації потенційних вразливостей системи та розробки відповідних заходів запобігання. Встановлення систем моніторингу, які можуть виявити ознаки надзвичайних ситуацій на ранніх стадіях, допомагає зменшити потенційні збитки та забезпечити оперативне реагування.

Після кожної надзвичайної ситуації необхідно проводити детальний аналіз її причин, ефективності дій персоналу та наявних процедур реагування. На основі цього аналізу слід вносити корективи у плани надзвичайних дій, поліпшувати процедури безпеки та організовувати додаткове навчання. Ефективне управління надзвичайними ситуаціями в ІТ-сфері вимагає всебічного підходу, який поєднує в собі як технічні, так і організаційні аспекти. Відповідальність за це лежить на всіх рівнях організації, починаючи з ІТ-фахівців і закінчуючи вищим керівництвом. Завдяки встановленню та дотриманню відповідних процедур можна мінімізувати ризики та забезпечити безперебійну роботу в умовах, що постійно змінюються.

ВИСНОВКИ

У машинному навчанні зменшення зміщення та дисперсії моделей є одним із ключових факторів, що визначають успіх процесу навчання. У літературі було доведено, що об'єднання результатів різних алгоритмів класифікації може зменшити помилку узагальнення без збільшення дисперсії моделі. Попереднє є ключовою суттю так званого ансамблевого навчання. Численні дослідницькі зусилля віддавали перевагу ансамблевому навчанню над навчанням однієї моделі в різних областях.

Основною перевагою ансамблевого навчання є об'єднання кількох окремих моделей для покращення ефективності прогнозування та отримання сильнішої моделі, яка перевершує їх. У літературі існує кілька ансамблевих методів для покращення алгоритмів класифікації. Основна відмінність між будь-якими двома ансамблевими методами полягає в навчанні базових моделей та способах їх об'єднання. Кілька дослідницьких зусиль впровадили ансамблеве навчання в моделі глибокого навчання, щоб виправити проблеми, що виникають під час процесу навчання моделей глибокого навчання. Зазвичай основною проблемою моделей глибокого навчання є те, що вони потребують багато знань та досвіду для налаштування оптимальних гіперпараметрів з метою досягнення глобального мінімуму помилки.

Однак пошук оптимальних гіперпараметрів вимагає виснажливої техніки в просторі пошуку, що, у свою чергу, стає нудним та трудомістким завданням. Таким чином, кілька дослідницьких зусиль застосовували глибоке ансамблеве навчання в багатьох галузях, і більшість цих зусиль зосереджені навколо простих ансамблевих методів. У цій статті наведено всебічний огляд різних стратегій ансамблевого навчання, особливо у випадку глибокого навчання. У статті також проілюстровано останні тенденції в ансамблевому навчанні, використовуючи кількісний аналіз кількох дослідницьких робіт. Крім того, у статті запропоновано різні фактори, що впливають на успіх ансамблевих методів, включаючи вибірку

навчальних даних, навчання базових моделей та методи об'єднання базових моделей.

Також у статтях обговорювалися переваги та недоліки кожного ансамблевого методу. Крім того, у статті детально представлено та представлено кілька дослідницьких зусиль, які використовували ансамблеве навчання в широкому спектрі областей, та класифіковано ці зусилля як традиційні моделі машинного або глибокого навчання як базові класифікатори. Варто зазначити, що ансамбль моделей глибокого навчання, що використовують прості методи усереднення, не є розумним вибором і дуже чутливий до упереджених базових моделей. З іншого боку, впровадження різноманітності в ансамблеве глибоке навчання може стати стійким до упереджених базових моделей. Різноманітності можна досягти шляхом навчання різних базових архітектур глибокого навчання на кількох вибірках даних. Однак різноманітність обмежується вартістю обчислень та наявністю відповідних даних для вибірки.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Deng, L., & Yu, D. (2014). Foundations and trends in signal processing: Deep learning—methods and applications.
2. Zhang, C., & Ma, Y. (2012). Ensemble machine learning (Vol. 144). New York: springer.
3. Dong, X., Yu, Z., Cao, W., Shi, Y., & Ma, Q. (2020). A survey on ensemble learning. *Frontiers of Computer Science*, 14(2), 241-258.
4. Haralabopoulos, G., Anagnostopoulos, I., & McAuley, D. (2020). Ensemble deep learning for multilabel binary classification of user-generated content. *Algorithms*, 13(4), 83.
5. Tasci, E., Uluturk, C., & Ugur, A. (2021). A voting-based ensemble deep learning method focusing on image augmentation and preprocessing variations for tuberculosis detection. *Neural Computing and Applications*, 33(22), 15541-15555.
6. Efron B. Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*. 1979. Vol. 7, No. 1. P. 1–26.
7. Breiman L. Bagging Predictors. *Machine Learning*. 1996. Vol. 24, No. 2. P. 123–140.
8. Breiman L. Random Forests. *Machine Learning*. 2001. Vol. 45, No. 1. P. 5–32.
9. Freund Y., Schapire R. E. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences*. 1997. Vol. 55, No. 1. P. 119–139.
10. Friedman J. H. Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics*. 2001. Vol. 29, No. 5. P. 1189–1232.
11. Polak J., Cook D. A Study on Student Performance, Engagement, and Experience with Kaggle InClass Data Challenges. *Journal of Statistics and Data Science Education*. 2021. Vol. 29, No. 1. P. 63–70. DOI: <https://doi.org/10.1080/10691898.2021.1892554>.

12. Martseniuk V. About the approach of solving machine learning problems integrated with data from open source systems of electronic medical records / Vasyl Martseniuk, Nazar Milian // *Scientific Journal of TNTU*. – Tern. : TNTU, 2019. – Vol 95. – No 3. – P. 105–115. – (Instrument-making and information-measuring systems).
13. Zadvornyi O., Pastukh O. Comparative analysis of machine learning algorithms for market capitalization time series forecasting // *Scientific Journal of TNTU, Ternopil*. 2025. Vol 119. No 3. P. 26–34.
14. Petrov Y., Pastukh O. Comparative analysis of MLP AND KAN neural network architectures in neurointerface technologies // *Scientific Journal of TNTU, Ternopil*. 2025. Vol 119. No 3. P. 107–114.
15. Tymoshchuk D., Yasniy O. Information technology for predicting the hysteresis behavior of shape memory alloys based on a stacking ensemble machine learning model // *Scientific Journal of TNTU, Ternopil*. 2025. Vol 119. No 3. P. 134–146.
16. Demchyk V., Yasniy O. Modelling of functional properties of shape memory alloy by machine learning methods // *Scientific Journal of TNTU, Ternopil*. 2025. Vol 119. No 3. P. 56–62.
17. Mohammed, A., & Kora, R. (2023). A comprehensive review on ensemble deep learning: Opportunities and challenges. *Journal of King Saud University-Computer and Information Sciences*, 35(2), 757-774.
18. Tur, G., Deng, L., Hakkani-Tür, D., & He, X. (2012, March). Towards deeper understanding: Deep convex networks for semantic utterance classification. In *2012 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 5045-5048). IEEE.
19. Palangi, H., Deng, L., & Ward, R. K. (2014, July). Recurrent deep-stacking networks for sequence classification. In *2014 IEEE China Summit & International Conference on Signal and Information Processing (ChinaSIP)* (pp. 510-514). IEEE.
20. Ortiz, A., Munilla, J., Gorriz, J. M., & Ramirez, J. (2016). Ensembles of deep learning architectures for the early diagnosis of the Alzheimer's disease. *International journal of neural systems*, 26(07), 1650025.

21. Tasci, E., Uluturk, C., & Ugur, A. (2021). A voting-based ensemble deep learning method focusing on image augmentation and preprocessing variations for tuberculosis detection. *Neural Computing and Applications*, 33(22), 15541-15555.
22. Xu, S., Liang, H., & Baldwin, T. (2016, June). Unimelb at semeval-2016 tasks 4a and 4b: An ensemble of neural networks and a word2vec based model for sentiment classification. In *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)* (pp. 183-189).
23. Mohammadi, A., Shaverizade, A., 2021. Ensemble deep learning for aspect-based sentiment analysis. *Int. J. Nonlinear Anal. Appl.* 12, 29–38.
24. Mohammed, A., & Kora, R. (2022). An effective ensemble deep learning framework for text classification. *Journal of King Saud University-Computer and Information Sciences*, 34(10), 8825-8837.
25. Liu, P., Han, S., Meng, Z., & Tong, Y. (2014). Facial expression recognition via a boosted deep belief network. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1805-1812).
26. Liu, P., Han, S., Meng, Z., & Tong, Y. (2014). Facial expression recognition via a boosted deep belief network. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1805-1812).
27. Livieris, I. E., Iliadis, L., & Pintelas, P. (2021). On ensemble techniques of weight-constrained neural networks. *Evolving Systems*, 12(1), 155-167.
28. Li, J., Chang, H., & Yang, J. (2015, March). Sparse deep stacking network for image classification. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 29, No. 1).
29. Haralabopoulos, G., Anagnostopoulos, I., & McAuley, D. (2020). Ensemble deep learning for multilabel binary classification of user-generated content. *Algorithms*, 13(4), 83.
30. Guo, P., Xue, Z., Mtema, Z., Yeates, K., Ginsburg, O., Demarco, M., ... & Antani, S. (2020). Ensemble deep learning for cervix image selection toward improving reliability in automated cervical precancer screening. *Diagnostics*, 10(7), 451.

31. Deng, L., & Platt, J. C. (2014). Ensemble deep learning for speech recognition. In Proc. Interspeech 2014 (pp. 1915-1919).
32. Deng, L., Tür, G., He, X., & Hakkani-Tür, D. (2012, December). Use of kernel deep convex networks and end-to-end learning for spoken language understanding. In SLT (pp. 210-215).
33. Codella, N. C., Nguyen, Q. B., Pankanti, S., Gutman, D. A., Helba, B., Halpern, A. C., & Smith, J. R. (2017). Deep learning ensembles for melanoma recognition in dermoscopy images. *IBM Journal of Research and Development*, 61(4/5), 5-1.
34. Alharbi, A., Kalkatawi, M., & Taileb, M. (2021). Arabic sentiment analysis using deep learning and ensemble methods. *Arabian Journal for Science and Engineering*, 46(9), 8913-8923.
35. Голінько В. І. Охорона праці в галузі інформаційних технологій: навч. посіб. / В. І. Голінько, М. Ю. Іконніков, Я. Я. Лебедєв; М-во освіти і науки України, Держ. вищий навч. закл. "Нац. гірн. ун-т". - Дніпропетровськ: НГУ, 2015. - 246 с.
36. Гандзюк М.П. Основи охорони праці: Підручник. 4-е вид./Гандзюк М.П., Желібо Є.П., Халімовський М.О. - Київ: Каревела, 2008. – 384с.
37. Техноекологія та цивільна безпека. Частина «Цивільна безпека»: Навчальний посібник; укл.: Стручок В. С. Тернопіль: ФОП Паляниця В.А., 2022. 150 с.
38. Безпека в надзвичайних ситуаціях. Методичний посібник для здобувачів освітнього ступеня «магістр» всіх спеціальностей денної та заочної (дистанційної) форм навчання / укл.: Стручок В. С. Тернопіль: ФОП Паляниця В. А., 2022. 156 с.
39. Умови праці працівників, які використовують у роботі персональні комп'ютери. Zolochiv.Net. URL: <https://zolochiv.net/umovy-pratsi-pratsivnykiv-iaki-vykorystovuiut-u-roboti-personal-ni-komp-iutery/> (дата звернення: 25.10.2024).

ДОДАТКИ

Тези доповіді

Міністерство освіти і науки України
Тернопільський національний технічний університет
імені Івана Пулюя
Маріборський університет (Словенія)
Технічний університет в Кошице (Словаччина)
Каунаський технологічний університет (Литва)
Львівський національний університет
імені Івана Франка
Гірничо-металургійна академія ім. Станіслава Сташиця (Польща)
Луцький національний технічний університет
Чернівецький національний університет
імені Юрія Федьковича
Вроцлавський економічний університет (Польща)
Університет технологій та економіки
імені Хелени Ходковської (Польща)
Донбаська державна машинобудівна академія



*Студентське наукове
товариство*



ІХ МІЖНАРОДНА

студентська науково - технічна конференція

**"ПРИРОДНИЧІ ТА ГУМАНІТАРНІ
НАУКИ. АКТУАЛЬНІ ПИТАННЯ"**

24-25 квітня 2026 р.

(збірник тез конференції)

Тернопіль 2026

Шмирко Р. ЦИФРОВА МОДУЛЯЦІЯ У СУЧАСНИХ СИСТЕМАХ РАДІОЗВ'ЯЗКУ	144
Шупа Д. БІОМЕДИЧНА ІНЖЕНЕРІЯ: СУТНІСТЬ, ЗАСТОСУВАННЯ ТА ПЕРСПЕКТИВИ РОЗВИТКУ	146
Яріш Б. КВАНТОВІ КОМП'ЮТЕРИ	148
Яцків О. НАСКІЛЬКИ ВАЖЛИВА ПРОФЕСІЯ РАДІОТЕХНІКА В НАШ ЧАС	150
Андрухов І. РОЗРОБКА ІНТЕЛЕКТУАЛЬНОГО ГОЛОСОВОГО АСИСТЕНТА НА БАЗІ ASTERISK ТА GEMINI LIVE API ДЛЯ ПРИЙМАЛЬНОЇ КОМІСІЇ ЗВО	151
Байдецька В. ІМІТАЦІЯ СВІДОМОСТІ У ВЕЛИКИХ МОВНИХ МОДЕЛЯХ(LLM) ТА ПРИЧИНИ ЇЇ СПРИЙНЯТТЯ ЛЮДИНОЮ	152
Бармак Р., Дегодюк І. МЕТОДОЛОГІЯ СПЕЦИФІКАЦІЙНО-ОРІЄНТОВАНОЇ РОЗРОБКИ ПРОГРАМНОГО ЗАБЕЗПЕЧЕННЯ ЗА ДОПОМОГОЮ ІНСТРУМЕНТІВ ШТУЧНОГО ІНТЕЛЕКТУ	154
Берестень М. РОЗРОБКА ПРОГРАМНОГО ЗАБЕЗПЕЧЕННЯ СЕРВІСУ УПРАВЛІННЯ ФІНАНСАМИ ПРИВАТНИХ ОСІБ З ВИКОРИСТАННЯМ ТЕХНОЛОГІЙ FLASK, REACT ТА БАНКІВСЬКИХ API	156
Бица Р. РОЗРОБКА WEB-ЗАСТОСУНКУ УПРАВЛІННЯ АВТОСАЛОНОМ	157
Білінський М. РОЗРОБКА ПРОГРАМНОГО ЗАБЕЗПЕЧЕННЯ ДЛЯ МОНІТОРИНГУ ЕНЕРГОСПОЖИВАННЯ КОМЕРЦІЙНИХ БУДІВЕЛЬ НА БАЗІ ТРИРІВНЕВОЇ АРХІТЕКТУРИ З	158
Боб О., Мага С., Лотоцький Д., Боднарчук І. ДО ПИТАННЯ ЯКОСТІ ДАНИХ В НАУКОВИХ ДОСЛІДЖЕННЯХ ТНТУ	159
Боднар Д. СИНЕРГІЯ МЕТОДІВ ОБРОБКИ ПРИРОДНОЇ МОВИ ТА ПОВЕДІНКОВОЇ АНАЛІТИКИ В ІНТЕЛЕКТУАЛЬНИХ СИСТЕМАХ МОНІТОРИНГУ	161

УДК 004.6:005.6

Боб О., Мага С., Лотоцький Д., Боднарчук І.

Тернопільський національний технічний університет імені Івана Пулюя

ДО ПИТАННЯ ЯКОСТІ ДАНИХ В НАУКОВИХ ДОСЛІДЖЕННЯХ ТНТУ

O.O. Bob, S.Yu. Maha, D.V. Lototskyi, I.O. Bodnarchuk

Ternopil Ivan Puluj National Technical University

ON THE ISSUE OF DATA QUALITY IN SCIENTIFIC RESEARCH AT TNTU

Задачею поточного дослідження було систематизувати в загальних рисах наукові публікації з афіліацією Тернопільського національного технічного університету імені Івана Пулюя на тему визначення, забезпечення та покращення якості даних.

Найбільша увага прикута саме до того, як забезпечити якість даних для «мозку» сучасних технологій – систем машинного навчання та штучного інтелекту. У ранніх, але важливих роботах [7] та [10] автори піднімають важливу проблему управління якістю та ефективності ML-систем, без чого загальна якість інформаційних систем суттєво знижується. Дослідження [7] крім теоретичних викладок пропонує також власний метод оцінювання якості даних на основі міжнародного стандарту ISO/IEC 25012, та описує власне програмне рішення для задач контролю якості даних. Робота [8] детально і покроково описує весь процес забезпечення якості для машинного навчання: від профілювання та очищення до стандартизації й усунення дублікатів.

Багато уваги приділяється також процесам підготовки даних. У публікації [6] підготовку та попередню обробку даних подано, як один з найважливіших кроків для підвищення їхньої якості перед подальшою обробкою прогностичними моделями. І це логічно, оскільки якість прогнозу буде високою, коли базові дані матимуть належну якість. У дослідженні [3] автори продовжують цю тему, аналізуючи, як саме параметри моделей впливають на якість класифікації, та розглядаючи методи очищення, нормалізації та масштабування даних.

Окремий частина знайдених робіт присвячена структурній надійності даних та їхній перевірці. На сьогодні ці характеристики даних мають критичну важливість. Наприклад, робота [5] містить огляд можливостей забезпечення цілісності даних в технологіях блокчейн у розподілених системах зберігання. Дуже практичний підхід продемонстровано в [1], де йдеться про використання Django ORM для управління базою даних, що буквально гарантує надійну цілісність даних при автоматизованому зборі наукових публікацій. Також публікація [4] описує реальну проблему: при розробці електронного кабінету абітурієнта ТНТУ виявилось, що однією з основних і найбільших проблем була саме складність отримання та верифікації персональних даних.

Окремої уваги заслуговують менш очевидні, проте не менш значущі напрями досліджень. Зокрема, у роботі [2] розглядається проблема впорядкування метаданих наукових документів: обговорюються підходи до підвищення їхньої повноти та точності шляхом автоматизованого збагачення з відкритих джерел. Дослідження [9], попри зосередженість на тестуванні програмного забезпечення, також торкається аналізу отриманих даних і систем відстеження дефектів – аспектів, що є невід'ємною складовою загального циклу забезпечення якості даних.

Таким чином, можна констатувати, що ТНТУ не лише декларує увагу до проблематики якості даних, а й володіє ґрунтовною дослідницькою базою у цій сфері. Вона охоплює як теоретико-методологічний рівень – зокрема, застосування моделі ISO/IEC 25012 – так і широкий спектр прикладних напрямів: від методів машинного навчання та технологій блокчейну до верифікації даних вступників.

Література

1. Юрчишин Д. І. Розробка інформаційної системи надання наукових сервісів з використанням бібліотеки Scrapy та ORM для взаємодії з базою даних : робота на здобуття кваліфікаційного ступеня бакалавра : спец. 121 - інженерія програмного забезпечення / наук. кер. М. Р. Петрик. Тернопіль : Тернопільський національний технічний університет імені Івана Пулюя, 2025. 59 с.
2. Сучков С. С. Система автоматичного формування блоку метаданих наукових метаданих наукових документів з використанням відкритих баз даних : робота на здобуття кваліфікаційного ступеня магістра : спец. 121 - інженерія програмного забезпечення / наук. кер. І. В. Бойко. Тернопіль : Тернопільський національний технічний університет імені Івана Пулюя, 2024. 77 с.
3. Микитишин А. Г. Застосування методів машинного навчання для класифікації даних в комп'ютеризованих системах керування / Андрій Григорович Микитишин, І. С. Дідич, Р. І. Яцишин // Тези XIII МНПК „Актуальні задачі сучасних технологій“, 11-12 грудня 2024 року. — Т. : ФОП Паляниця В. А., 2024. — С. 14–16. — (Нові матеріали, міцність і довговічність елементів конструкцій).
4. Карнаухов, О. К. (2024). Дослідження розробки електронного кабінету абітурієнта ТНТУ ім. І. Пулюя. Тези доповідей V міжнародної науково-практичної конференції учених та студентів "Цифрова економіка як фактор інновацій та сталого розвитку суспільства", 46-47.
5. Гладій В. В. Технології створення розподілених комп'ютерних систем зберігання даних на основі блокчейн : кваліфікаційна робота на здобуття освітнього ступеня магістр за спеціальністю "123 — комп'ютерна інженерія" / В. В. Гладій. — Тернопіль: ТНТУ, 2023. — 82 с.
6. Кучеренко О. А. Особливості передобробки даних для методів прогнозування / О. А. Кучеренко, О. О. Кучеренко // ІМСТТ, 13-14 грудня 2023 року. — Т. : ТНТУ, 2023. — С. 72. — (Інформаційні системи та технології, кібербезпека).
7. Журихін Ю. О. Методи забезпечення якості даних при проектуванні систем машинного навчання: автореферат дипломної роботи магістра за спеціальністю 123 «Комп'ютерна інженерія»/ Ю. О. Журихін – Тернопільський національний технічний університет імені Івана Пулюя – Тернопіль, ТНТУ, 2018. – 8 с.
8. Яцишин В. Процеси забезпечення якості даних при проектуванні систем машинного навчання / В. Яцишин, Ю. Журихін // Матеріали V науково-технічної конференції „Інформаційні моделі, системи та технології“, 1-2 лютого 2018 року. — Т. : ТНТУ, 2018. — С. 68. — (Секція 3. Комп'ютерні системи та мережі).
9. Чорновус, Р. М. Визначення якості тестування програмного забезпечення та аналіз отриманих даних. Матеріали конференції. Тернопіль: ТНТУ, 2017. URL: <http://elartu.tntu.edu.ua/handle/123456789/18917>.
10. Яцишин В. В. Оцінювання якості даних для систем машинного навчання / В. В. Яцишин, Ю. О. Журихін // Збірник тез доповідей VI Міжнародної науково-технічної конференції молодих учених та студентів „Актуальні задачі сучасних технологій“, 16-17 листопада 2017 року. — Т. : ТНТУ, 2017. — Том 2. — С. 196. — (Комп'ютерно-інформаційні технології та системи зв'язку).