НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ
імені ІГОРЯ СІКОРСЬКОГО»
ФАКУЛЬТЕТ БІОМЕДИЧНОЇ ІНЖЕНЕРІЇ
(повна назва інституту/факультету)

кафедра БІОМЕДИЧНОЇ КІБЕРНЕТИКИ
(повна назва кафедри)

«На правах рукопису»
УДК _____

«До захисту допущено»
В.о.завідувача кафедри БМК

_____ Світлана АЛХІМОВА
(підпис) (ініціали, ПРІЗВИЩЕ

"___" __*грудня*__ 2025р.

Магістерська дисертація

на здобуття ступеня магістра
за освітньо-професійною програмою *«Комп'ютерні технології в біології та медицині»*
зі спеціальності *122 «Комп'ютерні науки»*

на тему: **Система аналізу мутагенності хімічних сполук на основі методів машинного навчання**

Виконала: студентка **II** курсу, групи _ЗК-41мп__

ЄСИПЕНКО РУСЛАНА ВЯЧЕСЛАВІВНА _____
(прізвище, ім'я, по батькові) (підпис)

Науковий керівник: *ст. викл. каф. біомедичної кібернетики (БМК), Кисляк Сергій Володимирович* _____
(посада, науковий ступінь, вчене звання, прізвище, ім'я, по ініціали) (підпис)

Консультант з розділів магістерської дисертації:

_____ _____
(назва розділу) (посада, вчене звання, науковий ступінь, прізвище, ініціали) (підпис)

Рецензент: *доцент кафедри БМІ, к.б.н., доцент. Калашнікова Лариса Євгеніївна* _____
(посада, науковий ступінь, вчене звання, науковий ступінь, прізвище та ініціали) (підпис)

Засвідчую, що у цій магістерській дисертації немає запозичень з праць інших авторів без відповідних посилань.

Студентка _____
(підпис)

Київ – 2025 року

Національний технічний університет України

«Київський політехнічний інститут імені Ігоря Сікорського»

Факультет біомедичної інженерії

Кафедра біомедичної кібернетики

Рівень вищої освіти – другий (магістерський)

Спеціальність – 122 «Комп'ютерні науки»

Освітньо-професійна програма «Комп'ютерні технології в біології та медицині»

ЗАТВЕРДЖУЮ

В.о.завідувача кафедри БМК

_____ Світлана АЛХІМОВА

«_____» *листопада* 2025 р.

ЗАВДАННЯ
на магістерську дисертацію студентці

ЄСИПЕНКО РУСЛАНІ ВЯЧЕСЛАВІВНІ

(прізвище, ім'я, по батькові)

1. Тема дисертації **Система аналізу мутагенності хімічних сполук на основі методів машинного навчання**

науковий керівник дисертації

Кисляк Сергій Володимирович , старший викладач кафедри БМК

(прізвище, ім'я, по батькові, науковий ступінь, вчене звання)

затверджені наказом по університету від «03» *листопада* 2025 р. №***4751-с***

2. Термін подання студентом дисертації ***20-22 листопада 2025 року***

3. Об'єкт дослідження *Хімічні сполуки, представлені у форматі SMILES із відомими результатами тесту* Еймса

4. Вихідні дані *Система оцінки мутагенності хімічних сполук*

5. Перелік завдань, які потрібно розробити *Проаналізувати існуючі підходи до оцінки мутагенності та визначити їх обмеження; підготувати дані, що включають молекулярні дескриптори та класові характеристики хімічних сполук; розробити та навчити моделі машинного навчання для прогнозування мутагенності; провести оцінювання точності та ефективності системи на тестових даних, шнтегрувати розроблені моделі у єдину автоматизовану систем*

6. Орієнтовний перелік графічного (ілюстративного) матеріалу *16 рисунків, 20 таблиць, презентація з захисту МД на 17 слайдах*

7. Орієнтовний перелік публікацій *1 теза доповіді на міжнародній конференції, 1 стаття у журналі категорії А, 1 стаття у фаховому журналі категорії Б*

8. Консультанти розділів дисертації[1*] *не передбачено*

9. Дата видачі завдання ***28 серпня 2025р.***

Календарний план

№ з/п	Назва етапів виконання магістерської дисертації	Термін виконання етапів МД	Примітка
1	Отримати завдання на МД	До 28.08.2025	***виконано***
2	Завершення виконання практичної частина МД	До 25.10.2025	***виконано***
3	Завершення оформлення розділів МД (Вступ, Основні розділи та висновки до них, Загальні висновки, Список використаних джерел)	До 24.11.2025	***виконано***
4	Апробація та публікація результатів дослідження МД (отримати підтвердження про прийняття до друку)	До 01.12.2025	***виконано***
5	Перевірка МД науковим керівником	24.11.2025	***виконано***
6	Подання в електронному вигляді МД на перевірку нормоконтролера	До 01.12.2025	
7	Подання в електронному вигляді МД на перевірку подібності Strike Plagiarism com. Отримати позитивний звіт подібності.	До 05.12.2025	
8	Підготовка Експертної оцінки до звіту подібності.	До 10.12.2025	
9	Отримати відгук наукового керівника	10.12.2025	
10	Надати на кафедру пакет документів в паперовому та електронному вигляді (МД, відгук керівника, звіт подібності, Експертної оцінки до звіту подібності)	10-11.12.2025	
11	Отримати допуск до захисту МД в ЕК та направлення до рецензента (засідання кафедри)	Засідання кафедри	
12	Подання МД рецензенту. Отримати на надання рецензію до ЕК / на кафедру	До 19.12.2025	
13	Подання супровідного пакету документів по МД до захисту в ЕК[2]	До 19.12.2025	
14	Захист МД в ЕК	22 – 26.12.2025	

Студент _____ **Руслана ЄСИПЕНКО**
(підпис) (ім'я, ПРІЗВИЩЕ)

Науковий керівник _____ **Сергій КИСЛЯК**
(підпис) (ім'я, ПРІЗВИЩЕ)

Нормоконтролер _____ **Галина КОРНІЄНКО**
(підпис) (ім'я, ПРІЗВИЩЕ)

[1*] Якщо визначені консультанти. Консультантом не може бути зазначено наукового керівника магістерської дисертації.

[2] не пізніше ніж за один тиждень до затвердженої дати захисту МД в ЕК

РЕФЕРАТ

Магістерська дисертація за темою «Система аналізу мутагенності хімічних сполук на основі методів машинного навчання» виконана студенткою кафедри біомедичної кібернетики ФБМІ *Єсипенко Русланою Вячеславівною зі спеціальності 122 «Комп'ютерні науки» за освітньо-професійною програмою «Комп'ютерні технології в біології та медицині» та* складається зі: вступу; 5 розділів (*Поняття мутагенності та аналіз існуючих методів її оцінки, Вибір та обґрунтування засобів реалізації моделей машинного навчання, Розробка моделей машинного навчання для задачі класифікації, Програмна реалізація та методика роботи системи, Узагальнення результатів роботи*), розділу зі стартап проєкту, висновків до кожного з цих розділів; загальних висновків; списку використаних джерел, який налічує 46 джерел. Загальний обсяг роботи 98 сторінок.

Актуальність теми. Оцінка мутагенності хімічних сполук є важливим етапом у фармацевтичних, біомедичних та токсикологічних дослідженнях, оскільки саме мутагенні речовини можуть становити загрозу для здоров'я людини та довкілля. Традиційні лабораторні методи потребують значних ресурсів, часу та не дозволяють швидко аналізувати великі масиви даних. Використання методів машинного навчання відкриває можливість автоматизованого прогнозування мутагенності на основі структурних характеристик молекул, що значно підвищує ефективність первинного скринінгу та зменшує потребу у дорогих лабораторних дослідженнях. Саме тому тема магістерської дисертації є актуальною та важливою для сучасної науки й промисловості.

Мета і задачі дослідження.

Метою роботи є підвищення ефективності оцінки мутагенності хімічних сполук за допомогою методів машинного навчання та можливість отримання високих показників точності класифікації мутагенних та немутагенних сполук на основі молекулярних дескрипторів.

Сформульовано наступні **задачі**:

1. Проаналізувати існуючі підходи до оцінки мутагенності та визначити їх обмеження.
2. Підготувати дані, що включають молекулярні дескриптори та класові характеристики хімічних сполук.
3. Розробити та навчити моделі машинного навчання для прогнозування мутагенності.
4. Провести оцінювання точності та ефективності системи на тестових даних.
5. Інтегрувати розроблені моделі у єдину автоматизовану систему.

***Об'єкт дослідження*.** Хімічні сполуки, представлені у форматі SMILES із відомими результатами тесту Еймса.

Предмет дослідження. Прогнозування мутагенності хімічних сполук на основі молекулярних дескрипторів та методів машинного навчання.

Методи дослідження. Машинне навчання, випадковий ліс, градієнтний бустинг, багатошаровий перцептрон (нейронні мережі), рекурсивне виключення ознак з крос-валідацією (RFECV).

Апробація результатів дослідження Результати роботи проходили апробацію на 1 науковій конференції:

1. Єсипенко Р. В., Кисляк С. В., *In silico* моделі оцінки генотоксичності факторів навколишнього середовища. *Матеріали IX Міжнародної науково-практичної конференції «CURRENT CHALLENGES OF SCIENCE AND EDUCATION»*. 2024 р. Берлін, Німеччина – С. 50-53 (тези). ISBN 978-3-954753-05-5. URL: https://sci-conf.com.ua/ix-mizhnarodna-naukovo-praktichna-konferentsiya-current-challenges-of-science-and-education-6-8-05-2024-berlin-nimechchina-arhiv/ (дата звернення 01.11.2025)

Публікації. За результатами виконаної роботи були опублікована 1 наукова стаття в журналі категорії А та 1 наукова стаття в фаховому журналі категорії Б:

Kislyak, S., Dugan, O., Yesypenko, R., Starosyla, D. and Yalovenko, O. 2025. *In silico* the Ames Mutagenicity Predictive Model of Environment. Innovative Biosystems and Bioengineering. 9, 2 (May 2025), 42–52. DOI: https://doi.org/10.20535/ibb.2025.9.2.316239 (категорія А) (дата звернення 01.11.2025)

Кисляк С., Дуган О.., Єсипенко Р., Яловенко О. 2025. *In silico* моделі прогнозування мутагенності Еймса основних структурних класів ксенобіотиків на основі методу випадкового лісу. *Біомедична інженерія і технологія.* 2025. Т. 4, №.19. с. 48-62. URL: https://doi.org/10.20535/.2025.19.340327 (категорія Б) (дата звернення 20.11.2025).

Ключові слова. Мутагенність, хімічні сполуки, дескриптори, прогнозування, класифікація, машинне навчання, випадковий ліс, градієнтний бустинг, нейронна мережа, система.

Бібліографічний опис МД

Єсипенко, Р. В. Система аналізу мутагенності хімічних сполук на основі методів машинного навчання : магістерська дис. : 122 Комп’ютері науки / Єсипенко Руслана Вячеславівна. – Київ, 2025. – 120 с

ABSTRACT

Master's thesis on the topic «System for analyzing the mutagenicity of chemical compounds based on machine learning methods» was completed by Ruslana Yesypenko, a student of the Department of Biomedical Cybernetics of the Federal State Institute of Medical Sciences, specialty 122 "Computer Science" under the educational and professional program "Computer Technologies in Biology and Medicine" and consists of: an introduction; 5 sections (The concept of mutagenicity and analysis of existing methods for its assessment, Selection and justification of means for implementing machine learning models, Development of machine learning models for the classification task, Software implementation and methodology of the system, Generalization of work results), a section on the startup project, conclusions for each of these sections; general conclusions; a list of sources used, which includes 46 sources. The total volume of the work is 98 pages.

Relevance of the topic. Assessment of mutagenicity of chemical compounds is an important stage in pharmaceutical, biomedical and toxicological research, since it is mutagenic substances that can pose a threat to human health and the environment. Traditional laboratory methods require significant resources, time and do not allow for rapid analysis of large data sets. The use of machine learning methods opens up the possibility of automated prediction of mutagenicity based on the structural characteristics of molecules, which significantly increases the efficiency of primary screening and reduces the need for expensive laboratory studies. That is why the topic of the master's thesis is relevant and important for modern science and industry.

Purpose and objectives of the study.

The purpose of the work is to increase the efficiency of assessment of mutagenicity of chemical compounds using machine learning methods and the possibility of obtaining high accuracy rates of classification of mutagenic and non-mutagenic compounds based on molecular descriptors.

The following **tasks** are formulated:

1. To analyze existing approaches to assessment of mutagenicity and determine their limitations.

2. Prepare data including molecular descriptors and class characteristics of chemical compounds.

3. Develop and train machine learning models for mutagenicity prediction.

4. Evaluate the accuracy and efficiency of the system on test data.

5. Integrate the developed models into a single automated system.

Research object. Chemical compounds presented in SMILES format with known Ames test results.

Research subject. Prediction of mutagenicity of chemical compounds based on molecular descriptors and machine learning methods.

Research methods. Machine learning, random forest, gradient boosting, multilayer perceptron (neural networks), recursive feature elimination with cross-validation (RFECV).

Approbation of research results The results of the work were approbated at 1 scientific conference:

1. Yesipenko R. V., Kislyak S. V., *In silico* models for assessing the genotoxicity of environmental factors. *Proceedings of the IX International Scientific and Practical Conference "CURRENT CHALLENGES OF SCIENCE AND EDUCATION"*. 2024. Berlin, Germany – pp. 50-53 (abstracts). ISBN 978-3-954753-05-5. URL: https://sci-conf.com.ua/ix-mizhnarodna-naukovo-praktichna-konferentsiya-current-challenges-of-science-and-education-6-8-05-2024-berlin-nimechchina-arhiv/ (access date 01.11.2025)

Publications. Based on the results of the work, 1 scientific article was published in a category A journal and 1 scientific article in a category B journal:

Kislyak, S., Dugan, O., Yesypenko, R., Starosyla, D. and Yalovenko, O. 2025. *In silico* the Ames Mutagenicity Predictive Model of Environment. Innovative Biosystems and Bioengineering. 9, 2 (May 2025), 42–52. DOI:

https://doi.org/10.20535/ibb.2025.9.2.316239 (category A) (access date 01.11.2025)

Kislyak S., Dugan O.., Yesypenko R., Yalovenko O. 2025. *In silico* models for predicting the Ames mutagenicity of the main structural classes of xenobiotics based on the random forest method. Biomedical Engineering and Technology. 2025. Vol. 4, No. 19. pp. 48-62. URL: https://doi.org/10.20535/.2025.19.340327 (category B) (access date 11/20/2025).

Keywords. Mutagenicity, chemical compounds, descriptors, prediction, classification, machine learning, random forest, gradient boosting, neural network, system.

ЗМІСТ

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ, СИМВОЛІВ, ОДИНИЦЬ, СКОРОЧЕНЬ І ТЕРМІНІВ

DL – Deep Learning, глибоке навчання.

EL – Ensemble Learning, ансамблеве навчання.

GB – Gradient Boosting, градієнтний бустинг.

GPU – Graphics Processing Unit, графічний процесор.

KNN – k-nearest neighbor method, метод k-найближчих сусідів.

ML – Machine Learning, машинне навчання.

MLP – Multilayer Perceptron, багатошаровий персептрон.

NN – neural networks, нейроні мережі.

QSAR – це кількісний зв'язок "структура-активність" (Quantitative Structure-Activity Relationship), метод, який використовується для моделювання та прогнозування властивостей молекул на основі їхньої хімічної структури.

RF – Random forest, випадковий ліс.

RFECV – Recursive Feature Elimination with Cross-Validation, рекурсивне видалення ознак (RFE) з крос-валідацією (CV).

SVM – Support Vector Machines, метод опорних векторів.

ДНК – дезоксирибонуклеїнова кислота, це макромолекула, яка зберігає та передає генетичну інформацію в усіх живих організмах

ВСТУП

Оцінка мутагенності хімічних сполук є важливою складовою токсикологічної безпеки у фармацевтичній, хімічній, біотехнологічній та екологічній галузях. Традиційні *in vitro* та *in vivo* методи оцінки мутагенних ефектів забезпечують високу достовірність, але потребують значних фінансових і часових ресурсів та не дозволяють оперативно аналізувати великі масиви нових речовин. З огляду на зростання кількості хімічних сполук та тенденцію до зменшення експериментів на тваринах, особливої актуальності набувають *in silico* підходи до прогнозування мутагенного потенціалу.

Методи машинного навчання відкривають можливість побудови точних моделей прогнозування на основі молекулярних ознак. Проте більшість існуючих інструментів не враховує структурну класифікацію хімічних сполук, що знижує якість прогнозів на гетерогенних даних. Тому актуальною є розробка системи, яка поєднує машинне навчання, розрахунок молекулярних дескрипторів та адаптивні моделі для різних класів сполук забезпечуючи високу точність оцінки мутагенності.

У роботі застосовано алгоритми машинного навчання (RF, GB, NN), методи обчислення дескрипторів (RDKit, Mordred, PaDEL), автоматичний відбір ознак (RFECV) та програмну реалізацію з інтуїтивним інтерфейсом. Це дозволило створити систему, яка забезпечує автоматизовану, швидку та точну оцінку мутагенності хімічних сполук і може бути використана як ефективний інструмент у науковій та прикладній діяльності.

Мета і завдання роботи

Метою роботи є підвищення ефективності оцінки мутагенності хімічних сполук за допомогою методів машинного навчання.

Її досягнення передбачає вирішення наступних *завдань*:

1. Аналіз існуючіх підходів до оцінки мутагенності та визначення їх перваг та недоліків.
2. Підготовка даних для проведення *in-silico* моделювання мутагенності Еймса.
3. Розробка та навчання моделей машинного навчання для прогнозування мутагенності Еймса.
4. Оцінювання ефективності системи на тестових даних.
5. Інтегрування розроблених моделей у єдину автоматизовану систему.

Використані методи. Машинне навчання, випадковий ліс, градієнтний бустинг, багатошаровий перцептрон (нейронні мережі), рекурсивне виключення ознак з крос-валідацією (RFECV).

Отримані результати. Система «AmesScan» для оцінки мутагенності хімічних сполук з точністю понад 86%.

Апробація результатів роботи. За результатами виконаної роботи було опубліковано 1 матеріали на міжнародній конференції:

Єсипенко Р. В., Кисляк С. В., *In silico* моделі оцінки генотоксичності факторів навколишнього середовища. *Матеріали IX Міжнародної науково-практичної конференції «CURRENT CHALLENGES OF SCIENCE AND EDUCATION»*. 2024 р. Берлін, Німеччина – С. 50-53 (тези). ISBN 978-3-954753-05-5. URL: https://sci-conf.com.ua/ix-mizhnarodna-naukovo-praktichna-konferentsiya-current-challenges-of-science-and-education-6-8-05-2024-berlin-nimechchina-arhiv/ (дата звернення 01.11.2025) [1].

Публікації. За результатами виконаної роботи були опублікована 1 наукова стаття в журналі категорії А та 1 наукова стаття в фаховому журналі категорії Б:

1. Kislyak, S., Dugan, O., Yesypenko, R., Starosyla, D. and Yalovenko, O. 2025. *In silico* the Ames Mutagenicity *Predictive Model of Environment. Innovative Biosystems and Bioengineering. 9, 2 (May 2025), 42–52*. DOI: https://doi.org/10.20535/ibb.2025.9.2.316239 (категорія А) (дата

звернення 01.11.2025) [2].

2. Кисляк С., Дуган О.., Єсипенко Р., Яловенко О. 2025. *In silico* моделі прогнозування мутагенності Еймса основних структурних класів ксенобіотиків на основі методу випадкового лісу. Біомедична інженерія і технологія. 2025. Т. 4, №.19. с. 48-62. URL: https://doi.org/10.20535/.2025.19.340327 (категорія Б) (дата звернення 20.11.2025) [3].

Структура роботи

Магістерська дисертація за темою «Система аналізу мутагенності хімічних сполук на основі методів машинного навчання» виконана студенткою *Єсипенко Русланою Вячеславівною зі спеціальності 122 «Комп'ютерні науки» за освітньо-професійною програмою «Комп'ютерні технології в біології та медицині»,* побудована за класичним типом та викладена на 120 сторінках машинописного тексту. Вона складається з: вступу; 5 розділів (*Поняття мутагенності та аналіз існуючих методів її оцінки, Вибір та обґрунтування засобів реалізації моделей машинного навчання, Розробка моделей машинного навчання для задачі класифікації, Програмна реалізація та методика роботи системи, Узагальнення результатів роботи*), висновків до кожного з цих розділів; загальних висновків; списку використаних джерел, який налічує 46 джерел. В роботі представлено 16 рисунків і 20 таблиць.

РОЗДІЛ 1
ПОНЯТТЯ МУТАГЕННОСТІ ТА АНАЛІЗ ІСНУЮЧИХ МЕТОДІВ ЇЇ ОЦІНКИ

У цьому розділі розглянуто предметну область генетичної токсикології та сучасні методи оцінки мутагенності хімічних сполук. Проаналізовано перехід від традиційних експериментальних підходів *in vitro* та *in vivo* до високопродуктивних і обчислювальних методів, що забезпечують швидшу та більш достовірну оцінку мутагенного потенціалу речовин із мінімізацією експериментів на тваринах. Підкреслено, що сучасна генетична токсикологія базується на інтеграції біологічних даних з алгоритмами машинного навчання, що дозволяє глибше зрозуміти механізми дії генотоксичних агентів.

Особливу увагу приділено застосуванню моделей QSAR, методів машинного та глибокого навчання, які демонструють високу ефективність у прогнозуванні мутагенності Еймса на основі набору ознак, що представлений молекулярними дескрипторами. Використання таких алгоритмів, як випадковий ліс, метод опорних векторів і нейронні мережі, відкриває можливості для автоматизації процесу оцінки ризиків і створення інтелектуальних систем прогнозування.

1.1. Традиційні методи оцінки мутагенного потенціалу

Генетична токсикологія – це наука, що вивчає взаємодію між хімічними агентами та генетичним матеріалом організмів, зосереджуючись на мутаціях, хромосомних аберраціях, пошкодженні ДНК та пов'язаних з ними механізмах. Вона спрямована на оцінку генетичної шкоди, спричиненої хімічними речовинами, які можуть викликати гостру або хронічну токсичність. Хоча фундаментальні дослідження в середині 20-го століття

встановили ключові зв'язки між мутагенністю та канцерогенністю за допомогою таких аналізів, як тест Еймса, сучасна генетична токсикологія значно розвинулася у відповідь на зростаючі потреби в швидкому, точному та етичному тестуванні [4, 5]. До 1990-х років стало очевидним, що одних лише аналізів *in vitro* недостатньо для повного прогнозування канцерогенності для людини, що спонукало до переходу до інтегрованих стратегій, що поєднують підходи *in vitro*, *in vivo* та *in silico*. Ця інтеграція заклала основу для впровадження високопродуктивного скринінгу (HTS), обчислювальної токсикології та системної біології [5].

Генетична токсикологія має вирішальне значення для оцінки потенційних ризиків хімічних речовин та ліків для здоров'я людини та навколишнього середовища. Поява високопродуктивних технологій трансформувала цю галузь, забезпечуючи більш ефективні, економічно вигідні та етично обґрунтовані методи тестування генотоксичності. Вона використовує передові методи скринінгу, включаючи автоматизовані тест-системи *in vitro* та обчислювальні моделі, для швидкої оцінки генотоксичного потенціалу тисяч сполук одночасно.

Мутагенність і генотоксичність – тісно пов’язані, але не тотожні поняття. Генотоксичність охоплює будь-яке пошкодження генетичного матеріалу, включаючи розриви ДНК, хромосомні аберації, порушення репарації та зміни у клітинних механізмах, що можуть, але не обов’язково, призводити до мутацій. Мутагенність є різновидом генотоксичності і стосується виключно стабільних, успадкованих змін у ДНК, які передаються при поділі клітин. Усі мутагени є генотоксичними, проте не всі генотоксичні агенти є мутагенами, оскільки деякі з них можуть спричиняти пошкодження, що не фіксуються у вигляді мутацій. Ця різниця є принципово важливою при оцінці небезпеки хімічних сполук, адже визначає потенційний ризик довготривалих та спадкових ефектів.

Оскільки жоден окремий тест не може охопити всі відповідні генотоксичні кінцеві точки, рекомендується поєднання методів тестування *in*

vivo та *in vitro* для забезпечення комплексної оцінки. Ці тести пропонують раннє розуміння потенційного шкідливого впливу хімічних речовин у різних сферах, включаючи фармацевтичні препарати, косметику, агрохімікати, промислові сполуки, харчові добавки, природні токсини та наноматеріали. Раннє виявлення генотоксичних ризиків за допомогою згаданих методів тестування дозволяє знизити потенційний шкідливий вплив хімічних речовин, забезпечуючи безпечне впровадження нових сполук та одночасно охороняючи здоров'я людей і стан навколишнього середовища. Генетичні токсикологічні тести проводяться відповідно до суворих регуляторних рекомендацій, встановлених такими організаціями, як OECP (Організація економічного співробітництва та розвитку), ICH (Міжнародна рада з гармонізації) та EPA (Агентство з охорони навколишнього середовища). Ці рекомендації гарантують, що *in vivo* та *in vitro* методи оцінки генотоксичності є надійними, відтворюваними та відповідають стандартам належної лабораторної практики [5, 6].

In vitro тести виконуються в контрольованих лабораторних умовах із використанням ізольованих клітин, культур клітин ссавців або навіть субклітинних компонентів. Вони є одним із перших етапів оцінки мутагенного потенціалу сполук, оскільки забезпечують швидке та відносно недороге отримання даних. Крім того, застосування цих методів відповідає сучасним етичним вимогам, оскільки дозволяє зменшити використання лабораторних тварин. Найчастіше такі методи використовуються для початкового скринінгу великої кількості речовин.

In vivo тести проводяться на живих організмах (найчастіше на мишах або щурах) і є обов'язковим етапом підтвердження результатів, отриманих у клітинних системах. Вони надають більш комплексну інформацію про мутагенну чи генотоксичну дію, адже враховують ключові біологічні процеси: абсорбцію, розподіл, метаболізм, детоксикацію та репарацію ДНК. Серед найбільш розповсюджених методів оцінки генотоксичності виділяють мікроядерний тест *in vivo* та Comet assay, які дозволяють оцінити

пошкодження ДНК у клітинах кісткового мозку чи периферичної крові гризунів.

До найвідоміших підходів належать:

- Тест Еймса (bacterial reverse mutation test), який виявляє точкові мутації;
- Тест на хромосомні аберації, що дозволяє визначити структурні пошкодження хромосом у клітинах ссавців;
- Мікроядерний тест, який фіксує утворення мікроядер як індикатор хромосомних порушень;
- Кометний тест (single-cell gel electrophoresis), що використовується для оцінки пошкоджень ДНК на рівні окремих клітин.

Вищезгадані методи також відомі як короткострокові тести (Short-Term Tests, STT) і активно застосовуються ще з другої половини XX століття. Вони включають різноманітні підходи, зокрема тест Еймса для бактерій, аналіз хромосомних аберацій у клітинах ссавців та тест на мікроядра. Ключова перевага цих методів полягає у високій чутливості до виявлення потенційних мутагенів, що робить їх особливо корисними для попереднього скринінгу великої кількості хімічних сполук. Проте, оскільки короткострокові тести використовують ізольовані клітини або субклітинні системи, вони не враховують складні взаємодії між тканинами, органами та системними метаболічними процесами живого організму. Тому результати потребують додаткової валідації за допомогою *in vivo* досліджень або більш комплексних *in vitro* моделей.

Тест Еймса, розроблений у 1975 році, є швидким, чутливим та економічно ефективним методом оцінки мутагенності речовин. Він використовує гістидин-ауксотрофні штами *Salmonella typhimurium* , яким для росту потрібен гістидин із навколишнього середовища. Мутагенні речовини можуть повертати ці штами до прототрофного стану, що дозволяє ріст на середовищах без гістидину. Тест Еймса виявляє широкий спектр

генотоксичних канцерогенів та типів мутацій, включаючи зсуви рамки зчитування та заміни гетероциклічних основ. Однак його обмежена специфічність та складність інтерпретації позитивних результатів, а також відмінності між мікроорганізмами та ссавцями в генетичній складності та системах репарації ДНК, обмежують його використання в розробці ліків [5, 7].

Кометний аналіз та мікроядерний аналіз – це два широко використовувані методи виявлення генотоксичності завдяки їхній простоті, чутливості та здатності виявляти розриви ланцюга ДНК або втрату хромосом. Поєднання вищезгаданих методів дозволяє більш детально визначити механізми генотоксичної активності хімічних сполук, оскільки кожен з них виявляє пошкодження ДНК на різних рівнях. Кометний аналіз, запроваджений у 1984 році, виявляє розриви ланцюга ДНК шляхом візуалізації структур, подібних до хвоста комети, утворених фрагментованою ДНК, що мігрує до анода під час електрофорезу. Лужний кометний аналіз, більш чутливий варіант, виявляє ряд пошкоджень ДНК, включаючи дволанцюгові розриви та лужнолабільні ділянки, що робить його особливо корисним для ідентифікації генотоксичних агентів. Його переваги включають гнучкість, чутливість, низькі вимоги до клітин, швидке виконання та низьку вартість.

Мікроядерний аналіз виявляє хромосомні пошкодження шляхом ідентифікації мікроядер у цитоплазмі еритроцитів. Ці мікроядра, відкриті понад 100 років тому, вказують на втрату або порушення хромосом. Мікроядерний аналіз може оцінювати як кластогенні (руйнування хромосом), так і анеугенні (що впливають на кількість хромосом) ефекти [8].

Аналіз хромосомних аберацій додатково оцінює структурні аномалії в хромосомах, використовуючи клітинні лінії ссавців, такі як клітини яєчників китайського хом'яка, для ідентифікації агентів, що викликають хромосомні пошкодження. Аналіз обміну сестринськими хроматидами (SCE) та мікроядерний тест з блокуванням цитокінезу (CBMN) дозволяють отримати

додаткову інформацію про хромосомні зміни. Ці методи підвищують роздільну здатність у виявленні структурних аномалій хромосом і одночасно дають змогу оцінювати вплив тестованих агентів на різних стадіях клітинного циклу.

Окрім хромосомних пошкоджень, аналізи генних мутацій сприяють оцінці генотоксичності шляхом виявлення мутацій на рівні генів. Тест Еймса ідентифікує точкові мутації в бактеріях, тоді як аналіз лімфоми миші та аналіз гіпоксантин-гуанінфосфорибозилтрансферази виявляють генні мутації в клітинах ссавців. Аналіз мутацій трансгенних гризунів оцінює мутації *in vivo* як у соматичних, так і в зародкових клітинах, а аналіз Pig-a дозволяє виявляти мутації *in vivo* в клітинах крові. Методи для виявлення пошкодженя та репарації ДНК дають уявлення про генотоксичні впливи та клітинні реакції на молекулярному рівні. До таких методів відносяться аналіз γH2AX для виявлення дволанцюгових розривів ДНК, аналіз TUNEL для фрагментації, індукованої апоптозом, та аналіз позапланового синтезу ДНК (unscheduled DNA synthesis, UDS) для оцінки репарації ДНК. [5, 7].

Сучасна генетична токсикологія дедалі більше використовує інтегровані підходи, що об’єднують результати різних тестів і молекулярних аналізів для більш точної оцінки генотоксичних ефектів. Технології секвенування нового покоління дозволяють досліджувати геном із високою точністю, відкриваючи можливості для комплексної оцінки генотоксичності. Сучасні підходи використовують інтегровані *in vitro* моделі, такі як HepaRG, що дозволяють одночасно оцінювати кометний і мікроядерний аналізи разом із транскриптомічними змінами для глибшого розуміння механізмів дії тестованих речовин.

3D-моделі тканин вважаються одним із найпрогресивніших підходів для оцінки хімічної небезпеки, оскільки розширюють традиційні 2D-культури за біологічною релевантністю. Тривимірні моделі тканин відтворюють просторову організацію клітин і міжклітинного матриксу, що забезпечує функціонально наближені взаємодії між клітинами, дозволяючи

реалістично моделювати пошкодження ДНК, особливості репарації та проникнення хімічних агентів. До найпоширеніших моделей належать сфероїди, органоїди та біодруковані 3D-тканини, які дедалі частіше застосовують у генетичній токсикології завдяки їхній здатності більш точно відтворювати фізіологічні умови порівняно з традиційними 2D-моношаровими культурами.

Епігенетичні показники відображають зміни в регуляції генів, що не пов’язані зі змінами послідовності ДНК, і включають модифікації ДНК (наприклад, метилювання), модифікації гістонів, профілі некодуючих РНК та організацію хроматину. Включення цих показників у дослідження генотоксичності суттєво розширило можливості аналізу, дозволяючи оцінювати не лише класичні пошкодження ДНК, але й порушення регуляторних механізмів на молекулярному рівні. Сучасні методи аналізу епігенетичних змін відкривають доступ до глибшого розуміння механізмів дії генотоксичних чинників, а використання цих інструментів забезпечує більш комплексне бачення клітинних відповідей і підвищує інформативність традиційних тестів, що сприяє точнішому визначенню потенційних ризиків і небезпек.

1.2. Високопродуктивні та обчислювальні методи

Поява високопродуктивних технологій призвела до зміни парадигми в генетичній токсикології, відходячи від традиційних методів, таких як тестування на тваринах, до більш економічно ефективних та етичних підходів, таких як тестування *in vitro* та *in silico*. Цей перехід створив значені обсяги «великих даних», що дозволяє проводити високопродуктивний скринінг та ефективніший аналіз. Високопродуктивні методи використовують передові інструменти та платформи для скринінгу великої кількості сполук за допомогою кометного, мікроядерного та різних інших тестів (γ-H2AX-аналіз, бісульфітне секвенування, транскриптоміка,

мікрочиповий аналіз, візуалізація тощо) для виявлення генотоксичних властивостей, мутацій та пошкоджень ДНК. Ключові високопродуктивні технології для генетичної токсикології наведені у табл. 1.1 [5, 9, 10, 11, 12].

Таблиця 1.1 – Високопродуктивні технології для генетичної токсикології

Технології	Застосування	Переваги	Недоліки
SOS Хромотест	Бактеріальний колориметричний аналіз, що вказує на пошкодження ДНК або генотоксичність та легко адаптується для високотемпературної бактеріальної інфекції (HTS)	Простота, висока чутливість, швидші результати порівняно з тестом Еймса	Обмежена застосовність до певних типів генотоксичних агентів
Секвенування наступного покоління (NGS)	Аналіз мутацій по всьому геному, виявлення пошкоджень ДНК	Висока чутливість, повне покриття	Складність виявлення малочисленних соматичних мутацій, висока вартість, складний аналіз даних
Кількісна HTS	Скринінг сполук на генотоксичність та профілювання концентрації-відповіді	Швидке тестування тисяч сполук, зменшення кількості хибнонегативних результатів та економічна ефективність	Потрібна надійна інформатика для аналізу даних; мінливість оцінок потенції в різних профілях
Омічні технології (геноміка, транскриптоміка, протеоміка)	Комплексне вивчення генетичного матеріалу та клітинних реакцій	Дозволяє цілісне розуміння механізмів токсичності; швидке секвенування геному	Складність інтерпретації даних; вимагає інтеграції кількох наборів даних
Проточна цитометрія	Аналіз клітинного циклу, виявлення пошкоджень ДНК	Швидкий кількісний аналіз	Потрібне спеціальне фарбування та інструментарій

Кінець табл. 1.1

Технології	Застосування	Переваги	Недоліки
Візуалізація	Клітинна морфологія, візуалізація пошкоджень ДНК	Візуальне підтвердження, аналіз високого вмісту	Потрібне програмне забезпечення для аналізу зображень, можливість суб'єктивної інтерпретації
Високопродуктивні обчислювальні моделі	Прогнозне моделювання, інтеграція даних	Швидкий аналіз, виявлення потенційних небезпек	Точність моделі залежить від якості даних та складності моделі

Сучасні комп’ютерні технології, засновані на експертних правилах, статистичних залежностях та автоматизованих алгоритмах аналізу даних, відіграють ключову роль у прогнозуванні біологічної активності хімічних сполук. Системи, що поєднують статистичні методи, знання експертів та закономірності, виявлені в експериментальних наборах даних, дозволяють формувати обґрунтовані прогнози щодо потенційної небезпеки речовин та забезпечують високу відтворюваність прийнятих рішень.

Обчислювальні програми пропонують суттєві переваги, оскільки забезпечують швидкий скринінг великих бібліотек хімічних сполук, дозволяють працювати з високими обсягами структурних даних та не потребують значних фінансових чи лабораторних ресурсів. Такі системи дають змогу виявляти потенційно генотоксичні сполуки на ранніх етапах розробки лікарських засобів. Серед найбільш поширених підходів сьогодні домінують QSAR-моделі, системи на основі експертних правил, а також моделі машинного та глибокого навчання (DL), які забезпечують високу точність та масштабованість прогнозування.

Моделі QSAR (Quantitative Structure-Activity Relationship) є одним із найпоширеніших та найбільш обґрунтованих підходів до прогнозування токсикологічних властивостей хімічних сполук, зокрема їх мутагенного

потенціалу. Основною ідеєю QSAR є те, що біологічна активність молекули визначається її структурою, а отже може бути кількісно описана через систему формалізованих параметрів – молекулярних дескрипторів. Дескриптори відображають різні аспекти молекулярної будови та хімічної природи сполуки, включаючи топологічні, геометричні, електронні, енергетичні та квантово-хімічні характеристики. Серед найбільш інформативних дескрипторів традиційно розглядають молекулярну масу, ліпофільність (logP), площу полярної поверхні (PSA), зарядові розподіли, електронну густину, енергетичні параметри орбіталей, а також більш складні індекси форми та електронної доступності.

У контексті оцінки генотоксичності QSAR-моделі здійснюють ранжування та статистичний аналіз дескрипторів відповідно до їх кореляції з цільовою кінцевою точкою, такою як наявність мутагенного ефекту або ступінь ушкодження ДНК. Побудовані регресійні чи класифікаційні залежності дозволяють ідентифікувати критичні структурні фрагменти, що сприяють появі генотоксичності, та формувати передбачення для нових сполук без проведення додаткових експериментів. Такий підхід поєднує статистичну строгість із хімічною інтерпретованістю, що робить QSAR одним із ключових інструментів сучасної комп’ютерної токсикології.

В задачах бінарної класифікації оцінки мутагенності QSAR моделі можуть досягати точності зовнішньої валідації від 70% до 90%, залежно від набору даних та конкретної використовуваної моделі. Але зазвичай експериментальна точність є значено нижчою. Відповідно до наукової роботи [13] продуктивність окремих інструментів QSAR показала збалансовану точність у 57–73%, що є доволі посереднім показником [14].

Популярні платформи QSAR включають Toxtree, VEGA, OECD QSAR toolbox.

Toxtree – це відкрита програмна платформа для прогнозування токсичності хімічних сполук на основі правил і структурних ознак. Вона використовує набори експертних правил, які дозволяють оцінювати

потенційну мутагенність, канцерогенність, подразнювальну та інші токсичні властивості молекул, застосовуючи підхід дерева рішень. Toxtree підтримує інтеграцію з різними базами даних і надає прозорі пояснення для своїх прогнозів, що робить її корисним інструментом для швидкої первинної оцінки хімічних речовин у наукових та регуляторних дослідженнях.

VEGA – це платформа для кількісної оцінки структури-активності, що об’єднує різні моделі для прогнозування токсичності, екологічних ризиків та фізико-хімічних властивостей. Вона забезпечує користувачів механізмами перевірки надійності прогнозів та пояснює результати на основі структурних дескрипторів, що дозволяє більш обґрунтовано використовувати моделі у дослідженнях та регуляторному тестуванні.

OECD QSAR Toolbox – це програмна платформа, розроблена Організацією економічного співробітництва та розвитку (OECD) для систематизації та прогнозування токсичності хімічних сполук. Вона дозволяє класифікувати речовини, ідентифікувати структурні патерни, що пов’язані з токсичністю, та переносити дані з подібних речовин. Toolbox інтегрує великий обсяг експериментальних даних та QSAR-моделей, що сприяє стандартизованій і прозорій оцінці ризику для регуляторних та наукових цілей.

Основні особливості Toolbox:

1. Визначення відповідних структурних характеристик та потенційного механізму або способу дії цільової хімічної речовини.
2. Ідентифікація інших хімічних речовин, що мають такі ж структурні характеристики та/або механізм чи спосіб дії.
3. Використання існуючих експериментальних даних для заповнення прогалин у даних.

1.3. Використання моделей машинного навчання для аналізу мутагенності

Методи машинного навчання (ML) та глибокого навчання (DL) за останні роки зарекомендували себе як надзвичайно потужні інструменти для прогнозування мутагенності. Вони суттєво розширюють можливості класичних моделей QSAR, оскільки працюють з великими та різнорідними наборами даних, виявляючи складні нелінійні взаємозв'язки у структурних і функціональних характеристиках молекул. Зокрема, алгоритми навчання з учителем, такі як метод опорних векторів (SVM), випадковий ліс (RF) та метод k-найближчих сусідів (KNN), широко використовуються для класифікації хімічних речовин на генотоксичні та негенотоксичні за їхніми молекулярними дескрипторами [15, 16, 46].

Існують різні види кінцевих точок токсичності, які були передбачені *in silico*. Гостра пероральна токсичність, гепатотоксичність, кардіотоксичність, ендокринні порушення та мутагенність Еймса є одними з найбільш часто досліджуваних. Методи ML демонструють різну ефективність на різних наборах даних через різну складність, розподіл класів або охоплений хімічний простір, що ускладнює порівняння ефективності алгоритмів для різних кінцевих точок токсичності. Порівняння також ускладнюється різними метриками оцінки, визначення яких залежить головним чином від методу ML та використовуваної бази даних. [13, 17]

У межах магістерської дисертації зосереджено увагу на прогнозуванні мутагенності за тестом Еймса. Обрана кінцева точка генотоксичності дозволяє побудувати та оцінити моделі машинного навчання для класифікації хімічних сполук на мутагени та немутагени, що є критично важливим для швидкої оцінки потенційного генотоксичного ризику хімічних сполук.

Результати багатьох досліджень підтверджують ефективність методів ML та DL в контексті класифікації хімічних сполук за проявами

мутагенності. [38] Наприклад, відповідно до наукової роботи [18] застосування SVM та RF забезпечує точність прогнозування на рівні 80,5% та 83,4% відповідно, що перевищує показники багатьох традиційних підходів. Це свідчить про перспективність використання штучного інтелекту не лише як інструмента для автоматизації процесу оцінки генотоксичності, але й як засобу для відкриття нових біологічних закономірностей, які можуть бути корисними у фармацевтичних дослідженнях та токсикологічному скринінгу.

До популярних платформ, які використовують методи машинного навчання для прогнозування мутагенності хімічних сполук належать:

- ProTox 3.0 – віртуальної лабораторії для прогнозування токсичності малих молекул. Платформа використовує аналіз молекулярної подібності, оцінку внеску структурних фрагментів у токсичні ефекти та виявлення характерних повторюваних структурних мотивів. Крім того, система застосовує методи машинного навчання, зокрема перехресну валідацію CLUSTER на основі подібності фрагментів, формуючи загалом 61 модель для прогнозування різних кінцевих точок токсичності, таких як гостра токсичність, токсичність органів, токсикологічні кінцеві точки, молекулярні ініціативні події, метаболізм тощо. Модель прогнозування мутагенності базується на даних тесту Еймса і навчена на 6 156 сполуках, а тестовий набір містить 685 сполук. Збалансована точність прогнозів за 10-кратною перехресною перевіркою становить 84% [19].
- Вебсервер vNN-ADMET – це загальнодоступна онлайн-платформа для прогнозування властивостей абсорбції, розподілу, метаболізму, екскреції та токсичності та для створення нових моделей на основі методології змінного найближчого сусіда (vNN). Сервер містить 15 моделей прогнозування ADMET для швидкої та надійної оцінки деяких найважливіших властивостей, таких як цитотоксичність, мутагенність, кардіотоксичність, лікарська взаємодія, мікросомальна стабільність та лікарсько-індуковане ураження печінки. Модель прогнозування

мутагенності, розміщена на цьому сервері, демонструє показники точності від 54,4% до 85,4% залежно від відстані Танімото та коефіцієнту згладжування [20].

У бакалаврській роботі [2, 21] створювалися моделі машинного навчання для прогнозування мутагенності, результати загальної точності моделей на тестовому наборі даних становили від 79% до 86%. У магістерській дисертації ці показники будуть підвищені за рахунок розділення даних на хімічні класи, оскільки речовини з подібними структурними та фізико-хімічними властивостями можуть демонструвати схожі механізми дії. Такий підхід дозволить моделі краще виявляти специфічні закономірності всередині окремих груп сполук і, відповідно, забезпечити більш точне прогнозування їхньої мутагенної активності.

Зазвичай процес створення моделі охоплює кілька етапів: спочатку здійснюється збір та попередня обробка даних, потім визначається спосіб представлення молекул і підбираються відповідні алгоритми. Наступним кроком є навчання моделі та її оцінювання на частині даних (навчальна вибірка), після чого проводиться перевірка точності на тестовому наборі, з яким система не працювала раніше. На рис. 1.1 наведено узагальнену схему конвеєра для прогнозування мутагенного потенціалу із застосуванням методів машинного навчання.

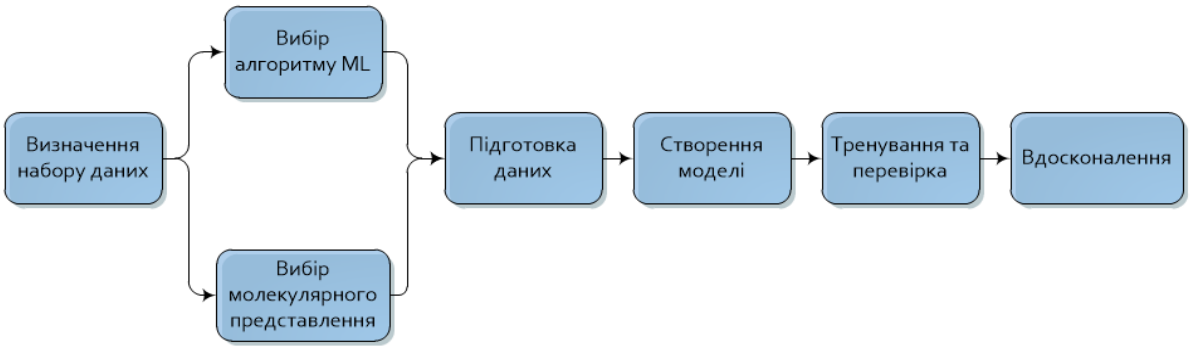


Рисунок 1.1 – Загальний конвеєр моделей машинного навчання

Однією з головних проблем, з якими сьогодні стикається обчислювальна токсикологія, є здатність отримати пояснення або розуміння виявлених токсичних реакцій. Методи машинного навчання, особливо моделі глибокого навчання, зазвичай розглядаються як «чорні скриньки»: вони можуть ефективно вирішувати складні задачі прогнозування, виявляти нелінійні взаємозв'язки між молекулярними ознаками та токсичними ефектами, проте часто не надають ясного пояснення, які саме структурні або хімічні фактори зумовили прогнозований результат. Це створює труднощі для інтерпретації результатів, особливо у регуляторних та наукових контекстах, де критично важливо не лише передбачити токсичність, а й зрозуміти механізм її виникнення. Через це сучасні дослідження активно інтегрують методи інтерпретованого машинного навчання, які дозволяють виділяти ключові дескриптори, оцінювати їхній внесок у прогноз та робити моделі більш зрозумілими, що, у свою чергу, дає змогу встановлювати причинно-наслідкові зв'язки між структурними характеристиками молекул і їх токсичними ефектами, підвищує надійність прогнозів і сприяє прийняттю обґрунтованих рішень у токсикологічній оцінці та розробці безпечних хімічних сполук.

Висновок з розділу 1

У даному розділі було розглянуто основні теоретичні та практичні аспекти генетичної токсикології, що є базовою основою для аналізу мутагенності хімічних сполук на основі методів машинного навчання. Детально проаналізовано традиційні методи дослідження, зокрема тести *in vitro* (тест Еймса, мікроядерний, кометний, аналіз хромосомних аберацій) та *in vivo*, які дозволяють оцінити мутагенний потенціал речовин на різних рівнях біологічної організації. Показано, що незважаючи на те, що ці методи залишаються важливими у токсикологічній оцінці, вони мають певні обмеження, пов'язані з витратами, тривалістю проведення та етичними

аспектами, що стимулює розвиток нових підходів на основі високопродуктивних і обчислювальних технологій.

Визначено сучасні тенденції у розвитку високопродуктивних технологій та обчислювальної токсикології, які забезпечують можливість аналізу великих масивів даних і швидкої ідентифікації потенційно небезпечних сполук. Висвітлено роль методів QSAR, а також моделей машинного та глибокого навчання, що дозволяють прогнозувати генотоксичність хімічних речовин за їх молекулярними дескрипторами. Застосування цих методів відкриває шлях до створення ефективних інтелектуальних систем, здатних зменшити потребу у *in vivo* та *in vitro* оцінці, підвищити точність прогнозування та прискорити процес виявлення мутагенного потенціалу хімічних сполук, які можуть нести ризиків для здоров’я людини та навколишнього середовища.

РОЗДІЛ 2
ВИБІР ТА ОБҐРУНТУВАННЯ ЗАСОБІВ РЕАЛІЗАЦІЇ МОДЕЛЕЙ МАШИННОГО НАВЧАННЯ

У цьому розділі представлено вибір та обґрунтування інструментів, використаних для розробки та реалізації моделей машинного навчання, призначених для прогнозування мутагенності хімічних сполук. Основна увага зосереджена на підборі програмного забезпечення, мови програмування, бібліотек та фреймворків, які забезпечують ефективну обробку даних, побудову моделей, навчання нейронних мереж і оцінювання результатів. Вибір середовища реалізації ґрунтується на його сумісності з сучасними підходами до моделювання, масштабованістю та здатністю інтегруватися з науковими інструментами аналізу даних.

2.1. Огляд застосованих алгоритмів для моделі оцінки генотоксичності

Для розробки моделей прогнозування мутагенності було застосовано алгоритми ML та DL, таких як SVM, RF, k- NN та нейронні мережі (NN). У науковій праці [22, с. 1963-1966] наведено алгоритми ML та DL, які використовувалися в описаних моделях прогнозування токсичності. На рис. 3.1 A представлено частоту використання алгоритмів ML та DL у моделях прогнозування токсичності, на рис. 2.1 B наведено продуктивність моделей на внутрішній та зовнішній валідаціях.

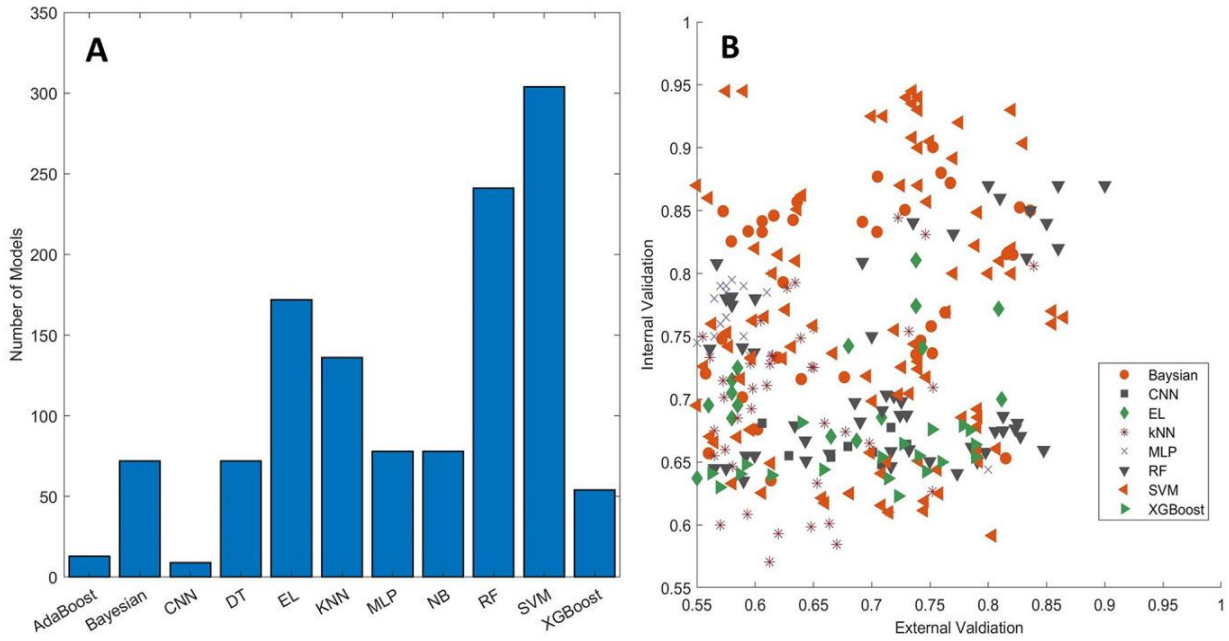


Рисунок 2.1 A – Частота використання алгоритмів ML та DL у моделях прогнозування токсичності; B – продуктивність моделей на внутрішній та зовнішній валідаціях [22]

Для моделей ML найчастіше використовуються SVM (Support Vector Machines), RF (Random Forest) та EL (Ensemble Learning), про які зареєстровано 304, 241 та 172 моделі відповідно. Для моделей DL широко використовуються алгоритми MLP та CNN, про які зареєстровано 78 та 9 моделей відповідно.

У межах магістерської дисертації для оцінки мутагенності використовувалися наступні методи: випадковий ліс, градієнтний бустинг та багатошаровий перцептрон (нейронна мережа). Вибір моделей обґрунтований кількома ключовими аспектами їх ефективності та практичності для вирішення задачі бінарної класифікації.

По-перше, Random Forest і Gradient Boosting добре зарекомендували себе на табличних даних з великою кількістю ознак, що представлені молекулярними дескрипторами. Вони здатні враховувати нелінійні та взаємозалежні зв’язки між структурними характеристиками молекул, що є критично важливим для правильної класифікації мутагенних та

немутагенних сполук. Крім того, обидві моделі дозволяють оцінювати важливість ознак, що сприяє покращенню інтерпретації результатів та виділенню ключових факторів, які впливають на мутагенну активність. Це важливо для підвищення прозорості моделей та формування гіпотез щодо механізмів проявів мутагенності.

По-друге, багатошаровий перцептрон (MLP) є типовою моделлю глибокого навчання, яка здатна моделювати складні, високовимірні взаємозв’язки між дескрипторами. Використання MLP дозволяє передбачати ймовірність мутагенності сполук на основі багатовимірних вхідних даних, що часто дає перевагу у випадках, коли лінійні або слабко нелінійні моделі не здатні захопити всі важливі закономірності. Завдяки здатності MLP працювати з великими обсягами даних і будувати складні функції, вона доповнює моделі ансамблевого типу, забезпечуючи комплексну оцінку і підвищуючи загальну точність прогнозів.

Поєднання цих трьох підходів дозволяє отримати комплексний підхід до прогнозування мутагенності, поєднуючи переваги інтерпретованих ансамблевих методів і потужних моделей глибокого навчання. Це забезпечує не лише високу точність класифікації, а й можливість аналізу важливості молекулярних ознак, що є цінним для подальших наукових та регуляторних застосувань.

2.2. Вибір мови програмування

Для реалізації програмного забезпечення, призначеного для побудови, тренування та розгортання моделей машинного та глибокого навчання, обрано мову програмування Python. При виборі мови програмування враховувалися такі критерії:

1) Наявність готових бібліотек і фреймворків;
2) Висока швидкість роботи через оптимізовані бібліотеки;
3) Простота написання коду і зручність розробки;

4) Велика спільнота і підтримка користувачів;
5) Легка інтеграція з іншими системами.

Популярність Python продовжує зростати, перевершуючи такі мови, як Java, C, C++ та C#. Станом на 2025 рік Python залишається найбільш затребуваною мовою програмування, згідно з оголошеннями про роботу в США, а також є найпоширенішою мовою на платформі GitHub. [23]

Згідно з рейтингом PYPL за 2025 рік, Python очолює список найпопулярніших мов програмування, що ілюструє рис. 2.2.

Rank	Change	Language	Share	1-year trend
1		Python	28.97 %	-0.5 %
2		Java	13.94 %	-1.5 %
3	↑	C/C++	10.54 %	+3.6 %
4	↑↑↑↑↑↑↑	Objective-C	7.05 %	+4.5 %
5	↓↓	JavaScript	6.33 %	-1.7 %
6		R	5.27 %	+0.6 %
7	↓↓	C#	3.96 %	-2.5 %
8	↓	PHP	3.19 %	-0.8 %

Рисунок 2.2 – Рейтинг мов програмування PYPL у 2025 році [24]

Вибір мови програмування Python для реалізації систем машинного та глибокого навчання обґрунтовується насамперед його багатою екосистемою спеціалізованих бібліотек і фреймворків, що забезпечують швидку та ефективну розробку алгоритмів. Python надає доступ до широкого спектру готових інструментів, серед яких scikit-learn для класичних моделей машинного навчання, TensorFlow та PyTorch для побудови глибоких

нейронних мереж, а також Keras, що забезпечує високорівневу абстракцію при побудові моделей. Використання цих бібліотек дозволяє розробникам зосередитися на проектуванні та оптимізації алгоритмів, мінімізуючи необхідність створювати базові обчислювальні функції з нуля. Це суттєво скорочує час розробки, знижує ймовірність помилок у коді та сприяє більшій надійності та відтворюваності результатів [25, 26].

Додатковою перевагою Python є наявність спеціалізованих бібліотек для роботи з хімічними структурами, таких як RDKit, Open Babel, DeepChem та MolVS, що дозволяє ефективно обробляти формати SMILES, генерувати молекулярні дескриптори та безпосередньо інтегрувати їх у моделі машинного та глибокого навчання. Це особливо важливо для завдань передбачення біологічної активності та токсичності хімічних сполук, де точність і структурна релевантність даних є критичними.

Ще одним важливим аспектом є продуктивність обчислень. Хоча Python є інтерпретованою мовою, більшість сучасних бібліотек машинного навчання реалізовані на високопродуктивних мовах програмування C/C++ і підтримують обчислення на графічних процесорах (GPU). Це дозволяє ефективно працювати з великими наборами даних, скорочуючи час навчання моделей, що особливо актуально при використанні глибоких нейронних мереж з великою кількістю параметрів [27].

Важливою перевагою є також широка інтеграція Python з іншими системами та технологіями. Мова підтримує взаємодію з різними базами даних, веб-сервісами та іншими мовами програмування, що спрощує розгортання навчальних моделей у виробничих середовищах. Крім того, стандартизовані формати обміну моделями, такі як ONNX (Open Neural Network Exchange), забезпечують переносимість моделей між різними платформами і середовищами виконання, що сприяє уніфікації робочих процесів та підвищує відтворюваність наукових результатів.

Незважаючи на численні переваги, Python має певні обмеження щодо швидкості обчислення та використання пам’яті порівняно з компільованими

мовами, такими як C++ чи Java. Інтерпретований характер мови знижує швидкість виконання числових обчислень, наприклад у задачах сортування або множення великих матриць. Однак ці обмеження можуть бути частково компенсовані використанням високопродуктивних бібліотек, паралельних обчислень та обчислень на GPU.

Таким чином, враховуючи велику кількість бібліотек та фреймворків, ефективність обчислень, зручність розробки, підтримку спільноти та можливості інтеграції з іншими системами, Python є оптимальною мовою програмування для реалізації програмного забезпечення, призначеного для побудови, навчання та розгортання моделей машинного та глибокого навчання. Недоліки мови не є критичними і можуть бути подолані за допомогою сучасних інструментів, що робить Python логічним вибором для сучасних науково-дослідних та прикладних проєктів у галузі обчислювальної токсикології, біоінформатики та хімічного моделювання.

2.3. Випадковий ліс

Випадковий ліс – це поширений алгоритм машинного навчання, який був заснований Лео Брейманом та Адель Катлер, він поєднує результати кількох дерев рішень для отримання єдиного результату. Випадковий ліс використовується для вирішення як задач класифікації, так і задач регресії. Кожне дерево навчається на випадковій підмножині даних, а для кожного розгалуження вибирається випадкова підмножина ознак. Прогнозування здійснюється шляхом агрегації результатів усіх дерев: для задач класифікації – за допомогою голосування більшості, для регресії – шляхом обчислення середнього значення. [28]

Основні переваги та недоліки даного алгоритму представлені в табл. 2.1 нижче.

Таблиця 2.1 – Переваги та недоліки випадкового лісу

Переваги	Недоліки
Висока точність прогнозування	Зниження інтерпретованості моделі
Стійкість до перенавчання	Високі вимоги до пам'яті та часу обчислень
Можливість оцінки важливості ознак	Можливі упередження у оцінці важливості ознак
Гнучкість у роботі з різними типами даних	Складність налаштування гіперпараметрів

Для реалізації моделей випадкового лісу у даній роботі обрано бібліотеку scikit-learn, оскільки вона є перевіреним стандартом у класичному машинному навчанні, забезпечує високу стабільність і ефективність при роботі з середніми за розміром наборами даних, а також надає широкий набір готових інструментів для підготовки даних, крос-валідації, оцінки точності моделей та оптимізації гіперпараметрів. Крім того, scikit-learn відрізняється простотою інтеграції з іншими бібліотеками Python, такими як pandas і NumPy, що забезпечує гнучкість у обробці молекулярних дескрипторів та дозволяє швидко створювати прототипи та відлагоджувати моделі.

Існує декілька підходів для кількісного визначення важливості ознак у моделях випадкового лісу, кожен із яких ґрунтується на різних принципах оцінювання внеску окремих дескрипторів у якість прогнозу. Одним із найпоширеніших методів є mean decrease impurity (MDI), або важливість за зменшенням нечистоти. Цей підхід оцінює, наскільки використання певної ознаки сприяє розщепленню вузлів у дереві рішень, зменшуючи міру нечистоти (наприклад, ентропію або індекс Джині). Чим більший сумарний внесок ознаки у зниження нечистоти по всіх деревах ансамблю, тим вищою є її важливість. MDI є ефективним і швидким методом, проте він має певні обмеження, зокрема схильність завищувати важливість ознак із великою кількістю унікальних значень.

Другим поширеним методом є mean decrease accuracy (MDA). Цей підхід ґрунтується на аналізі зміни точності моделі після випадкового

перемішування значень окремої ознаки в тестовому наборі. Якщо перемішування призводить до суттєвого погіршення точності, ознака вважається важливою для прогнозу. MDA має перевагу над MDI, оскільки менш чутливий до масштабів ознак і здатний точніше оцінювати реальну прогностичну цінність дескрипторів. Водночас він є більш обчислювально затратним, оскільки потребує повторного оцінювання моделі для кожної ознаки.

Ще одним потужним методом є recursive feature elimination (RFE), який реалізує ітеративну стратегію зворотного відбору ознак. На кожному кроці модель навчається на повному наборі наявних дескрипторів, після чого видаляється найменш значуща ознака відповідно до певного критерію (наприклад, ваги моделі або важливості за деревами). Процес повторюється доти, доки не буде досягнуто оптимальної кількості ознак. RFE дозволяє не лише оцінити важливість окремих дескрипторів, але й сформувати мінімальний інформативний набір, що забезпечує високу точність прогнозу та зменшує ризик перенавчання. У контексті прогнозування мутагенності цей метод є особливо корисним, оскільки дозволяє усунути надлишкові або корельовані молекулярні дескриптори, підвищуючи інтерпретованість та стабільність моделей. [29]

Вищеперераховані методи дозволяють визначити, які ознаки найбільше впливають на прогнозування моделі, що є корисним для розуміння механізмів прийняття рішень та оптимізації моделі для кращої точності та скорочення розмірності даних.

2.4. Градієнтний бустинг

Градієнтний бустинг (Gradient Boosting) – це потужний ансамблевий метод машинного навчання, який поєднує слабкі моделі (зазвичай дерева рішень) для побудови сильної прогностичної моделі. Процес навчання полягає у послідовному додаванні нових моделей, кожна з яких коригує

помилки попередніх, мінімізуючи функцію втрат за допомогою градієнтного спуску. Цей підхід дозволяє ефективно моделювати складні нелінійні залежності в даних [30].

Сучасні реалізації градієнтного бустингу, такі як XGBoost, LightGBM та CatBoost, відрізняються за стратегіями побудови дерев, оптимізацією навчання та підтримкою роботи з категоріальними ознаками, що безпосередньо впливає на продуктивність та ефективність моделей. Порівняння основних видів градієнтного бустингу наведено в табл. 2.2 нижче.

Таблиця 2.2 – Порівняння основних видів градієнтного бустингу

Критерій	XGBoost	LightGBM	CatBoost	Коментар
Тип дерев	Класичні CART, рівноважні по глибині	Листова оптимізація (Leaf-wise)	CART та Ordered Boosting	LightGBM часто швидший завдяки листовій стратегії
Прискорення навчання	Паралельне навчання дерев, оптимізація кешу	Глибока оптимізація Leaf-wise, histogram-based	Pipelined computation, CPU/GPU	LightGBM зазвичай швидший на великих наборах даних
Обробка категоріальних змінних	Не підтримує без кодування (one-hot)	Потрібне кодування (one-hot, label encoding)	Вбудована обробка категоріальних змінних	CatBoost вигідний для задач з великою кількістю категорій
Регуляризація	L1, L2, early stopping, subsample	L1, L2, bagging, feature fraction	L2, Bayesian smoothing	Усі методи підтримують регуляризацію для уникнення перенавчання

Кінець табл. 2.2

Критерій	XGBoost	LightGBM	CatBoost	Коментар
Обробка відсутніх значень	Вбудована	Вбудована	Вбудована	Усі фреймворки підтримують автоматичну обробку NaN
Підтримка GPU	Так, для прискорення навчання	Так, підтримка CUDA	Так, обмежена	Для швидкого навчання великих моделей GPU часто критично важлвий
Популярність та використання	Широко у фінансовому та біомедичному секторі	Широко у промислових проектах з великими даними	Часто у задачах з категоріальними ознаками (маркетинг, психологія)	Вибір залежить від типу даних та обмежень ресурсів

Серед сучасних реалізацій градієнтного бустингу XGBoost вирізняється високою стабільністю та точністю прогнозування завдяки поєднанню класичних дерев рішень з ефективними методами регуляризації, такими як L1 та L2, а також можливістю ранньої зупинки навчання. Він демонструє ефективну роботу на числових та змішаних наборах даних середнього розміру, забезпечує контроль перенавчання та надає гнучкі механізми налаштування параметрів моделі. Крім того, XGBoost підтримує паралельне навчання дерев та оптимізовану роботу з пам’яттю, що дозволяє зменшити час навчання без втрати точності. Також XGBoost є найбільш популярним методом бустингу в світі за даними Google Trends (див. рис. 2.3).

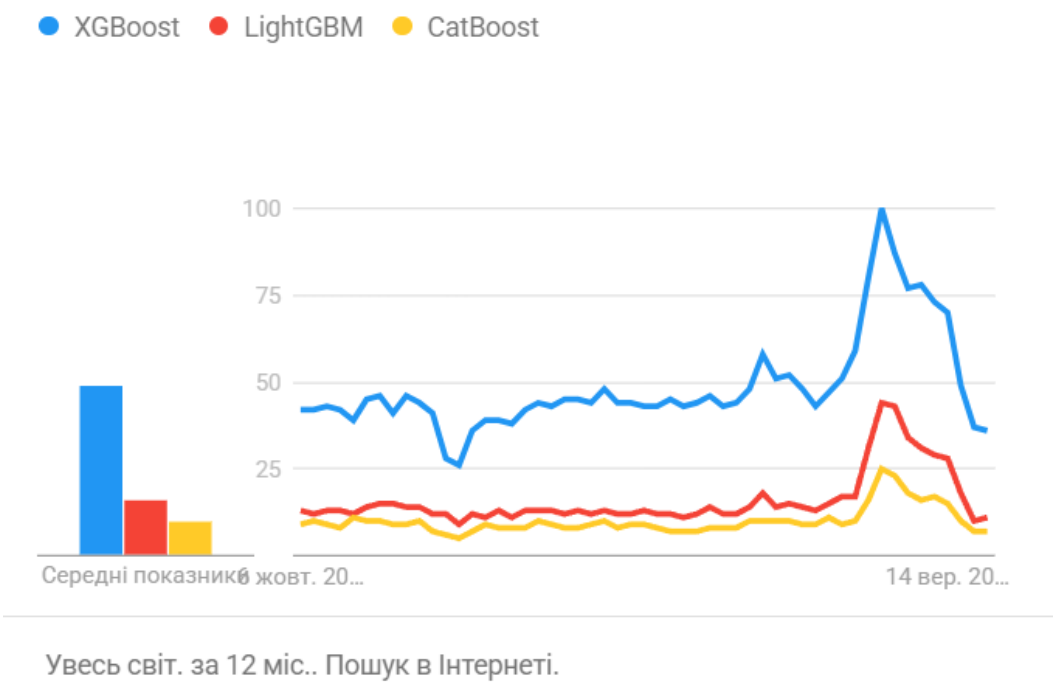


Рисунок 2.3 – Глобальна популярність видів бустингу протягом останнього року

Для реалізації моделей градієнтного бустингу у даній роботі обрано XGBoost, оскільки цей фреймворк поєднує високу точність прогнозування з ефективним використанням обчислювальних ресурсів, підтримує паралельне навчання дерев та оптимізовану роботу з пам’яттю, що особливо важливо при аналізі великої кількості хімічних дескрипторів. Крім того, XGBoost надає розширені механізми регуляризації (L1, L2, рання зупинка), що дозволяє зменшити ризик перенавчання, а також забезпечує гнучке налаштування гіперпараметрів, що робить його надійним інструментом для побудови точних та стабільних моделей прогнозування мутагенності.

2.5. Нейронна мережа

Нейронні мережі сьогодні є одним із найпотужніших інструментів для задач прогнозування біологічних ефектів хімічних сполук, зокрема токсичності та генотоксичності. Вони дозволяють знаходити нелінійні

залежності між структурними ознаками молекул і біологічною активністю, поєднуючи великі набори дескрипторів, послідовностей SMILES або графових представлень молекул у єдину модель прогнозу. Це робить їх придатними для автоматизованих систем скринінгу, що дозволяє ефективно виявляти потенційно небезпечні сполуки та мінімізувати ресурси на експериментальне тестування, що є ключовою метою даної роботи [22, 25]

У програмних застосунках використовуються різні типи нейронних мереж, кожен з яких має свої сильні та слабкі сторони:

Багатошарові персептрони (Feed-forward / DNN) – класичні щільно-зв'язані мережі, які ефективно працюють з наборами попередньо обчислених дескрипторів або фрагментних fingerprints (наприклад, ECFP, MACCS).

Згорткові нейронні мережі (CNN) для рядкових представлень (SMILES) – CNN та 1D-CNN підходи можуть опрацьовувати послідовності SMILES як «мову молекули», вилучаючи локальні патерни, що відповідають функціональним підструктурам. Такі підходи показали добрі результати в гібридних моделях для прогнозу канцерогенности й генотоксичності.

Графові нейронні мережі (GNN) – моделюють молекулу як граф (атоми – вершини, зв’язки – ребра) і навчаються безпосередньо на топології та атрибутах вузлів/ребер. Сучасні варіанти (включно з еквіваріантними GNN, що враховують 3D-геометрію) демонструють найкращі результати для прогнозу біологічних властивостей, оскільки здатні захоплювати просторово-хімічні взаємодії, релевантні для механізмів токсичної дії.

У магістерській дисертації для побудови моделі прогнозування обрано багатошаровий персептрон (Multilayer Perceptron, MLP) як оптимальний тип нейронної мережі. Це рішення зумовлене характером вхідних даних – числовими молекулярними дескрипторами, які відображають структурні, електронні, геометричні та топологічні властивості сполук. Для таких даних не потрібна складна просторово-графова обробка, тому використання архітектур на кшталт CNN чи GNN не є необхідним. MLP, у свою чергу,

ефективно працює саме з табличними наборами ознак, дозволяючи моделі виявляти нелінійні взаємозв’язки між дескрипторами та цільовою змінною.

При реалізації нейронних мереж у практичній системі важливі два підходи: швидкість розробки для дослідницьких експериментів і стабільність/масштабованість для продакшн-розгортання. Серед найбільш поширених фреймворків – TensorFlow і PyTorch, їх порівняння неведене в табл. 2.3.

PyTorch – це фреймворк з відкритим кодом для глибокого навчання, розроблений Лабораторією досліджень штучного інтелекту (FAIR) компанії Facebook. Він широко використовується для машинного навчання (ML) та програм глибокого навчання, включаючи комп'ютерний зір, обробку природної мови (NLP) та навчання з учителем. Він надає гнучкий спосіб побудови та навчання нейронних мереж за допомогою динамічного графа обчислень [26].

TensorFlow – це фреймворк машинного навчання з відкритим кодом, розроблений Google у 2015 році. Він широко використовується для глибокого навчання, складних обчислень та масштабних застосувань машинного навчання в різних галузях. Він застосовується для застосувань машинного навчання в охороні здоров'я, фінансах та автономних системах.

Таблиця 2.3 – Порівняння TensorFlow і PyTorch

Критерій	**TensorFlow**	**PyTorch**	**Коментар**
Тип фреймворку	Структурований фреймворк для глибокого навчання із низькорівневим API та високорівневим API Keras	Динамічний фреймворк для глибокого навчання із “define-by-run” підходом	TensorFlow більше нагадує традиційний фреймворк з графами обчислень, PyTorch – гнучкий у динамічних обчисленнях

Кінець табл. 2.3

Критерій	TensorFlow	PyTorch	Коментар
Тип фреймворку	Структурований фреймворк для глибокого навчання із низькорівневим API та високорівневим API Keras	Динамічний фреймворк для глибокого навчання із “define-by-run” підходом	TensorFlow більше нагадує традиційний фреймворк з графами обчислень, PyTorch – гнучкий у динамічних обчисленнях
API високого рівня	Keras інтегровано, спрощує створення моделей	Torch.nn, Torchvision для комп’ютерного зору	Keras робить TensorFlow зручним для початківців, PyTorch більш «ручний» для кастомізації
Продуктивність на GPU/TPU	Висока, підтримка NVIDIA GPU та TPU (Google Cloud TPU)	Висока, оптимізація під CUDA	TensorFlow має підтримку TPU, PyTorch підтримує GPU через CUDA
Популярність у наукових дослідженнях	За даними Google Scholar (2024): ~358 000 публікацій	За даними Google Scholar (2024): ~242 000 публікацій	TensorFlow історично лідирує, проте PyTorch швидко набирає популярність у академічних дослідженнях
Популярність у промисловості	Використовується у Google, Intel, IBM; широкий стек для продакшн (TensorFlow Serving)	Використовується у Facebook, Microsoft, OpenAI; популярний у R&D	TensorFlow зручний для розгортання моделей у продуктивних системах, PyTorch – у прототипуванні та дослідженнях
Візуалізація та моніторинг	TensorBoard – вбудоване рішення	TensorBoard через плагін, також WandB, Visdom	TensorFlow має нативний інструмент, PyTorch потребує додаткової інтеграці

Для оцінки актуальності та популярності різних фреймворків для глибокого навчання проведено аналіз запитів у Google Trends за останні 5 років. Це дозволяє виявити, які технології частіше використовуються дослідниками та практиками у глобальному та локальному масштабі, а також оцінити їх тренди в Україні та світі. (див. рис. 2.4 та рис.2.5).

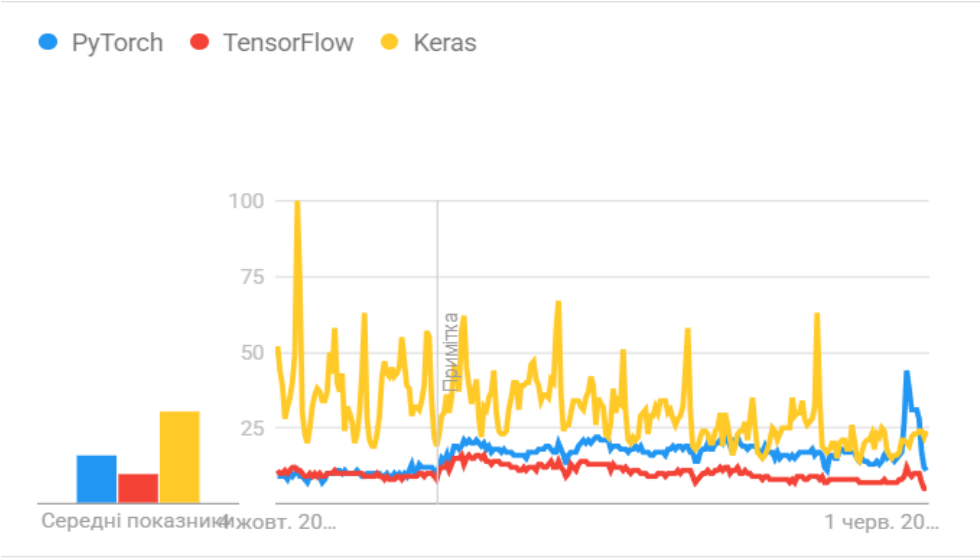


Рисунок 2.4 – Глобальна популярність фреймворків (5 років)

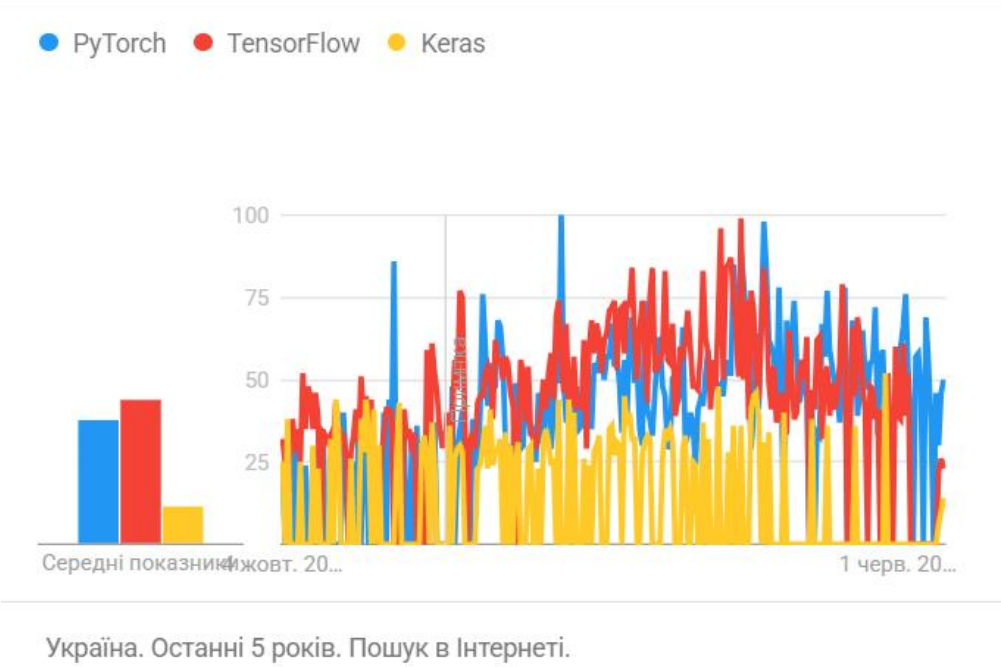


Рисунок 2.5 – Популярність фреймворків в Україні (5 років)

Аналіз популярності фреймворків показує, що TensorFlow залишається найпопулярнішим інструментом для розробки моделей глибокого навчання в Україні, що обумовлює його вибір для реалізації нейронної мережі у даній роботі. Крім того, TensorFlow забезпечує надійну інтеграцію з Keras, підтримку GPU/TPU та можливості для розгортання моделей у продуктивному середовищі, що є критично важливим для створення автоматизованої системи оцінки генотоксичності.

Висновок до розділу 3

У даному розділі було розглянуто підходи, алгоритми та інструменти, що застосовуються для створення моделей аналізу мутагенності хімічних сполук із використанням методів машинного та глибокого навчання. На основі проведеного аналізу було обґрунтовано вибір моделей, реалізованих у даній роботі. Для побудови системи прогнозування використано три підходи: метод випадкового лісу, градієнтного бустингу та штучну нейронну мережу типу багатошарового персептрона. Таке поєднання моделей забезпечує високу точність класифікації та дозволяє враховувати як прості, так і складні нелінійні залежності між молекулярними дескрипторами.

Для реалізації програмного забезпечення обрано мову програмування Python, що обумовлено її популярністю у сфері машинного навчання, наявністю великої кількості спеціалізованих бібліотек і фреймворків, а також простотою розробки та інтеграції з іншими системами. Python надає розробнику доступ до таких бібліотек, як scikit-learn для класичних моделей машинного навчання, TensorFlow і Keras для побудови та навчання нейронних мереж, а також RDKit і DeepChem для обробки хімічних структур і генерації дескрипторів.

У розділі також розглянуто особливості реалізації кожного з вибраних алгоритмів. Для випадкового лісу обґрунтовано застосування бібліотеки scikit-learn як стабільного інструменту для побудови моделей середнього

розміру та оцінки важливості ознак. Для градієнтного бустингу обрано фреймворк XGBoost, який поєднує високу точність із ефективним використанням ресурсів і розвиненими механізмами регуляризації. Для реалізації нейронної мережі типу MLP використано TensorFlow у поєднанні з Keras, що забезпечує зручність побудови моделей, підтримку GPU і можливість масштабування.

РОЗДІЛ 3
РОЗРОБКА МОДЕЛЕЙ МАШИННОГО НАВЧАННЯ ДЛЯ ЗАДАЧІ КЛАСИФІКАЦІЇ

Розділ містить опис процесу розробки та тестування моделей машинного навчання, які є основою системи аналізу мутагенності хімічних сполук, зокрема, формування навчальної вибірки, аналіз хімічних класів та відбір релевантних молекулярних дескрипторів, що становлять основу для побудови якісного алгоритмічного рішення. Послідовно розглянуто етапи підготовки даних, методи скорочення простору ознак та створення різних типів моделей – випадкового лісу, градієнтного бустингу та нейронних мереж.

3.1. Формування навчальної бази даних

Моделі ML та DL навчаються з використанням відомих експериментальних даних для вивчення взаємозв'язків між хімічними структурами та кінцевими точками токсичності в навчальних хімічних речовинах. Тому якість експериментальних даних, що використовуються для навчання моделей ML та DL, важлива для надійності розроблених моделей прогнозування токсичності. Порівняно з великими наборами даних з мільярдами або навіть трильйонами даних в аналізі зображень, розмір даних для галузі прогнозування мутагенності зазвичай відносно невеликий через високу вартість та час, необхідний для проведення токсикологічних експериментів. На рис. 3.1 показано розподіл розмірів наборів даних, які використовувалися для розробки моделей машинного навчання для прогнозування різних типів токсичності. Найбільшим набором даних є набір даних про цитотоксичність, який містить 62 655 сполук, та кілька наборів

даних містять понад 10 000 сполук. Більшість наборів даних містять близько 1 000 хімічних речовин.

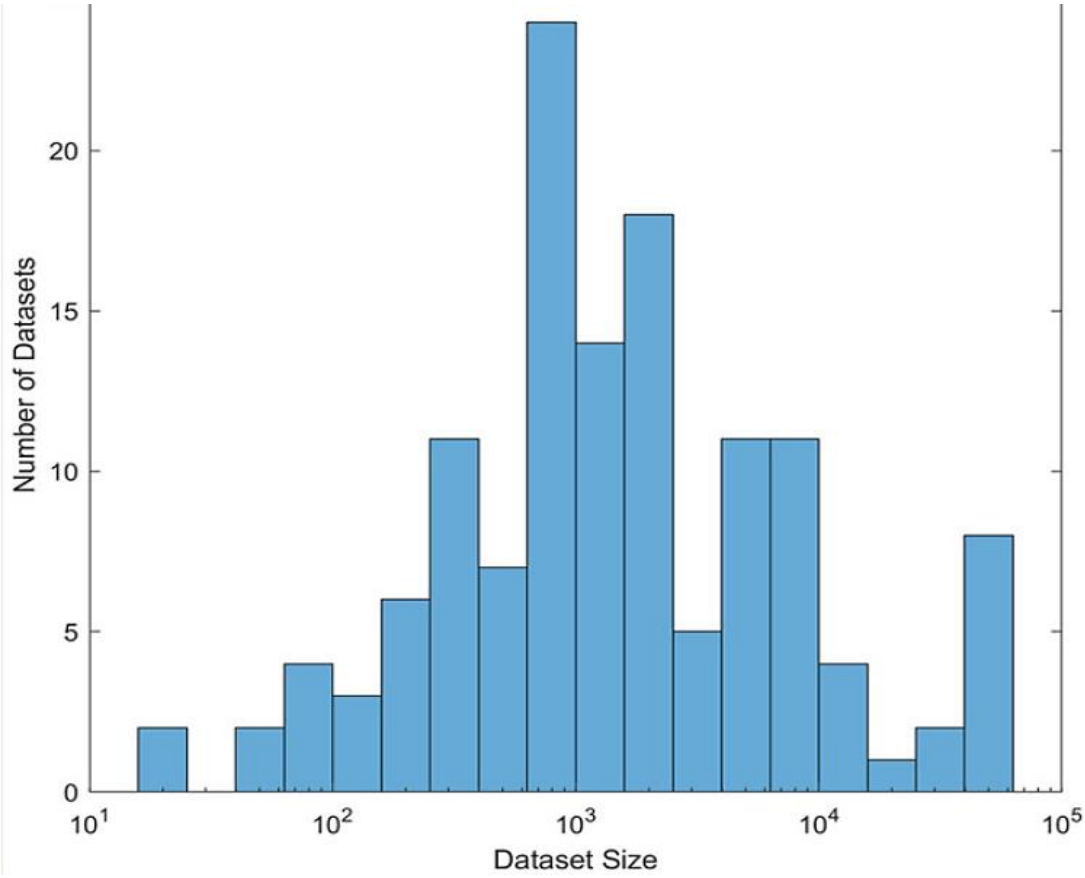


Рисунок 3.1 – Гістограма розмірів наборів даних, що використовуються в розробці моделей машинного навчання для прогнозування токсичності [22]

База даних, використана в даній роботі, була сформована самостійно на основі кількох відомих наборів даних, що широко застосовуються в дослідженнях мутагенності хімічних сполук:

- Набір даних Hansen – один із найвідоміших і найбільш об’ємних відкритих датасетів, створений у 2009 році. Він містить 6 513 хімічних сполук, для яких визначено результати тесту Еймса – стандартного біологічного методу для виявлення мутагенних властивостей речовин. Цей набір охоплює широкий спектр органічних сполук різних класів і є базовим джерелом для розробки моделей QSAR, спрямованих на прогнозування мутагенності [31];

- Набір даних Kazius/Bursi – сформований у 2005 році, містить 4 337 хімічних сполук із чітко визначеними мітками мутагенності. Він є одним із перших систематизованих наборів для оцінки генотоксичного потенціалу речовин і використовується як еталон для валідації алгоритмів машинного навчання. Цей датасет забезпечує збалансоване представлення мутагенних і немутагенних сполук, що робить його особливо цінним для класифікаційних завдань [32];
- Набір даних EFSA – сучасніший набір, створений у 2016 році в межах досліджень, пов’язаних із безпечністю харчових добавок, пестицидів та інших речовин. Він містить 695 сполук, для яких наявна перевірена інформація щодо мутагенності. Незважаючи на менший обсяг, цей набір є важливим джерелом актуальних і високоякісних даних, які відображають сучасні стандарти токсикологічної оцінки [33];
- База даних мікотоксинів – включає 4 257 мікотоксинів, для яких наявна перевірена інформація щодо мутагенності, класифікованих за хімічними структурами у 170 категорій, з використанням інформації про назви, ізомерні SMILES, PubChem CID та інші ідентифікатори [34]. Мікотоксини – це вторинні метаболіти, що виробляються певними нитчастими грибами. Вони можуть накопичуватися у продуктах харчування та передаватися через харчові ланцюги, що створює ризик для здоров’я людини через їх потенційну мутагенність.

Після видалення повторюваних елементів база даних, на основі 4 різних наборів, складалася з 8 454 хімічних сполук, для яких було ідентифіковано їх структурний клас за допомогою веб-сервісу ClassyFire [35]. На рис. 3.2 представлено розподіл даних за класами хімічних сполук.

Рисунок 3.2 – Розподіл даних за класами хімічних сполук

У моделюванні мутагенності за допомогою методів машинного навчання врахування класів хімічних сполук є критично важливим, оскільки різні структурні групи мають відмінні механізми взаємодії з біологічними мішенями. Аліфатичні, ароматичні, гетероциклічні чи гомоциклічні молекули демонструють різну електронну будову, ступінь кон'югації, наявність гетероатомів та типи зв'язків, що визначають їх реакційну здатність і потенціал утворення мутагенних метаболітів. Ці фактори впливають на набір релевантних дескрипторів, а отже – на взаємодії, які модель здатна виявити. Побудова окремих моделей для кожного структурного класу дозволяє алгоритмам ефективно виявляти специфічні для даної групи молекулярні закономірності в даних, необхідні для вирішення задачі бінарної класифікації.

Окрема модель, створена на повному датасеті, навпаки, дозволяє визначити закономірності на рівні всіх класів, спільні для більшості органічних молекул, та оцінити загальну здатність алгоритму прогнозувати

мутагенність у змішаних наборах даних. Порівняння моделей для основних структурних класів з глобальною моделлю дає змогу виявити, які структурні групи є найбільш складними для прогнозування, які дескриптори відіграють ключову роль у межах конкретного класу. Такий підхід не лише підвищує точність прогнозів, але й сприяє кращому розумінню механізмів хімічної мутагенності для різних типів органічних сполук.

Таким чином, для побудови моделей було використано чотири окремі класи хімічних сполук, а також узагальнений клас, який охоплює всю базу даних. Це дозволило здійснити комплексний аналіз даних, забезпечити оцінку ефективності моделей як у межах окремих класів, так і для всієї сукупності сполук, а також провести порівняння отриманих результатів. Детальний опис класів наведено в розділі 3.1.1 нижче.

3.1.1. Класи хімічних сполук

У сучасній хімії класифікація органічних сполук є фундаментальним етапом у дослідженнях їхніх властивостей, реакційної здатності та біологічної активності. У контексті оцінки мутагенності така класифікація має особливе значення, оскільки різні структурні типи сполук по-різному взаємодіють з ДНК та клітинними ферментними системами, утворюючи реактивні метаболіти або проміжні продукти.

Найпростіші органічні сполуки – це ті, що складаються лише з двох елементів: вуглецю та водню. Ці сполуки називаються вуглеводнями. Представлені в базі даних вуглеводні поділяються на два типи: аліфатичні вуглеводні та ароматичні вуглеводні [36].

Аліфатичні вуглеводні – це вуглеводні, що базуються на ланцюгах атомів С. Існує три типи аліфатичних вуглеводнів. Алкани – це аліфатичні вуглеводні лише з одним ковалентним зв'язком. Алкени – це вуглеводні, що містять принаймні один подвійний зв'язок, а алкіни – це вуглеводні, що містять потрійний зв'язок. Зрідка трапляються аліфатичні вуглеводні з кільцем атомів С; ці вуглеводні називаються циклоалканами (або

циклоалкенами, або циклоалкінами). Аліфатичні сполуки можуть бути лінійними або циклічними [36].

Аліфатичні ациклічні сполуки – це вуглеводні або їх похідні, в яких вуглецевий скелет не утворює замкненого кільця (тобто відкриті ланцюги, прямих або розгалужених типів). Вони можуть бути насиченими (алкани) або ненасиченими (алкени, алкіни) і включають широкий спектр хімічних властивостей залежно від замісників та ненасиченості [37].

Аліфатичні гетероциклічні сполуки – це органічні молекули, у структурі яких присутні кільця з одним або кількома гетероатомами (наприклад, киснем, азотом або сіркою), але без системи спряжених подвійних зв’язків, характерної для ароматичних структур. На відміну від ароматичних гетероциклів, аліфатичні мають насичену або частково ненасичену природу, що визначає їхню високу реакційну здатність і схильність до хімічних перетворень. До цієї групи належать такі сполуки, як тетрагідрофуран, піролідин, азиридин, морфолін та інші. Їхні властивості залежать від розміру кільця, типу гетероатома та ступеня насиченості, що впливає на стабільність та можливість утворення реакційно активних проміжних форм.

У біологічному та токсикологічному контексті аліфатичні гетероциклічні сполуки є важливими завдяки своїй участі в біосинтетичних процесах і фармакологічній активності, але деякі представники цього класу можуть проявляти генотоксичні властивості. Зокрема, азиридинові та епоксидні структури здатні утворювати ковалентні аддукти з ДНК, спричиняючи мутації або хромосомні пошкодження.

Ароматичні вуглеводні мають спеціальне шестичленне кільце, яке називається бензольним кільцем. Електрони в бензольному кільці мають особливі енергетичні властивості, що надають бензолу фізичних та хімічних властивостей, що значно відрізняються від алканів. Спочатку термін «ароматичний» використовувався для опису цього класу сполук, оскільки вони були особливо ароматними. Однак у сучасній хімії термін

«ароматичний» позначає наявність шестичленного кільця, яке надає молекулі різних та унікальних властивостей [37].

Ароматичні гетероциклічні сполуки – це циклічні системи, в яких кольцевий ланцюг містить принаймні один гетероатом (найчастіше N, O або S) і які мають ароматичну π-кон'юговану систему, що підпорядковується правилу Гюккеля (4n + 2 π-електронів). Завдяки цьому вони здобувають підвищену стабільність, планарність і специфічні електронні властивості. У класичних підручниках з гетероциклічної хімії описується, що структури на кшталт піридину, фурану, тіофену, індолу тощо є типовими прикладами таких систем.[38]

З точки зору токсикології й генетичного ризику, багато ароматичних гетероциклів (особливо гетероароматичні аміни, Heterocyclic Aromatic Amines, HAAs) виявляють мутагенні й канцерогенні властивості, зумовлені їх метаболічною активацією до реактивних проміжків, які можуть утворювати аддукти з ДНК. Наприклад, у дослідженнях на клітинних лініях людини було показано, що HAA-сполуки індукують мікронуклеї, як ознаку хромосомних ушкоджень.[39] Також дослідження середовищ довкілля вказують, що гетероароматичні поліциклічні сполуки (hetero-PAHs) можуть бути генотоксичними у мікронуклеусових тестах (наприклад, на клітинах печінки риб) – що підкреслює їх екологічний ризик [40].

Ароматичні гомоциклічні сполуки – це органічні молекули, в яких кільцеві системи складаються виключно з вуглецевих атомів і мають ароматичні властивості. Класичним прикладом є бензен та конденсовані ароматичні вуглеводні, такі як нафталін, антрацен, фенантрен. Ці сполуки характеризуються стабільною π-електронною системою, планарною геометрією і високою електронною делокалізацією, що визначає їхню хімічну інертність у багатьох реакціях і специфічну реакційну здатність у заміщувальних процесах.

3.1.2. Набори дескрипторів

У сучасних дослідженнях токсикології та хімічної безпеки важливу роль відіграють молекулярні дескриптори – кількісні характеристики, що описують структуру та властивості молекул. Вони є основою для створення моделей кількісних структурно-активних взаємозв'язків (QSAR), які дозволяють передбачати біологічну активність, зокрема мутагенність, без проведення дорогих та тривалих експериментів. У цьому підрозділі розглянуто три популярні інструменти для генерації молекулярних дескрипторів: PaDEL, RDKit та Mordred, які були використані в роботі.

PaDEL-Descriptor – це програмне забезпечення з відкритим кодом, яке дозволяє обчислювати понад 1800 молекулярних дескрипторів та 12 типів молекулярних відбитків (fingerprints), включаючи CDK, Estate, MACCS, PubChem та інші. Воно підтримує обробку великих наборів молекул і є популярним інструментом у QSAR-дослідженнях [41].

RDKit – це потужний набір інструментів для хімічної інформатики, який надає можливість обчислювати широкий спектр молекулярних дескрипторів, включаючи 1D, 2D та 3D характеристики. RDKit інтегрується з мовою програмування Python, що забезпечує гнучкість у створенні та налаштуванні QSAR-моделей [42].

Mordred – це швидкий та ефективний калькулятор молекулярних дескрипторів, здатний обчислювати понад 1800 2D та 3D дескрипторів. Він вільно доступний через GitHub. Mordred можна легко встановити та використовувати в інтерфейсі командного рядка, як веб-додаток або як високогнучкий пакет Python на всіх основних платформах (Windows, Linux та macOS). Проте слід враховувати, що Mordred на даний момент не підтримує Python 3.13, що може призводити до конфліктів версій при використанні нових середовищ [43].

У зв’язку з цим у магістерський дисертації автоматично обчислюються лише дескриптори RDKit, що гарантує стабільну роботу програми на сучасних версіях Python без конфліктів бібліотек.

У табл. 3.1 представлено узагальнюючу інформацію щодо основних характеристик інструментів для обчислення дескрипторів.

Таблиця 3.1 – Порівняння PaDEL, RDKit та Mordred

Характеристика	PaDEL-Descriptor	RDKit	Mordred
Кількість дескрипторів	1800+	200+	1800+
Типи дескрипторів	1D, 2D, 3D, fingerprints	1D, 2D, 3D	2D, 3D
Підтримка мов програмування	Java, командний рядок, Python	Python	Python, командний рядок
Швидкість обчислень	Середня	Висока	Дуже висока
Ліцензія	Відкрита	Відкрита	Відкрита

3.2. Підготовка даних

У межах дослідження здійснено ретельну підготовку набору даних, що є критично важливим етапом у побудові надійної та інтерпретованої моделі машинного навчання для прогнозування мутагенних ефектів хімічних сполук.

На початковому етапі дані було завантажено у середовище Python із використанням бібліотеки pandas.

Цільова змінна mutagenity, яка приймає дискретні значення, була виділена в окремий вектор Y, а ознаки (features) для моделювання – у матрицю X. Для забезпечення коректної подальшої обробки даних було здійснено усунення пропущених та нескінченних значень. Усі значення NaN, inf та -inf були замінені на нульові. Це дозволило уникнути збоїв при застосуванні статистичних методів та забезпечити стабільність обчислень без втрати даних. Додатково видалено ознаки, що не мають варіативності, тобто ті, у яких усі значення були однаковими для всіх спостережень. Такі змінні

не несуть корисної інформації та можуть лише збільшувати розмірність простору ознак.

Для зменшення мультиколінеарності між ознаками було застосовано фільтрацію за допомогою кореляційної матриці. Обчислено матрицю парних коефіцієнтів Пірсона для всіх ознак. Далі розглядалася лише верхня трикутна частина цієї матриці, щоб уникнути дублювання пар. Ознаки, які мали коефіцієнт кореляції понад 0,95 з будь-якою іншою ознакою, вважалися надлишковими та були видалені (видалялася одна з двох ознак). Такий підхід дозволяє зберегти незалежність ознак та підвищити стійкість моделі до перенавчання.

Після зменшення кількості ознак було проведено нормалізацію. Спершу дані були трансформовані за допомогою методу QuantileTransformer з параметром *output_distribution='uniform'*, що приводить розподіл кожної ознаки до рівномірного. Цей метод зменшує вплив викидів та дозволяє краще працювати з моделями, чутливими до розподілу даних. Додатково до цього застосовано масштабування за допомогою StandardScaler, який центрує кожну ознаку навколо нуля та нормалізує її до одиничного стандартного відхилення. Це забезпечує сумісність масштабів між різними ознаками та покращує збіжність при навчанні моделей.

З метою усунення дисбалансу класів, який є типовим у задачах мутагенності, застосовано техніку штучного надсемплінгу за методом SMOTE. Цей метод генерує нові синтетичні зразки меншості на основі існуючих, зберігаючи просторову структуру даних. Балансування класів дозволяє уникнути упередженості моделі у бік переважаючого класу та забезпечити об'єктивне оцінювання ефективності класифікатора. Розподіл класів хімічних сполук за мутагенністю наведено на рис. 3.3.

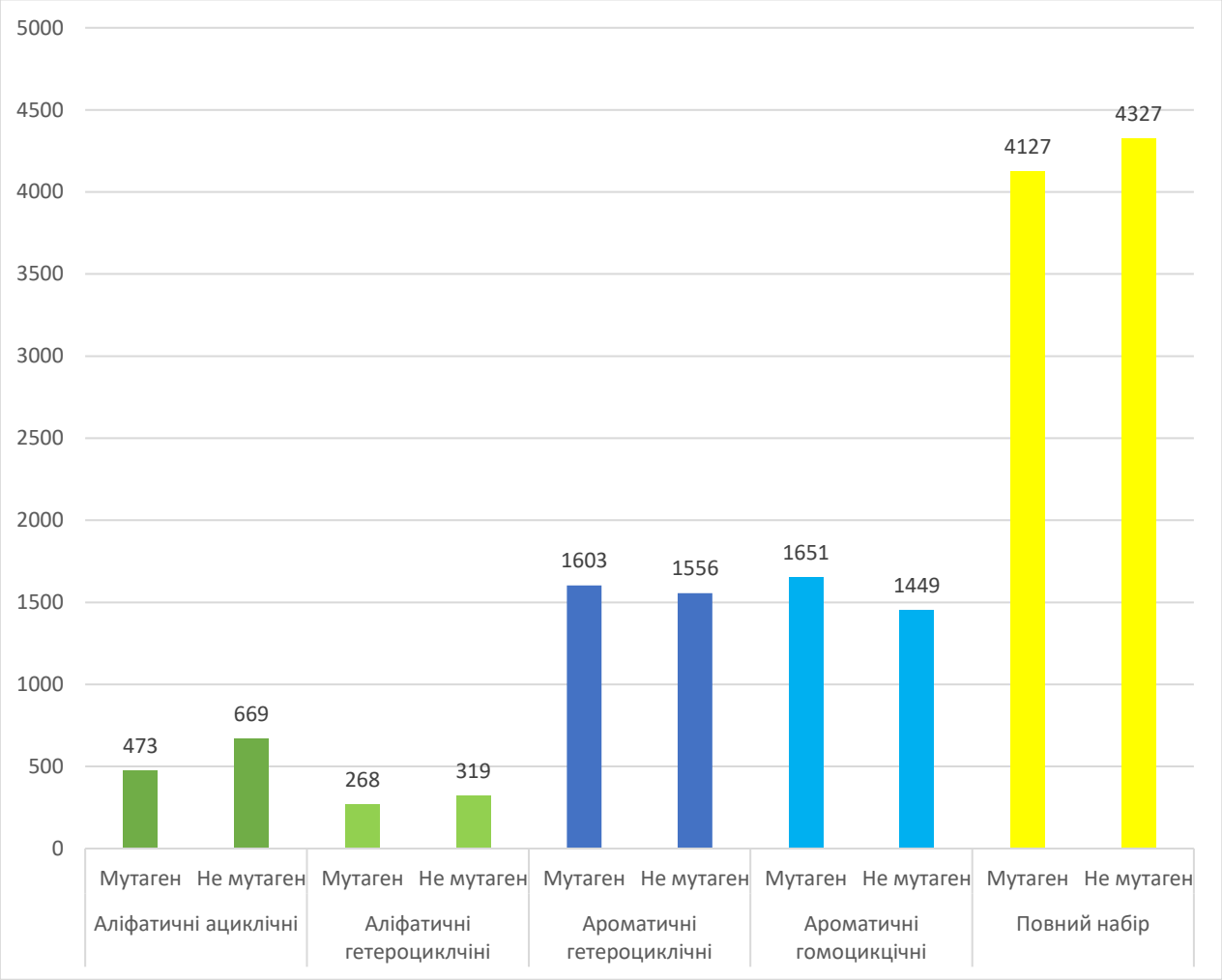


Рисунок 3.3 – Розподіл класів хімічних сполук за мутагенністю

Як бачимо, в межах окремих класів нема суттєвих дисбалансів між мутагенами та не мутагенами, тому застосування SMOTE не призведе до генерації великої кількості синтетичних зразків.

Після балансування дані було поділено на навчальну та тестову вибірки у співвідношенні 80:20. У науковій паці [44] автори рекомендують за наявності достатньої кількості даних розглядати можливість виділення підмножини (наприклад, 10% елементів даних, вибраних випадковим чином) як резервного набору, щоб використовувати його як альтернативний тестовий набір для підтвердження висновків. Для цього з тестової підвибірки випадково було відібрано 10% від загальної кількості елементів, які надалі використовуються як незалежна екзаменаційна вибірка. Важливо, що ці

зразки виключаються з тренувальної та тестової вибірки, тобто жодна з вибірок не перетинається між собою, що гарантує коректність та чистоту експерименту. Тобто загальне співвідношення між вибірками становить 80:10:10.

Екзаменаційна вибірка виконує роль остаточної перевірки моделі після навчання і налаштування, тобто дозволяє оцінити узагальнюючу здатність моделі на нових даних. Її використання особливо важливе у випадках, коли модель проходить додаткові етапи оптимізації.

Таким чином, підхід включає три рівні оцінювання якості моделі:

1. Крос-валідація– для внутрішнього налаштування моделі та оцінки стабільності;
2. Тестування на тестовій вибірці – для проміжної незалежної перевірки;
3. Перевірка на екзаменаційній вибірці – для остаточної оцінки.

3.3. Вибір найбільш інформативних дескрипторів

Відбір дескрипторів – це процес вибору ознак, що мають найтісніші взаємозв'язки з цільовою змінною. Присутність неінформативних ознак призводить до зниження точності багатьох моделей, особливо лінійних.

Відбір ознак забезпечує три основні переваги, а саме: зменшення перенавчання, підвищення точності та скорочення часу навчання, це було підтверджено у бакалаврській роботі[21, 2].

У роботі для відбору ознак застосовано метод рекурсивного виключення ознак (Recursive Feature Elimination, RFE). Алгоритм RFE передбачає ітеративне навчання моделі на повному наборі ознак та оцінку їх значущості; після цього виключаються один або кілька найменш значущих дескрипторів, і процедура повторюється до досягнення заданої кількості найінформативніших ознак. Основним обмеженням цього методу є

необхідність ручного визначення параметра n_features_to_select, який контролює кінцеву кількість обраних ознак.

Для автоматизації цього процесу використовується метод RFECV, який інтегрує RFE з перехресною валідацією. У RFECV кількість вибраних ознак визначається автоматично шляхом оцінки продуктивності моделі для різних конфігурацій відбору ознак на заданих розбиттях перехресної перевірки (параметр cv). Показники ефективності моделі агрегуються по всіх складках, після чого обирається така кількість ознак, яка забезпечує максимальне значення метрики «accuracy ». Завдяки цьому метод RFECV дозволяє не лише оптимізувати перелік дескрипторів, але й підвищити стабільність та точність побудованих моделей [45].

3.4. Створення моделей випадкового лісу

Спочатку здійснювалося навчання моделей класифікації без попереднього відбору ознак, з метою отримання базових результатів для подальшого порівняння з оптимізованими підходами. Для розв’язання задачі класифікації було обрано алгоритм Random Forest Classifier, який є ансамблевим методом, що поєднує велику кількість дерев рішень, підвищуючи точність і стійкість до перенавчання.

Гіперпараметри моделі, зокрема кількість дерев у лісі та максимальна кількість листків у кожному дереві, обиралися емпірично на основі багаторазового тестування моделей, що дозволило адаптувати архітектуру до особливостей конкретної вибірки.

Модель було навчено на тренувальній підвибірці, після чого проведено прогнозування ймовірностей для об'єктів тестової множини. Для перевірки стабільності моделі та її узагальнюючої здатності була застосована п’ятикратна крос-валідація із використанням метрики точності «accuracy ». За результатами кожної ітерації крос-валідації обчислювалися значення точності, а також середнє значення цієї метрики, що дозволило оцінити

варіативність моделі при зміні підвибірок. Початкові результати моделей випадкового лісу та їх параметри наведено в табл. 3.2.

Таблиця 3.2 – Результати моделей випадкового лісу до здійснення скорочення набору дескрипторів [3]

Набір даних	**Клас**	**Кількість дерев**	**Макс кількість листків**	**Середня точність (5-на перехресні перевірка)**	**Площа під ROC кривою**
PaDell	Аліфатичні ациклічні	500	300	83,44%	0,89
	Аліфатичні гетеромоно (полі) циклчіні	100	200	84,51%	0,92
	Ароматичні гетеромоно (полі) циклчіні	650	200	85,61%	0,93
	Ароматичні гомомоно (полі) циклічні	650	200	84,29%	0,9
	Повний набір	300	600	84,62%	0,91
RdKit	Аліфатичні ациклічні	200	200	85,05%	0,95
	Аліфатичні гетеромоно (полі) циклчіні	100	100	81,57%	0,9
RdKit	Ароматичні гетеромоно (полі) циклчіні	400	300	86,04%	0,93
	Ароматичні гомомоно (полі) циклічні	300	400	84,17%	0,91
	Повний набір	350	500	84,62%	0,91
Mordered	Аліфатичні ациклічні	250	250	84,65%	0,86
	Аліфатичні гетеромоно (полі) циклчіні	200	200	83,92%	0,89
	Ароматичні гетеромоно (полі) циклчіні	550	350	85,76%	0,93
	Ароматичні гомомоно (полі) циклічні	400	300	84,32%	0,92
	Повний набір	350	500	84,66%	0,91

Після цього було здійснене зменшення кількості дескрипторів за допомогою методу RFECV. Початкова кількість дескрипторів в наборах даних така:

- PaDell – 1 444 дескриптори;
- RdKit – 196 дескрипторів;
- Mordered – 1 613 дескрипторів.

Результати моделей випадкового лісу після оптимізації наведено в табл. 3.3.

Таблиця 3.3 – Результати моделей випадкового лісу після здійснення скорочення набору дескрипторів [3]

Набір даних	**Клас**	**Кількість дескрипторів**	**Точність (5-на перехресні перевірка)**	**Площа під ROC кривою після**	**Точність моделі на тестовому у наборі**	**F1 міра на тестовому у наборі**
PaDell	Аліфатичні ациклічні	612	84,54%	0,89	81,33%	81,33%
	Аліфатичні гетеромоно (полі) циклчіні	49	86,27%	0,93	83,59%	83,61%
	Ароматичні гетеромоно (полі) циклчіні	649	86,08%	0,93	86,54%	85,33%
	Ароматичні гомомоно (полі) циклічні	274	84,70%	0,9	82,62%	81,57%
	Повний набір	289	84,89%	0,91	83,63%	83,98%
RdKit	Аліфатичні ациклічні	94	85,05%	0,95	88,36%	89,71%
	Аліфатичні гетеромоно (полі) циклчіні	85	81,76%	0,9	81,43%	78,69%
	Ароматичні гетеромоно (полі) циклчіні	139	86,15%	0,93	88,99%	88,82%
	Ароматичні гомомоно (полі) циклічні	87	84,47%	0,91	84,33%	84,51%
	Повний набір	145	84,73%	0,91	84,20%	84,48%

Кінець табл.3.3

Набір даних	Клас	Кількість дескрипторів	Точність (5-на перехресні перевірка)	Площа під ROC кривою після	Точність моделі на тестовому наборі	F1 міра на тестовому наборі
Mordered	Аліфатичні ациклічні	322	84,56%	0,87	79,61%	75,59%
	Аліфатичні гетеромоно (полі) циклчіні	218	84,12%	0,9	81,37%	82,15%
	Ароматичні гетеромоно (полі) циклчіні	675	85,69%	0,93	85,63%	85,71%
	Ароматичні гомомоно (полі) циклічні	416	84,55%	0,93	84,90%	84,37%
	Повний набір	776	84,66%	0,91	82,96%	83,53%

Таким чином скорочення кількості дескрипторів дозволило збільшити точність на крос-влідації в середньому на 0,8% для наборів PaDell, 0,14% на наборах RdKit та 0,05% на наборах Mordered.

Результати оптимізованих моделей випадкового лісу на екзаменаційній вибірці наведено в табл. 3.4.

Таблиця 3.4 – Результати оптимізованих моделей випадкового лісу на екзаменаційній вибірці

Набір даних	Клас	Accuracy	Precision	Recall	Specificity	F1 Score	Площа під ROC кривою тест
PaDell	Аліфатичні ациклічні	81,06%	81,67%	77,78%	84,06%	79,67%	0,88
	Аліфатичні гетеромоно (полі) циклчіні	84,48%	87,10%	84,38%	84,62%	85,71%	0,91
	Ароматичні гетеромоно (полі) циклчіні	87,30%	90,00%	85,71%	89,12%	87,80%	0,93
	Ароматичні гомомоно (полі) циклічні	85,16%	87,58%	83,23%	87,25%	85,35%	0,93
	Повний набір	85,56%	87,00%	84,60%	86,59%	85,78%	0,92

Кінець табл. 3.4.

Набір даних	Клас	Accuracy	Precision	Recall	Specificity	F1 Score	Площа під ROC кривою тест
RdKit	Аліфатичні ациклічні	89,47%	94,12%	88,89%	90,48%	94,43%	0,96
	Аліфатичні гетеромоно (полі) циклчіні	81,03%	85,86%	84,62%	78,12%	80,00%	0,93
	Ароматичні гетеромоно (полі) циклчіні	89,21%	91,30%	88,02%	90,54%	89,63%	0,94
	Ароматичні гомомоно (полі) циклічні	82,62%	81,88%	81,33%	83,13%	81,61%	0,91
	Повний набір	85,68%	86,32%	85,31%	86,06%	85,81%	0,93
Mordered	Аліфатичні ациклічні	80,70%	74,51%	80,85%	80,60%	77,55%	0,85
	Аліфатичні гетеромоно (полі) циклчіні	89,66%	91,30%	84,00%	93,94%	87,50%	0,95
	Ароматичні гетеромоно (полі) циклчіні	86,76%	89,86%	81,58%	91,41%	85,52%	0,94
	Ароматичні гомомоно (полі) циклічні	87,10%	87,41%	85,03%	88,96%	86,21%	0,95
	Повний набір	85,44%	86,65%	84,00%	86,90%	85,30%	0,93

Аналіз показників прогностичної здатності моделей машинного навчання для оцінки генотоксичності хімічних сполук із використанням різних наборів дескрипторів (PaDEL, RDKit та Mordred) свідчить про високу ефективність підходу.

Для PaDEL-дескрипторів точність класифікації коливалася від 81,06% для аліфатичних ациклічних сполук до 87,30% для ароматичних гетеромономоноциклічних сполук. Значення F1-score та площі під ROC-кривою (AUC) також демонструють високий рівень прогнозування, зокрема для ароматичних класів (F1 ≈ 87-88%, AUC ≈ 0,93). Повний набір сполук показав збалансовані показники точності (Accuracy 85,56%) та F1-score

(85,78%), що підтверджує здатність моделей узагальнювати знання на різні класи сполук.

У випадку RDKit-дескрипторів найвищу точність спостерігали для аліфатичних ациклічних (Accuracy 89,47%) та ароматичних гетеромономоноциклічних сполук (Accuracy 89,21%), із високим рівнем F1-score (≈ 89-94%) та AUC (0,94-0,96). Це свідчить про особливу ефективність RDKit-дескрипторів для класифікації деяких специфічних хімічних класів. Загальна продуктивність для повного набору склала Accuracy 85,68%, що практично не відрізняється від PaDEL, демонструючи надійність прогнозів при використанні різних дескрипторів [2].

Для Mordred-дескрипторів точність класифікації також залишалася високою (80,70-89,66%), зокрема для аліфатичних гетеромономоноциклічних сполук (Accuracy 89,66%, AUC 0,95). Показники F1-score та специфічності підтверджують здатність моделей до збалансованого розпізнавання як позитивних, так і негативних прикладів. Повний набір демонструє збалансовану продуктивність (Accuracy 85,44%, F1-score 85,30%, AUC 0,93).

3.5. Створення моделей градієнтного бустингу

Модель будується на основі алгоритму XGBClassifier. Це реалізація градієнтного бустингу над деревами рішень, яка добре працює для задач класифікації з табличними даними. Вказані основні гіперпараметри: кількість дерев, максимальна глибина дерева, швидкість навчання та метрика логарифмічної втрати. Навчання проводиться без використання label_encoder, оскільки класи вже представлені у вигляді чисел.

Модель спочатку тренується на всіх ознаках. Її якість оцінюється за допомогою 5-кратної крос-валідації. Початкові результати моделей градієнтного бустингу та їх параметри наведено в табл. 3.5.

Таблиця 3.5 – Результати моделей градієнтного бустингу до здійснення скорочення набору дескрипторів

Набір даних	Клас	Кількість дерев	Макс глибина	Learning rate	Середня точність (5-на перехресні перевірка)	Площа під ROC кривою
PaDell	Аліфатичні ациклічні	200	5	0,1	84,94%	0,9
	Аліфатичні гетеромоно (полі) циклчіні	200	7	0,1	85,10%	0,9
	Ароматичні гетеромоно (полі) циклчіні	200	6	0,1	85,84%	0,92
	Ароматичні гомомоно (полі) циклічні	200	7	0,1	85,00%	0,93
	Повний набір	200	7	0,1	84,62%	0,93
RdKit	Аліфатичні ациклічні	200	7	0,05	84,30%	0,94
	Аліфатичні гетеромоно (полі) циклчіні	200	5	0,1	81,96%	0,89
	Ароматичні гетеромоно (полі) циклчіні	150	7	0,05	85,57%	0,94
	Ароматичні гомомоно (полі) циклічні	200	7	0,1	84,44%	0,91
RdKit	Повний набір	200	7	0,1	84,11%	0,91
Mordered	Аліфатичні ациклічні	200	5	0,1	84,56%	0,87
	Аліфатичні гетеромоно (полі) циклчіні	250	7	0,1	84,12%	0,87
	Ароматичні гетеромоно (полі) циклчіні	200	5	0,1	85,41%	0,92
	Ароматичні гомомоно (полі) циклічні	200	7	0,1	84,70%	0,92
	Повний набір	200	7	0,1	85,09%	0,92

Методом RFECV відібрано найважливіші дескриптори, після чого модель треновано повторно. Результати наведено в табл. 3.6.

Таблиця 3.6 – Результати моделей градієнтного бустингу після здійснення скорочення набору дескрипторів

Набір даних	Клас	Кількість дескрипторів	Точність (5-на перехресні перевірка)	Площа під ROC кривою після	Точність моделі на тестовому наборі	F1 міра на тестовому наборі
PaDell	Аліфатичні ациклічні	154	86,16%	0,91	79,61%	78,01%
	Аліфатичні гетеромоно (полі) циклчіні	345	85,49%	0,92	82,86%	85,00%
	Ароматичні гетеромоно (полі) циклчіні	409	86,11%	0,92	86,85%	87,01%
	Ароматичні гомомоно (полі) циклічні	385	85,38%	0,93	84,90%	84,08%
	Повний набір	204	85,15%	0,93	85,21%	85,46%
RdKit	Аліфатичні ациклічні	74	84,30%	0,95	87,01%	88,76%
RdKit	Аліфатичні гетеромоно (полі) циклчіні	56	82,35%	0,89	82,86%	82,35%
	Ароматичні гетеромоно (полі) циклчіні	89	85,69%	0,95	87,77%	87,50%
	Ароматичні гомомоно (полі) циклічні	62	84,66%	0,91	84,33%	84,24%
	Повний набір	100	84,66%	0,91	84,20%	84,85%
Mordered	Аліфатичні ациклічні	465	84,84%	0,87	80,26%	80,77%
	Аліфатичні гетеромоно (полі) циклчіні	170	85,69%	0,87	81,43%	80,00%
	Ароматичні гетеромоно (полі) циклчіні	397	85,69%	0,94	87,46%	86,90%
	Ароматичні гомомоно (полі) циклічні	374	84,81%	0,92	86,32%	85,71%
	Повний набір	320	85,24%	0,92	85,10%	85,03%

Таким чином скорочення кількості дескрипторів дозволило збільшити точність на крос-влідації в середньому на 0,56% для наборів PaDell, 0,26% на наборах RdKit та 0,48% на наборах Mordered.

Результати оптимізованих моделей градієнтного бустингу на екзаменаційній вибірці наведено в табл. 3.7.

Таблиця 3.7 – Результати оптимізованих моделей градієнтного бустингу на екзаменаційній вибірці

Набір даних	**Клас**	**Accuracy**	**Precision**	**Recall**	**Specificity**	**F1 Score**	**Площа під ROC кривою тест**
PaDell	Аліфатичні ациклічні	83,33%	80,00%	84,62%	82,26%	82,24%	0,9
	Аліфатичні гетеромоно (полі) циклчіні	84,48%	76,00%	86,36%	83,33%	80,85%	0,88
	Ароматичні гетеромоно (полі) циклчіні	86,35%	86,53%	87,18%	85,53%	86,35%	0,93
	Ароматичні гомомоно (полі) циклічні	85,16%	84,05%	87,26%	83,01%	85,62%	0,9
	Повний набір	85,92%	86,44%	86,24%	85,57%	86,34%	0,93
RdKit	Аліфатичні ациклічні	89,47%	89,06%	91,94%	86,54%	90,48%	0,96
	Аліфатичні гетеромоно (полі) циклчіні	82,76%	79,31%	85,19%	80,65%	82,14%	0,88
	Ароматичні гетеромоно (полі) циклчіні	85,71%	88,00%	83,02%	88,46%	85,44%	0,92
	Ароматичні гомомоно (полі) циклічні	84,84%	82,47%	86,39%	83,44%	84,39%	0,9
	Повний набір	86,51%	87,02%	85,58%	87,23%	86,40%	0,87
Mordered	Аліфатичні ациклічні	78,95%	74,47%	74,47%	82,09%	74,47%	0,87
	Аліфатичні гетеромоно (полі) циклчіні	81,03%	83,33%	80,65%	81,48%	81,97%	0,9
	Ароматичні гетеромоно (полі) циклчіні	86,03%	85,26%	86,36%	85,71%	85,81%	0,92
	Ароматичні гомомоно (полі) циклічні	84,19%	84,91%	84,38%	84,00%	84,64%	0,92
	Повний набір	86,15%	87,16%	86,58%	85,68%	86,87%	0,94

Для PaDEL-дескрипторів точність класифікації (Accuracy) коливалася від 83,33% для аліфатичних ациклічних сполук до 86,35% для ароматичних гетеромоно (полі) циклчіних. Показники F1-score та площі під ROC-кривою (AUC) свідчать про стабільну роботу моделей для всіх класів, зокрема для повного набору сполук F1-score склав 86,34%, а AUC – 0,93, що підтверджує здатність моделей до узагальнення знань на різні класи сполук.

У випадку RDKit-дескрипторів моделі продемонстрували найвищу точність для аліфатичних ациклічних сполук (Accuracy 89,47%) та ароматичних гетеромоно (полі) циклчіних (Accuracy 85,71%), із високими значеннями f1-score (≈ 85-90%) та AUC (0,92-0,96). Для повного набору сполук точність склала 86,51%, F1-score – 86,40%, що вказує на високу ефективність RDKit-дескрипторів для різних хімічних класів.

Для Mordred-дескрипторів найкращі результати спостерігалися для ароматичних гетеромоно (полі) циклчіних сполук (Accuracy 86,03%, AUC 0,92), при цьому точність для аліфатичних класів була дещо нижчою (≈79–82%). Повний набір демонструє збалансовані показники (Accuracy 86,15%, F1-score 86,87%, AUC 0,94), що підтверджує стабільну роботу моделей на різних класах.

3.6. Створення нейронних мереж

Розроблена модель належить до класу багатошарових перцептронів (Multilayer Perceptron, MLP) – це тип штучних нейронних мереж прямого поширення (feedforward neural networks), які складаються з кількох повнозв’язних шарів (dense layers) без зворотних зв’язків. Мережа призначена для бінарної класифікації, тобто для прогнозування однієї категоріальної змінної (генотоксичність: 1 – генотоксична, 0 – негенотоксична). Подібні архітектури ефективно використовуються у хімоінформатиці для прогнозування біоактивності та токсичності сполук за їх молекулярними дескрипторами.

Архітектура моделі складається з вхідного шару, трьох прихованих шарів та вихідного шару. Схематично її можна зобразити так (рис. 3.4):

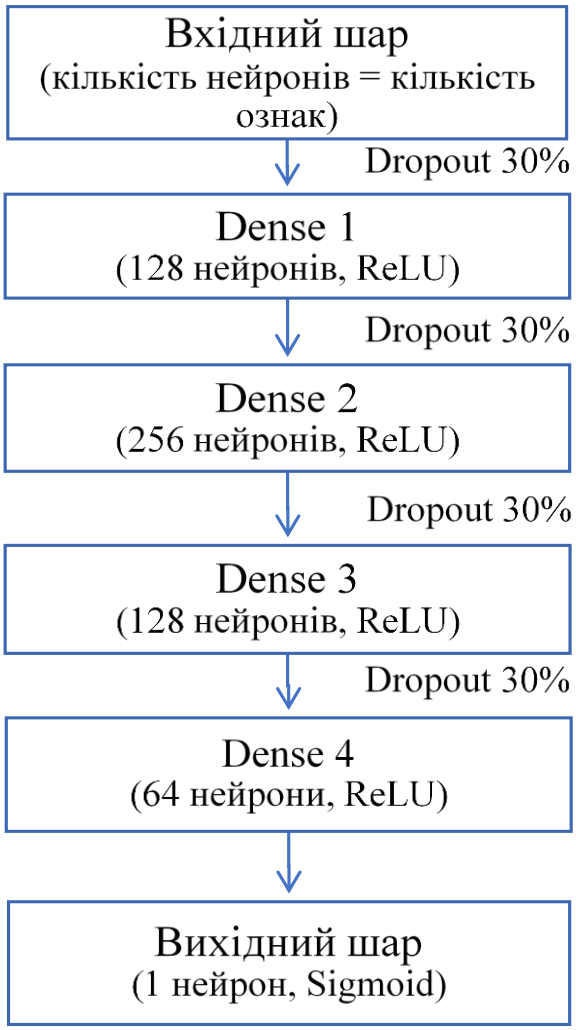


Рисунок 3.4 – Архітектура нейронної мережі

Вхідний шар приймає багатовимірний вектор дескрипторів, який описує фізико-хімічні властивості молекул. Кількість нейронів у цьому шарі відповідає кількості відібраних ознак після попереднього відбору. Кожен вхідний нейрон передає своє значення далі у приховані шари без змін.

Для моделювання нелінійних залежностей між молекулярними ознаками застосовано три приховані шари з кількістю нейронів 128, 256, 128

та 64. Вони використовують функцію активації ReLU, визначену відповідно до формули 3.1.

$$f(x) = \max(0; x) \tag{3.1}$$

де

$f(x)$ – вихідне значення нейрона після застосування функції активації ReLU,

x – вхідне значення нейрона,

Ця функція дозволяє уникнути проблеми зникнення градієнта, що є типовою для сигмоїдних або тангенсових активацій у глибоких мережах.

Однією з головних проблем глибоких нейронних мереж є перенавчання – ситуація, коли модель добре відтворює закономірності навчальної вибірки, але не здатна узагальнювати знання на нові дані. Для боротьби з цим явищем у створеній архітектурі застосовано L2-регуляризацію, яка додає до функції втрат штраф за великі значення вагових коефіцієнтів, що змушує модель утримувати ваги ближчими до нуля, тим самим зменшуючи складність моделі. Математично модифікована функція втрат має вигляд (формула 3.2):

$$L = L_{binary\_crossentropy} + \lambda \sum_{i=1}^{n} w_i^2 \P \tag{3.2}$$

де

L – загальна функція втрат,

$L_{binary\_crossentropy}$ – основна функція втрат для бінарної класифікації,

w – вагові коефіцієнти моделі,

λ – коефіцієнт регуляризації, що визначає силу штрафу (у цій моделі λ=0,001).

Таким чином, регуляризація не дозволяє окремим вагам набувати надмірно великих значень, що зменшує залежність моделі від окремих ознак і підвищує її стійкість до шуму у даних.

Іншим важливим засобом запобігання перенавчанню, використаним у розробленій нейронній мережі, є Dropout-регуляризація. Dropout – це стохастичний метод, який під час кожної ітерації навчання випадково "вимикає" певний відсоток нейронів у шарі разом з усіма їхніми вхідними та вихідними зв’язками. У даній моделі застосовано Dropout з імовірністю 0,3, тобто 30% нейронів випадково деактивуються при кожному проході даних через мережу. Цей процес змінюється на кожній епосі, завдяки чому модель фактично «вчиться» на різних підмножинах нейронів.

Вихідний шар складається з одного нейрона з сигмоїдною функцією активації, яка перетворює значення у діапазон (0; 1) та інтерпретується як ймовірність належності сполуки до класу генотоксичних. Кінцеве рішення приймається за стандартним порогом 0,5: якщо вихід ≥ 0.5 – сполука класифікується як генотоксична, інакше – негенотоксична.

Мережа оптимізується за допомогою бінарної крос-ентропії (binary cross-entropy) – стандартної функції втрат для задач бінарної класифікації, що мінімізує розбіжність між передбаченими ймовірностями та фактичними мітками класів.

Однією з ключових складових процесу навчання нейронної мережі є оптимізатор – алгоритм, який керує оновленням вагових коефіцієнтів з метою мінімізації функції втрат. Від правильного вибору оптимізатора залежить швидкість збіжності, стабільність навчання та якість фінальної моделі.

У розробленій системі було використано оптимізатор Nadam, що поєднує механізми адаптивної зміни швидкості навчання, характерні для Adam, із моментум-прискоренням Нестерова, притаманним класичним градієнтним методам. Adam є одним із найпопулярніших оптимізаторів у глибокому навчанні завдяки поєднанню адаптивної швидкості навчання для кожного параметра та використання моментів градієнта, що прискорює збіжність. Однак Nadam вдосконалює Adam шляхом додавання прискорення за методом Нестерова, що дозволяє враховувати «очікуване» положення ваг

перед оновленням, тим самим забезпечуючи більш точний напрямок руху в просторі параметрів.

Під час експериментів розглядалися також альтернативні алгоритми – Adam, RMSprop та SGD з моментом. Хоча Adam продемонстрував швидку збіжність на початкових епохах, він мав тенденцію до стабілізації у локальних мінімумах, що призводило до незначного зниження точності на тестовій вибірці. RMSprop показав стабільність, але потребував ручного налаштування швидкості навчання та часто демонстрував нижчу швидкість збіжності. Класичний SGD, попри простоту реалізації, виявився менш ефективним через повільне навчання і залежність від чутливих гіперпараметрів.

Таким чином, обраний оптимізатор Nadam забезпечив найкращий компроміс між швидкістю навчання, стабільністю збіжності та здатністю узагальнювати результати, що підтверджується вищими показниками точності та F1-міри на валідаційній вибірці.

Навчання відбувалося протягом 30 епох при розмірі батчу 32.

Моделі спочатку тренуються на всіх ознаках. Її якість оцінюється за допомогою 5-кратної крос-валідації. Проте, на відміну від класичних моделей машинного навчання, таких як логістична регресія, SVM чи дерева рішень, для нейронних мереж метод не використовується. Це пояснюється тим, що RFECV передбачає поетапне видалення ознак і повторне навчання моделі на кожному кроці, що є ресурсоємним у випадку глибоких нейронних мереж, де навчання потребує значних обчислювальних потужностей. Крім того, нейронні мережі самі мають властивість автоматичного виявлення найінформативніших ознак через механізм оптимізації ваг, тому додаткова рекурсивна елімінація не є доцільною.

З огляду на це, для побудови оптимізованої нейронної мережі та забезпечення можливості порівняння з моделями випадкового лісу та градієнтного бустингу було використано скорочений перелік дескрипторів, який попередньо визначався на основі результатів моделей градієнтного

бустингу. Ці моделі продемонстрували високу стабільність і точність у відборі найзначущіших змінних, що мають найбільший вплив на генотоксичність сполук. Отриманий набір ознак слугував вхідними даними для оптимізованої нейронної мережі, що дозволило зменшити розмірність вхідного простору, знизити ризик перенавчання та прискорити процес навчання моделі без втрати якості прогнозування. Результати нейронних мереж до та після скорочення вхідного набору даних наведено в табл. 3.8.

Таблиця 3.8 – Результати нейронних мереж до та після здійснення скорочення набору дескрипторів

Набір даних	**Клас**	**Середня точність (5-на перехресні перевірка)**	**Площа під ROC кривою**	**Кількість дескрипторів у скороченому наборі**	**Середня точність (5-на перехресні перевірка)**	**Площа під ROC кривою**
PaDell	Аліфатичні ациклічні	80,89%	0,93	154	82,39%	0,95
	Аліфатичні гетеромоно (полі) циклчіні	82,16%	0,95	345	82,78%	0,97
	Ароматичні гетеромоно (полі) циклчіні	83,42%	0,92	409	84,63%	0,82
	Ароматичні гомомоно (полі) циклічні	82,09%	0,90	385	82,17%	0,91
	Повний набір	82,51%	0,90	204	83,35%	0,91
RdKit	Аліфатичні ациклічні	85,51%	0,90	74	85,89%	0,91
	Аліфатичні гетеромоно (полі) циклчіні	82,55%	0,91	56	83,14%	0,92
	Ароматичні гетеромоно (полі) циклчіні	83,93%	0,91	89	84,56%	0,93
	Ароматичні гомомоно (полі) циклічні	82,35%	0,89	62	83,04%	0,89
	Повний набір	82,59%	0,91	100	82,85%	0,91

Кінець табл. 3.8

Набір даних	Клас	Середня точність (5-на перехресні перевірка)	Площа під ROC кривою	Кількість дескрипторів у скороченому наборі	Середня точність (5-на перехресні перевірка)	Площа під ROC кривою
Mordered	Аліфатичні ациклічні	82,53%	0,92	465	82,92%	0,93
	Аліфатичні гетеромоно (полі) циклчіні	82,37%	0,94	170	82,96%	0,96
	Ароматичні гетеромоно (полі) циклчіні	83,89%	0,92	397	84,13%	0,92
	Ароматичні гомомоно (полі) циклічні	82,58%	0,91	374	83,15%	0,92
	Повний набір	82,94%	0,90	320	83,06%	0,90

Таким чином скорочення кількості дескрипторів дозволило збільшити точність на крос-влідації в середньому на 0,85% для наборів PaDell, 0,51% на наборах RdKit та 0,38% на наборах Mordered. Результати нейронних мереж після оптимізації на тестовому наборі даних наведено в табл. 3.9.

Таблиця 3.9 – Результати нейронних мереж після здійснення скорочення набору дескрипторів на тестовому наборі даних

Набір даних	Клас	Точність моделі на тестовому наборі	F1 міра на тестовому наборі
PaDell	Аліфатичні ациклічні	88,81%	87,80%
	Аліфатичні гетеромоно (полі) циклчіні	93,55%	93,55%
	Ароматичні гетеромоно (полі) циклчіні	85,40%	84,89%
	Ароматичні гомомоно (полі) циклічні	83,64%	83,13%
	Повний набір	83,35%	82,65%
RdKit	Аліфатичні ациклічні	82,93%	84,56%
	Аліфатичні гетеромоно (полі) циклчіні	86,15%	85,71%
	Ароматичні гетеромоно (полі) циклчіні	87,27%	86,10%

Кінець табл. 3.9

Набір даних	Клас	Точність моделі на тестовому наборі	F1 міра на тестовому наборі
RdKit	Ароматичні гомомоно (полі) циклічні	82,73%	82,67%
	Повний набір	83,83%	83,71%
Mordered	Аліфатичні ациклічні	88,06%	88,41%
	Аліфатичні гетеромоно (полі) циклчіні	92,31%	92,75%
	Ароматичні гетеромоно (полі) циклчіні	86,34%	86,25%
	Ароматичні гомомоно (полі) циклічні	84,89%	85,12%
	Повний набір	84,03%	83,14%

Результати оптимізованих нейронних мереж на екзаменаційній вибірці наведено в табл. 3.10.

Таблиця 3.10 – Результати оптимізованих нейронних мереж на екзаменаційній вибірці

Набір даних	Клас	Accuracy	Precision	Recall	specificity	F1 Score	Площа під ROC кривою тест
PaDell	Аліфатичні ациклічні	87,12%	92,86%	80,00%	94,03%	85,95%	0,93
	Аліфатичні гетеромоно (полі) циклчіні	93,65%	93,33%	93,33%	93,94%	93,33%	0,98
	Ароматичні гетеромоно (полі) циклчіні	86,56%	85,71%	87,34%	85,80%	86,52%	0,93
	Ароматичні гомомоно (полі) циклічні	84,85%	84,48%	86,47%	83,13%	85,47%	0,91
	Повний набір	84,41%	86,67%	82,17%	86,76%	84,36%	0,92

Кінець табл. 3.10

Набір даних	Клас	Accuracy	Precision	Recall	specificity	F1 Score	Площа під ROC кривою тест
RdKit	Аліфатичні ациклічні	84,96%	85,00%	82,26%	87,32%	83,61%	0,93
	Аліфатичні гетеромоно (полі) циклчіні	87,30%	87,50%	87,50%	87,10%	87,50%	0,93
	Ароматичні гетеромоно (полі) циклчіні	86,25%	87,06%	87,06%	85,33%	87,06%	0,92
	Ароматичні гомомоно (полі) циклічні	83,08%	82,07%	86,78%	78,98%	84,36%	0,89
	Повний набір	84,62%	84,24%	83,21%	85,90%	83,72%	0,92
Mordered	Аліфатичні ациклічні	87,12%	90,91%	80,65%	92,86%	85,47%	0,93
	Аліфатичні гетеромоно (полі) циклчіні	87,30%	80,56%	96,67%	78,79%	87,88%	0,93
	Ароматичні гетеромоно (полі) циклчіні	88,75%	89,40%	87,10%	90,30%	88,24%	0,93
	Ароматичні гомомоно (полі) циклічні	86,36%	87,28%	86,78%	85,90%	87,03%	0,92
	Повний набір	84,28%	86,47%	83,04%	85,64%	84,72%	0,91

Для PaDEL-дескрипторів точність класифікації коливалася від 84,85% для ароматичних гомомоно (полі) циклічних сполук до 93,65% для аліфатичних гетеромоно (полі) циклічних. Показники f1-score ≈ 85-93% та площі під ROC-кривою (AUC) на рівні 0,91–0,98 свідчать про стабільну роботу моделей для всіх класів сполук. Для повного набору даних accuracy

склала 84,41%, f1-score – 84,36%, а AUC – 0,92, що підтверджує здатність моделі ефективно узагальнювати знання для різних типів хімічних структур.

У випадку RdKit-дескрипторів моделі продемонстрували найвищу точність для аліфатичних гетеромоно (полі) циклічних сполук (accuracy = 87,30%) та ароматичних гетеромоно (полі) циклічних (accuracy = 86,25%), із значеннями f1-score ≈ 84-87% і AUC (0,89-0,93). Для повного набору сполук точність становила 84,62%, f1-score – 83,72%, а AUC – 0,92, що свідчить про високу ефективність моделей на основі RdKit-дескрипторів, що демонструють високу збалансованість між чутливістю та специфічністю.

Для Mordred-дескрипторів найкращі результати спостерігалися для ароматичних гетеромоно (полі) циклічних сполук (accuracy = 88,75%, AUC = 0,93), тоді як для аліфатичних гетеромоно (полі) циклічних сполук точність становила 87,30% при найвищому показнику recall 96,67%. Повний набір демонструє збалансовані показники (accuracy = 84,28%, f1-score = 84,72%, AUC = 0,91), що підтверджує стабільну роботу моделі та високу узагальнювальну здатність при використанні Mordred-дескрипторів.

3.7. Автоматичні налаштування системи

Зі створених 45 моделей було відібрано 5 найкращих, які будуть використані як автоматичні налаштування системи. Результати найкращих моделей наведено в таблиці 3.11 нижче.

Таблиця 3.11 – Результати найкращих моделей

Клас	Модель	Набір дескрипторів	Accuracy	Precision	Recall	Specificity	F1 Score
Аліфатичні ациклічні	Випадковий ліс	RdKit	89,47%	94,12%	88,89%	90,48%	94,43%
Аліфатичні гетероциклічні	Нейронна мережа	PaDEL	93,65%	93,33%	93,33%	93,94%	93,33%

Кінець табл. 3.11

Клас	Модель	Набір дескрипторів	Accuracy	Precision	Recall	Specificity	F1 Score
Ароматичні гетероциклічні	Випадковий ліс	RdKit	89,21%	91,30%	88,02%	90,54%	89,63%
Ароматичні гомоциклічні	Випадковий ліс	Mordered	87,10%	87,41%	85,03%	88,96%	86,21%
Повний набір	Градієнтний бустинг	Mordered	86,15%	87,16%	86,58%	85,68%	86,87%

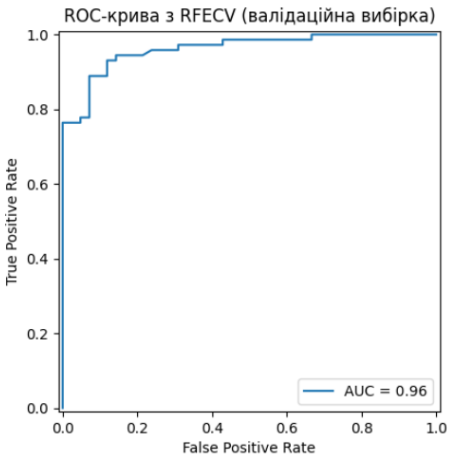
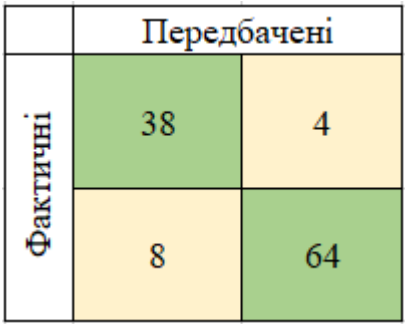
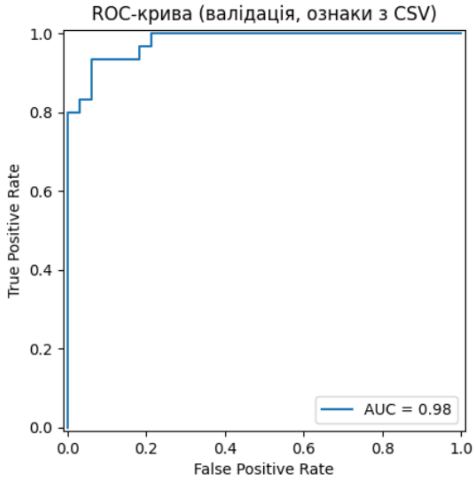
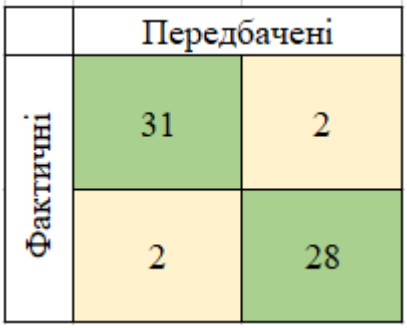
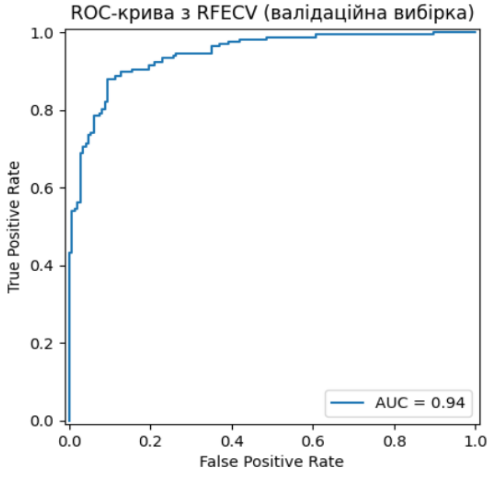
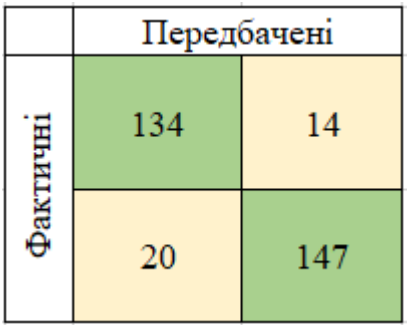
Аналіз отриманих результатів демонструє, що моделі, побудовані з урахуванням хімічних класів, забезпечують вищу точність порівняно з моделлю, натренованою на повному змішаному датасеті. Це підтверджується тим, що всі чотири класифікаційні моделі показали Accuracy у межах 87,10–93,65%, тоді як глобальна модель на повному наборі даних досягла лише 86,15%. Найвищу продуктивність продемонструвала нейронна мережа для аліфатичних гетероциклічних сполук (Accuracy 93,6). Крім того, моделі на окремих класах показали більш збалансовані метрики Precision, Recall та Specificity, що означає кращу здатність коректно відрізняти мутагени від не-мутагенів саме в межах структурно схожих сполук.

Помітно, що модель, отримана на повному датасеті поступається майже всім, моделям, побудованим з врахуванням основних класів хімічних сполук не лише за точністю, а й за узгодженістю ключових показників. Це можна пояснити високою варіативністю структур у великій змішаній вибірці, яка ускладнює виявлення чітких взаємозв’язків. Натомість моделі, створені для окремих класів – аліфатичних ациклічних, аліфатичних гетероциклічних, ароматичних гетероциклічних та ароматичних гомоциклічних – працюють у межах більш однорідних даних, що дозволяє алгоритмам точніше знаходити специфічні закономірності. У підсумку це підтверджує ефективність

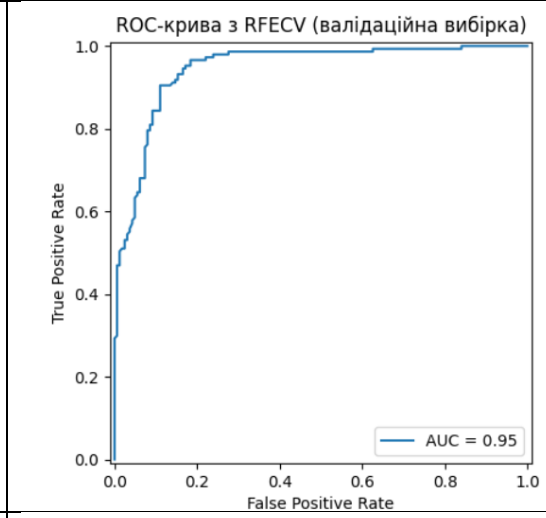
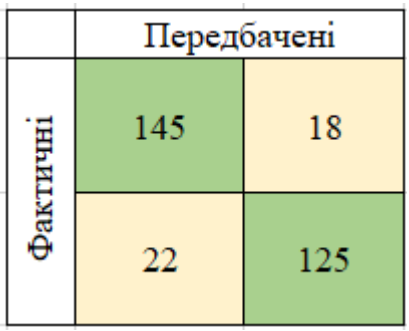
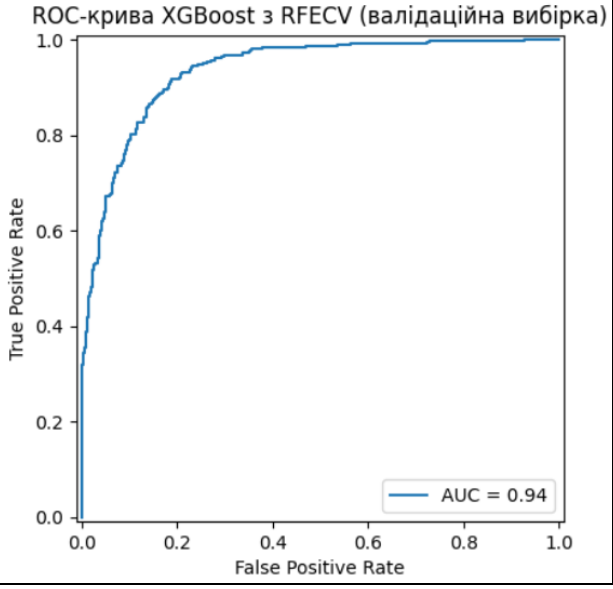
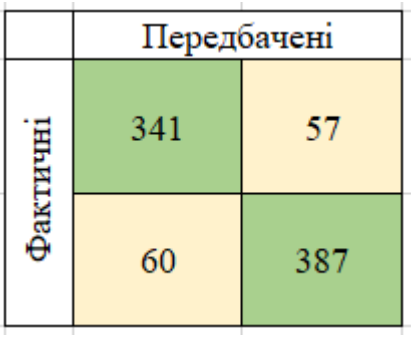
методології, яка пропонує побудову окремих моделей для основних структурних класів хімічних сполук.

У табл. 3.12 наведено ROC-криві та матриці плутанини на екзаменаційній вибірці для відібраних найкращих моделей.

Таблиця 3.12 – ROC-криві та матриці плутанини для найкращих моделей

Клас	ROC-крива	Матриця плутанини
Аліфатичні ациклічні		
Аліфатичні гетероциклчіні		
Ароматичні гетероциклчіні		

Кінець табл. 3.12

Клас	ROC-крива	Матриця плутанини
Ароматичні гомоциклічні		
Повний набір		

Висновок до розділу 3

У межах розділу було здійснено реалізацію моделей для системи для оцінки мутагенності хімічних сполук. Основна увага була зосереджена на побудові навчальної бази даних, розробці, оптимізації, навчанні та тестуванні моделей.

Для навчання моделей сформовано репрезентативну базу даних. Після очищення та уніфікації даних отримано вибірку з понад восьми тисяч хімічних сполук, які були класифіковані за допомогою сервісу ClassyFire відповідно до їхніх структурних особливостей. Це дозволило не лише

провести моделювання для узагальненого набору, а й оцінити ефективність алгоритмів у межах окремих класів сполук, що підвищило точність і гнучкість прогнозування.

На основі підготовлених даних побудовано та протестовано моделі машинного навчання, серед яких випадковий ліс, градієнтний бустинг та нейронна мережа. За результатами проведеного тестування було встановлено, що ефективність моделей машинного навчання залежить від структурного класу хімічних сполук та типу використаних молекулярних дескрипторів. Для кожної групи сполук визначено оптимальну комбінацію алгоритму та дескрипторного набору, яка забезпечила найвищі показники точності класифікації. Зокрема:

- для аліфатичних ациклічних сполук найкращі результати продемонструвала модель випадкового лісу, побудована на основі дескрипторів RDKit, із точністю прогнозування 89,47%;
- для аліфатичних гетероциклічних сполук найвищу ефективність показала нейронна мережа, навчена на дескрипторах PaDEL, із точністю 93,65%;
- для ароматичних гетероциклічних сполук найкращою виявилася модель випадкового лісу з дескрипторами RDKit, яка досягла точності 89,21%;
- для ароматичних гомоциклічних сполук оптимальною стала модель випадкового лісу, побудована з використанням дескрипторів Mordred, що забезпечила точність 87,10%;
- для повного набору даних, який охоплює всі структурні класи сполук, найкращий результат показала модель градієнтного бустингу з дескрипторами Mordred, що досягла точності 86,15%.

РОЗДІЛ 4
ПРОГРАМНА РЕАЛІЗАЦІЯ ТА МЕТОДИКА РОБОТИ СИСТЕМИ

У цьому розділі представлено програмну реалізацію системи аналізу мутагенності хімічних сполук та методику її роботи. Основна увага приділяється опису побудови функціональної структури програмного забезпечення, алгоритмів обробки даних та інтеграції моделей машинного навчання для прогнозування мутагенності. Також розглянуто розробку інтерфейсу користувача та базові аспекти захисту інформації, що забезпечують зручність використання системи та безпечне зберігання даних. У розділі буде детально показано, як програмна реалізація дозволяє автоматизувати процес оцінки мутагенності хімічних сполук, від завантаження вхідних даних до отримання результатів прогнозування.

4.1. Розробка інтерфейсу користувача

Розроблення користувацького інтерфейсу для системи оцінки мутагенності здійснювалось із використанням фреймворку Streamlit, який є сучасним інструментом для створення інтерактивних веб-застосунків на основі Python. Такий підхід дозволив забезпечити швидку інтеграцію алгоритмів машинного навчання із зручним графічним інтерфейсом без необхідності використання складних вебтехнологій (HTML, CSS, JavaScript). Streamlit забезпечує автоматичне оновлення інтерфейсу при зміні коду, що сприяє швидкому прототипуванню та спрощує розгортання системи.

Інтерфейс побудовано у вигляді веб-додатку, який запускається локально або може бути розгорнутий на сервері. Основна структура містить головне вікно для вибору файлів та відображення результатів і бічну панель

(sidebar) для вибору параметрів. Зображення головної сторінки представлено на рис. 4.1.

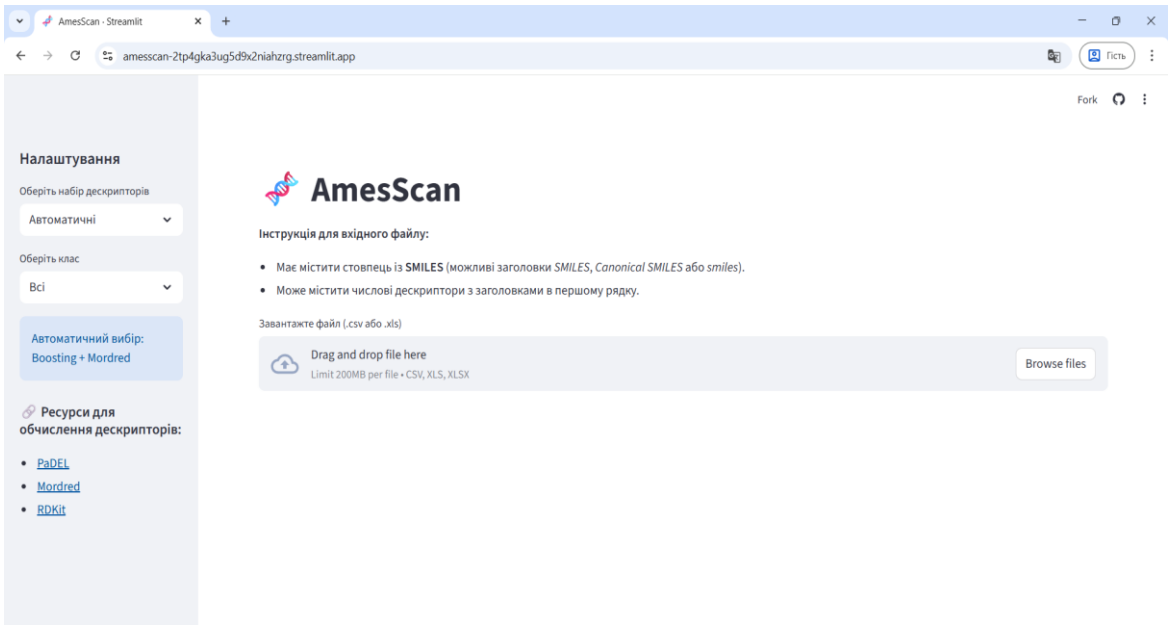


Рисунок 4.1 – Головна сторінка веб-серверу

У бічній панелі користувач може обрати тип дескрипторів (PaDEL, RDKit, Mordred або автоматичний вибір), модель машинного навчання (Random Forest, Boosting, Neural Network), а також клас хімічних сполук (аліфатичні ациклічні, аліфатичні гетероциклічні тощо). Така структура забезпечує гнучкість і можливість швидкої зміни конфігурацій моделі.

Система передбачає інтелектуальний механізм автоматичного вибору оптимальної моделі для кожного класу сполук на основі попередньо визначених правил. Наприклад, для ароматичних гетероциклічних сполук автоматично застосовується модель Random Forest із RDKit-дескрипторами, а для аліфатичних гетероциклічних – нейронна мережа з PaDEL-дескрипторами. Такий підхід мінімізує потребу користувача у спеціальних знаннях у галузі хімоінформатики та підвищує точність прогнозування. При виборі автоматичних налаштувань користувач бачить, яка модель та набір дескрипторів використовується (див. рис. 4.2), це дозволяє користувачу коректно завантажити вхідний файл.

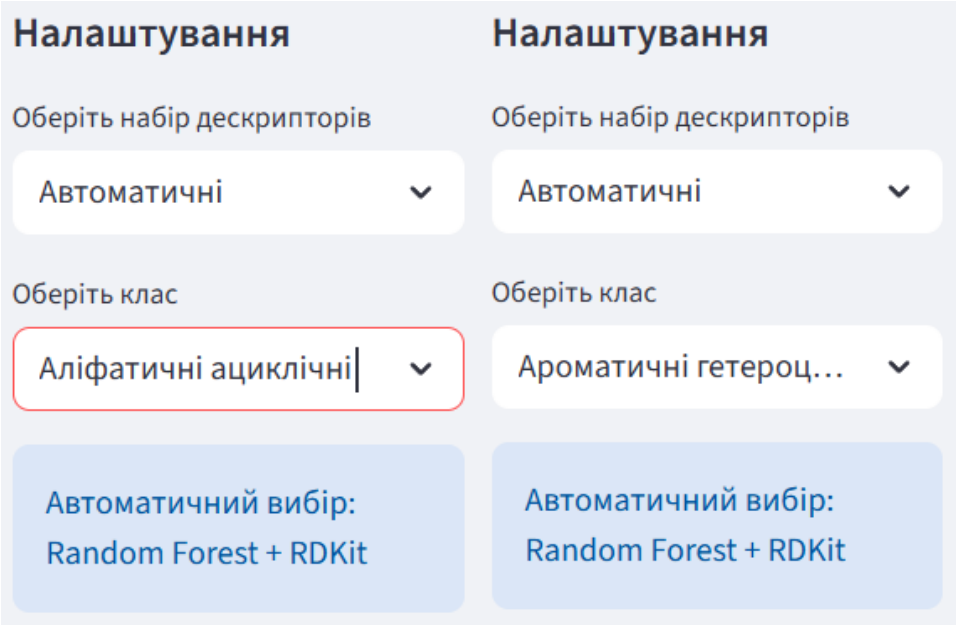


Рисунок 4.2 – Відображення автоматичних налаштувань

Інтерфейс дозволяє користувачу завантажити файл з SMILES-нотаціями та дескрипторами (формати .csv, .xls, .xlsx). Фіксування форматів файлів запобігає вибору некоректного типу файлу. Для зручності користувача також додано посилання на онлайн-сервіси Galaxy, які дозволяють обчислювати дескриптори за допомогою PaDEL, Mordred або RDKit, які використовує система.

Система також підтримує автоматичне обчислення дескрипторів RDKit для молекул. Якщо користувач завантажує файл, який містить лише SMILES-нотації, і при цьому обрані автоматичні налаштування, де набором дескрипторів є RDKit, або якщо користувач самостійно обирає RDKit як набір дескрипторів, у інтерфейсі з’являється спеціальна кнопка для автоматичного розрахунку цих дескрипторів. Після натискання кнопки програма обробляє кожну молекулу, визначає хімічну структуру та обчислює всі доступні RDKit-дескриптори, додаючи їх до початкового набору даних. У разі виявлення некоректних SMILES-нотацій користувач отримує попередження про невдалі розрахунки для окремих молекул, що дозволяє вчасно виправити помилки або скористатися альтернативними сервісами для обчислення дескрипторів. Після завершення обчислень дані з дескрипторами

готові для запуску прогнозування, що робить процес підготовки даних для моделі більш швидким і автоматизованим, зменшуючи потребу у ручному обчисленні та підвищуючи зручність використання системи. (див. рис. 4.3)

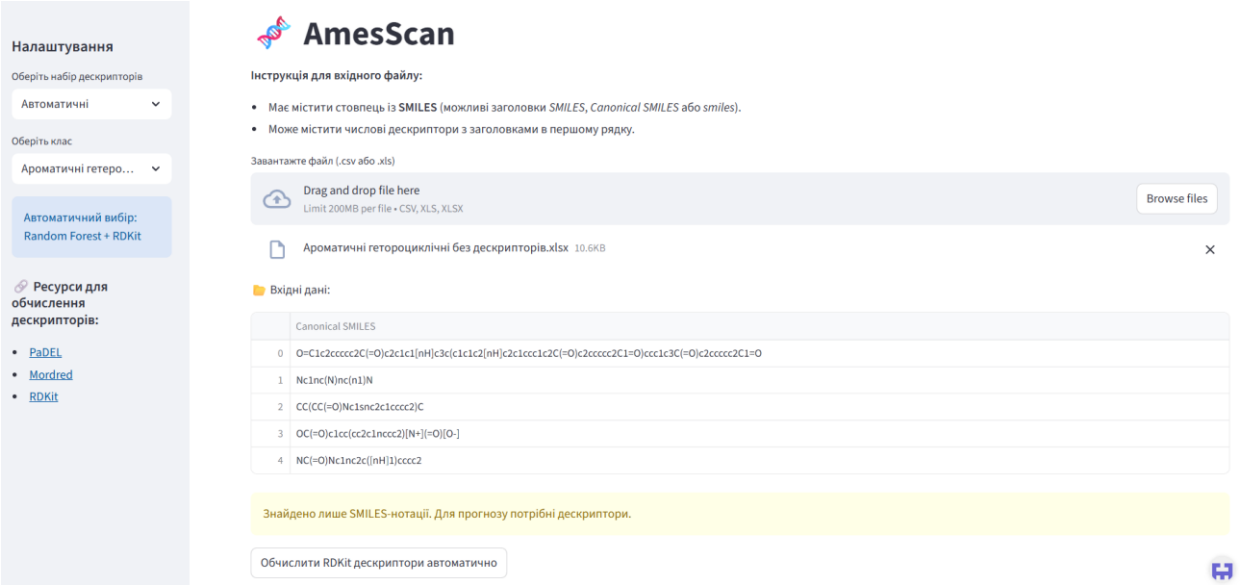


Рисунок 4.3 – Автоматичний розрахунок дескрипторів RDKit

Якщо дескриптори не RDKit кнопка не з’являється (див. рис. 4.4).

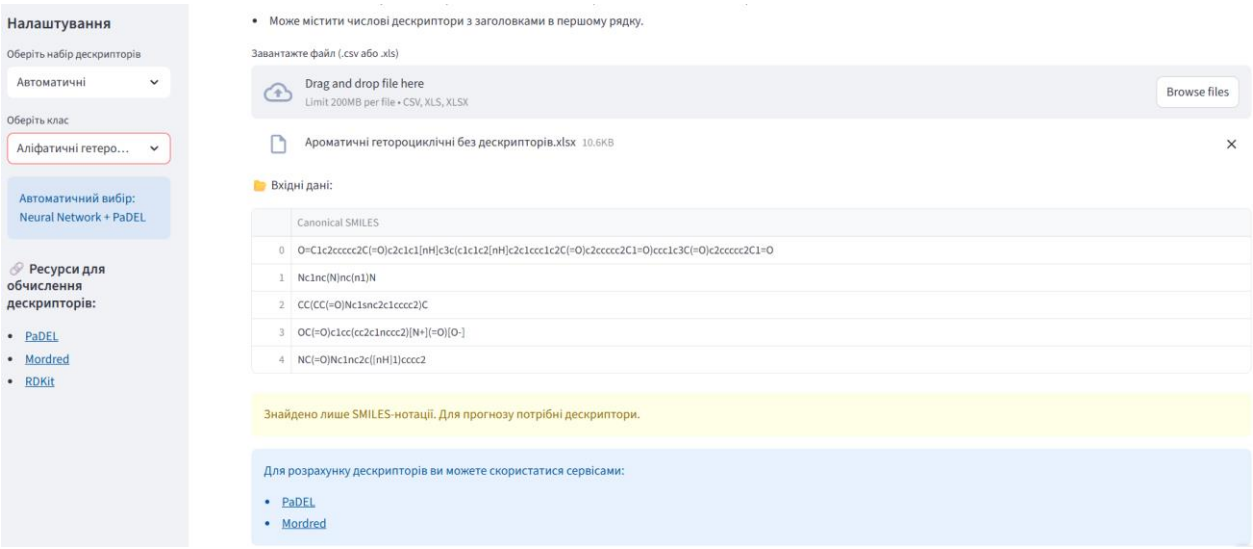
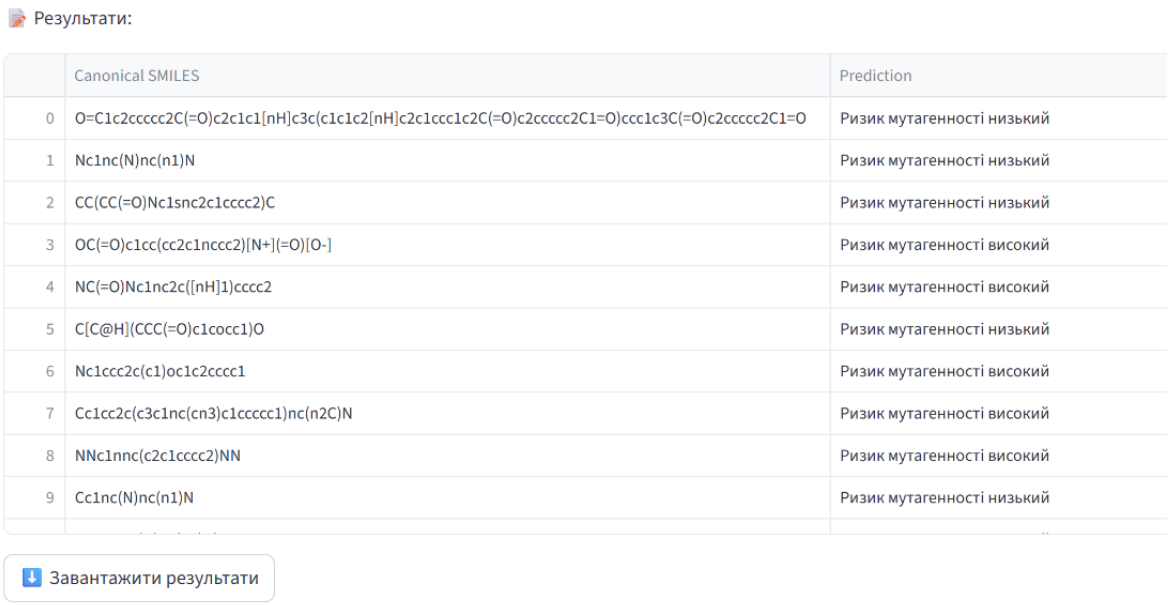


Рисунок 4.4 – Посилання на сервери для обчислення дескрипторів

Результати прогнозу подаються у вигляді таблиці, яка містить SMILES-нотації молекул та відповідний результат оцінки генотоксичності. (див. рис. 4.5). Оскільки модель не гарантує стовідсоткової точності, результати класифікуються у два рівні ризику: «Ризик мутагенності високий» та «Ризик мутагенності низький».


Результати:

	Canonical SMILES	Prediction
0	O=C1c2ccccc2C(=O)c2c1c1[nH]c3c(c1c1c2[nH]c2c1ccc1c2C(=O)c2ccccc2C1=O)ccc1c3C(=O)c2ccccc2C1=O	Ризик мутагенності низький
1	Nc1nc(N)nc(n1)N	Ризик мутагенності низький
2	CC(CC(=O)Nc1snc2c1cccc2)C	Ризик мутагенності низький
3	OC(=O)c1cc(cc2c1nccc2)[N+](=O)[O-]	Ризик мутагенності високий
4	NC(=O)Nc1nc2c([nH]1)cccc2	Ризик мутагенності високий
5	C[C@H](CCC(=O)c1cocc1)O	Ризик мутагенності низький
6	Nc1ccc2c(c1)oc1c2cccc1	Ризик мутагенності високий
7	Cc1cc2c(c3c1nc(cn3)c1ccccc1)nc(n2C)N	Ризик мутагенності високий
8	NNc1nnc(c2c1cccc2)NN	Ризик мутагенності високий
9	Cc1nc(N)nc(n1)N	Ризик мутагенності низький

Завантажити результати

Рисунок 4.5 – Результати прогнозу

У системі реалізовано обробники помилок, які перевіряють коректність вхідного файлу та його вміст: визначається, чи файл не порожній, чи містить необхідні стовпці з дескрипторами для вибраної моделі (див. рис. 4.6), а у разі їх відсутності користувачу виводяться зрозумілі повідомлення з пропозицією скористатися сервісами для обчислення дескрипторів. Крім того, додано обробку помилок при завантаженні файлу та під час виконання прогнозу, що дозволяє уникнути аварійного завершення роботи програми та інформує користувача про проблеми з даними або моделлю.

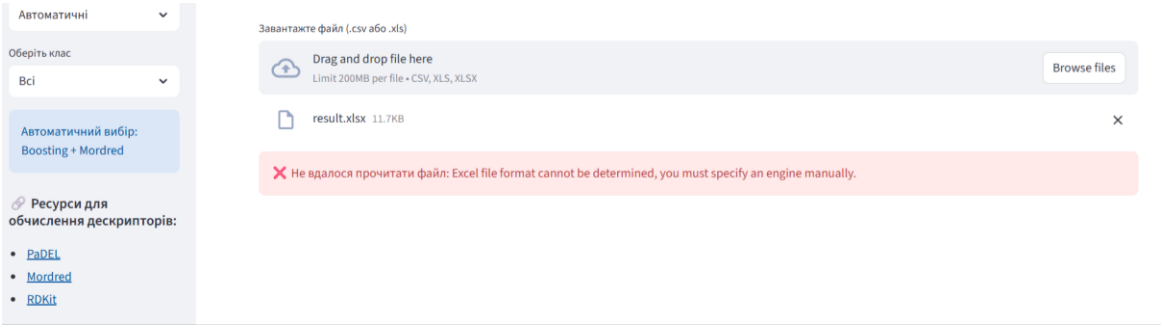


Рисунок 4.6 – Приклад обробки помилок

4.2. Захист інформації

У процесі розробки системи оцінки мутагенності хімічних сполук питання захисту інформації було враховано на кількох рівнях, хоча система не передбачає складних механізмів шифрування чи управління доступом до даних. Основна увага була зосереджена на забезпеченні конфіденційності та безпеки використання вхідних та вихідних даних, а також на захисті розроблених моделей машинного навчання.

По-перше, в системі реалізовано обмеження доступу до внутрішніх компонентів моделі. Користувач взаємодіє зі системою через веб-інтерфейс, який дозволяє завантажувати файли з молекулярними даними, обирати класи хімічних сполук та дескриптори, а також запускати прогноз. Проте користувач не має можливості безпосередньо переглядати внутрішню структуру моделей, їх параметри або навчальні дані. Це забезпечує конфіденційність алгоритмів, оскільки архітектура нейронних мереж, налаштування Random Forest або Boosting моделей залишаються прихованими. Подібний підхід ускладнює несанкціоноване копіювання або модифікацію моделей, а також забезпечує базовий рівень захисту інтелектуальної власності, оскільки навчальні ваги та структури моделей зберігаються на сервері або в локальному сховищі, недоступному користувачу.

По-друге, обробка даних користувача здійснюється без збереження персональних або ідентифікаційних даних, що мінімізує ризики витоку особистої інформації. Система приймає на вхід файли з SMILES-нотаціями та дескрипторами молекул, обробляє їх у пам’яті програми та повертає результати прогнозу у вигляді таблиці. Вихідні файли формуються лише після активної дії користувача (натискання кнопки завантаження), тому дані не зберігаються постійно на сервері або у програмних каталогах за замовчуванням. Такий підхід дозволяє контролювати потік інформації та знижує ризик випадкової або зловмисної втрати даних.

У перспективі можливе розширення захисних механізмів системи. Наприклад, можна інтегрувати автентифікацію користувачів, шифрування файлів моделей, обмеження доступу до сервісів обчислення дескрипторів або логування дій користувача для аудиту безпеки. Такі доповнення дозволять підвищити надійність системи в умовах використання у комерційних або промислових дослідженнях, де вимоги до безпеки та збереження інтелектуальної власності є критичними.

Висновок до розділу 4

Під час практичної частини роботи було розроблено програмне забезпечення для аналізу мутагенності хімічних сполук, що реалізує алгоритми машинного навчання та забезпечує автоматизацію процесу обробки та оцінки даних. Було створено функціональний інтерфейс користувача на основі фреймворку Streamlit, що дозволяє запускати систему як веб-додаток. Інтерфейс забезпечує можливість завантаження файлів із дескрипторами, вибору методу оцінки, отримання прогнозів і відображення результатів у табличній формі. Передбачено можливість збереження результатів у файл, що робить систему зручною для практичного застосування в лабораторних, дослідницьких і промислових умовах. Інтерфейс є інтуїтивно зрозумілим, адаптивним до різних розмірів вікна та

візуально привабливим завдяки використанню графічних елементів і кольорового оформлення.

Також розглянуто аспекти захисту інформації. Доступ користувачів до моделей машинного навчання обмежено, що запобігає несанкціонованому копіюванню або зміні їх архітектури. Система не зберігає персональних даних, а обробка здійснюється лише на рівні хімічних дескрипторів і прогнозних результатів. Це гарантує відповідність вимогам конфіденційності та безпеки при роботі з дослідницькими чи корпоративними даними.

Розроблена система підтверджує ефективність обраного підходу та слугує основою для подальшого розвитку, зокрема для інтеграції нових моделей і розширення функціоналу.

РОЗДІЛ 5
УЗАГАЛЬНЕННЯ РЕЗУЛЬТАТІВ РОБОТИ

У цьому розділі здійснюється узагальнення результатів виконаної магістерської дисертації, зокрема проведено детальний аналіз та порівняння отриманих результатів з існуючими аналогами та бакалаврською роботою.

5.1 Порівняння системи з результатами бакалаврської роботи

В межах бакалаврської роботи [21] була розроблена локальна система оцінки мутагенності хімічних сполук, яка базувалася на PaDEL-дескрипторах та використовувала моделі лінійної регресії, випадкового лісу та нейронної мережі. Навчання здійснювалося на повному наборі даних, що містив близько 8 тисяч сполук, а відбір важливих дескрипторів проводився вручну на основі коефіцієнтів регресії та методів Random Forest (mean decrease impurity та permutation feature importance). Основним обмеженням цього підходу було ручне визначення кількості ознак, що залишаються, що могло впливати на точність та стабільність моделей [1].

У магістерській дисертації запропонована система значно вдосконалена: вона реалізована як веб-сервер, який враховує класи хімічних сполук, що дозволяє підвищити точність прогнозування та адаптувати модель під різні групи сполук. Використання методу RFECV для автоматичного відбору дескрипторів дозволило зменшити кількість ознак і уникнути ручних налаштувань, що робить процес моделювання більш ефективним і надійним.

Порівняльний аналіз показує, що магістерська система забезпечує більш точне прогнозування завдяки врахуванню структурних класів сполук. У бакалаврській роботі моделі навчалися на повному наборі даних без поділу на класи, що обмежувало можливості адаптації алгоритмів до різних типів

сполук і призводило до усередненої точності 80–86%. У магістерській системі спеціалізація моделей під конкретні класи хімічних сполук дозволила досягти вищих результатів: для аліфатичних ациклічних сполук точність випадкового лісу на дескрипторах RDKit склала 89,47%, для аліфатичних гетероциклічних – 93,65% за нейронною мережею на PaDEL-дескрипторах, а для інших класів сполук також спостерігалося підвищення точності завдяки спеціалізованим моделям та оптимальному підбору дескрипторів.

Крім того, для кожного класу сполук було визначено оптимальний перелік дескрипторів, що забезпечує високу точність моделей та дозволяє проводити детальний аналіз причин мутагенності конкретних сполук. Такий підхід дає змогу не лише прогнозувати мутагенні властивості, але й досліджувати, які структурні характеристики та молекулярні ознаки впливають на прояв генотоксичних ефектів, що підвищує наукову цінність системи та сприяє прийняттю обґрунтованих рішень у хіміко-біологічних дослідженнях.

Отже, порівняльний аналіз демонструє, що запропонована система не лише забезпечує вищу точність прогнозування, але й пропонує більш гнучкий та автоматизований підхід до обробки даних та прогнозування мутагенності хімічних сполук. Вона перевершує локальну бакалаврську систему за функціональністю, адаптивністю та ефективністю моделей, що підтверджує практичну цінність розробки.

5.2 Порівняння системи з аналогами

Для оцінки ефективності розробленої системи доцільно провести порівняння з існуючими аналогами у сфері прогнозування токсичності та мутагенності хімічних сполук. Було виділено два найбільш релевантні сервіси: ProTox 3.0 та веб-сервер vNN-ADMET, їх опис наведено в розділі 1.3 вище.

Розроблена в межах магістерської дисертації система відрізняється від існуючих аналогів ключовими особливостями, які підвищують її ефективність у прогнозуванні мутагенності хімічних сполук. Основною перевагою є врахування структурних класів сполук, що дозволяє спеціалізувати моделі для різних груп та досягати вищої точності прогнозування. Крім того, застосування методу RFECV для автоматичного відбору дескрипторів забезпечує оптимальний набір ознак для кожного класу, підвищує стабільність моделей та дозволяє аналізувати структурні причини мутагенності конкретних сполук.

Система демонструє високі показники точності: для аліфатичних гетероциклічних сполук точність складає 93,65%, для аліфатичних ациклічних – 89,47%, для інших класів сполук також спостерігається покращення результатів у порівнянні з існуючими платформами. Основним недоліком є те, що система прогнозує лише мутагенність, тоді як деякі аналоги охоплюють ширший спектр кінцевих точок токсичності.

Нижче наведено порівняльну таблицю ключових характеристик розробленої системи та аналогів (див. табл. 5.1).

Таблиця 5.1 – Порівняльна таблиця розробленої системи та аналогів

Характеристика	**Розроблена система «AmesScan»**	**ProTox 3.0**	**vNN-ADMET**
Врахування класів сполук	Так	Ні	Ні
Точність прогнозу мутагенності	86–93% (залежно від класу)	84%	54–85%
Розмір датасету	8 454	6 156	6 512
Доступність	Онлайн	Онлайн	Онлайн
Прогнозовані властивості	Лише мутагенність	Різні токсикологічні кінцеві точки	Різні токсикологічні кінцеві точки
Аналіз причин мутагенності	Обмежений	Обмежений	Обмежений

Висновок до розділу 5

У розділі було виконано порівняння розробленої системи з існуючими аналогами, що дозволило оцінити її конкурентоспроможність та визначити ключові переваги й обмеження. Аналіз показав, що врахування структурних класів хімічних сполук та застосування оптимального відбору дескрипторів забезпечують значне підвищення точності прогнозування мутагенності порівняно з поширеними платформами ProTox 3.0 та vNN-ADMET. Розроблена система демонструє більш стабільні результати для різних груп хімічних сполук, що робить її ефективним інструментом для наукових досліджень та попереднього скринінгу речовин.

Разом із тим встановлено, що основним недоліком системи є орієнтація виключно на прогнозування мутагенності, тоді як аналоги охоплюють ширший спектр токсикологічних властивостей. Це визначає потенційний напрям подальшого розвитку – розширення функціональності системи шляхом включення інших кінцевих точок токсичності. Загалом проведене порівняння підтверджує, що запропонований підхід є перспективним і має суттєві переваги в контексті мутагенності, адаптивності та практичної цінності

РОЗДІЛ 6
СТАРТАП-ПРОЄКТ ЗА ТЕМОЮ МАГІСТЕРСЬКОГО ДОСЛІДЖЕННЯ

У розділі розглянуто стартап-проєкт, заснований на розробленій у межах магістерської дисертації системі «AmesScan», що призначена для аналізу мутагенності хімічних сполук із застосуванням методів машинного навчання. Актуальність такого проєкту зумовлена зростаючою потребою ринку у швидких, надійних і економічно ефективних інструментах токсикологічного скринінгу, здатних частково замінити традиційні лабораторні дослідження.

6.1. Назва проєкту

Система аналізу мутагенності хімічних сполук на основі методів машинного навчання «AmesScan»

6.2. Короткий опис проєкту

«AmesScan» – це інноваційна система для оцінки мутагенності хімічних сполук, яка поєднує сучасні досягнення хімоінформатики та машинного навчання. Її мета – забезпечити швидкий, точний та економічно ефективний спосіб передбачення потенційного мутагенного ризику речовин, що є критично важливим у фармацевтиці, хімічній промисловості, аграрному секторі та екології.

Основою системи є набір спеціалізованих моделей машинного навчання, що аналізують структурні та класові характеристики молекул. Вони враховують як загальні закономірності, так і специфічні властивості

різних хімічних класів, що дозволяє досягати високої точності прогнозування навіть для нових або слабо вивчених сполук.

Система буде реалізована як зручний веб-сервіс із можливістю інтеграції в робочі процеси науково-дослідних лабораторій та промислових підприємств.

6.1. Бізнес-модель

6.1.1. Цінність продукту

Цінність продукту «AmesScan» полягає у поєднанні наукової точності, високої ефективності та практичної користі для широкого кола користувачів. Система дозволяє значно скоротити потребу у дорогих та тривалих лабораторних дослідженнях, надаючи швидку оцінку мутагенного потенціалу хімічних сполук на основі їхньої молекулярної структури.

Однією з ключових переваг є можливість виявлення потенційно небезпечних речовин на ранніх етапах розробки, що зменшує ризики фінансових втрат та правових ускладнень для компаній. Крім того, точність прогнозів підвищується завдяки врахуванню класових особливостей хімічних сполук, що робить «AmesScan» особливо цінним для фармацевтичних і хімічних компаній, які працюють з різними групами речовин.

Таким чином, основними цінностями стартап-проєкту є:

1. Суттєве скорочення часу на проведення первинного аналізу без потреби в лабораторних експериментах.
2. Зниження фінансових витрат на токсикологічні дослідження, особливо на ранніх етапах розробки речовин.
3. Висока точність прогнозів, завдяки врахуванню класових особливостей хімічних сполук, що забезпечує надійність результатів.
4. Рання ідентифікація потенційно небезпечних речовин запобігає їх потраплянню у виробничі ланцюги.
5. Зменшення потреби в тестуванні на тваринах завдяки використанню ін-силіко методів.

6. Можливість використання як окремого веб-сервісу або як частини корпоративної IT-інфраструктури.

Унікальність стартапу «AmesScan» полягає в інтеграції інтелектуальних моделей машинного навчання з диференційованим підходом до хімічних класів сполук, що дозволяє автоматично обирати оптимальні дескриптори для кожної групи речовин та забезпечувати високоточну, адаптивну і науково обґрунтовану оцінку їх мутагенності.

6.1.2. Сегмент споживачів

Професійний ринок наукових досліджень і розробок, тому що науковці, фармацевтичні компанії, токсикологи та спеціалісти з хімічної безпеки постійно шукають ефективні способи оцінки безпечності нових речовин із мінімальними затратами часу та ресурсів, і «AmesScan» дозволяє це реалізувати завдяки автоматизованому аналізу мутагенності.

Що болить? – Невизначеність і високі ризики у процесі розробки нових хімічних сполук, складність і висока вартість експериментальних досліджень, необхідність дотримання жорстких регуляторних норм без втрати темпів інновацій.

Що об’єднує? – Бажання мати надійний, швидкий і науково обґрунтований інструмент, який дозволяє прогнозувати токсичність ще до лабораторного тестування, оптимізувати дослідницький процес і зменшити залежність від тестів на тваринах.

Як вони звикли купувати?

- оформити підписку на онлайн-доступ до системи «AmesScan» (SaaS-модель);
- інтегрувати модуль у корпоративну хімоінформаційну систему (B2B-рішення);
- замовити консалтингову оцінку мутагенності як послугу (outsourcing-модель).

6.1.3. Канали збуту

Головний канал збуту – B2B-продажі через прямі контакти з підприємствами. Основним джерелом доходу для системи буде прямий продаж індустріальним клієнтам, зацікавленим у швидкій та достовірній оцінці мутагенності. Це можуть бути:

- Фармацевтичні компанії
- Хімічні виробники
- Аграрно-хімічні корпорації
- Контрактні дослідницькі організації (CROs)
- Екологічні агентства та регуляторні органи

Продаж здійснюється через особисті зустрічі, участь у конференціях, демонстрації MVP-продукту, вебінари тощо. Це дозволяє налагодити довгострокові стосунки, надати індивідуальні умови ліцензування й інтеграції.

Іешими каналами збуту є:

SaaS-платформа - веб-сервіс із підписною моделлю, що дозволяє клієнтам з науково-дослідних установ, університетів та малих лабораторій самостійно користуватися системою «AmesScan» онлайн з можливістю інтеграції через API.

Співпраця з компаніями, що розробляють лабораторні інформаційні системи або надають токсикологічні консалтинг послуги, що дозволить вбудувати «AmesScan» у ширші екосистеми досліджень і регуляторного аналізу.

Наукові гранти та тендери - залучення фінансування через державні або міжнародні наукові гранти й участь у тендерах на інноваційні рішення може стати стабільним джерелом доходу та розширення впливу на ринку.

6.1.4. Взаємодія з споживачами

На початкових етапах розвитку стартапу залучення користувачів буде зосереджене на академічному та науково-дослідному середовищі, зокрема –

через участь у наукових конференціях, студентських хакатонах, спеціалізованих семінарах і презентаціях в університетах. Важливу роль відіграватимуть особисті контакти, тематичні спільноти в соцмережах, а також просування проєкту через наукові публікації або короткі огляди на популярних платформах для молодих дослідників.

У подальшому планується залучення клієнтів через галузеві конференції, наукові форуми, публікації у профільних журналах, а також через цільову цифрову рекламу та SEO-оптимізований сайт. Особливу увагу буде приділено демонстраціям продукту (онлайн/офлайн), безкоштовним пробним періодам та кейсам, що показують ефективність моделі для різних галузей.

Підтримка буде реалізована у зручному та доступному форматі: через email або месенджери, з оперативною відповіддю напряму від розробників. Планується створення короткого гіда або навчального відео, що пояснює, як користуватися системою. Користувачі також зможуть надсилати фідбек прямо через інтерфейс системи, що допоможе у швидкому вдосконаленні продукту.

6.1.5. Дохід

Стартап «AmesScan» передбачає отримання доходу за допомогою моделі підписки (SaaS), коли користувачі – здебільшого наукові установи, дослідницькі лабораторії або малі інноваційні підприємства – сплачують щомісячну або річну плату за доступ до онлайн-сервісу. Також передбачається можливість разових платежів за окремі оцінки або розширені функції системи, наприклад, генерацію детального звіту чи інтеграцію через API. У перспективі можуть з’явитися індивідуальні ліцензії для великих компаній із персональними умовами користування та підтримки.

6.1.6. Ключові види діяльності

- Розробка моделей машинного навчання для точного прогнозування мутагенності хімічних сполук.

- Збір, обробка та оновлення хімічних і токсикологічних даних для покращення аналітики.
- Створення та технічна підтримка веб-сервісу, через який користувачі отримують доступ до системи.
- Забезпечення зв’язку з користувачами, надання консультацій та оперативне реагування на зворотний зв’язок.
- Просування проєкту через наукові заходи, соцмережі та встановлення партнерських контактів.

6.1.7. Ключові ресурси

Матеріальні ресурси:

- Сервери– для обробки даних, навчання моделей і роботи веб-сервісу.
- База даних – для збереження інформації про хімічні сполуки та результати аналізу.
- Комп’ютери та обладнання – для роботи команди.
- Веб-сайт і інтерфейс – через який користувачі взаємодіють із системою.
- Інтелектуальні ресурси:
- Моделі машинного навчання – що передбачають мутагенність речовин.
- Алгоритми для аналізу молекул – що враховують їхню структуру та клас.
- Власна база знань – зібрана під час розробки та тестування системи.
- Бренд і концепція «AmesScan» – унікальна ідея, назва, логотип, інтерфейс.

Людські ресурси:

- Фахівець з машинного навчання – створюють і навчають моделі.
- Хімік (токсиколог) – допомагають правильно інтерпретувати результати.
- Програміст – розробляють веб-сервіс та інтегрують моделі.

- Керівник проєкту – координує команду.
- Фахівець із просування та продажів – шукає клієнтів і партнерів.

Фінансові ресурси:

- Початкові інвестиції або гранти – кошти для старту проєкту: купівля обладнання, оплата хостингу, початкова розробка.
- Витрати на розробку – зарплати команді (програмісти, дата-сайентисти, хіміки), оплата ліцензій, хмарних сервісів.
- Маркетинг і просування – реклама, участь у виставках, створення презентацій, вебінарів, просування в інтернеті.
- Юридичні та адміністративні витрати – реєстрація компанії, захист інтелектуальної власності, ліцензії, договори.
- Операційні витрати – підтримка веб-сайту, сервіси для клієнтів, оновлення моделей, резервне копіювання даних.
- Майбутні джерела доходу – підписка на сервіс, оплата за кожен аналіз, корпоративні ліцензії, інтеграції з лабораторіями.

6.1.8. Ключові партнери

Для успішного запуску та розвитку «AmesScan» важливими партнерами є наукові установи, хімічні та фармацевтичні компанії, а також постачальники хімоінформатичних ресурсів. Наприклад, стратегічним партнером може стати Європейське хімічне агентство (ECHA), яке надає доступ до великих баз токсикологічних даних через REACH. Також важливими є платформи як PubChem (від Національного інституту здоров'я США) і ChEMBL (від Європейського біоінформатичного інституту), які забезпечують відкриті дані для навчання моделей.

Іншими потенційними партнерами можуть бути фармацевтичні компанії, наприклад, Arterium або Farmak в Україні, які можуть використовувати систему для попередньої оцінки нових сполук. Також доцільно співпрацювати з університетами (наприклад, Київський

національний університет імені Тараса Шевченка, Львівський національний медичний університет) для проведення спільних досліджень, тестування моделей та підготовки кадрів.

6.1.9. Витрати

В табл. 6.1 наведено основні статті витрат на основні продукти в межах стартап-проєкту.

Таблиця 6.1 – Витрати на основні продукти в межах стартап-проєкту

Стаття витрат / доходу	Щомісячна підписка (грн)	Річна підписка (грн)
Повна собівартість продукту	800	8 000
Прибуток стартапу	200	2 000
Оптова ціна виробника	1 000	10 000
Акцизні збори	0	0
ПДВ (20%)	200	2 000
Оптова відпускна ціна	1 200	12 000
Посередницька надбавка	300	3 000
– з них: Витрати посередника	100	1 000
– з них: Прибуток посередника	200	2 000
ПДВ посередника (20%)	60	600
Оптова ціна закупки торгівельною організацією	1 560	15 600
Торгівельна надбавка	300	3 000
– з них: Витрати торгівельної організації	150	1 500
– з них: Прибуток торгівельної організації	150	1 500
ПДВ торгівельної організації (20%)	60	600
Роздрібна ціна реалізації (для кінцевого клієнта)	1 920 грн / міс	18 000 грн / рік

Знижка на річну підписку по відношенню до щомісячної:

1 920 × 12 = 23 040 → економія клієнта 5 040 грн на рік (~22%)

Це дозволяє стимулювати довгострокові підписки та формує стабільний грошовий потік для стартапу.

6.1.10. Споживчі властивості товару

Система «AmesScan» для покупця має головну споживчу властивість:

У результаті використання системи «AmesScan» за призначенням, споживач (науковець, дослідник або підприємство) отримує можливість

швидко та достовірно оцінити мутагенний ризик хімічних сполук без необхідності в дорогих і тривалих лабораторних тестах.

Використовуючи сучасні моделі машинного навчання, система аналізує структурні та класові характеристики молекул, що дозволяє:

- скоротити до 90% час на попередню токсикологічну оцінку;
- зменшити витрати на первинні дослідження до 70%;
- підвищити точність прогнозу ризику навіть для нових або рідковивчених речовин.

Це робить «AmesScan» незамінним інструментом для фармацевтичних компаній, агропідприємств, хімічних виробництв і екологічних лабораторій, які прагнуть відповідати сучасним вимогам сталого хімічного дизайну, регуляторних норм та економії ресурсів.

6.1.11. Дослідження ринку

На світовому ринку рішень для оцінки мутагенності спостерігається стабільне зростання через посилення регуляторних вимог, зростання екологічної обізнаності та попиту на безпечні фармакологічні й хімічні продукти. Автоматизовані системи оцінки токсичності з використанням *in silico, in vitro* та AI-технологій є одним з найперспективніших напрямів розвитку. У цьому контексті «AmesScan» є проривним рішенням, адже забезпечує комплексну автоматизовану оцінку мутагенності із урахуванням класових характеристик хімічних сполук – чого не пропонує жодна з існуючих систем. Таким чином, на момент запуску прямих повноцінних аналогів не існує, що робить «AmesScan» ноу-хау у сфері біоінформатики та токсикологічного скринінгу.

На ринку рішень для оцінки мутагенності присутня низка гравців, які частково перекривають функціональність стартапу «AmesScan», проте не забезпечують його рівня автоматизації та врахування класових характеристик хімічних сполук. Список можливих конкурентів наведено в табл. 6.2 нижче.

Таблиця 6.2 – Список можливих конкурентів

№ з/п	Назва підприємства конкурента	Короткий опис діяльності	Короткий опис послуги/продукту	Схожість зі стартапом
1	WuXi AppTec	Провідна CRO-компанія, що надає послуги з оцінки безпеки хімічних сполук	Генетична токсикологія, включаючи *in vitro* та *in vivo* тести, (Q)SAR моделювання	Прямий конкурент: надає комплексні послуги з оцінки мутагенності
2	Medicilon	Дослідницька компанія з Китаю, що спеціалізується на доклінічних дослідженнях	Генотоксикологічні тести, включаючи швидкий скринінг та цитотоксичність	Прямий конкурент: пропонує подібні послуги з оцінки мутагенності
3	Molecular Biology Resources (MBR)	Постачальник молекулярно-біологічних продуктів та послуг	Інгібіторні тести на ДНК-полімерази для оцінки мутагенності	Опосередкований конкурент: пропонує специфічні тести, але не повний спектр послуг
5	Alvascience (AlvaDesc)	Розробник програмного забезпечення для хемоінформатики	AlvaDesc – інструмент для розрахунку молекулярних дескрипторів та QSAR-моделювання	Опосередкований конкурент: фокусується на хемоінформатиці, але може бути використаний для оцінки мутагенності
6	Cytel	Постачальник статистичного програмного забезпечення для клінічних досліджень	East Horizon – платформа для моделювання клінічних досліджень	Опосередкований конкурент: фокусується на статистичному аналізі, але може бути використаний для оцінки безпеки

Основними прямими конкурентами виступають такі компанії, як WuXi AppTec та Medicilon. Вони пропонують широкий спектр послуг – від *in vitro* та *in vivo* тестів до програмного забезпечення для управління мутагенними дослідженнями. Проте, їхні рішення вимагають значної участі користувача, а також не реалізують концепцію автоматизованої класифікації хімічних речовин за класами.

Опосередкованими конкурентами є Molecular Biology Resources, Alvascience (AlvaDesc) та Cytel. Вони спеціалізуються на вузьких нішах – молекулярній біології, хемоінформатиці або статистичному аналізі – і можуть лише частково бути використані для оцінки мутагенності.

Таким чином, на момент аналізу ринку прямих аналогів «AmesScan» з повним спектром можливостей не існує, що дає проєкту унікальну позицію для виходу на глобальний ринок.

6.1.12. Маркетингова стратегія просування

«AmesScan» – новий революційний продукт на існуючому ринку тестування токсичності. Стратегія просування базується на таких етапах:

1. Науково-експертне позиціонування, за допомогою публікацій у міжнародних фахових журналах (наприклад, Toxicological Sciences, Mutation Research).
2. Запуск пілотних проєктів з науковими установами та фармкомпаніями.
3. Цільова реклама через LinkedIn, ResearchGate, профільні платформи для біотех-інновацій.

6.1.13. Елементи фінансового плану

Опис бізнес-проєкту

«AmesScan» – це наукоємний стартап, спрямований на розробку автоматизованої системи для оцінки мутагенності хімічних речовин з урахуванням їхньої структурно-класової належності. Проєкт має на меті впровадження програмного продукту, який дозволяє пришвидшити процес оцінки безпеки речовин, знизити потребу у дорогих лабораторних тестах, а також забезпечити відповідність регуляторним вимогам фармацевтичної та хімічної індустрії.

Опис продукту

Програмний комплекс «AmesScan» поєднує алгоритми хемоінформатики, машинного навчання та структурно-аналітичної класифікації, що дозволяє проводити *in silico* оцінку мутагенності речовин.

Продукт стане в пригоді фармацевтичним компаніям, хімічним лабораторіям, університетам та регуляторним органам. Перевагою є швидкість, доступність, автоматизованість процесу, а також локалізація системи для українських користувачів.

Маркетинг та продаж

В Україні сегмент інструментів для оцінки токсичності є малорозвиненим, але потреба в подібному продукті зростає – як у рамках євроінтеграційних процесів, так і через розширення ринку біотехнологій. Основними каналами просування будуть:

- Співпраця з університетами та НДІ
- Презентації на наукових форумах, конференціях
- Інтернет-маркетинг через професійні платформи (LinkedIn, ResearchGate)
- Інтеграція через партнерські CRO-компанії
- Пілотне тестування в лабораторіях

Фінансовий план

Фінансовий план на 12 місяців наведено в таблиці 6.3.

Таблиця 6.3 – Список можливих конкурентів

Стаття витрат	Орієнтовна сума (грн)
Зарплата для 6 осіб команди (в середньому 30 000 грн/міс)	2 160 000 грн
Закупівля обладнання (сервер, ноутбуки, периферія)	150 000 грн
Ліцензії ПЗ, хмарні сервіси, доступ до баз даних	100 000 грн
Розробка MVP (прототипу) та тестування	150 000 грн
Резерв на непередбачувані витрати	100 000 грн
Всього:	2 660 000,00

6.1.14. Резюме

Коли ми говоримо про безпеку хімічних речовин – косметики, ліків, харчових добавок чи засобів побутової хімії – найперше, що спадає на думку, це відповідність базовим нормативам безпеки. Але майже ніхто не

замислюється: а що, якщо речовина не викликає миттєвого отруєння, але повільно і системно змінює ДНК наших клітин?

Мова йде про мутагенність – приховану загрозу, яка не завжди виявляється класичними лабораторними методами. Більше того, більшість існуючих систем тестування покладаються на застарілі *in vitro* чи *in vivo* дослідження, які не враховують найголовнішого – структурного класу сполуки, її схожості з речовинами, що вже продемонстрували мутагенний ефект.

Система аналізу мутагенності хімічних сполук на основі методів машинного навчання «AmesScan» – це перша в Україні автоматизована система для оцінки мутагенності, яка враховує структурно-класову належність хімічних сполук. Ми пропонуємо інтелектуальний програмний продукт, що не просто аналізує потенційну токсичність, а класифікує сполуки за їх біохімічними профілями, надаючи точні прогнози, які можуть замінити частину лабораторних тестів.

Сьогодні, коли швидкість розробки нових ліків, агрохімікатів та матеріалів критично важлива, «AmesScan» стає незамінним інструментом. Це – рішення, що дозволяє компаніям:

- Економити мільйони на лабораторних випробуваннях,
- Прискорювати вихід нових продуктів на ринок,
- Забезпечувати відповідність сучасним стандартам регуляторних органів (OECD, EMA, FDA),
- Та головне – захищати здоров’я мільйонів людей, запобігаючи використанню потенційно небезпечних сполук.

Висновок до розділу 6

У цьому розділі було сформовано концепцію стартап-проєкту, що базується на розробленій системі «AmesScan». Проведений аналіз ринку, визначення цінності продукту, споживачів, каналів збуту, ключових ресурсів

та витрат підтвердили перспективність впровадження технології у промислові та наукові процеси. Система демонструє конкурентні переваги завдяки високій точності прогнозування, врахуванню структурних класів сполук, доступності у вигляді веб-сервісу та можливості інтеграції в корпоративні IT-рішення. «AmesScan» має потенціал стати комерційно успішним продуктом, що водночас забезпечує реальну користь для галузей, де рання оцінка мутагенного ризику є критично важливою.

ЗАГАЛЬНІ ВИСНОВКИ

У магістерській дисертації проведено комплексне дослідження проблеми прогнозування мутагенності хімічних сполук та створено програмну систему «AmesScan», що реалізує сучасні методи машинного навчання для *in silico* оцінювання мутагенного потенціалу хімічних сполук.

Було проаналізовано традиційні та сучасні методи оцінки мутагенності, включаючи експериментальні тести, високо-продуктивні методи та *in silico* підходи. Показано, що алгоритми ML/DL значно підвищують ефективність первинної оцінки мутагенного потенціалу хімічних речовин, скорочують потребу в дороговартісних і тривалих експериментах та підвищують точність прогнозування.

Сформовано датасет із 8 454 сполук, який базується на кількох відкритих базах токсикологічних даних. Для моделювання використано молекулярні дескриптори RDKit, Mordred та PaDEL, що забезпечило різноманітність структурних ознак які використовувалися при моделюванні.

Розроблено та оптимізовано 45 моделей машинного навчання для основних структурних класів сполук. Використано ансамблеві методи випадкового лісу та градієнтного бустингу, а також нейронну мережу. Завдяки застосуванню RFECV сформовано оптимальні набори дескрипторів, що забезпечили високу узагальнюючу здатність моделей.

Досягнуто високих показників точності прогнозування, порівняно з наявними аналогами та класичними методами оцінки мутагенності, зокрема:

- 93,65% – для аліфатичних гетероциклічних сполук,
- 89,47% – для аліфатичних ациклічних,
- 89,21% – для ароматичних гетероциклічних,
- 87,10% – для ароматичних гомоциклічних,
- 86,15% – на повному наборі даних.

Реалізовано програмну систему «AmesScan», яка забезпечує автоматичне обчислення дескрипторів RdKit, бінарну класифікацію хімічних сполук з використанням різних моделей ML для прогнозування мутагенних ефектів, зручний графічний інтерфейс користувача, модуль збереження результатів та базові механізми захисту інформації.

Проведено порівняння розробленої системи з існуючими аналогами (ProTox 3.0, vNN-ADMET). Показано, що «AmesScan» перевершує їх за точністю прогнозування та забезпечує можливість прогнозування мутагенності з урахуванням структурних класів хімічних сполук.

Розроблено стартап-концепцію, яка демонструє комерційний потенціал системи. Представлено бізнес-модель, сегменти споживачів, ринковий аналіз, маркетингову стратегію та елементи фінансового планування. Показано, що система є перспективною для фармацевтичної, хімічної, біотехнологічної та екологічної галузей.

Результати роботи проходили апробацію на 1 науковій конференції:

Єсипенко Р. В., Кисляк С. В., *In silico* моделі оцінки генотоксичності факторів навколишнього середовища. *Матеріали IX Міжнародної науково-практичної конференції «CURRENT CHALLENGES OF SCIENCE AND EDUCATION»*. 2024 р. Берлін, Німеччина – С. 50-53 (тези). ISBN 978-3-954753-05-5. URL: https://sci-conf.com.ua/ix-mizhnarodna-naukovo-praktichna-konferentsiya-current-challenges-of-science-and-education-6-8-05-2024-berlin-nimechchina-arhiv/ (дата звернення 01.11.2025) [1].

За результатами виконаної роботи були опублікована 1 наукова стаття в журналі категорії А та 1 наукова стаття в фаховому журналі категорії Б:

Kislyak, S., Dugan, O., Yesypenko, R., Starosyla, D. and Yalovenko, O. 2025. *In silico* the Ames Mutagenicity Predictive Model of Environment. *Innovative Biosystems and Bioengineering*. 9, 2 (May 2025), 42–52. DOI: https://doi.org/10.20535/ibb.2025.9.2.316239 (категорія А) (дата звернення 01.11.2025) [2].

Кисляк С., Дуган О.., Єсипенко Р., Яловенко О. 2025. *In silico* моделі прогнозування мутагенності Еймса основних структурних класів ксенобіотиків на основі методу випадкового лісу. *Біомедична інженерія і технологія.* 2025. Т. 4, №.19. с. 48-62. URL: https://doi.org/10.20535/.2025.19.340327 (категорія Б) (дата звернення 20.11.2025) [3].

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1 Єсипенко Р. В., Кисляк С. В., *In silico* моделі оцінки генотоксичності факторів навколишнього середовища. *Матеріали IX Міжнародної науково-практичної конференції «CURRENT CHALLENGES OF SCIENCE AND EDUCATION»*. 2024 р. Берлін, Німеччина – С. 50-53 (тези). ISBN 978-3-954753-05-5. URL: https://sci-conf.com.ua/ix-mizhnarodna-naukovo-praktichna-konferentsiya-current-challenges-of-science-and-education-6-8-05-2024-berlin-nimechchina-arhiv/ (дата звернення 01.11.2025).

2 Kislyak, S., Dugan, O., Yesypenko, R., Starosyla, D. and Yalovenko, O. 2025. *In silico* the Ames Mutagenicity Predictive Model of Environment. *Innovative Biosystems and Bioengineering*. 9, 2 (May 2025), 42–52. DOI: https://doi.org/10.20535/ibb.2025.9.2.316239 (категорія А) (дата звернення 01.09.2025)

3 Кисляк С., Дуган О., Єсипенко Р., Яловенко О. 2025. *In silico* моделі прогнозування мутагенності Еймса основних структурних класів ксенобіотиків на основі методу випадкового лісу. Біомедична інженерія і технологія. 2025. Т. 4, №.19. с. 48-62. URL: https://doi.org/10.20535/.2025.19.340327 (дата звернення 20.11.2025).

4 Unravelling the molecular mechanism of mutagenic factors impacting human health / K. Goyal та ін. *Environmental science and pollution research*. 2021. URL: https://doi.org/10.1007/s11356-021-15442-9 (дата звернення 01.09.2025).

5 Alnasser S. M. Revisiting the approaches to DNA damage detection in genetic toxicology: insights and regulatory implications. *BioData mining*. 2025. Т. 18, № 1. URL: https://doi.org/10.1186/s13040-025-00447-8 (дата звернення 01.09.2025).

6 Benigni R. Structure−Activity relationship studies of chemical mutagens and

carcinogens: mechanistic investigations and prediction approaches. *Chemical reviews*. 2005. Т. 105, № 5. С. 1767–1800. URL: https://doi.org/10.1021/cr030049y (дата звернення: 01.09.2025).

7 The various aspects of genetic and epigenetic toxicology: testing methods and clinical applications / N. Ren та ін. *Journal of translational medicine*. 2017. Т. 15, № 1. URL: https://doi.org/10.1186/s12967-017-1218-4 (дата звернення 01.09.2025).

8 Fenech M. The *in vitro* micronucleus technique. *Mutation research/fundamental and molecular mechanisms of mutagenesis*. 2000. Т. 455, № 1-2. С. 81–95. URL: https://doi.org/10.1016/s0027-5107(00)00065-8 (дата звернення: 01.09.2025).

9 Sukumaran S. K., Ranganatha R., Chakravarthy S. High-throughput approaches for genotoxicity testing in drug development: recent advances. International journal of high throughput screening. 2016. С. 1. URL: https://doi.org/10.2147/ijhts.s70362 (дата звернення: 16.09.2025).

10 Dahui Q. Next-generation sequencing and its clinical application. *Cancer biology & medicine*. 2019. Т. 16, № 1. С. 4–10. URL: https://doi.org/10.20892/j.issn.2095-3941.2018.0055 (дата звернення: 16.11.2025).

11 Quality control of quantitative high throughput screening data / K. R. Shockley та ін. *Frontiers in genetics*. 2019. Т. 10. URL: https://doi.org/10.3389/fgene.2019.00387 (дата звернення: 16.09.2025).

12 Jia X., Wang T., Zhu H. Advancing computational toxicology by interpretable machine learning. *Environmental science & technology*. 2023. URL: https://doi.org/10.1021/acs.est.3c00653 (дата звернення: 16.09.2025).

13 An evaluation of existing QSAR models and structural alerts and development of new ensemble models for genotoxicity using a newly compiled experimental dataset / P. Pradeep та ін. *Computational Toxicology*. 2021. Т. 18. С. 100167. URL:

https://doi.org/10.1016/j.comtox.2021.100167 (дата звернення 03.09.2025).

14 QSARtuna: an automated QSAR modeling platform for molecular property prediction in drug design / L. Mervin та ін. *Journal of chemical information and modeling*. 2024. URL: https://doi.org/10.1021/acs.jcim.4c00457 (дата звернення: 15.09.2025).

15 Abbod M., Safaie N., Gholivand K. Genetic algorithm multiple linear regression and machine learning-driven QSTR modeling for the acute toxicity of sterol biosynthesis inhibitor fungicides. *Heliyon*. 2024. Т. 10, № 16. С. e36373. URL: https://doi.org/10.1016/j.heliyon.2024.e36373 (дата звернення: 15.09.2025).

16 Suprapto S. Comparative analysis of preprocessing methods for molecular descriptors in predicting anti-cathepsin activity. *South african journal of chemical engineering.* 2023. URL: https://doi.org/10.1016/j.sajce.2023.11.001 (дата звернення: 15.09.2025).

17 Cavasotto C. N., Scardino V. Machine learning toxicity prediction: latest advances by toxicity end point. *ACS omega*. 2022. URL: https://doi.org/10.1021/acsomega.2c05693 (дата звернення: 16.09.2025).

18 Quantitative structure–activity relationship models for genotoxicity prediction based on combination evaluation strategies for toxicological alternative experiments / X. Yang та ін. *Scientific reports*. 2021. Т. 11, № 1. URL: https://doi.org/10.1038/s41598-021-87035-y (дата звернення: 08.09.2025).

19 ProTox-3.0 - Prediction of TOXicity of chemicals. ProTox-3.0 - Prediction of TOXicity of chemicals. URL: https://tox.charite.de/protox3/index.php?site=models (дата звернення: 10.10.2025).

20 VNN-ADMET. vNN-ADMET. URL: https://vnnadmet.bhsai.org/vnnadmet/buildclassresult.xhtml?myjob=8377a8eb9e68e7d3aa902342620d5472f3ae106c8a0dfce68879bb1a8f36666b (дата звернення: 10.10.2025).

21 Системи оцінки мутагенних ефектів методами машинного навчання. *DSpace :: ELAKPI :: Репозитарій КПІ ім. Ігоря Сікорського*. URL: https://ela.kpi.ua/handle/123456789/69754. (дата звернення 09.10.2025).

22 Review of machine learning and deep learning models for toxicity prediction / W. Guo та ін. Experimental biology and medicine. 2023. URL: https://doi.org/10.1177/15353702231209421 (дата звернення 08.09.2025).

23 2025's 5 most in-demand machine learning languages. Learn Web Development, Data Analytics, *Cybersecurity & UX/UI Design* | Ironhack. URL: https://www.ironhack.com/us/blog/2023-s-5-most-in-demand-machine-learning-languages?utm_source=chatgpt.com (дата звернення 17.09.2025).

24 PYPL PopularitY of Programming Language index. Page Redirection. URL: https://pypl.github.io/PYPL.html (дата звернення 22.09.2025).

25 Türkmen G., Sezen A., Şengül G. Comparative analysis of programming languages utilized in artificial intelligence applications: features, performance, and suitability. *International journal of computational and experimental science and engineering*. 2024. Т. 10, № 3. URL: https://doi.org/10.22399/ijcesen.342 (дата звернення 17.09.2025).

26 Johns R. PyTorch vs tensorflow for your python deep learning project – real python. Python Tutorials – Real Python. URL: https://realpython.com/pytorch-vs-tensorflow/?utm_source=chatgpt.com (дата звернення 22.09.2025).

27 Hasan A. F., Jabba S. F., Raheem F. S. Comparative analysis of four programming languages for machine learning. *Ingénierie des systèmes d information*. 2025. Т. 30, № 6. С. 1437–1445. URL: https://doi.org/10.18280/isi.300603 (дата звернення 22.09.2025).

28 IBM. What is random forest? | IBM. IBM. URL: https://www.ibm.com/think/topics/random-forest?utm_source=chatgpt.com (дата звернення 17.09.2025).

29 Feature selection. scikit-learn. URL: https://scikit-

learn.org/stable/modules/feature_selection.html#rfe (дата звернення 17.09.2025).

30 Wizards D. S. Understanding the gradient boosting algorithm. Medium. URL: https://medium.com/@datasciencewizards/understanding-the-gradient-boosting-algorithm-9fe698a352ad (дата звернення 17.09.2025).

31 Benchmark data set for *In silico* prediction of ames mutagenicity / K. Hansen та ін. *Journal of chemical information and modeling*. 2009. Т. 49, № 9. С. 2077–2081. URL: https://doi.org/10.1021/ci900161g (дата звернення: 20.09.2025).

32 Kazius J., McGuire R., Bursi R. Derivation and validation of toxicophores for mutagenicity prediction *Journal of medicinal chemistry*.. 2005. Т. 48, № 1. С. 312–320. URL: https://doi.org/10.1021/jm040835a (дата звернення: 20.09.2025).

33 Guidance on the establishment of the residue definition for dietary risk assessment. *EFSA journal*. 2016. Т. 14, № 12. URL: https://doi.org/10.2903/j.efsa.2016.4549 (дата звернення: 20.09.2025).

34 MicotoXilico: an interactive database to predict mutagenicity, genotoxicity, and carcinogenicity of mycotoxins / J. Tolosa та ін. *Toxins*. 2023. Т. 15, № 6. С. 355. URL: https://doi.org/10.3390/toxins15060355 (дата звернення 26.09.2025).

35 Run Classification - ClassyFire. Run Classification - ClassyFire. URL: http://classyfire.wishartlab.com/ (дата звернення 26.09.2025).

36 LibreTexts. 9.1: аліфатичні вуглеводні. *LibreTexts - Ukrayinska*. URL: https://ukrayinska.libretexts.org/%D0%A5%D1%96%D0%BC%D1%96%D1%8F/%D0%92%D1%81%D1%82%D1%83%D0%BF%D0%BD%D0%B8%D0%B9%2C_%D0%BA%D0%BE%D0%BD%D1%86%D0%B5%D0%BF%D1%82%D1%83%D0%B0%D0%BB%D1%8C%D0%BD%D0%B8%D0%B9_%D1%82%D0%B0_%D0%97%D0%9E%D0%91_%D0%A5%D1%96%D0%BC%D1%96%D1%8F/%D0%A5%D1%96%D0%BC%D1%96%D1%8F_%D0%B4%D0%BB%D1%8F_%D0%B7%D0%BC%D1%96%D0

%BD%D0%B8_%D1%87%D0%B0%D1%81%D1%96%D0%B2_(Hill_%D1%96_McCreary)/09%3A_%D0%9E%D1%80%D0%B3%D0%B0%D0%BD%D1%96%D1%87%D0%BD%D0%B0_%D1%85%D1%96%D0%BC%D1%96%D1%8F/9.02%3A_%D0%90%D0%BB%D1%96%D1%84%D0%B0%D1%82%D0%B8%D1%87%D0%BD%D1%96_%D0%B2%D1%83%D0%B3%D0%BB%D0%B5%D0%B2%D0%BE%D0%B4%D0%BD%D1%96 (дата звернення 02.10.2025).

37 LibreTexts. 16.2: вуглеводні. LibreTexts - Ukrayinska. URL: https://ukrayinska.libretexts.org/%D0%A5%D1%96%D0%BC%D1%96%D1%8F/%D0%92%D1%81%D1%82%D1%83%D0%BF%D0%BD%D0%B8%D0%B9%2C_%D0%BA%D0%BE%D0%BD%D1%86%D0%B5%D0%BF%D1%82%D1%83%D0%B0%D0%BB%D1%8C%D0%BD%D0%B8%D0%B9_%D1%82%D0%B0_%D0%97%D0%9E%D0%91_%D0%A5%D1%96%D0%BC%D1%96%D1%8F/%D0%A5%D1%96%D0%BC%D1%96%D1%8F_%D0%B4%D0%BB%D1%8F_%D0%B7%D0%BC%D1%96%D0%BD%D0%B8_%D1%87%D0%B0%D1%81%D1%96%D0%B2_(Hill_%D1%96_McCreary)/09%3A_%D0%9E%D1%80%D0%B3%D0%B0%D0%BD%D1%96%D1%87%D0%BD%D0%B0_%D1%85%D1%96%D0%BC%D1%96%D1%8F/9.02%3A_%D0%90%D0%BB%D1%96%D1%84%D0%B0%D1%82%D0%B8%D1%87%D0%BD%D1%96_%D0%B2%D1%83%D0%B3%D0%BB%D0%B5%D0%B2%D0%BE%D0%B4%D0%BD%D1%96 (дата звернення 02.10.2025).

38 Alvarez-Builla J., Barluenga J. Heterocyclic compounds: an introduction. Wiley-VCH - Home. URL: https://application.wiley-vch.de/books/sample/3527332014_c01.pdf?utm_source=chatgpt.com (дата звернення 12.09.2025).

39 Genotoxic effects of heterocyclic aromatic amines in human derived hepatoma (HepG2) cells / S. Knasmuller та ін. Mutagenesis. 1999. Т. 14, № 6. С. 533–540. URL: https://doi.org/10.1093/mutage/14.6.533 (дата звернення 12.09.2025).

40 Genotoxicity of heterocyclic pahs in the micronucleus assay with the fish liver cell line RTL-W1 / M. Brinkmann та ін. *PLoS ONE*. 2014. Т. 9, № 1. С. e85692. URL: https://doi.org/10.1371/journal.pone.0085692 (дата звернення 12.09.2025).

41 PaDEL-Descriptor. URL: http://yapcwsoft.com/dd/padeldescriptor/ (дата звернення 02.10.2025).

42 RDKit. RDKit. URL: https://www.rdkit.org/ (дата звернення 09.10.2025).

43 Mordred: a molecular descriptor calculator / H. Moriwaki та ін. *Journal of cheminformatics*. 2018. Т. 10, № 1. URL: https://doi.org/10.1186/s13321-018-0258-y (дата звернення 09.10.2025).

44 Chicco D., Jurman G. The ABC recommendations for validation of supervised machine learning results in biomedical sciences. Frontiers in big data. 2022. Т. 5. URL: https://doi.org/10.3389/fdata.2022.979465 (дата звернення 26.09.2025).

45 Rfecv. scikit-learn. URL: https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.RFECV.html (дата звернення 02.10.2025).

46 Cavasotto C. N., Scardino V. Machine learning toxicity prediction: latest advances by toxicity end point. *ACS omega*. 2022. URL: https://doi.org/10.1021/acsomega.2c05693 (дата звернення 08.09.2025).