

ХЕРСОНСЬКИЙ НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ
ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ ТА ДИЗАЙНУ
КАФЕДРА ІНФОРМАТИКИ І КОМП'ЮТЕРНИХ НАУК

Пояснювальна записка

до дипломної магістерської роботи

на тему:

**Використання консолідованого інформаційного ресурсу для розробки
регресійних моделей дермальної абсорбції хлорорганічних пестицидів**

Виконав: студент 6 курсу, групи 6КІ
спеціальності 122 «Комп'ютерні науки»

Захарченко Є. Д.

Керівник: Литвиненко В.І.

Рецензент: Г.В. Рудакова

Факультет	Інформаційних технологій та дизайну
Кафедра	Інформатики і комп'ютерних наук
Рівень вищої освіти	<u>другий (магістерський) рівень</u>
Галузь підготовки	<u>12 «Інформаційні технології»</u> (шифр і назва)
Освітньо-професійна програма	<u>Консолідована Інформація</u> (назва)
Спеціальність	<u>122 «Комп'ютерні науки»</u> (шифр і назва)

ЗАТВЕРДЖУЮ

Завідувач кафедри ІКН,
професор

_____ В.І. Литвиненко

«_____» _____ 2024 року

З А В Д А Н Н Я НА ДИПЛОМНУ РОБОТУ СТУДЕНТА

Захарченко Євгенія Дмитровича

(прізвище, ім'я, по батькові)

Тема роботи: Використання консолідованого інформаційного ресурсу для розробки регресійних моделей дермальної абсорбції хлорорганічних пестицидів.

1. Керівник роботи Литвиненко Володимир Іванович, доктор технічних наук, професор кафедри інформатики і комп'ютерних наук

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

затверджені наказом ХНТУ від «25».09.2024 р. № № 499-с

2. Строк подання студентом роботи

3. Вихідні дані до роботи _____

4. Зміст розрахунково-пояснювальної записки (перелік питань, які потрібно розробити): Вступ. Розділ 1 Огляд попередніх досліджень. Розділ 2. Теоретична частина. Визначення та опис предметної частини. Розділ 3. Практична частина. Аналіз отриманих даних та результатів.

5. Перелік графічного матеріалу (з точним зазначенням обов'язкових креслень)
Малюнків – 42.

Таблиць – 42.

Формул – 16.

7. Консультанти розділів роботи

Розділ	Прізвище, ініціали та посада консультанта	Підпис, дата	
		завдання видав	завдання прийняв

8. Дата видачі завдання 25.09.2024

КАЛЕНДАРНИЙ ПЛАН

№ з/п	Назва етапів дипломної роботи	Строк етапів роботи	Прим.
1	Обрання теми диплому, створення плану роботи та написання вступу дипломної роботи.	27.09.2024 – 03.10.2024	
2	Пошук інформації в літературних джерелах, та огляд попередніх досліджень.	05.10.2024 – 10.10.2024	
3	Складання теоретичної частини, на основі зібраної інформації.	12.10.2024 – 16.10.2024	
4	Підготовка даних за методом відбору ознак	18.10.2024 – 30.10.2024	
5	Проведення регресії з необхідними параметрами	1.11.2024 – 14.11.2024	
6	Отримання результатів дослідження	16.11.2024 – 23.11.2024	
7	Написання висновків до розділів	24.11.2024 – 30.11.2024	
8	Оформлення звіту до дипломної роботи, корегування та підготовка до захисту.	31.11.2024 – 08.12.2024	

Студент _____
(підпис)

Є.Д. Захарченко
(прізвище та ініціали)

Керівник роботи _____
(підпис)

В.І Литвиненко
(прізвище та ініціали)

РЕФЕРАТ

Дана кваліфікаційна магістерська робота присвячена розробці регресійних моделей дермальної абсорбції хлорорганічних пестицидів. У пояснювальній записці до роботи викладено результати досліджень, які структуровані у 3 розділах. Документ містить 42 малюнки, 42 таблиці, 16 формул.

Під час виконання дослідження були відібрані деструктори, використані методи відбору ознак, з подальшою регресією та аналізом результатів. Для роботи було використано мову програмування R з відповідними бібліотеками для використання методів відбору ознак, також було використано програмне забезпечення WEKA для проведення регресії з необхідними параметрами. Отримані результати було збережено у таблиці MS Excel, а далі дані було використано для побудови графіків з подальшим аналізом даних.

Ключові слова: методи машинного навчання, методи відбору ознак, регресія.

ABSTRACT

This master's qualification thesis is dedicated to the development of regression models for dermal absorption of chlorinated pesticides. The explanatory note accompanying the thesis presents the research results structured into three sections. The document includes 42 figures, 42 tables, and 16 formulas.

During the study, destructors were selected, feature selection methods were applied, followed by regression and result analysis. The programming language R, along with its relevant libraries for feature selection methods, was utilized for this work. Additionally, WEKA software was employed to perform regression with the necessary parameters. The obtained results were saved in an MS Excel table and subsequently used to create graphs for further data analysis.

СКОРОЧЕННЯ ТА УМОВНІ ПОЗНАЧЕННЯ

CMIM	Conditional Mutual Information Maximization
MIM	Mutual Information Maximization
JMIM	Joint Mutual Information Maximization
MRMR	Minimum Redundancy Maximum Relevancy
NJMIM	Normalized Joint Mutual Information
JMI	Joint Mutual Information
DISR	Double Input Symmetrical Relevance

Зміст	
Вступ.....	9
Розділ 1: Огляд літератури	13
Огляд попередніх досліджень.....	13
Висновок до розділу 1	23
Розділ 2: Опис предметної частини.....	25
2.1. Постановка проблеми.....	25
2.2. Опис даних.....	27
2.3. Методи відбору ознак.....	29
2.3.1. Conditional Mutual Information Maximisation	30
2.3.2. Maximum Relevancy Minimum Redundancy (mRMR)	32
2.3.3. Joint Mutual Information	34
2.3.4. DISR.....	35
2.3.5. JMIM та NJMIM	37
2.4. Методи машинного навчання	40
2.4.1. Багатошарова перцептронна мережа (MLP).....	40
2.4.2. К-ближчих сусідів (k-NN)	43
2.4.3. Дерева рішень	44
2.4.4. SVR.....	46
2.4.5. Random Forest.....	46
Висновок до розділу 2.....	48
Розділ 3: Аналіз отриманих даних і результатів.....	49
3.1. CMIM	49
3.2. DISR	59
3.3. JMI.....	69
3.4. JMIM.....	80
3.5. MIM	91

3.6. MRMR	102
3.7. NJMIM.....	113
Висновок до розділу 3	124
Висновок до роботи	126
Перелік використаних джерел	129

Вступ

Розробка регресійних моделей дермальної абсорбції хлорорганічних пестицидів на основі молекулярних дескрипторів є важливим напрямом у науці, пов'язаним з необхідністю оцінки впливу токсичних речовин на людину та навколишнє середовище. Це актуально не тільки у зв'язку з історичним застосуванням хлорорганічних пестицидів, але і з їхньою присутністю в навколишньому середовищі. Хлорорганічні пестициди, такі як ДДТ, альдрін, гептахлор, хлордан та багато інших, були широко застосовані у сільському господарстві та боротьбі зі шкідниками протягом кількох десятиліть. Незважаючи на їх заборону або суворе обмеження в більшості країн, ці речовини мають високу стійкість до розкладання і продовжують перебувати в ґрунтах, водоймах, а також жирових тканинах тварин і людини.

Основний шлях надходження хлорорганічних пестицидів в організм включає дермальну абсорбцію, що особливо актуально для працівників сільського господарства і людей, які безпосередньо контактують із забрудненими поверхнями, ґрунтом або водою. У звичайних умовах шкірний бар'єр ефективно захищає організм від багатьох шкідливих речовин, проте при роботі з високотоксичними та ліпофільними сполуками, такими як хлорорганічні пестициди, цей бар'єр може виявитися недостатньо ефективним. Розуміння процесів дермальної абсорбції стає критично важливим для оцінки можливих ризиків для здоров'я та розробки методів їхньої мінімізації. Важливо враховувати, що процеси абсорбції залежать від багатьох факторів: це не тільки фізико-хімічні властивості самої речовини, але й стан шкіри, температура, вологість навколишнього середовища, наявність інших речовин на шкірі (наприклад, мастил або розчинників), які можуть змінювати проникність шкірного бар'єру.

Проблема створення моделей дермальної абсорбції полягає у складності проведення експериментальних досліджень, які є дорогими, трудомісткими та потребують використання *in vivo* та *in vitro* методів, часто із застосуванням лабораторних тварин. Не лише обмежує можливості дослідження великої кількості сполук, а й викликає питання з погляду етики. У зв'язку з цим виникає потреба в альтернативних методах, що дозволяють передбачати ступінь абсорбції дермальної без проведення великої кількості експериментів. Одним із таких методів є використання регресійних моделей, побудованих на основі молекулярних дескрипторів — параметрів, які кількісно описують структуру молекули та її фізико-хімічні властивості.

Молекулярні дескриптори - це набір числових значень, які характеризують молекулу з різних сторін: її розмір, форму, ступінь полярності, наявність специфічних функціональних груп, розподіл заряду, ліпофільність (часто вимірюється через логарифм коефіцієнта розподілу октанол-вода, $\log P$) та багато інших параметрів. Всі ці властивості можуть впливати на здатність речовини проникати через шкірний бар'єр. Наприклад, дескриптори, що описують ліпофільність речовини ($\log P$), мають безпосереднє відношення до процесу абсорбції через роговий шар шкіри, який значною мірою складається з ліпідів. Ліпофільні сполуки мають тенденцію легше проникати через цей бар'єр, тоді як гідрофільні речовини зустрічають більший опір. Розмір і форма молекули також мають значення: великі молекули з більш громіздкими структурами можуть важче проходити через міжклітинні простори, ніж невеликі та компактні молекули.

Використання регресійних моделей дозволяє виявляти залежності між дескрипторами та ступенем абсорбції, що може бути корисним для передбачення абсорбції нових сполук. Такі моделі допомагають економити час і кошти, які знадобилися б для проведення великої кількості експериментів, і сприяють зниженню необхідності використання тварин у дослідженнях. Більш того,

застосування молекулярних дескрипторів робить ці моделі універсальними та придатними для адаптації до різних умов впливу, оскільки характеристики молекул можна враховувати у широкому діапазоні значень параметрів, таких як температура, вологість та індивідуальні особливості шкірного бар'єру.

Сучасні методи машинного навчання та штучного інтелекту дозволяють розширити можливості побудови регресійних моделей, зробивши їх більш точними та здатними враховувати нелінійні залежності. Алгоритми машинного навчання, такі як градієнтний бустинг, випадкові ліси та нейронні мережі, здатні виявляти складні взаємозв'язки між дескрипторами та абсорбцією, що покращує точність передбачень. Це особливо важливо у випадках, коли речовини мають складні молекулярні структури, які фізико-хімічні властивості важко описати простими лінійними моделями. Машинне навчання дозволяє будувати моделі, які не тільки враховують існуючі дані, але й адаптуються в міру появи нових експериментальних результатів, покращуючи свою точність та універсальність.

Регресійні моделі на основі молекулярних дескрипторів можуть також враховуватися в оцінці нормативного впливу та використовувати регуляторні органи при розробці стандартів безпеки. З їх допомогою можна більш обґрунтовано визначати допустимі рівні впливу хімічних речовин та розробляти заходи щодо їх обмеження. Це також сприяє розвитку міжнародних норм і стандартів, пов'язаних з безпечним застосуванням хімічних речовин, що є особливо важливим для глобальних ринків, де хімічні речовини можуть бути вироблені в одній країні, а використані в іншій.

Крім того, такі моделі дозволяють проводити більш глибокий аналіз взаємодії факторів, що впливають на абсорбцію, та можуть використовуватись для оцінки індивідуальних ризиків. Наприклад, з урахуванням стану здоров'я конкретної людини, наявності ушкоджень шкіри, віку чи навіть генетичних особливостей можна робити точніші прогнози впливу хлорорганічних

пестицидів. Це відкриває можливості для персоналізованої медицини та розробки специфічних заходів профілактики та лікування.

Таким чином, актуальність розробки регресійних моделей дермальної абсорбції хлорорганічних пестицидів на основі їх дескрипторів обумовлена необхідністю точної та оперативної оцінки ризику, оптимізацією процесу регуляторної оцінки хімічних речовин, зниженням витрат на експерименти та покращенням розуміння механізмів дермальної абсорбції токсичних сполук. Цей напрямок досліджень робить значний внесок у розвиток передбачуваної токсикології, екологічної безпеки та захисту здоров'я людини.