

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

Пояснювальна записка
до кваліфікаційної магістерської роботи

на тему

**ВИКОРИСТАННЯ МЕТОДІВ ГЛИБИННОГО НАВЧАННЯ ДЛЯ
РОЗВ'ЯЗАННЯ ЗАДАЧ КЛАСИФІКАЦІЇ МОЛЕКУЛЯРНИХ ДАНИХ
КОНСОЛІДОВАНОГО РЕСУРСУ ОНКОЛОГІЧНИХ ХВОРИХ.**

Виконав:
студент 6 курсу, група 6КІ
спеціальність 122 «Комп'ютерні науки»

Шкарупа Є.М.

Керівник: Литвиненко В.І.

Рецензент: Райко Г.О.

Херсон - 2020

Факультет
Кафедра
Рівень вищої освіти
Галузь підготовки

Освітньо-професійна програма

Спеціальність

Інформаційних технологій та дизайну
Інформатики і комп'ютерних наук
другий (магістерський) рівень
12 «Інформаційні технології»
(шифр і назва)

Консолідована інформація
(назва)
122 «Комп'ютерні науки»
(шифр і назва)

ЗАТВЕРДЖУЮ

Завідувач кафедри ІКН,
професор

_____ В.І. Литвиненко

«_____» _____ 2020 року

З А В Д А Н Н Я **НА КВАЛІФІКАЦІЙНУ РОБОТУ СТУДЕНТА**

Шкарупа Єгор Михайлович

(прізвище, ім'я, по батькові)

1. Тема роботи: Використання методів глибинного навчання для розв'язання
задач класифікації молекулярних даних консолідованого ресурсу
онкологічних хворих

керівник роботи Литвиненко Володимир Іванович, доктор технічних наук,
професор

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

затверджені наказом ХНТУ від «07» вересня 2020 року №461-с.

2. Строк подання студентом роботи 04.12.2020 р.

3. Вихідні дані до роботи Матеріали та результати, отримані під час
проходження переддипломної практики, методичні вказівки, технічна
література

4. Зміст розрахунково-пояснювальної записки (перелік питань, які потрібно
розробити): Вступ. Аналітичний огляд предметної області. Ознайомлення з R.
Огляд пакетів h2o та keras. Порівняння програмних засобів h2o та keras.
Класифікація даних консолідованого ресурсу онкологічних хворих. Висновки.

5. Перелік графічного матеріалу (з точним зазначенням обов'язкових креслень)
В роботі наведено:

Рисунки - 43

Таблиці - 4

6. Консультанти розділів роботи

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

7. Дата видачі завдання _____

КАЛЕНДАРНИЙ ПЛАН

№ з/п	Назва етапів дипломного роботи	Строк виконання етапів роботи	Примітка
1	Збір інформації. Аналіз існуючих аналогів.	14.10.2020-16.10.2020	Виконано
2	Вивчення базових рекомендаційних алгоритмів та вибір варіанту для розробки	17.10.2020-25.10.2020	Виконано
3	Аналіз сервісів що використовують РС	26.10.2020-15.11.2020	Виконано
4	Розробка алгоритму та структури системи. Розробка програми	16.11.2020-29.12.2020	Виконано
5	Тестування програми та отримання результатів. Оформлення роботи.	30.12.2020-04.12.2020	Виконано

Студент _____ Є.М. Шкарупа
(підпис) (прізвище та ініціали)

Керівник роботи _____ В.І. Литвиненко
(підпис) (прізвище та ініціали)

ЗМІСТ

РЕФЕРАТ.....	6
ПЕРЕЛІК УМОВНИХ СКОРОЧЕНЬ.....	7
ВСТУП.....	8
РОЗДІЛ 1. АНАЛІТИЧНИЙ ОГЛЯД ПРЕДМЕТНОЇ ОБЛАСТІ. ОБГРУНТУВАННЯ МЕТОДІВ ВИРІШЕННЯ ПОСТАВЛЕНОЇ ЗАДАЧІ	Ошибка! Закладка не определена.
1.1 Проблема онкологічних захворювань.....	Ошибка! Закладка не определена.
1.2 Мікрочіпи. Методи роботи мікрочіпів.....	Ошибка! Закладка не определена.
1.2.1 Модифікація мікрочіпів та методи формування ДНК-мікрочіпу	Ошибка! Закладка не определена.
1.2.2 Методи обробки даних ДНК мікрочіпу для визначення експресій генів.....	Ошибка! Закладка не определена.
1.2.3 Фонова корекція даних мікрочіпу ДНК.....	Ошибка! Закладка не определена.
1.2.4 Методи нормалізації інтенсивностей світла мікрочіпу ДНК	Ошибка! Закладка не определена.
1.2.5 Методи РМ-корекції та сумаризації інтенсивностей світла мікрочіпу ДНК.....	Ошибка! Закладка не определена.
1.3 Огляд консолідованого ресурсу даних онкологічних хворих	Ошибка! Закладка не определена.
1.4 Глибинне навчання	Ошибка! Закладка не определена.
Висновки до першого розділу.....	Ошибка! Закладка не определена.
РОЗДІЛ 2. ТЕОРЕТИЧНА ЧАСТИНА. ОЗНАЙОМЛЕННЯ З R. ОГЛЯД ПАКЕТІВ H2O ТА KERAS.....	Ошибка! Закладка не определена.
2.1 Мова програмування R.....	Ошибка! Закладка не определена.

2.2 Інтегрована середовище розробки RStudio	Ошибка!	Закладка	не определена.
2.3 Пакет глибокого навчання «H2O»	Ошибка!	Закладка	не определена.
2.4 Пакет глибокого навчання «Keras»	Ошибка!	Закладка	не определена.
Висновки до другого розділу	Ошибка!	Закладка	не определена.
РОЗДІЛ 3. ПРАКТИЧНА ЧАСТИНА. ПОРІВНЯННЯ ПРОГРАМНИХ ЗАСОБІВ H2O ТА KERAS. КЛАСИФІКАЦІЯ ДАНИХ КОНСОЛІДОВАНОГО РЕСУРСУ ОНКОЛОГІЧНИХ ХВОРИХ.....			
3.1 Запропонована методологія			14
3.1.1 Масштабування ознак.....			15
3.1.2 Зниження розмірів			15
3.1.3 Класифікація даних мікрочіпів на основі глибокого вивчення			16
3.2 Експериментальні дослідження і аналіз отриманих результатів			22
3.2.1 Описання використаного набору даних			22
3.2.2. Робота з пакетом keras.....			23
3.2.3 Результати проведення обчислюваних експериментів			29
3.2.4. Результати використання пакета H2O			34
3.2.5. Дослідження отриманих результатів			38
3.3 Результати порівнянних експериментів роботи з пакетами «keras» та «H2O».			42
Висновки до третього розділу.....			47
ВИСНОВКИ	Ошибка!	Закладка	не определена.
СПИСОК ВИКОРИСТАНИХ ЛІТЕРАТУРНИХ ТА ІНШИХ ДЖЕРЕЛ	Ошибка!	Закладка	не определена.
Додаток А. Лістинг програмного забезпечення	Ошибка!	Закладка	не определена.
Додаток Б. Графічний результат роботи програмного коду	Ошибка!	Закладка	не определена.
Додаток В. Штучний інтелект	Ошибка!	Закладка	не определена.

Додаток Г. Нейронні мережі..... Ошибка! Закладка не определена.

Додаток Д. Машинне навчання..... Ошибка! Закладка не определена.

Шкарупа Є.М. 6К1

РЕФЕРАТ

В даній дипломній роботі вібудується дослідження методів глибинного навчання шляхом класифікації даних онкологічних хворих. В якості набору даних для класифікації використовувались дані пацієнтів дитячої дослідницької лікарні Св. Джуда, хворих на гострий лімфобластний лейкоз (ГЛЛ) у дітей. Класифікація була виконана двома програмними засобами для роботи з глибинним навчанням - програмними пакетами «Keras» та «H2O» на мові програмування «R». Отримані в результаті досліджень дані використовувались для оцінки роботи кожного з програмних пакетів з ціллю надання порівняльної характеристики кожному з пакетів та вибору найбільш зручного і функціонального програмного продукту для роботи з задачами глибинного навчання.

ABSTRACT

This thesis focuses on the study of deep learning methods by classifying the data of cancer patients. Data from St. Jude Children's Research Hospital patients with acute lymphoblastic leukemia (ALL) in children were used as a data set for classification. The classification was performed by two software tools for working with deep learning - software packages "Keras" and "H2O" in the programming language "R". The data obtained as a result of research were used to evaluate the performance of each of the software packages in order to provide a comparative description of each of the packages and select the most convenient and functional software product for working with deep learning tasks.

ПЕРЕЛІК УМОВНИХ СКОРОЧЕНЬ

Скорочення, термін, позначення	Пояснення
ОС	Операційна система
БД	База даних
ПЗ	Програмне забезпечення
Рис	Рисунок
ШІ	Штучний інтелект
ЕОМ	Електронна обчислювальна машина
НМ	Нейронна мережа
ГЛЛ	Гострий лімфобластний лейкоз
ANN	Artificial neural network
API	Application programming interface
RBM	Restricted Boltzmann Machine Обмежена машина Больцмана
DBN	Deep Belief Network Мережа глибинних вірувань
DNN	Deep Neural Network Глибинна нейронна мережа

ВСТУП

Сучасне суспільство наразі знаходиться в епосі неперервного та рішучого розвитку інформаційних технологій, з року в рік темпи розвитку технологій тільки зростають та будуть зростати й надалі. Для галузі штучного інтелекту наразі життєво необхідним є поглиблене навчання технології глибинного навчання, завдяки якій комп'ютерні пристрої мають змогу навчитися виконувати окремі задачі за допомогою обробки даних і виявленні в них певних закономірностей. Більшість алгоритмів глибинного навчання представляють собою різноманітні види нейронних мереж. Нейронні мережі являються системою, що складається з нейронів та навчається не тільки на заданих алгоритмах, а ще й на основі особистого досвіду.

З появою машинного навчання перед комп'ютерами відкрилася можливість підступитися до завдань, що вимагають знань про реальний світ, і приймати рішення, що здаються суб'єктивними. Глибинне навчання - це окремий випадок машинного навчання, що дозволяє досягти більшої ефективності і гнучкості за рахунок подання світу у вигляді ієрархії вкладених концепцій, в якій кожна концепція визначається в термінах більш простих концепцій, а більш абстрактні уявлення обчислюються в термінах менш абстрактних.

Класифікація раку є однією з основних дослідних тем в медицині. Дані мікрочіпів раку, як правило, містять невелику кількість зразків, які мають велику кількість рівнів експресії генів в якості ознак, однак, робить класифікацію досить складною. Використовуючи алгоритм глибокого навчання, заснований на багатошаровому перцептроні, ми показуємо, що для класифікації раку можна отримати результати класифікації, по крайній мере, такі ж гарні, як опубліковані результати.

Рак є гетерогенний комплекс хвороб з вираженими біологічними особливостями, які призводять до відмінностей в реакції на різні методи

лікування і клінічні наслідки. Пухлини зі схожими клінічними і патологічними проявами можуть мати різну поведінку. Аналіз експресії генів сфокусований на визначенні більш докладних біологічних характеристик для поліпшення стратифікації ризику пацієнта і забезпечення найбільшої ймовірності користі і найменшої токсичності від конкретного способу лікування.

Точна діагностика раку життєво важлива для успішного застосування специфічних методів лікування. За останнє десятиліття класифікація раку покращилася, все ще існує потреба в повністю автоматизованому і менш суб'єктивному методі діагностики раку. З огляду на характер даних мікрочіпів раку, які зазвичай складаються з декількох сотень зразків з тисячами генів в якості ознак, аналіз повинен бути проведений ретельно. Існує безліч підходів і процедур, спрямованих на виявлення груп генів, сильно пов'язаних з певним типом раку.

Сучасний стан обробки біомедичних систем може бути вдосконалено за умови розробки ефективних і дієвих алгоритмів глибокого навчання. Концепція глибокого навчання виникла в результаті досліджень штучних нейронних мереж. Хорошим прикладом моделей з глибокої архітектурою є багат шаровий перцептрон з безліччю прихованих шарів. В останні роки методи глибокого навчання застосовуються для вирішення широкого кола проблем. У багатьох з цих додатків алгоритми, засновані на глибокому навчанні, перевершили колишню продуктивність.

Глибинне навчання використовує класифікатори для оцінки корисності підмножин ознак, які повинні забезпечувати кращу класифікаційну точність для відповідних класифікаторів, оскільки ознаки вибираються відповідно до їхнього внеску в класифікаційну точність класифікаторів.

Аналіз даних, отриманих за допомогою мікрочіпів, являє собою досить складну задачу в зв'язку з притаманною природі даних складністю, пов'язаною з більш високою розмірністю, меншим розміром вибірки, незбалансованим кількістю класів, гучної структурою даних і більшої дисперсією значень ознак. Це призвело до меншої точності класифікації та проблеми перенасичення. У

даній роботі розроблений метод глибинного зворотного зв'язку для класифікації даних про рак мікрочіпів в наборі класів для подальшої діагностики. В кожному пакеті для набору даних була використана індивідуальна архітектура глибокої нейронної мережі з різними параметрами. Проблеми малого розміру зразка і розмірності вирішуються за допомогою відомого методу зменшення розмірності, а саме аналізу основних компонентів. Значення характеристик нормалізуються різними способами, і пропонований підхід валідується на спеціальній (тестувальній) частині даних із набору даних по раку мікрочіпів. Для вимірювання втрат використовується категоріальна крос-ентропія, а для оптимізації розглядається алгоритм «ADAM». Продуктивність пропонованого підходу оцінюється з використанням точності класифікації, помилок класифікації, відклику, f-вимірювання, логарифмічних втрат, і матриці плутанини.

В даному дипломному проєкті на тему «Використання методів глибинного навчання для розв'язання задач класифікації молекулярних даних консолідованого ресурсу онкологічних хворих» відбувається порівняння двох програмних продуктів для роботи з глибинним навчанням «H2O» та «Keras» шляхом класифікації даних мікрочіпів раку онкологічних хворих.

Правильний науковий підхід до глибокого навчання в поєднанні з особливостями генного мікрочіпа має потенціал для захоплення зразків експресії генів, що може підвищити ефективність класифікації. Методи машинного навчання, такі як глибинне навчання, в поєднанні з новими молекулярними методами, обіцяють поліпшення діагностики, меншу вірогідність ризику рецидивів, ретельний відбір терапії та ідентифікація цілей, що беруть участь в карконогенезі і функції пухлинних клітин, що призводить до відкриття селективних лікарських засобів.

Актуальність теми. Актуальність роботи полягає в гіперболічному масштабі розвитку технологій глибинного навчання, завдяки чому з'явилася можливість використовувати методи глибинного навчання для вирішення найбільш актуальних питань сучасного людства. Правильно виконана

класифікація даних онкологічних хворих може не тільки призвести до точного ідентифікування відомих прогностично важливих підтипів лейкемії, але і може ще більше підвищити здатність оцінювати ризик невдалої терапії пацієнта. Крім того, профілювання експресії генів дозволяє точно ідентифікувати деяких пацієнтів, які піддаються високому ризику неефективності традиційних терапевтичних підходів. Можливо припустити, що в найближчому майбутньому профілювання експресії стане важливим діагностичним інструментом для оцінки онкологічних хворих пацієнтів.

Також, актуальним завданням є порівняння існуючих програмних засобів для роботи з глибинним навчанням, в результаті популяризації теми глибинного навчання на ринку послуг з'явилося багато таких засобів, серед яких слід виділити продукти «H2O» та «Keras», в даній роботі класифікація даних онкологічних хворих буде здійснюватися на кожному з цих засобів для надання їм всебічної порівняльної характеристики та вибору найбільш зручного та функціонального програмного продукту для роботи з задачами глибинного навчання.

Метою дипломної роботи є розробка інформаційної технології консолідованого інформаційного глибинного навчання шляхом для розв'язання задач класифікації молекулярних даних онкологічних хворих з використанням пакетів «H2O» та «Keras».

Для досягнення поставленої мети в магістерській роботі поставлені та розв'язані **наступні задачі**:

- проаналізувати існуючі технології, моделі та програмні засоби глибинного навчання їх переваги і недоліки;
- проаналізувати проблемну область класифікації молекулярних даних онкологічних хворих з використанням глибинного навчання та клас задач які в ній вирішуються;
- розробити метод глибинного навчання для розв'язання задач класифікації молекулярних даних;