Форма № Н-9.02

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
“КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ ІМ. ІГОРЯ СІКОРСЬКОГО ”
Факультет електроніки
Кафедра електронної інженерії

До захисту допущено
Завідувач кафедри
_____ В. І. Тимофєєв
“___”_____ 20___ р.

Дипломна робота

освітнього рівня «бакалавр»
за спеціальністю 153 мікро- та наносистемна техніка

на тему «Машинне навчання для діагностики захворювань за даними секвенування мікробіому кишківника людини»

Виконала студентка 4 курсу, групи ДМ-92
Кузьмінська Дарія Вячеславівна _____
(прізвище, ім’я, по батькові) (підпис)

Керівник к.т.н., доц. Іванько К. О._____ _____
(посада, вчене звання, науковий ступінь, прізвище та ініціали) (підпис)

Консультант _____ _____ _____
(назва розділу) (вчені ступінь та звання, прізвище, ініціали) (підпис)

Рецензент к.т.н., доц. Коваль В.М._____ _____
(посада, вчене звання, науковий ступінь, прізвище та ініціали) (підпис)

Засвідчую, що у цій дипломній роботі немає запозичень з праць інших авторів без відповідних посилань.
Студент _____
(підпис)

Київ - 2023 року

Форма № Н-9.01

Національний технічний університет України
“Київський політехнічний інститут імені Ігоря Сікорського”

Факультет електроніки

Кафедра електронної інженерії
Освітній рівень «бакалавр»

за спеціальністю 153 мікро- та наносистемна техніка

ЗАТВЕРДЖУЮ

Завідувач кафедри

_____ В. І. Тимофєєв

“___”_____ 20___ р.

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

Кузьмінська Дарія Вячеславівна

(прізвище, ім’я, по батькові)

1. Тема роботи «Машинне навчання для діагностики захворювань за даними секвенування мікробіому кишківника людини»_____
_____

керівник роботи доц., к.т.н., доц. Іванько Катерина Олегівна,
(прізвище, ім’я, по батькові, науковий ступінь, вчене звання)

затверджені наказом по університету від “30” травня 2023 року № 2063-с

2. Строк подання студентом роботи __09.06.2023_____

3. Вихідні дані до роботи: масив з ймовірними назвами захворювань для відповідних послідовностей ДНК, які були завантажені.

4. Зміст дипломної роботи (перелік питань, які потрібно розробити): аналіз даних секвенування мікробіому кишківника людини за допомогою машинного навчання.

5. Перелік графічного (ілюстративного) матеріалу (із зазначенням

обов'язкових креслень, плакатів, презентацій тощо): рисунки, веб-інтерфейс, презентація.

6. Консультанти розділів роботи

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

7. Дата видачі завдання 15.04.2023

КАЛЕНДАРНИЙ ПЛАН

№ з/п	Назва етапів виконання дипломної роботи	Строк виконання етапів роботи	Примітка
1.	Літературний огляд матеріалу стосовно методів та засобів секвенування.	17.04.2023-23.04.2023	
2.	Порівняння технологій приладів біонанопорового секвенування Oxford Nanopore та секвенування за допомогою синтезу Illumina.	23.04.2023-30.04.2023	
3.	Огляд літератури стосовно досліджень мікробіому та його зв'язку зі здоров'ям людини.	01.05.2023-14.05.2023	
4.	Аналіз методів дослідження сигналів нанопорового секвенування, а також даних секвенування мікробіому людини.	08.05.2023-14.05.2023	
5.	Опанування роботи з даними секвенування ДНК у FASTQ, SFF форматах.	15.05.2023-21.05.2023	
6.	Огляд літератури щодо методів машинного навчання.	15.05.2023-21.05.2023	
7.	Розробка програмного забезпечення для розпізнавання захворювань людини за аналізом	15.05.2023-02.06.2023	

	даних секвенування мікробіому кишківника з використанням методів машинного навчання.		
8.	Аналіз ДНК мікробіому кишківника людини та вибір двох найефективніших моделей.	22.05.2023-04.06.2023	
9.	Розробка веб-інтерфейсу для користування навченими моделями машинного навчання для розпізнавання захворювань людини.	25.05.2023-06.06.2023	
10.	Оформлення дипломної роботи.	29.05.2023-08.06.2023	

Студент _____ Кузьмінська Д.В.

(підпис) (прізвище та ініціали)

Керівник роботи _____ Іванько К.О.

(підпис) (прізвище та ініціали)

РЕФЕРАТ

Дипломна робота складається з 4 розділів включаючи вступ та висновки, має 18 ілюстрацій, 5 таблиць, 8 додатків, 34 джерела. Загальний обсяг роботи – 69 сторінок.

МАШИННЕ НАВЧАННЯ ДЛЯ ДІАГНОСТИКИ ЗАХВОРЮВАНЬ ЗА ДАНИМИ СЕКВЕНУВАННЯ МІКРОБІОМУ КИШКІВНИКА ЛЮДИНИ

Дана дипломна робота сфокусована на аналізі даних секвенування ДНК мікробіому кишківника людини. Аналіз був реалізований за допомогою методів машинного навчання, його результати можна використовувати мікробіологам для діагностики захворювань кишківника людини.

У першому розділі було розглянуто поняття секвенування ДНК та методи, які використовуються для цього. Також було розглянуто поняття мікробіому та його зв’язок зі здоров’ям людини.

Другий розділ містить опис роботи з секвенованими даними. Показано вид сигналів, які передаються під час секвенування та подальшу обробку даних перед їх аналізом. Також було описано як проводиться аналіз даних секвенування ДНК мікробіому.

У третьому розділі міститься розробка скрипту для аналізу, яка була виконана за допомогою мови програмування Python, використовуючи методи машинного навчання (бібліотека scikit-learn).

Також було розроблено веб-інтерфейс, який містить навчені моделі, які показали найефективніші результати під час дослідження. Даним інтерфейсом можуть користуватись мікробіологи для діагностики таких захворювань як: хвороба Крона, діабет другого типу та синдром подразненого кишечника.

Точність проведеного аналізу залежить від даних, які будуть використовуватись для діагностування. Для аналізу трьох-класового пакету даних точність досліджень 80-90%. Для аналізу чотирьох-класового пакету даних – 75%.

ABSTRACT

The bachelor's diploma work consists of 4 sections including introduction and conclusions, has 18 illustrations, 5 tables, 8 applications, 34 sources. The total amount of work reachs 69 pages.

MACHINE LEARNING FOR DISEASE DIAGNOSTICS BASED ON HUMAN INTESTINAL MICROBIOME SEQUENCE DATA

This bachelor's diploma work focuses on the analysis of DNA sequencing data of the human gut microbiome. The analysis was realized using machine learning methods, and its results can be used by microbiologists to diagnose human gut diseases.

In the first section, the concept of DNA sequencing and the methods used for this purpose was discussed. The concept of the microbiome and its relationship to human health was also discussed.

The second section describes how to work with sequenced data. It shows the type of signals transmitted during sequencing and the subsequent processing of data before its analysis. It also describes how the microbiome DNA sequencing data is analyzed.

The third section contains the development of the analysis script, which was performed with the Python programming language using machine learning methods (scikit-learn library).

Also a web interface that contains trained models that showed the most effective results of classification was developed. This interface can be used by microbiologists to diagnose diseases such as Crohn's disease, type 2 diabetes, and irritable bowel syndrome.

The accuracy of the analysis depends on the data that will be used for diagnosis. For the analysis of a three-class data package, the accuracy of the the classification is 80-90%. For the analysis of a four-class data package - 75%.

ЗМІСТ

СКОРОЧЕННЯ ТА УМОВНІ ПОЗНАКИ

ДНК – дезоксирибонуклеїнова кислота

РНК – рибонуклеїнова кислота

Crohn`s, IBD (Overview of Inflammatory Bowel Disease) – захворювання Крона

CSS (Cascading Style Sheets) – каскадні таблиці стилів

ddNTP – дидезоксинуклеотід

FASTQ – текстовий формат даних

HMP (Human Microbiome Project) – проект Мікробіом людини

HTML (HyperText Markup Language) – мова гіпертекстової розмітки

IBS (Irritable bowel syndrome) – синдром подразненого кишечника

NGS (Next-Generation Sequencing) – Технології секвенування наступного покоління

PKL (Python Pickle) – двійковий формат Python

SFF (Structured Fax File) – структурований факсимальний файл

T2D (Type 2 diabetes) – діабет другого типу

pH – водневий показник

ВСТУП

Людське здоров'я виступає як один з ключових аспектів, над удосконаленням якого вчені безперервно працюють. Сучасні методи терапевтичних заходів та профілактики недугів набувають все більшої популярності та за ефективністю не поступаються відомим методам. Отже, дослідження мікробіому людини, взаємодії ДНК мікробіому та здоров'я людини є однією з перспективних областей діагностування та лікування різноманітних захворювань.

Машинне навчання стає все більш актуальним і широко використовується в різних сферах людської діяльності. Не є виключенням і медицина, зокрема дослідження мікробіому кишківника людини. Для аналізу великих об'ємів даних, отриманих внаслідок секвенування мікробіому, використовуються різноманітні методи машинного навчання.

Ці алгоритми допомагають не тільки краще зрозуміти складні взаємодії між різними видами мікроорганізмів, але й визначити їх вплив на здоров'я людини. При цьому їх ефективність нерідко перевершує традиційні методи аналізу. Тому застосування машинного навчання для вивчення мікробіому кишківника людини відкриває нові перспективи в діагностиці та лікуванні різноманітних захворювань.

1 МЕТОДИ СЕКВЕНУВАННЯ ДНК ТА МІКРОБІОМ ЛЮДИНИ

1.1. Секвенування ДНК

1.1.1. Класифікація методів секвенування ДНК

Секвенування ДНК – це один з методів молекулярної біології, що використовується для визначення послідовності нуклеотидів (A – аденін, G – гуанін,
C – цитозин, T – тимін) в певному фрагменті ДНК. Методи секвенування почали свій розвиток у XX столітті.

Перший метод секвенування був розроблений британським біохіміком, двічі лауреатом Нобелівської премії з хімії – Фредеріком Сегнером та його колегами наприкінці 1970-х років. Також цей метод називають методом обриву ланцюга. Під час реплікації ДНК продовженням ланцюга є приєднання до нуклеозидтрифосфату 3-гідроксильної групи останнього нуклеотиду зростаючого ланцюга. Але якщо нуклеозидтрифосфат буде замінено на дидезоксинуклеотід, то нуклеотид зростаючого ланцюга не зможе утворити зв’язок, таким чином синтез ДНК буде зупинений. Позначивши ddNTP різними флуоресцентними барвниками, було визначено послідовність нуклеотидів у молекулі ДНК. Даний метод секвенування є методом першого покоління. Але він відносно трудомісткий та дорогий у реалізації, тому наразі він використовується достатньо рідко.

Секвенування другого покоління, яке має назву технології секвенування наступного покоління з’явилися в середині 2000-х років. На відміну від першого покоління, ці методи характеризуються високопродуктивним секвенуванням, що дає можливість одночасно секвенувати мільйони фрагментів ДНК. До NGS відноситься технологія секвенування Illumina, ще відома як технологія секвенування шляхом синтезу.

Наступним поколінням є третє покоління секвенування, до якого відноситься технологія секвенування Oxford Nanopore. Дана платформа може генерувати довгі зчитування, чим вирізняється з поміж інших поколінь. За допомогою довгого

зчитування краще отримуються складні геноми, що дозволяє проводити більш глибокі дослідження ДНК.

1.1.2 Секвенування Illumina

Технологія секвенування Illumina була розроблена компанією Illumina, Inc і вперше комерційно реалізована в 2006 році. Це високопродуктивне секвенування, яке має високу точність та відносно низьку вартість бази. Існує три платформи секвенування Ilumina: HiSeq, MiSeq і NovaSeq [3]. Вони мають різну пропускну здатність та довжину зчитування, що зручно для виконання різних задач.

Розглянемо процес секвенування Illumina. Його можна поділити на такі складові:

1. Підготовка бібліотеки: фрагментування ДНК та додавання до кінців адаптерів, що являють собою коротку молекулу ДНК і містять послідовності, необхідні для наступних етапів.
2. Генерація кластерів: підготовлені фрагменти ДНК прикріплюють до твердої поверхні, яка зазвичай покрита комплементарними послідовностями до адаптерів, що дозволяє ДНК зв’язуватись; потім прикріплені фрагменти піддаються мостовій ампліфікації – процес отримання кластерів, які складаються з локалізованих груп ідентичних фрагментів ДНК.
3. Секвенування шляхом синтезу: нуклеотиди помічають флуоресцентними барвниками та послідовно додають в ланцюги ДНК.
4. Отримання даних: флуоресцентні сигнали з помічених нуклеотидів виявляються та записуються.

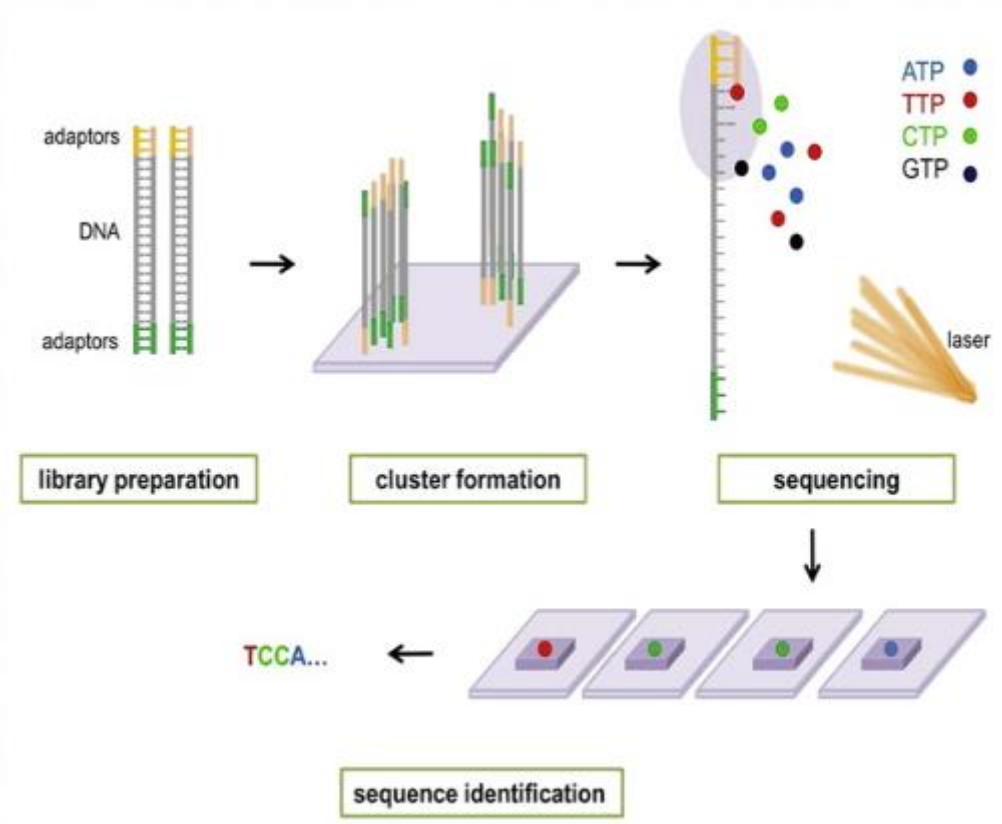

Рисунок 1.1 – Процес секвенування Illumina

Джерело: [7]

Секвенування Illumina має високу пропускну здатність, точність та економічну ефективність, тому воно широко використовується в сучасних дослідженнях.

1.1.3 Секвенування Oxford Nanopore

Oxford Nanopore секвенування було вперше комерційно реалізоване у 2014 році компанією Oxford Nanopore Technologies. Ця технологія використовує нанопори для аналізу окремих молекул ДНК. Нанопори – це маленький білковий канал, вставлений у синтетичну мембрану, діаметр якого розрахований для проходження одно-ланцюгової молекули ДНК.

Коли молекула ДНК захоплюється і рухається крізь нанопору, провідність каналу G_c зменшується, викликаючи падіння вимірюваного іонного струму. Ця зміна провідності і, як наслідок, зміна струму та його тривалості використовуються для характеристики захопленої нанопорою молекули ДНК. Секвенування ДНК за допомогою нанопори використовується з метою ідентифікації послідовності нуклеотидів ДНК на основі змін іонного струму крізь нанопору.

Патч-кламп – метод електрофізіології, який дозволяє ізолювати фрагмент клітинної мембрани з іонними каналами, задавати необхідну різницю потенціалів крізь цей фрагмент клітинної мембрани, створювати по обидві сторони мембрани середовище з певним іонним складом і вимірювати в цих контрольованих умовах струм іонних каналів. В експерименті з постійною керуючою напругою V_c зміна іонного струму I_c свідчить про зміну провідності каналу G_c:

$$V_c = \frac{I_i}{G_c}$$

таким чином, про зміщення значень провідності G_c сигналізує зміна іонного струму I_i. Крізь нанопору проходить 400 баз нуклеотидів за секунду. На Рисунку 1.2 показаний зріз білкового каналу крізь який проходить одно-ланцюгова молекула ДНК та відбувається вимірювання іонного струму.

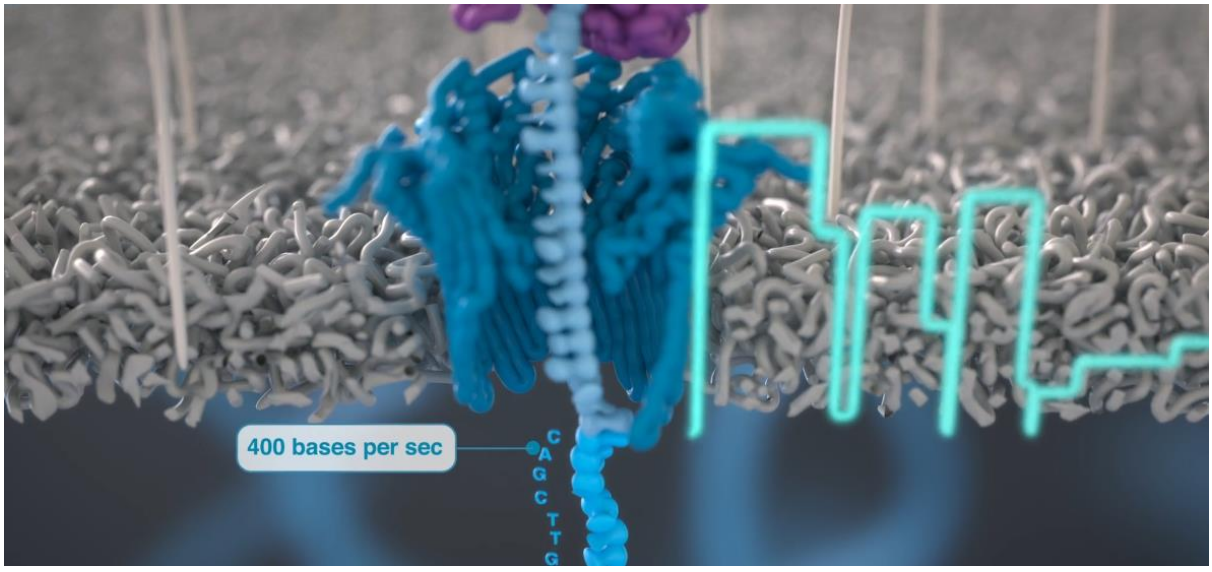

Рисунок 1.2 – Зріз білкового каналу

Джерело: [4]

За рахунок такого швидкого процесу визначення послідовності молекул ДНК, можна генерувати дані в режимі реального часу. На Рисунку 1.4 показано весь процес секвенування за допомогою нанопор.

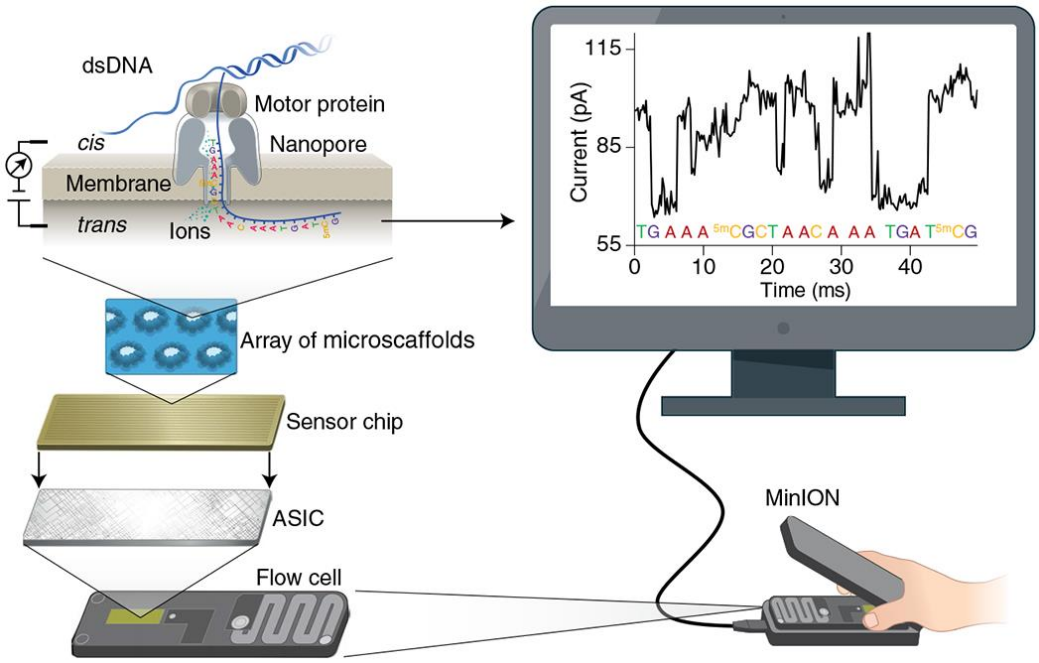

Рисунок 1.3 – Принцип нанопорового секвенування

Джерело: [6]

Важливою особливістю даної технології є генерування довгих послідовностей, що є корисним для дослідження повторюваних ділянок геному, виявлення структурних варіацій та отримання складних геномів. Також не менш важливим є можливість проведення прямого секвенування, тобто без необхідності ампліфікації чи тривалої підготовки зразків, що зменшує зміщення, внесені під час підготовки, та збільшує ймовірність виявлення модифікованих основ. Секвенатори Oxford Nanopore відносно невеликі, тому є більш портативними, ніж інші платформи секвенування, це дає змогу виконувати секвенування у віддаленому режимі.

Дана технологія має широке застосування в різних галузях, що є великим плюсом, але також є певні обмеження. Рівень помилок вищий у порівнянні з іншими методами секвенування. Також вартість бази даної технології є вищою, ніж технологій другого покоління. Але незважаючи і на мінуси, метод секвенування Oxford Nanopore продовжує розвиватись та вдосконалюватись.

1.2 Мікробіом людини та його зв’язок зі здоров’ям людини

1.2.1 Поняття мікробіому

Мікробіом – це сукупність мікроорганізмів, включаючи бактерії, віруси, гриби та інші мікроорганізми, які живуть у певному середовищі. До мікробіому людини належать безліч мікроорганізмів, які дуже тісно пов’язані з її здоров’ям. Ці мікроорганізми взаємодіють між собою і з людським тілом, впливаючи на різні фізіологічні процеси та сприяють належному функціонуванню імунної системи людини, метаболізму, травленню.

1.2.2 Дослідження мікробіому людини, проект “Human Microbiome Project”

Проект “Human Microbiome Project” був започаткований Національним інститутом здоров’я в США у 2007 році, його мета полягала в дослідженні мікробів, які є в організмі людини та їх впливу на здоров’я та хвороби людини[9].

Проводилось дослідження чотирьох різних ділянок людського тіла, а саме: шкіра, ротова порожнина, кишківник та репродуктивна система. Метою було зібрати базу мікроорганізмів, які знаходяться в людському тілі, визначити їх генетичний склад та вплив на здоров’я людини. У проекті було використано передові технології секвенування ДНК для створення мікробної геномної бази. Вивчаючи генетичну інформацію мікробіому, дослідники хотіли ідентифікувати види мікробів, їх взаємодію та виявити потенційні кореляції з конкретними станами здоров’я людини.

HMP мав два основні компоненти: проекти “Reference Genome Sequencing” та “Metagenomic Sequencing”. Проект Reference Genome Sequencing передбачав секвенування геномів культивованих мікроорганізмів, які представляють мікробіом людини. Ця довідкова база даних стала основою для розуміння генетичного різноманіття мікробіому.

Проект Metagenomic Sequencing був зосереджений на аналізі молекул ДНК, які були безпосередньо вилучені зі зразків, без необхідності культивування мікроорганізмів у лабораторії. Це дозволило охарактеризувати мікробні спільноти в їх природному стані.

Отримані дані значно розширили знання про мікробіом людини та його зв’язок зі здоров’ям і хворобами. Проект дав уявлення про різноманітність і динаміку мікробних спільнот у різних частинах тіла.

HMP проклав шлях до подальших досліджень і став каталізатором дослідження мікробіомів у всьому світі. Він підкреслив важливість мікробіома для здоров’я людини та заклав основу для подальшого вивчення терапевтичного та діагностичного потенціалу маніпулювання мікробіомом.

1.2.3 Функції та класифікація мікробіому людини

Основними з багатьох функцій мікробіому є:

1. Захист організму людини від шкідливих мікроорганізмів та сполук
2. Підтримка гомеостазу
3. Покращення процесів регулювання життєво важливих функцій

Мікробіом людини розглядається в межах 5 ключових місць Рисунок 1.5, з яких було взято матеріал для дослідження проектом HMP.

Рисунок 1.4 – Ключові місця мікробіому людини

Шкіра – найбільший орган тіла людини, де проживають найрізноманітніші види мікробів. Мікробіом шкіри бере участь у підтримці бар'єрної функції шкіри, є важливим захисником від патогенів і регулятором імунних реакцій. Склад мікробіому шкіри залежить від навколишнього середовища, гігієни та віку людини. Одними з найрозповсюджених бактерій на шкірі людини є: стафілокок, стрептокок та пропіоновокислі бактерії. Більшість з вказаних бактерій не патогенні та є частиною нормального мікробіому шкіри людини. Але патогенні види даних бактерій можуть викликати гнійні зараження, пневмонію, харчові отруєння, акне та інші запальні захворювання шкіри[13].

Незважаючи на те, що шкіра є найбільшим органом, але найбільш дослідженим на даний час є мікробіом кишківника. Він відіграє важливу роль у травленні, засвоєнні поживних речовин з їжі, метаболізмі та формуванні імунної системи людини. Даний мікробіом налічує тисячі видів різних бактерій, наявність певних з яких може зробити подальші прогнозування різних захворювань і попередити можливі ризики для здоров'я людини. Наприклад, бактерія Helicobacter pylori – це кислотостійка бактерія і може перебувати у шлунку людини ще з ранніх років її життя, за даними дослідників, ця бактерія присутня у мікробіомі майже у

половини людей, що населяють нашу планету. Спочатку ця бактерія впливає безсимптомно на здоров'я, тому без проведення спеціальних досліджень, таких як аналіз калу на антиген Helicobacter pylori, дихальний тест або аналіз крові на антитіла до Helicobacter pylori, неможливо виявити наявність бактерії в шлунку людини. Ця бактерія розповсюджується при прямому контакті, зазвичай може передаватись через їжу та воду. У більшості випадків інфікування відбувається в ранньому дитинстві. Ключовими причинами зараження можуть бути: проживання в оселях з великою кількістю родичів, відсутність постійного водопостачання, постійний контакт з носіями хелікобактерної інфекції. Через досить тривалий час можуть з'явитись такі симптоми як: біль і печія в шлунку, нудота, здуття живота, втрата маси тіла. Ці і ще більш критичні симптоми можуть сигналізувати про серйозні важкі захворювання шлунково-кишкового тракту людини. А саме, це може бути як і виразкова хвороба шлунка, запалення слизової оболонки шлунка, так і рак шлунка, від якого, не почавши вчасно лікуватись, людина може померти за два роки. Загалом бактерія Helicobacter pylori є одним з ключових чинників, що сприяють розвитку деяких типів раку. Дякуючи австралійським дослідникам Баррі Дж. Маршаллу і Робіну Уоррену, які відкрили бактерію Helicobacter pylori і розшифрували її роль у розвитку гастриту та виразкової хвороби і отримали за це Нобелівську премію [10]. Наразі людство може превентивно вилучити бактерії з організму, щоб запобігти розвитку таких ускладнень як рак шлунку. Отже, на цьому прикладі можна зрозуміти, що дослідження мікробіому шлунково-кишкового тракту є дуже важливим і несе за собою безліч позитивних чинників для більш глибокого аналізу даних, аби застерегти людей від ще більшого радіусу захворювань пов'язаних не тільки зі шлунком, а і з усім організмом людини. Оскільки мікробіом кишківника пов'язаний з такими захворюваннями як запальні захворювання кишківника, хворобою Крона, діабетом другого типу, ожирінням, метаболічними розладами та навіть розладами психічного здоров'я.

Мікробіом порожнини рота включає в себе бактерії, грибки, віруси та інші мікроорганізми, які проживають в роті, а саме в середині та ззовні на яснах, зубах, слині і язику. Даний мікробіом теж грає неабияку роль в здоров'ї людини, оскільки

він допомагає розщеплювати частинки їжі, підтримує здоровий баланс у ротовій порожнині та захищає від шкідливих бактерій. Дислабанс мікробіома ротової порожнини можна наглядно спостерігати у людей з карієсом на зубах, захворюванням ясен і неприємного запаху з рота. Отже, дослідження даного мікробіому важливе для вибору якісних препаратів як для підтримання балансу мікробіому, так і для лікування захворювань порожнини рота.

Мікробіом носа – це сукупність бактерій, які живуть в носовій порожнині, допомагаючи підтримувати здорову дихальну систему, виконуючи функцію захисту від патогенів та впливаючи на імунну відповідь організму. Бактерії взаємодіють з епітелієм всередині носа та виробляють антимікробні речовини. Склад даного мікробіому може залежати від генетики, чинників з навколишнього середовища та гігієни. Дисбаланс мікробіома носової порожнини може викликати синусит, закладеність носа, респіраторні інфекції. Дослідження мікробіома носа важливе, оскільки розуміючи його взаємодію з органами дихання, можна розробити нові методи лікування та знизити підвищену чутливість організму людини до будь-якого алергену індивідуально[12].

Мікробіом репродуктивної системи ще додатково поділяється на жіночий та чоловічий. До нього відносяться мікроби, які знаходяться в репродуктивних шляхах. Більше досліджується вагінальний мікробіом, оскільки він відіграє вирішальну роль у репродуктивному здоров’ї жінки. Бактерії, які населяють жіночу репродуктивну систему підтримують кислий pH, запобігаючи розмноженню шкідливих бактерій, захищають від різних інфекцій і підтримують репродуктивну здатність. Зміна у жіночому репродуктивному мікробіомі може бути причиною захворювань таких, як інфекція сечовивідних шляхів, бактеріальний вагіноз, дріжджова інфекція. Менш дослідженим є мікробіом чоловічої репродуктивної системи, включно з уретрою та передміхуровою залозою. Але дисбаланс даного мікробіому теж впливає на репродуктивне здоров’я чоловіків. На мікробіом репродуктивної системи як у жінок, так і в чоловіків, можуть впливати гормональні зміни, сексуальна активність, гігієна, застосування антибіотиків, дієта, спосіб життя, який може в свою чергу залежати від регіону проживання. Даний мікробіом

є динамічною екосистемою, склад якої може бути різний в залежності від сукупності вище зазначених чинників. Дослідження мікробіому репродуктивної системи є достатньо важливими для здоров'я людства. Саме вагінальний мікробіом жінок, оскільки на відміну від чоловіків, він має певні тонкощі і від нього більше залежить продовження роду. Розуміння функцій та складу вагінального мікробіому дає можливість визначити дії, які можуть сприяти кращим результатам та веденню вагітності в цілому для матері та плоду. Дані щодо впливу мікробіому на запліднення можуть допомогти при розробці новітніх методів втручання задля покращення лікування безпліддя. Можливість проаналізувати мікробіом на індивідуальному рівні забезпечить прогнозування лікування чи інфекційного захворювання, чи безпліддя індивідуально, враховуючи всі можливі ризики для конкретної особи з індивідуальним складом мікробіому. Дослідження все ще розвиваються, але перспективи великі, особливо для здоров'я жінок. Розуміючи найменші деталі мікробіому репродуктивної системи, медицина може вийти на новий рівень в діагностиці та терапевтичному втручанні.

1.2.4 Взаємодія та вплив мікробіому на здоров'я людини

Після досліджень основних мікробіомів, вчені почали все глибше вивчати вплив мікробіому на здоров'я і хвороби людини. Почались дослідження потенційних застосувань науки про мікробіоми у лікуванні та профілактиці. Наразі можна окреслити такі галузі досліджень:

1. Зв'язок кишківник-мозок.

Досліджуючи мікробіом кишківника виникає все більше доказів двонаправленої системи зв'язку між кишківником та мозком, відомої як вісь кишківник-мозок. Мікробіом може впливати на роботу мозку та поведінку людини, а саме на її настрій, реакції та стрес. В свою чергу також тривога, депресія,

постійний стрес впливають на дисбаланс мікробіому кишківника, що може призвести до багатьох захворювань шлунково-кишкового тракту.

2. Метаболічні розлади.

Як раніше було розглянуто, мікробіом допомагає перетравленню їжі, вилученню із неї поживних речовин та енергії. Отже, дисбаланс мікробіому кишківника пов’язаний з такими метаболічними розладами як ожиріння, діабет другого типу, метаболічний синдром. Тому вчені займаються розробкою стратегій маніпулювання мікробіомом для того щоб заздалегідь виявити, не допустити та вилікувати дані розлади.

3. Вплив на імунну систему.

Мікробіом допомагає модулюватись та розвиватись імунітету людини, тим самим захищаючи організм від шкідливих патогенів і зберігаючи толерантність до корисних та нешкідливих речовин. Аутоімунні захворювання, алергія, астма – це захворювання які пов’язані з дисбалансом мікробіома людини, який вплинув на імунітет. На прикладі алергії – це підвищена чутливість до будь-якого алергену, що при нормальному балансі мікробіому не відбувається. Отже, дослідивши способи маніпулювання мікробіомом, ймовірно в майбутньому дані захворювання не будуть турбувати людство[12].

4. Індивідуальна медицина.

Вчені досліджують вплив індивідуальних варіацій в мікробіомі на метаболізм ліків і реакцію на лікування. Проаналізувавши специфіку мікробіома пацієнта, можна оптимізувати його лікування задля мінімізації побічних ефектів.

5. Зв’язок мікробіома з конкретними захворюваннями.

Численні аналізи мікробіомів дали змогу виявити зв’язок мікробіома з конкретним захворюванням людини. Наприклад, зміни в мікробіомі кишківника пов’язуються з такими захворюваннями як синдром подразненого кишківника, діабетом другого типу, захворюванням Крона, запальними захворюваннями

кишечника, колоректальним раком. Тому наразі проводяться дослідження, які допоможуть зрозуміти які саме зміни в ДНК мікробіома можуть відповідати конкретному захворюванню.

6. Терапевтичне застосування.

Одним з новітніх терапевтичних застосувань досліджень стало маніпулювання мікробіомом задля лікування захворювань. Одним з винайдених та перспективних методів є трансплантація фекальної мікробіоти – це перенесення фекального матеріалу від здорового донора до хворого для відновлення балансу в мікробіомі. Позитивна реакція була в лікуванні інфекцій товстого кишківника. Також досліджується такий метод для лікування інших захворювань, таких як запальні захворювання кишківника та метаболічні розлади.

Наразі ця сфера інтенсивно розвивається, тому знання про специфічність та взаємодію мікробіому з організмом людини слід більше поглиблювати.

1.3 Висновки до першого розділу

Знання про мікробіом може врятувати не одне життя, тому що мікробіом та здоров’я людини – це тісно пов’язані між собою поняття. Вище була приведена лише частина прикладів щодо цієї сфери досліджень. Вивчаючи та аналізуючи отримані дані, вчені прагнуть розробити новітні технології діагностики та інноваційні підходи профілактики та лікування хвороб, пов’язаних з усім організмом людини. Також варто зазначити про індивідуальність, яку забезпечує ця сфера досліджень: персоналізовані втручання та рекомендації щодо дієти та способу життя, а не стандартний рецепт для всіх людей з певним захворюванням. Не менш важливим є раннє виявлення

захворювань, щоб, за допомогою індивідуальних ліків, запобігти розвитку хвороби.

Галузь досліджень мікробіома швидко розвивається і має позитивні наслідки для здоров'я людини. Розгадуючи тонкощі мікробіома, вчені прагнуть революціонізувати лікування пацієнтів та профілактику здорового населення.

2 ОБРОБКА ТА АНАЛІЗ ДАНИХ СЕКВЕНУВАННЯ ДНК

2.1 Аналіз методів дослідження сигналів нанопорового секвенування

Нанопорове секвенування — це технологія секвенування ДНК третього покоління, яка має певні переваги порівняно з попередніми поколіннями, включаючи збір даних у реальному часі, можливість тривалого зчитування та потенціал для секвенування однієї молекули та прямого РНК. Принцип секвенування нанопор полягає в тому, що одна молекула ДНК або РНК переміщується крізь нанопору, а зміни в іонному струмі крізь пору, викликані унікальною формою кожного типу нуклеотиду, виявляються та використовуються для визначення послідовності.

Існує декілька основних методів, які зазвичай використовуються в аналізі сигналів нанопорового секвенування:

1. Попередня обробка зареєстрованого сигналу: це початковий крок, на якому необроблений електричний сигнал обробляється, щоб зробити його придатним для аналізу. Мета полягає в тому, щоб зменшити шум і зміщення.
2. Сегментація: цей крок передбачає поділ обробленого сигналу на сегменти, кожен з яких представляє нуклеотид.
3. Визначення нуклеотидних основ за сигналом струму крізь нанопору: це процес перетворення сигналу від секвенатора нанопор у рядок нуклеотидних основ.
4. Вирівнювання та визначення відхилень від еталонних геномів: коли послідовність визначена, її можна вирівняти за еталонним геномом. Тоді варіанти (відмінності від посилання) можна викликати за допомогою засобів виклику варіантів, таких як Nanopolish або Medaka.
5. Контроль якості та виправлення помилок: як і всі методи секвенування, нанопорове секвенування має помилки. Це можуть бути

помилки вставки, видалення або заміни. Методи виправлення помилок можуть бути використані для виправлення цих помилок або за допомогою послідовності з кількох зчитувань тієї самої молекули, або за допомогою інформації з високоякісного еталонного геному. Метрики контролю якості, такі як Q-оцінки, можна розрахувати, щоб дати вказівку на надійність базових викликів. Прикладом реалізації Q-оцінки може бути порівняння з еталоном: можна порівняти отримані послідовності з ним і обчислити Q-оцінку на основі відповідності та відмінностей між ними. Для цього після підрахунку збігів, незбігів та вставок використовується дана формула:

$$Q\text{-оцінка} = \frac{\text{збіги} - \text{незбіги}}{\text{збіги} + \text{незбіги} + \text{вставки}}$$

6. Аналіз прямого секвенування РНК: однією з унікальних особливостей секвенування нанопор є можливість безпосередньо секвенувати РНК і виявляти модифіковані нуклеотиди, що дозволяє аналізувати транскрипти РНК. Для цього потрібні спеціальні інструменти для обробки сигналів і визначення послідовності нуклеотидів, а також для виявлення й аналізу модифікацій.

2.2 Робота з даними секвенування ДНК у FASTQ, FAST5, SFF форматах

Формат FASTQ є стандартним виводом секвенування другого покоління, приклад якого наведений на рисунку 2.1.

```
@SEQ_ID
ACGTGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGC
+
BBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB
```

Рисунок 2.1 – Приклад файлу формату FASTQ

На Рисунку 2.1 показано невеликий приклад файлу формату FASTQ. Перший рядок після «@» вказує ідентифікатор послідовності. Другий рядок містить послідовність ДНК. У третьому рядку знак «+» існує як роздільник. А четвертий рядок містить, відповідні до другого рядку, показники якості. В більш реальних файлах FASTQ часто розміщуються декілька записів один за одним для різних послідовностей ДНК[15].

Показниками якості виступають звичайні букви і символи. Оцінки Q кодуються як символи з системи кодів ASCII (Рисунок 2.2). Дана система містить в собі відповідність від 0 до 127, де кожне число відповідає певним літерам, цифрам та іншим існуючим символам. У форматі FASTQ використовується оцінювання від 0, що відповідає символу ‘!’, до 93 – ‘~’. Отже, щоб розшифрувати показник якості можна відняти 33 від значення із системи кодів ASCII символу якості з формату FASTQ. На прикладі з Рисунку 2.1 якість ‘B’ відповідає 33 із 93 оцінок якості Q.

ASCII TABLE

Decimal	Hexadecimal	Binary	Octal	Char
0	0	0	0	[NULL]
1	1	1	1	[START OF HEADING]
2	2	10	2	[START OF TEXT]
3	3	11	3	[END OF TEXT]
4	4	100	4	[END OF TRANSMISSION]
5	5	101	5	[ENQUIRY]
6	6	110	6	[ACKNOWLEDGE]
7	7	111	7	[BELL]
8	8	1000	10	[BACKSPACE]
9	9	1001	11	[HORIZONTAL TAB]
10	A	1010	12	[LINE FEED]
11	B	1011	13	[VERTICAL TAB]
12	C	1100	14	[FORM FEED]
13	D	1101	15	[CARRIAGE RETURN]
14	E	1110	16	[SHIFT OUT]
15	F	1111	17	[SHIFT IN]
16	10	10000	20	[DATA LINK ESCAPE]
17	11	10001	21	[DEVICE CONTROL 1]
18	12	10010	22	[DEVICE CONTROL 2]
19	13	10011	23	[DEVICE CONTROL 3]
20	14	10100	24	[DEVICE CONTROL 4]
21	15	10101	25	[NEGATIVE ACKNOWLEDGE]
22	16	10110	26	[SYNCHRONOUS IDLE]
23	17	10111	27	[END OF TRANS. BLOCK]
24	18	11000	30	[CANCEL]
25	19	11001	31	[END OF MEDIUM]
26	1A	11010	32	[SUBSTITUTE]
27	1B	11011	33	[ESCAPE]
28	1C	11100	34	[FILE SEPARATOR]
29	1D	11101	35	[GROUP SEPARATOR]
30	1E	11110	36	[RECORD SEPARATOR]
31	1F	11111	37	[UNIT SEPARATOR]
32	20	100000	40	[SPACE]
33	21	100001	41	!
34	22	100010	42	"
35	23	100011	43	#
36	24	100100	44	$
37	25	100101	45	%
38	26	100110	46	&
39	27	100111	47	'
40	28	101000	50	(
41	29	101001	51	)
42	2A	101010	52	*
43	2B	101011	53	+
44	2C	101100	54	,
45	2D	101101	55	-
46	2E	101110	56	.
47	2F	101111	57	/
48	30	110000	60	0
49	31	110001	61	1
50	32	110010	62	2
51	33	110011	63	3
52	34	110100	64	4
53	35	110101	65	5
54	36	110110	66	6
55	37	110111	67	7
56	38	111000	70	8
57	39	111001	71	9
58	3A	111010	72	:
59	3B	111011	73	;
60	3C	111100	74	<
61	3D	111101	75	=
62	3E	111110	76	>
63	3F	111111	77	?
64	40	1000000	100	@
65	41	1000001	101	A
66	42	1000010	102	B
67	43	1000011	103	C
68	44	1000100	104	D
69	45	1000101	105	E
70	46	1000110	106	F
71	47	1000111	107	G
72	48	1001000	110	H
73	49	1001001	111	I
74	4A	1001010	112	J
75	4B	1001011	113	K
76	4C	1001100	114	L
77	4D	1001101	115	M
78	4E	1001110	116	N
79	4F	1001111	117	O
80	50	1010000	120	P
81	51	1010001	121	Q
82	52	1010010	122	R
83	53	1010011	123	S
84	54	1010100	124	T
85	55	1010101	125	U
86	56	1010110	126	V
87	57	1010111	127	W
88	58	1011000	130	X
89	59	1011001	131	Y
90	5A	1011010	132	Z
91	5B	1011011	133	[
92	5C	1011100	134	\
93	5D	1011101	135	]
94	5E	1011110	136	^
95	5F	1011111	137	_
96	60	1100000	140	`
97	61	1100001	141	a
98	62	1100010	142	b
99	63	1100011	143	c
100	64	1100100	144	d
101	65	1100101	145	e
102	66	1100110	146	f
103	67	1100111	147	g
104	68	1101000	150	h
105	69	1101001	151	i
106	6A	1101010	152	j
107	6B	1101011	153	k
108	6C	1101100	154	l
109	6D	1101101	155	m
110	6E	1101110	156	n
111	6F	1101111	157	o
112	70	1110000	160	p
113	71	1110001	161	q
114	72	1110010	162	r
115	73	1110011	163	s
116	74	1110100	164	t
117	75	1110101	165	u
118	76	1110110	166	v
119	77	1110111	167	w
120	78	1111000	170	x
121	79	1111001	171	y
122	7A	1111010	172	z
123	7B	1111011	173	{
124	7C	1111100	174	|
125	7D	1111101	175	}
126	7E	1111110	176	~
127	7F	1111111	177	[DEL]

Рисунок 2.2 – Система кодів ASCII

Джерело[26]

Розмір файлів FASTQ може сягати кілька мегабайт або гігабайт.[16]

Формат FAST5 є стандартним виводом нанопорового секвенування. На відміну від FASTQ, даний формат є двійковим[15]. Тому формат FAST5 дозволяє зберігати не тільки послідовності ДНК а і сигнал нанопори.

Робота з даними секвенування складається з декількох етапів:

1. Попередня обробка:

- Зчитування: дані у форматі SFF потрібно конвертувати у формат FASTQ.

- Адаптер і обрізка якості: обрізання послідовності адаптерів і бази низької якості з з прочитань послідовності нуклеотидів.

2. Контроль якості:

- Оцінка якості зчитування: відбувається за допомогою таких інструментів як FastQC.

- Фільтрація низькоякісних зчитувань: видалення даних зчитувань послідовності ДНК з низькими середніми показниками якості або короткою довжиною на основі оцінки якості.

3. Вирівнювання та складання:

- Вирівнювання: вирівнювання зчитування з еталонним геномом або транскриптомом за допомогою вирівнювачів. Цей крок допомагає ідентифікувати геномне розташування або транскрипт, з якого походить кожне зчитування.

- Складання: якщо наявні невідповідні зчитування або виконується складання заново, можна відновити оригінальну послідовність ДНК із зчитувань, що накладаються.

4. Виклик і аналіз варіантів:

- Виклик варіантів: ідентифікація генетичних варіантів з вирівняних зчитувань. Цей крок допомагає виявити геномні варіації між секвенованим зразком і еталонним геномом.

- Анотація фрагментів, відмінних від еталонних: додавання до виявлених відмінних фрагментів інформації про їхні функціональні ефекти, популяційні частоти та асоціації захворювань.

Для всіх пунктів існують певні програмні засоби і конвеєри, які можуть оптимізувати аналіз залежно від характеристик даних та мети аналізу.

Для аналізу за допомогою мови Python дані формату FASTQ потрібно конвертувати, оскільки буде використовуватись лише частина файлу, де вказана послідовність ДНК. В даній роботі використовувалося конвертування у формат PKL.

Python Pickle (PKL) – це двійковий формат Python, який використовується для серіалізації об’єктів.

Щоб перетворити дані ДНК з формату FASTQ в PKL використовується бібліотека ‘biopython’, яка має інструменти для роботи з біологічними послідовностями. Приклад скрипта для перетворення представлено у Додатку А. Вказаний скрипт виконує зчитування файлу FASTQ, витягнення послідовностей ДНК та збереження їх у вказаний вихідний файл формату PKL.

2.3 Методи машинного навчання

Найбільший успіх у сфері алгоритмів отримали методи машинного навчання, метою роботи яких є автоматизація процесів ухвалення рішень за допомогою узагальнення відомих знань. В цій роботі було використано машинне навчання з вчителем, що являє собою навчання випробувальної моделі на вже відомих прикладів («стимул-реакція») з метою подальшого визначення моделлю «реакціїї» для майбутніх невідомих «стимулів». Даний алгоритм можна використовувати далі з новими даними і отримувати певний відсоток правильних відповідей. Точність алгоритму залежить від правильності обрання методу машинного навчання для поточної задачі, а також від встановлених параметрів моделі.

Під час роботи з даними секвенування ДНК мікробіому кишківника людини для діагностики захворювань буде використано 4 методи машинного навчання: метод k-найближчих сусідів (KNeighbors clasifier), адаптивне підвищення (Adaboost classifier), дерево рішень (Decision Tree classifier) та випадковий ліс (Random Forest classifier). Далі розглянуто основні їх особливості.

2.3.1 Метод k-найближчих сусідів

Це тип навчання на основі екземплярів або метод неузагальнюючого навчання. Він використовується для класифікації завдань у машинному навчанні та працює за достатньо простим принципом:

1. Визначення метрики для відстані: це може бути евклідова відстань або будь-який інший тип метрики відстані, який відповідає даним. Як приклад, відстань між двома точками P та Q в n-мірному евклідовому просторі буде дорівнювати

$$d(P,Q) = \sqrt{((Q_1 - P_1)^2 + (Q_2 - P_2)^2 + \cdots + (Q_n - P_n)^2}$$

2. Обчислення відстані від нової точки (тої, для якої потрібно передбачити мітку) до всіх існуючих точок у вхідному наборі даних для навчання.

3. Сортування відстані та вибір `k` найближчих сусідів.

4. Кожен із `k` сусідів має право голосувати за прогнозований клас нової точки. У разі рівності вибирається клас найближчого сусіда серед рівних класів.

2.3.2 Дерево рішень

Це контрольований метод навчання, який використовується для завдань класифікації та регресії. Метою дерева рішень є створення моделі, яка передбачає значення цільової змінної шляхом вивчення простих правил прийняття рішень, виведених із характеристик даних.

Принцип роботи:

1. Створення кореневого вузла: дерево рішень починається з кореневого вузла, який містить увесь набір даних. Потім кореневий вузол розбивається на дочірні вузли.

2. Розділення: у кожному вузлі рішення приймається на основі функції. Мета цієї функції полягає у поділі даних таким чином, щоб підмножини були максимально чистими, тобто кожна підмножина повинна містити точки даних, які належать переважно до одного класу.

3. Обрізання: включає зрізання гілок дерева для спрощення моделі та покращення її здатності до узагальнення невидимих даних.

4. Лицевий вузол: кінцеві вузли, які передбачають результат (класифікацію або регресію). Коли нова точка даних дотримується правил прийняття рішень від кореневого вузла до кінцевого вузла. Значення або клас цього кінцевого вузла є прогнозом моделі для нової точки даних.

2.3.3 Випадковий ліс дерев рішень

Випадковий ліс дерев рішень – це модель, яка підбирає кілька класифікаторів дерева рішень для різних підвибірок набору даних і використовує усереднення для підвищення точності прогнозування та контролю за переобладнанням, використовується для завдань класифікації.

Принцип роботи:

1. Початкова вибірка: випадкові ліси починаються зі створення кількох підмножин вхідного набору даних, випадкового вибору із заміною. Кожна з цих підмножин використовується для навчання окремого дерева рішень.

2. Побудова дерева: для кожної підмножини будується дерево рішень. Але на відміну від стандартного дерева рішень, де на кожному вузлі всі функції оцінюються, щоб знайти найкращий поділ, у випадкових лісах для поділу на кожному вузлі розглядається лише випадкова підмножина функцій. Це вносить більше різноманітності в ліс і робить модель «міцнішою».

3. Передбачення: під час прогнозування, для класифікації вибирається клас із більшістю голосів з усіх дерев.

2.3.4 AdaBoost класифікатор

Це алгоритм машинного навчання, який використовується як класифікатор. При використанні з навчанням дерева рішень інформація, зібрана на кожному етапі алгоритму, подається в алгоритм вирощування дерева таким чином, що наступні дерева, як правило, зосереджуються на прикладах, які складніше класифікувати.

Принцип роботи:

1. Ініціалізація ваги: алгоритм починає з призначення однакових ваг усім прикладам навчання.

2. Навчання слабких учнів: для кожної ітерації модель намагається правильно класифікувати навчальні дані, базуючись на тому, наскільки добре він працював під час ітерації, алгоритм призначає вищі ваги спостереженням, які йому не вдалося правильно класифікувати.

3. Оновлення вагових коефіцієнтів: після навчання кожного класифікатора ваги перерозподіляються. Неправильно класифіковані бали отримують вищу вагу.

4. Остаточний прогноз: Остаточний прогноз є зваженою сумою прогнозів, зроблених попередніми класифікаторами.

3 РОЗРОБКА ПРОГРАМНОГО ЗАБЕЗПЕЧЕННЯ ДЛЯ РОЗПІЗНАВАННЯ ЗАХВОРЮВАНЬ ЛЮДИНИ ЗА АНАЛІЗОМ ДАНИХ СЕКВЕНУВАННЯ МІКРОБІОМУ КИШКІВНИКА

3.1 Використання методів машинного навчання

3.1.1 Підготовка даних для аналізу

На виході після секвенування ДНК мікробіому людини отримуються дані формату FASTQ, приклад на Рисунку 3.1

```
@MM123:002:FC123AB:3:2208:3330:9840 2:Y:18:ATCACG
AGGATACTAGCATAGATACCCTAGATAGTCATAGATCATGATAGGGAGATCTA
+
IJJJJJJIIIIIJIIIIIFFFEEEEEDDDDDDCABBBBB@@00))))*(*&%!
```

Рисунок 3.1 – Приклад даних формату FASTQ

Джерело[17]

Щоб далі працювати з даними було відокремлено послідовності ДНК від ідентифікатора та показників якості. Для цього було конвертовано дані формату FASTQ у формат PKL. Щоб отримати список послідовностей формату PKL використовується модуль ‘pickle’.

Для аналізу зчитаних послідовностей було використано метод k-mer. K-mer – це підпослідовності довжиною k заданої послідовності ДНК[18]. На Рисунку 3.2 представлено приклад поділу певної послідовності на k-mer з довжиною 4.

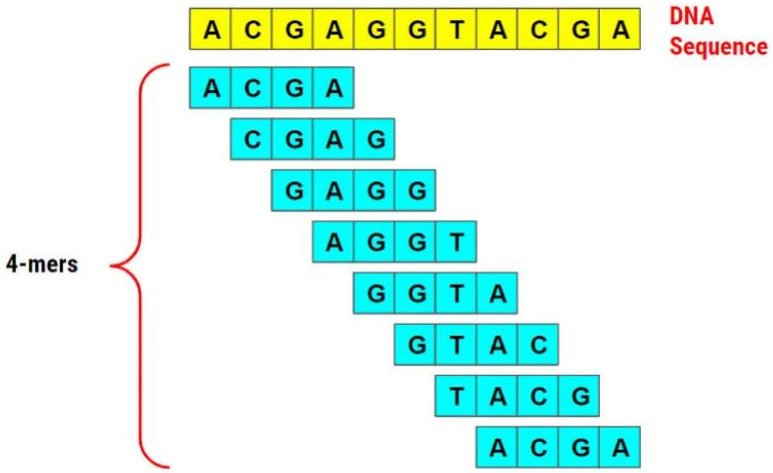

Рисунок 3.2 – Приклад розбиття послідовності на k-mer

Джерело[18]

Приклад генерування k-mer з послідовності ДНК формату PKL представлено у Додатку Б. Довжина k-mer налаштовується залежно від конкретної задачі. Приклад вихідних даних з поділом послідовності ДНК з k-mer = 6 приведений на Рисунку 3.3.

['ACGAGT CGAGTA GAGTAG AGTAGA GTAGAC TAGACT AGACTA GACTAG ACTAGC CTAGCA TAGCAT AGCATA GCATAA CATAAT ATAATT TAATTT AATTTC ATTTCC TTTCCG TTCCGT TCCGTT CCGTTT CGTTTG GTTTGG']

Рисунок 3.3 – Вигляд масиву з k-mer = 6

Після такого поділу було використано методи обробки природньої мови для задачі дослідження мікробіому. Обробка природної мови надає можливість комп’ютерам здатності розуміння тексту[20]. Було використано модель Bag of Words – просту техніку обробки природної мови, основною ідеєю якої є перетворення фрагменту тексту в числову форму, рахуючи скільки разів конкретне слово або словосполучення зустрічалось у тексті. Але важливо зазначити, що дана модель не враховує порядок або структуру слів у тексті.

Зв’язок між k-mers і моделлю Bag of Words полягає в їхньому підході до виділення ознак і представлення текстових даних. Обидва методи спрямовані на

охоплення важливих характеристик тексту, але на різних рівнях деталізації. K-mers вловлюють локальні шаблони, розглядаючи суміжні символи або слова, а модель Bag of Words представляє текст як набір слів і підраховує їх частоти. Тому в даній роботі було використано обидва методи.

Довжина зчитувань ДНК відрізняється, отже, і кількість k-mer буде відрізнятись. Але модель Bag of Words має можливість аналізу текстових даних з різною довжиною. Для врахування цієї довжини нормується частота появу k-mer до кількості k-mer у поточному зчитуванні ДНК, даний термін в обробці природної мови називається ‘term frequency’

$$TF = \frac{N_k}{N_{kt}}$$

Модель Bag of Words була створена за допомогою модуля CountVectorizer (бібліотека ‘sklearn’), який рахує кількість слів, які зустрічаються в файлі, в досліджуваному випадку це кількість певних послідовностей k-mer (скрипт представлений у Додатку В). На Рисунку 3.4 зображено результат скрипту, в якому можна побачити матрицю, стовпці якої вказують на унікальну послідовність нуклеотидів, а рядки на певну послідовність ДНК, аналіз якої проводився.

	AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT
0	8	7	9	14	13	10	5	4	9	6	22	8	8	10	8	5
1	15	7	10	9	10	7	10	10	11	11	4	7	5	13	8	9
2	8	22	20	16	23	27	24	20	23	24	22	22	11	21	26	28
3	3	8	4	10	13	8	12	11	5	12	10	11	4	17	11	7
4	10	4	9	7	11	15	13	2	6	18	9	5	3	4	8	22

Рисунок 3.4 – Приклад використання моделі Bag of Words для k-mer = 2

3.1.2 Початок роботи з обробленими даними

В даній роботі досліджувалися дані секвенування мікробіому кишківника для умовно здорових людей, пацієнтів із синдромом подразненого кишківника,

хворобою Крона та діабетом другого типу. Таким чином, було поставлено задачу мультикласової класифікації на 4 класи.

Для початку аналізу було створено DataFrame з бібліотеки ‘pandas’ з бітової послідовності файлу PKL за допомогою бібліотеки ‘pickle’. Також для зручності вибору необхідного файлу було використано бібліотеку ‘tkinter’ для виклику системного вікна вибору файлу. Для зручності аналізу послідовності ДНК та відповідні до них класи було записано в масиви.

Отримані дані було розбито на два набори, навчальний і тестовий, а також було додатково їх перемішано, що дало змогу одразу перевірити модель не завантажуючи інші файли з даними.

Для початку роботи з моделями було використано дані з 12000 послідовностей для кожного класу, k-mer = 5 та парами класів: ‘Healthy’ та ‘Chrons’, ‘Healthy’ та ‘IBS’, ‘Healthy’ та ‘T2D’.

Оцінка якості класифікації була проведена за допомогою показника ‘accuracy’. Це найбільш інтуїтивно зрозумілий показник, порівняно з іншими можливими, являє собою співвідношення правильно зроблених класифікацій до загальної кількості. Визначається за даною формулою

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

де TP, FP, FN та TN це кількість істинно позитивних, хибно-позитивних, хибно-негативних та істинно негативних результатів.

3.1.3 Класифікатор на основі методу k-найближчих сусідів

Було проведено аналіз секвенованих даних ДНК з довжиною k-mer = 5 для 2-класової класифікації даних послідовностей умовно здорової людини та з хворобою за допомогою класифікатору на основі методу k-найближчих сусідів.

Отримані результати якості класифікації методом k-найближчих сусідів для пар здоровий та хвороба подано у таблиці 3.1.

Таблиця 3.1 – Результат аналізу методом k-найближчих сусідів

Хвороба	Тестові дані	Навчальні дані
Crohn`s	68.08%	100.00%
IBS	66.60%	100.00%
T2D	59.27%	94.98%
Середнє значення	64.65%	

Результати аналізу таблиці 3.1 свідчать про те, що найточніше вдалось діагностувати хворобу Крона. Але показник 68% є недостатнім для використання цього методу в подальших дослідженнях. Отже, дана модель не підходить для поставленої задачі.

3.1.4 AdaBoost класифікатор

Було проведено аналіз даних секвенування ДНК з довжиною k-mer = 5 за допомогою AdaBoost класифікатору для 2-класової класифікації даних з послідовностями умовно здорової людини та з хворобою. Отримані результати якості класифікації методом AdaBoost подано у таблиці 3.2.

Таблиця 3.2 – Результати аналізу методом AdaBoost

Захворювання	Тестові дані	Навчальні дані
Crohn`s	85.35%	85.58%
IBS	81.88%	82.56%
T2D	99.95%	99.67%
Середнє значення	89.06%	

Результати аналізу таблиці 3.2 свідчать про те, що найточніше вдалось діагностувати діабет другого типу, оскільки відповідний показник вказує майже на 100% правильність того, що модель визначає послідовність ДНК мікробіому людини, яка має хворобу діабет другого типу при 2-класовій класифікації.

Найнижчий відсоток правильності вибору припадає на синдром подразненого кишківника.

3.1.5 Класифікатор на основі дерев рішень

Було проведено аналіз секвенованих даних ДНК з довжиною k-mer = 5 для 2-класової класифікації даних послідовностей умовно здорової людини та з хворобою за допомогою класифікатору на основі дерев рішень. Отримані результати якості класифікації методом дерев рішень подано у таблиці 3.3.

Таблиця 3.3 – Результати аналізу методом дерево рішень

Захворювання	Тестові дані	Навчальні дані
Crohn`s	86.35%	100.00%
IBS	82.19%	100.00%
T2D	97.35%	100.00%
Середнє значення	88.63%	

Результати аналізу таблиці 3.2 свідчать про те, що найбільший відсоток правильних рішень модель зробила для захворювання діабет другого типу, з-поміж усіх інших. Найменший відсоток – для синдрому подразненого кишечника.

3.1.6 Класифікатор на основі випадкового лісу дерев рішень

Було проведено аналіз секвенованих даних ДНК з довжиною k-mer = 5 для 2-класової класифікації даних послідовностей умовно здорової людини та з хворобою за допомогою класифікатору на основі випадкового лісу дерев рішень.

Отримані результати якості класифікації методом випадкового лісу дерев рішень подано у таблиці 3.4.

Таблиця 3.4 – Результати аналізу методом випадковий ліс

Захворювання	Тестові дані	Навчальні дані
Crohn`s	85.58%	98.69%
IBS	81.31%	98.42%
T2D	99.06%	99.89%
Середнє значення	88.65%	

Даний метод показав найкращі результати щодо діагностування хвороби діабет другого типу, тобто практично ідеальний результат визначення, а ось щодо інших хвороб – показники не такі високі. Найменший відсоток правильного діагностування спостерігається для синдрому подразненого кишківника.

3.1.7 Висновки до 3.1

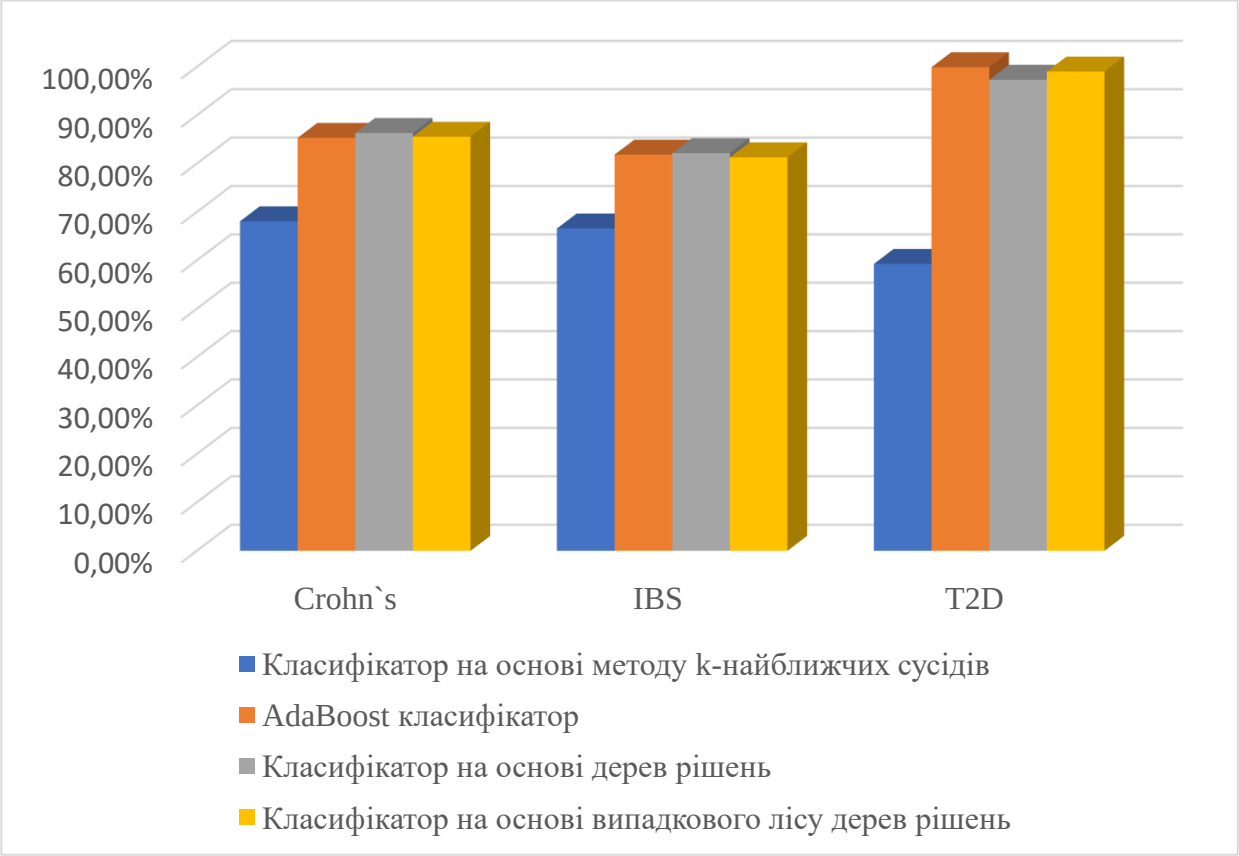

Рисунок 3.5 – Якісний розподіл застосування методів машинного навчання до діагностування захворювань

Було проведено аналіз 2-класової класифікації чотирма різними методами машинного навчання для даних з k-mer = 5.

Дослідження показало, що метод k-найближчих сусідів для аналізу даних ДНК мікробіому не підходить, про що свідчить достатньо велика похибка.

Модель на основі AdaBoost класифікатору найточніше визначила діабет другого типу (99.95%) з-поміж інших моделей. Але якісні показники для інших захворювань не є найкращими(захворювання Крона 85.35%, синдром подразненого кишечника 81.88%).

Методи дерево рішень та випадковий ліс мають високий результат точності для всіх проаналізованих захворювань.

Отже, для подальшого дослідження було обрано два методи: випадковий ліс та дерево рішень.

Також варто зазначити, що найкращий показник діагностування мало захворювання діабет другого типу, а найгірший – синдром подразненого кишківника.

3.2 Дослідження секвенованих даних ДНК за допомогою 3-класової та 4-класової класифікації

В подальшому дослідженні було використано два методи машинного навчання: випадковий ліс та дерево рішень. Але Також для отримання більш точного результату була проведена модифікація параметрів запропонованих методів.

Дерево рішень має такі параметри як:

1. criterion{“gini”, “entropy”, “log_loss”}, default=”gini”

Критерій визначає якість поділу. Часто використовуваними параметрами є “gini” (коефіцієнт Джінні) та “entropy” (ентропія).

2. splitter{“best”, “random”}, default=”best”

Стратегія, яка використовується для вибору поділу на вузлі: “best”для вибору найкращого розподілу та “random” – для найкращого випадкового поділу.

3. max_depth: int, default=None

Максимальна глибина обмежує глибину дерева рішень. Менше значення обмежує глибину, а більше – дозволяє дереву «рости глибше».

4. min_samples_split: int or float, default=2

Мінімальна кількість зразків, необхідних для розбиття внутрішнього вузла.

5. min_samples_leaf: int or float, default=1

Мінімальна кількість зразків, необхідних для листкового вузла. Це може мати ефект згладжування моделі, особливо в регресії.

6. random_state: int, RandomState instance or None, default=None

Контроль випадковості оцінювача.

Випадковий ліс має такі параметри:

1. n_estimators: int, default=100

Кількість дерев у лісі.

2. criterion{“gini”, “entropy”, “log_loss”}, default=”gini”

Критерій визначає якість поділу.

3. max_depth: int, default=None

Максимальна глибина обмежує глибину дерева рішень.

4. min_samples_split: int or float, default=2

Мінімальна кількість зразків, необхідних для розбиття внутрішнього вузла.

5. min_samples_leaf: int or float, default=1

Мінімальна кількість зразків, необхідних для листкового вузла. Це може мати ефект згладжування моделі, особливо в регресії.

6. min_weight_fraction_leaf: float, default=0.0

Мінімальна зважена частка загальної суми ваг.

Існують також інші параметри, але вони не будуть задіяні у цій роботі.

Розглянемо ближче параметр ‘criterion’, який є як у методі Дерево рішень так і у Випадковий ліс. Найбільш використовуваними є параметрами є “gini” (домішка Джині) та “entropy” (ентропія).

Вчений Клод Шеннон винайшов концепцію ентропії, яка вимірює домішку вхідного набору даних. Дерево рішень використовує приріст інформації, що в свою чергу є зменшенням ентропії[24]. Ентропія H(S) є мірою невизначеності набору даних S

$$H(S) = -\sum_{x \in X} P(x_i)\, log_2 P(x_i)$$

де S це набір даних, X це множина класів в наборі даних, P(x) відношення числа елементів у класі x до числа елементів у множині S

Коли H(S)=0, тоді у множині S всі елементи належать до одного класу. Ентропія обчислюється для кожного атрибута, на кожній ітерації атрибут з найменшою ентропією використовується для розбиття множини S.

Критерій Джині є мірою, наскільки часто випадково вибраний елемент з набору неправильно позначається, якщо він випадково позначається відповідно до розподілу міток у підмножині[24]. Критерій Джині визначається з домішки Джині, яка дорівнює

$$Gini(p) = 1 - \sum_{i=1}^{N} p_i^2$$

де p це набір даних, N це кількість класів, p_i частота класу i в наборі даних

Індекс Джині розглядає двійкове розбиття для кожного атрибута. Якщо двійкове розбиття на атрибуті A розбиває дані p на p1 та p2, тоді індекс Джині дорівнюватиме

$$Ginni_A(p) = \frac{|p1|}{|p|} Ginni(p1) + \frac{|p2|}{|p|} Ginni(p2)$$

Під час аналізу даних секвенування ДНК буде використано дані параметри, щоб збільшити відсоток якісного діагностування хвороб.

Під час досліджень ефективності зміни даних параметрів класифікаторів для 3-класової та 4-класової ідентифікації захворювань кишківника за даними

секвенування ДНК мікробіому, будуть використані такі дані: 20000 послідовностей для кожного класу, k-mer = 5. Для 3-класової класифікації було використано дані умовно здорової людини, з хворобою Крона та діабетом другого типу. Для 4-класової класифікації було використано дані умовно здорової людини, з хворобою Крона, діабетом другого типу та синдромом подразненого кишківника. Порівняння якості класифікації різними методами представлено у таблиці 3.5.

Таблиця 3.5 – Результати аналізу методів дерево рішень та випадковий ліс

Методи	4 класи	3 класи
Дерево рішень, зміна параметрів	71.69%	84.82%
Дерево рішень, без змін	71.19%	84.63%
Випадковий ліс, зміна параметрів	75.22%	88.67%
Випадковий ліс, без змін	78.04%	90.67%

Аналіз даних таблиці 3.5 свідчить про те, що зміна параметрів вплинула позитивно для аналізу за допомогою класифікатора на основі дерев рішень та негативно для класифікатору на основі випадкового лісу дерев рішень. Отже, для подальшого аналізу краще використовувати ці параметри тільки для моделі на основі дерев рішень.

Далі проведено аналіз для даних послідовностей з k-mer від 2 до 10 з 4-класовою класифікацією. Для k-mer від 8 до 10 були використані дані з 10000 послідовностей для кожного класу, оскільки через велику довжину k-mer не вистачало оперативної пам’яті середньостатистичного ПК.

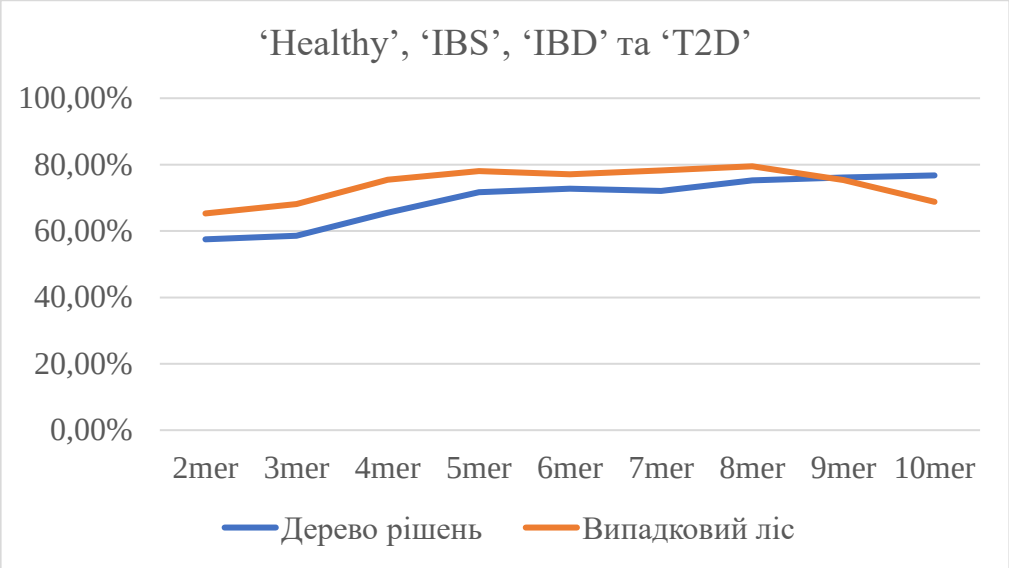

Рисунок 3.6 – Графік розподілу якості методів відносно k-mer для 4 класів

Аналіз рисунка 3.6 свідчить про те, що метод дерево рішень має менший відсоток правильності вибору, ніж метод випадковий ліс для k-mer від 2 до 8. Але там, де довжина зростає до 9 та 10, метод випадковий ліс має більше похибку, а дерево рішень – навпаки, прямує до своєї максимальної якості.

Далі було проведено аналіз даних з такою ж самою довжиною k-mer, але тепер вже для трьох класів, Проведемо даний аналіз для всіх можливих варіацій, тільки залишаючи кожен раз клас ‘Healthy’.

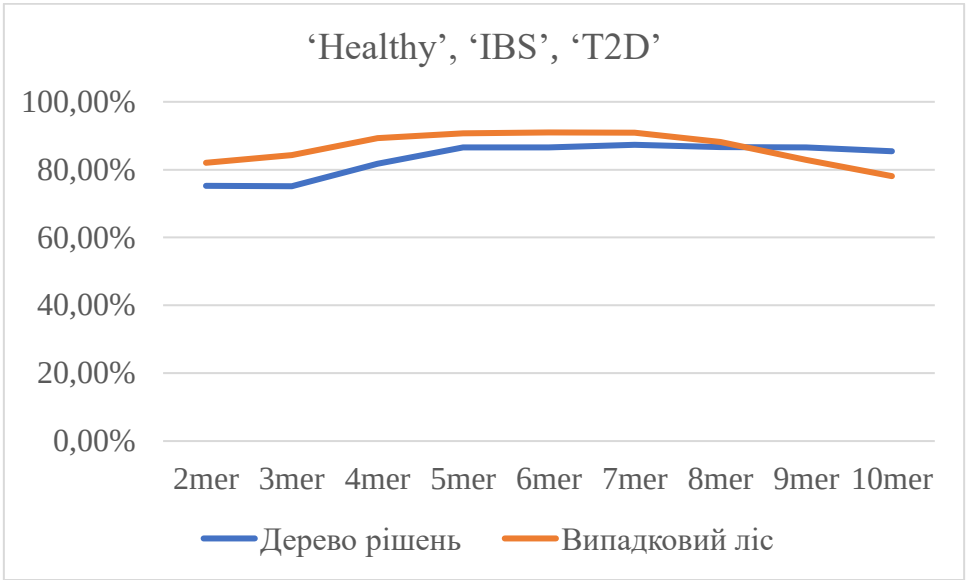

Рисунок 3.7 – Графік розподілу якості методів відносно k-mer для класів ‘Healthy’, ‘IBS’, ‘T2D’

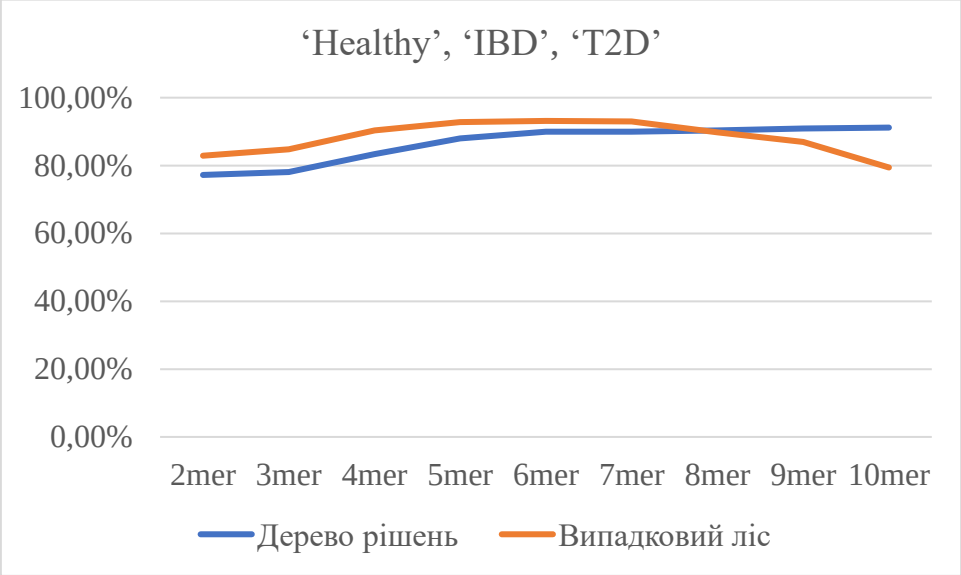

Рисунок 3.8 – Графік розподілу якості методів відносно k-mer для класів ‘Healthy’, ‘IBD’, ‘T2D’

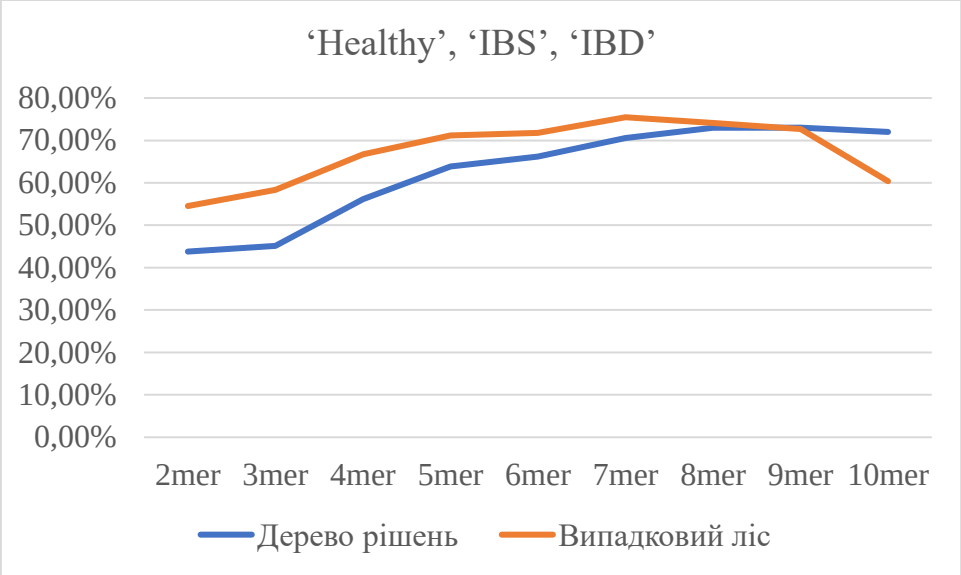

Рисунок 3.9 – Графік розподілу якості методів відносно k-mer для класів ‘Healthy’, ‘IBS’, ‘IBD’

Проаналізувавши рисунки 3.7-3.9 можна побачити, що, як і для даних з чотирма класами, для трьох класів метод дерево рішень має менший відсоток правильності вибору, ніж метод випадковий ліс. Але це відбувається для k-mer від 2 до 7. Починаючи з довжини 8 відсоток правильно прийнятих рішень методу

випадковий ліс зменшується, а метод дерево рішень або прямує до своєї максимальної якості, або не знижує відсоток правильності, порівняно з аналізом даних з довжиною k-mer 6-7.

3.3 Висновки до третього розділу

У результаті дослідження було проаналізовано багато відмінних між собою даних. Досліджені секвеновані послідовності відрізнялись як за довжиною k-mer, так і за різною кількістю та варіацією класів захворювань.

Початково було використано спершу 4 методи машинного навчання, а саме: метод k-найближчих сусідів, AdaBoost, дерево рішень та випадковий ліс. Перший аналіз відбувався на даних секвенування ДНК мікробіому кишківника з довжиною k-mer = 5, двома класами, та кількістю послідовностей 12000 для кожного класу. Варіаціями двох класів в цих даних були послідовності здорової людини та з хворобою Крона, діабетом другого типу та синдромом подразненого кишківника, тобто відповідно три різні файли.

Після проведення аналізу було відібрано два метода, які показали найкращі результати в якості діагностування захворювань кишківника людини: дерево рішень (86-97%) та випадковий ліс (85-99%). Найгірший результат показав метод k-найближчих сусідів – 65-68% правильності прийнятих рішень. Метод AdaBoost показав гарний результат під час виявлення діабету другого типу – 99,95%, але для інших захворювань показники були менші, ніж обраних моделей, тому метод AdaBoost було теж відхилено через нестійкість до різних захворювань.

Далі було проведено аналіз даних секвенування ДНК мікробіому кишківника з довжиною k-mer від 2 до 10, чотирма і трьома класами, та кількістю послідовностей 20000 для кожного класу. Однак для k-mer від 8 до 10 були використані дані з 10000 послідовностей для кожного класу, оскільки через велику довжину k-mer не вистачало оперативної пам’яті середньостатистичного ПК.

Після проведення аналізу за допомогою двох обраних методів машинного навчання, були отримані результати щодо якості вибору даних для аналізу та щодо якості самих моделей. Проаналізувавши результати для k-mer від 2 до 10, можна зробити висновки, що краще розбивати дані на довжину 5-7 k-mer та використовувати метод випадковий ліс. Хоча метод дерево рішень показав гарні результати під час аналізу даних з довжиною k-mer від 8 до 10, однак максимальний відсоток правильності прийнятих рішень був для довжини 5-7 k-mer, отриманий методом випадковий ліс.

Для діагностування захворювань під час аналізу даних з 4-класовою класифікацією, а саме, здоровий, хвороба Крона, діабет другого типу та синдром подразненого кишківника, максимальним відсотком якості даної класифікації було наближення до 80%. Знизивши кількість класів до трьох, було помічено збільшення правильності до 90%. Для набору даних з класами здоровий, захворювання Крона та діабет другого типу, метод випадковий ліс мав найбільшу точність – 93,14% для k-mer = 7. Для масиву тих же вхідних даних метод дерево рішень мав свій максимальний результат у 91,18% для k-mer = 10.

Однак навіть зменшивши кількість класів до трьох, провівши аналіз всіх можливих варіацій співставлення захворювань до даних здорової людини, було виявлено досить низький, на фоні інших аналізів, відсоток діагностування захворювань для даних з класами здоровий, синдром подразненого кишківника та хвороби Крона. Максимальним відсотком правильності прийняття рішень було всього 75,46%. Висновком цього може бути схожість послідовностей ДНК для цих двох захворювань. Отже, для подальшого використання цього аналізу мікробіологами, має бути попередження про можливість плутанини у виявленні хвороби Крона на синдрому подразненого кишківника.

Отже, для використання навчених моделей у веб-інтерфейсі було обрано: моделі класифікатору на основі методу випадковий ліс дерев рішень для діагностування захворювань кишківника з 3-класовою та 4-класовою класифікацією для даних з довжиною k-mer = 6. Для 3-класової класифікації якість

вибору становитиме 75-93%, в залежності від видів хвороб, які класифікуються. Для 4-класової класифікації точність вибору становитиме близько 80%.

4 РОЗРОБКА ВЕБ-ЗАСТОСУНКУ

4.1 Розробка веб-інтерфейсу зі збереженими моделями машинного навчання

Після проведеного аналізу за допомогою методів машинного навчання було збережено найефективніші моделі, які дають змогу діагностувати захворювання кишківника людини.

Було відібрано чотири моделі, дані секвенування ДНК мікробіому людини було поділено на k-mer з довжиною 6, по яким можна визначати захворювання кишківника людини в таких комбінаціях:

1. T2D, IBS, IBD.
2. T2D, IBS.
3. T2D, IBD.
4. IBS, IBD.

За допомогою веб-інтерфейсу було візуалізовано аналіз секвенованих даних ДНК мікробіому кишківника людини з використанням збережених моделей.

В якості веб-сервера буде використаний популярний мікро фреймворк Flask, який забезпечив легку і надійну роботу веб-інтерфейсу.

Далі було проведено інтеграцію скрипту (див. Додаток Г) з веб-сервером (див. Додаток Д). Для розмітки сторінки за основу було взято BootStrap5 – набір інструментів, призначений для створення веб-застосунків, який в собі містить шаблони CSS та HTML.

Для початку користування веб-застосунку користувачу необхідно перейти за посиланням. Далі сторінка браузера матиме вигляд, як це показано на Рисунку 4.1.

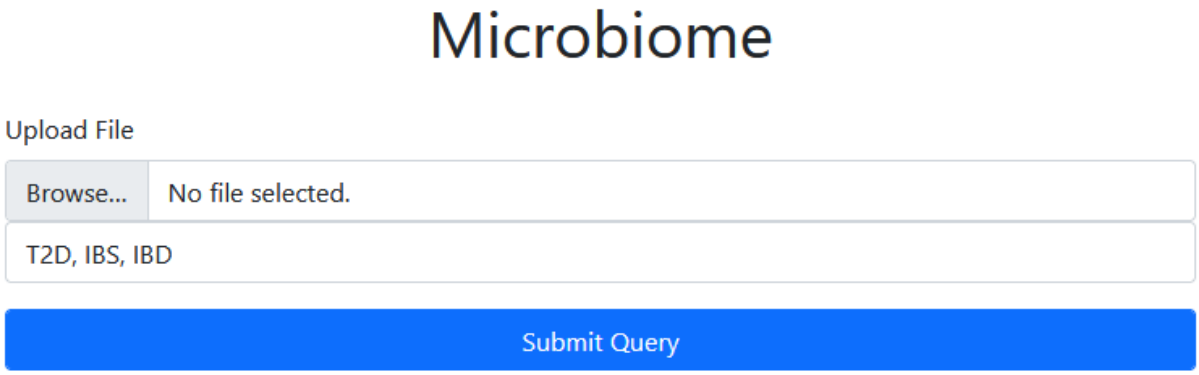

Рисунок 4.1 – Веб-форма для обробки файлу

Після цього користувачу потрібно обрати файл секвенування ДНК мікробіому кишківника людини з розширенням PKL та довжиною k-mer = 6 потім – комбінацію захворювань, які будуть класифікуватись Рисунок 4.2.

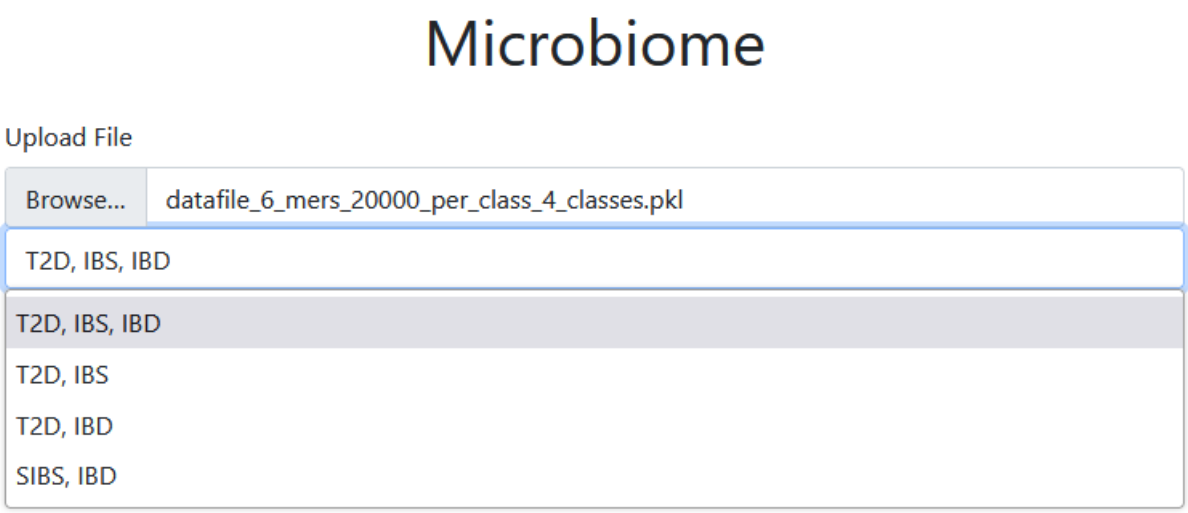

Рисунок 4.2 – Вікно вибору доступних комбінацій захворювань

Після натискання кнопки ‘Submit Query’, застосунок починає аналіз заданих секвенованих послідовностей ДНК. Через деякий проміжок часу з’являється вікно з вибором шляху, де можна зберегти отриманий вихідний файл.

Вихідний файл являє собою текстовий файл, в якому міститься масив з ймовірними назвами захворювань для відповідних послідовностей ДНК, які були завантажені. Приклад такого файлу представлений на Рисунку 4.3.

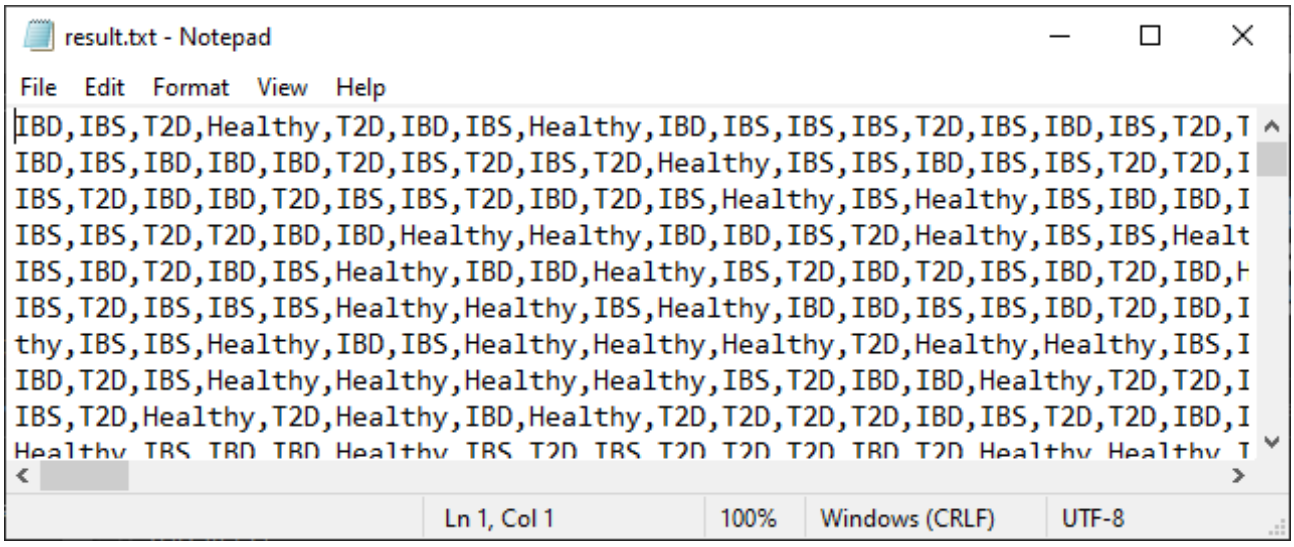

Рисунок 4.3 – Приклад вихідного файлу з результатами дослідження

Таким чином користувачі отримують можливість проводити аналіз засобами веб-браузера.

5 ВИСНОВКИ

Під час виконання даної роботи було проведено дослідження методів секвенування. Найсучаснішими та найбільш використовуваними є метод секвенування Illumina та Oxford Nanopore секвенування. Було розглянуто поняття мікробіому та його вплив на здоров'я людини. Мікробіом дуже впливає на стан та здоров'я людини і навіть невеликі дисбаланси в ньому можуть призвести до тяжких захворювань в майбутньому, тому вчені прямують до вивчення мікробіому людини на рівні, коли можна буде ідентифікувати захворювання на дуже ранніх етапах та визначати індивідуальний підхід до кожного пацієнта. На шляху до цього була створена і дана робота також.

Було проведено аналіз даних, які видаються на виході секвенування ДНК мікробіому кишківника людини. Було розглянуто їх типи та методи подальшої обробки цих даних. Одним з найрозповсюджених типів даних які видає секвенатор є FASTQ. У даній роботі було розглянуто як взаємодіяти з файлами даного типу, як проводити аналіз ДНК послідовностей, які він в собі вміщує.

Було розглянуто такі методи аналізу текстових даних як K-mer та модель Bag of Words. Ці методи розроблені для аналізу ключових характеристик тексту, але при цьому вони відрізняються за глибиною деталізації. K-mers виявляють локальні паттерни, оглядаючи сусідні символи або слова, тоді як модель Bag of Words розглядає текст як колекцію слів і обраховує їх вживання. Тому в цьому дослідженні було застосовано обидва методи.

Було досліджено дані секвенування мікробіому кишківника для умовно здорових людей, пацієнтів із синдромом подразненого кишківника, хворобою Крона та діабетом другого типу. Таким чином, було поставлено задачу мультикласової класифікації на 4 класи.

Після обробки секвенованих даних було використано методи машинного навчання такі як: метод k-найближчих сусідів, метод дерево рішень, метод випадковий ліс дерев рішень та AdaBoost класифікатор. Спочатку було проведено

2-класову класифікацію даних умовно здорової людини в парі з даними людини з: захворюванням Крона, діабетом другого типу та синдромом подразненого кишківника. Після проведення даного аналізу було визначено два найточніших класифікатори: класифікатор на основі методу дерев рішень із середньою точністю 2-класової класифікації 88.63% та класифікатор на основі методу випадковий ліс дерев рішень із середньою точністю 2-класової класифікації 88.65%. Отже, для подальшого дослідження було обрано ці два класифікатора.

Було проведено дослідження при зміні параметрів класифікаторів, яке дало позитивний результат для класифікатора на основі методу дерев рішень та негативний для класифікатора на основі методу випадковий ліс дерев рішень. Отже, було прийнято рішення продовжувати аналіз змінивши тільки параметри criterion="entropy" та random_state=2 для класифікатору на основі методу дерев рішень.

Далі була проведена 4-класова та 3-класова класифікація секвенованих даних з довжиною k-mer від 2 до 10 за допомогою класифікатору на основі методу дерев рішень та класифікатору на основі випадкового лісу дерев рішень. Максимальним відсотком якості 4-класової класифікації було наближення до 80%, 3-класової класифікації – 90%. Але навіть для 3-класової класифікації було виявлено досить невеликий у порівнянні з іншими аналізами відсоток якості діагностики хвороб (75,46%) для даних з класами умовно здорова людина, з синдромом роздратованого кишечника та хворобою Крона. Одним з висновків може бути подібність мікробіому для цих двох захворювань. Таким чином, при подальшому використанні цього аналізу мікробіологами, слід бути готовим до потенційної невизначеності в диференціації хвороби Крона і синдрому роздратованого кишечника. Отже, для веб-застосунку було обрано моделі класифікатору на основі методу випадковий ліс дерев рішень для діагностування захворювань кишківника з 3-класовою та 4-класовою класифікацією. Для аналізу було обрано довжину k-mer = 6, оскільки вона містить баланс часу, який витрачається для класифікації та якості самої класифікації. Для 3-класової класифікації точність вибору становитиме 75-93%, в

залежності від видів хвороб, які класифікуються. Для 4-класової класифікації точність вибору становитиме близько 80%.

Було розроблено веб-застосунок зі збереженими моделями вказаних вище класифікаторів, який дає можливість мікробіологам та іншим користувачам проводити аналіз засобами веб-браузера. Вихідними даними є файл з масивом з ймовірними назвами захворювань для відповідних послідовностей ДНК, які були завантажені.

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ

1. Wikipedia [Електронний ресурс] : Режим доступу до ресурсу : https://en.wikipedia.org/
2. "Introduction to Genetic Analysis" by Anthony J.F. Griffiths, Susan R. Wessler, et al.
3. Секвенування Illumina [Електронний ресурс] : Режим доступу до ресурсу : https://www.illumina.com/
4. Секвенування Oxford Nanopore[Електронний ресурс] : Режим доступу до ресурсу : https://nanoporetech.com/
5. National human genome institute [Електронний ресурс] : Режим доступу до ресурсу : https://www.genome.gov/
6. National library of medicine [Електронний ресурс] : Режим доступу до ресурсу : https://pubmed.ncbi.nlm.nih.gov/34750572/
7. Science direct [Електронний ресурс] : Режим доступу до ресурсу : https://www.sciencedirect.com/topics/immunology-and-microbiology/illumina-dye-sequencing
8. National institute of health [Електронний ресурс] : Режим доступу до ресурсу : https://commonfund.nih.gov/hmp
9. The Human Microbiome Project [Електронний ресурс] : Режим доступу до ресурсу : https://www.nature.com/articles/nature06244
10. Широбоков В. П., Димент Г. С., Янковський Д. С. Мікробіом у фізіології людини. – 2018.
11. Scientific American [Електронний ресурс] : Режим доступу до ресурсу : https://www.scientificamerican.com/article/an-endangered-species-in/
12. National library of medicine: The human microbiome [Електронний ресурс] : Режим доступу до ресурсу : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7306068/
13. Grice E. A., Segre J. A. The skin microbiome. Nature Reviews Microbiology.

14. Мюллер А., Гвидо С. Введение в машинное обучение с помощью Python //Руководство для специалистов по работе с данными. СПб.: ООО «Альфа-книга. – 2017.
15. Sequence File Formats [Електронний ресурс] : Режим доступу до ресурсу : https://timkahlke.github.io/LongRead_tutorials/APP_FORM.html#:~:text=The%20FAST5%20format%20is%20the,with%20a%20normal%20text%20editor.
16. Illumina knowledge [Електронний ресурс] : Режим доступу до ресурсу : https://knowledge.illumina.com/software/general/software-general-reference_material-list/000002211
17. Zymo research: FASTQ format [Електронний ресурс] : Режим доступу до ресурсу : https://zymoresearch.eu/blogs/blog/fastq-file-format
18. Bioinformatics [Електронний ресурс] : Режим доступу до ресурсу : https://medium.com/swlh/bioinformatics-1-k-mer-counting-8c1283a07e29
19. Zeng, Tao, Xiangtian Yu, and Zhangran Chen. "Applying artificial intelligence in the microbiome for gastrointestinal diseases: a review." Journal of Gastroenterology and Hepatology 36.4 (2021): 832-840
20. IBM [Електронний ресурс] : Режим доступу до ресурсу : https://www.ibm.com/topics/natural-language-processing#:~:text=Natural%20language%20processing%20(NLP)%20refers,same%20way%20human%20beings%20can.
21. Marlicz, Wojciech, George Koulaouzidis, and Anastasios Koulaouzidis. "Artificial Intelligence in Gastroenterology—Walking into the Room of Little Miracles." Journal of Clinical Medicine 9.11 (2020): 3675.
22. Pandas documentation URL [Електронний ресурс] : Режим доступу до ресурсу : https://pandas.pydata.org/docs/
23. SciKit learn user guide URL [Електронний ресурс] : Режим доступу до ресурсу : https://scikit-learn.org/stable/user_guide.html
24. Datacamp [Електронний ресурс] : Режим доступу до ресурсу : https://www.datacamp.com/tutorial/decision-tree-classification-python

25. Grinberg, M. (2018). Flask Web Development: Developing Web Applications with Python. United States: O'Reilly Media.
26. ASCII table [Електронний ресурс] : Режим доступу до ресурсу : https://commons.wikimedia.org/wiki/File:ASCII-Table-wide.svg
27. Dhaliwal, Jasbir, et al. "Accurate classification of pediatric colonic inflammatory bowel disease subtype using a random forest machine learning classifier." Journal of Pediatric Gastroenterology and Nutrition 72.2 (2021): 262-269.
28. Ketkar, N. (2017). Deep Learning with Python: A Hands-on Introduction. United States: Apress.
29. Nearest-neighbor Methods in Learning and Vision: Theory and Practice. (2005). Great Britain: MIT Press.
30. Garreta, R., Moncecchi, G. (2013). Learning Scikit-learn: Machine Learning in Python. United Kingdom: Packt Publishing.
31. Harrison, M. (2016). Learning the Pandas Library: Python Tools for Data Munging, Analysis, and Visual. Germany: CreateSpace Independent Publishing Platform.
32. Shaw, Z. (2013). Learn Python the Hard Way: A Very Simple Introduction to the Terrifyingly Beautiful World of Computers and Code. Великобритания: Pearson Education.
33. Maimon, Oded Z., and Rokach, Lior. Data Mining with Decision Trees: Theory and Applications. Сингапур, World Scientific, 2008.
34. Mardis, Elaine R. "DNA sequencing technologies: 2006–2016." Nature protocols 12.2 (2017): 213-218.
35. Jain, Miten, et al. "The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community." Genome biology 17 (2016): 1-11

ДОДАТОК А

Скрипт для перетворення даних ДНК з формату FASTQ в PKL

```
from Bio import SeqIO
import pickle

# Input and output filenames

fastq_file = "input.fastq"
pkl_file = "output.pkl"

# Read FASTQ file and convert to Python list of sequences

sequences = []
with open(fastq_file, "r") as handle:
    for record in SeqIO.parse(handle, "fastq"):
        sequences.append(record.seq)

# Save sequences to PKL file

with open(pkl_file, "wb") as handle:
    pickle.dump(sequences, handle)
```

ДОДАТОК Б

Скрипт для генерування k-mer

```
import pickle
from kmer import kmers

with open('output.pkl', 'rb') as handle:
    sequences = pickle.load(handle)

k = 5  # Set the desired k-mer length
kmer_list = []

for sequence in sequences:
    kmer_list.extend(kmers(sequence, k))
```

ДОДАТОК В

Скрипт для початку аналізу файлів формату PKL

```
import pickle
import pandas as pd
from sklearn.feature_extraction.text import CountVectorizer
import tkinter as tk
from tkinter import filedialog



def get_file_path():
    root = tk.Tk()
    root.withdraw()  # this will hide the main tkinter window
    file_path = filedialog.askopenfilename()  # this will open the system dialog
    return file_path



if __name__ == '__main__':
    with open(get_file_path(), 'rb') as f:
        file = pickle.load(f)

    df = pd.DataFrame(file)

    dna_text = list(df['seq'])
    dna_class = list(df['class'])

    for item in range(len(dna_text)):
        dna_text[item] = ' '.join(dna_text[item])

    # Creating the Bag of Words model using CountVectorizer()
```

```
cv = CountVectorizer(ngram_range=(1, 1), lowercase=False)
T_dna = cv.fit_transform(dna_text)
T_dna_tabl = pd.DataFrame(T_dna.toarray(), columns=cv.get_feature_names_out())
print(T_dna_tabl.head())
```

ДОДАТОК Г

Скрипт веб-інтерфейсу з використанням збережених моделей машинного навчання

```
import os
import pickle
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.neighbors import KNeighborsClassifier
from sklearn.model_selection import train_test_split
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier
from io import BytesIO

def calculate(model_type: int, file: BytesIO):
    data = pickle.load(file)
    file.close()

    df = pd.DataFrame(data)

    models = ['random_forest_model_1.pkl', 'random_forest_model_2.pkl', 'random_forest_model_3.pkl', 'random_forest_model_4.pkl']
    if model_type > (len(models)-1):
        return list()
    with open(models[model_type], 'rb') as f:
        loaded_model = pickle.load(f)

    dna_text = list(df['seq'])
```

```
for item in range(len(dna_text)):
        dna_text[item] = ' '.join(dna_text[item])

cv = CountVectorizer(ngram_range=(1, 1), lowercase=False)  # lowercase = True
T_dna = cv.fit_transform(dna_text)
forest_predictions = loaded_model.predict(T_dna)
return list(forest_predictions)
```

ДОДАТОК Ґ

Скрипт для веб-сервера

```
import os
from flask import Flask, render_template, request, send_file
from io import BytesIO, StringIO

from worker import calculate

app = Flask(__name__)

def create_file_in_memory(data_array):
    # Join the array elements with commas
    data_string = ','.join(data_array)

    # Create a file-like object in memory
    file_obj = BytesIO(data_string.encode('utf-8'))
    file_obj.seek(0)

    return file_obj


@app.route("/", methods=['GET', 'POST'])
def main():

    if request.method == 'POST':
        data = BytesIO(request.files.get('file', None).read())
        a = request.form
        model_type = int(request.form.get('model_type', 0))
        text_array = calculate(model_type,data)
        text_file = create_file_in_memory(text_array)
```

```
        return send_file(text_file, as_attachment=True, download_name='result.txt')

    return render_template('index.html')

@app.route('/how are you')
def hello():
    return 'I am good, how about you?'

if __name__ == "__main__":
    app.run(host="0.0.0.0", port=8080)
```

ДОДАТОК Д

Розмітка сторінки в форматі HTML

```
<!DOCTYPE html>
<html>
<head>
   <title>Microbiome</title>
   <link href="static/assets/dist/css/bootstrap.min.css" rel="stylesheet">
</head>
<body>
   <div class="container py-5">
      <h1 class="mb-4 text-center">Microbiome</h1>
      <div class="row justify-content-center">
         <div class="col-lg-8">
            <form action="/" method="POST" enctype="multipart/form-data">
               <div class="mb-3">
                  <label for="file" class="form-label">Upload File</label>
                  <input type="file" class="form-control" name="file" id="file">
                  <select class="form-control" name="model_type">
                     <option value="0">T2D, IBS, IBD</option>
                     <option value="1">T2D, IBS</option>
                     <option value="2">T2D, IBD</option>
                     <option value="3">IBS, IBD</option>
                  </select>
               </div>
               <div class="mb-3">
                  <input type="submit" class="form-control btn btn-primary">
               </div>
            </form>
         </div>
```

```
        </div>
        {% if text %}
        <div id="data" class="row justify-content-center">
            <div class="col-lg-8">
                <div class="my-4">{{ text | safe }}</div>
                <img src="static/temp_plot.png" class="img-fluid">
            </div>
        </div>
        {% endif %}
    </div>
</body>
</html>
```