Хмельницький національний університет
Факультет інформаційних технологій
Кафедра комп'ютерних наук

КВАЛІФІКАЦІЙНА РОБОТА БАКАЛАВРА

на тему *Метод визначення діагнозу за текстовим описом симптомів NLP-засобами*

Галузь знань *12 – Інформаційні технології*
Шифр і назва галузі знань

Спеціальність *122 – Комп'ютерні науки*
Шифр і назва спеціальності

Освітня програма *Комп'ютерні науки*
Назва освітньої програми

Виконала: *студентка групи КН-20-2* _____ *Юлія КОЗЕНКО*
Курс, група виконавця | Підпис | Ім'я, ПРІЗВИЩЕ

Керівник: *викладач каф. КН* _____ *Валерія КЛІМЕНКО*
Науковий ступінь, посада | Підпис | Ім'я, ПРІЗВИЩЕ

Нормоконтроль: *к.т.н., доц. каф. КН* _____ *Руслан БАГРІЙ*
Науковий ступінь, посада | Підпис | Ім'я, ПРІЗВИЩЕ

До захисту допускаю:
Зав. кафедри КН, д.т.н., професор _____ *Олександр БАРМАК*
Підпис | Ім'я, ПРІЗВИЩЕ

18 червня 2024 р.

Хмельницький 2024

ХМЕЛЬНИЦЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
Факультет *інформаційних технологій*
Кафедра *комп'ютерних наук*
Освітній ступінь *бакалавр*
Галузь знань *12 – Інформаційні технології*
Спеціальність *122 – Комп'ютерні науки*

ЗАТВЕРДЖУЮ
Завідувач кафедри *комп'ютерних наук*
_____
(підпис)
д.т.н., професор Олександр БАРМАК
«*16*» *лютого* *2024* року

ЗАВДАННЯ
НА КВАЛІФІКАЦІЙНУ РОБОТУ БАКАЛАВРА

1. Тема кваліфікаційної роботи бакалавра: *«Метод визначення діагнозу за текстовим описом симптомів NLP-засобами»*

2. Завдання видано студентці *Юлії КОЗЕНКО*
(Ім'я, прізвище)

3. Керівник роботи *викладач кафедри КН Валерія КЛІМЕНКО*
(посада, ім'я, прізвище)

4. Затверджено наказом університету від «*15*» *лютого* *2024* р. № *8*

5. Дата видачі завдання студенту: «*16*» *лютого* *2024* р.

6. Зміст пояснювальної записки (перелік задач) та вихідні дані:

Метою кваліфікаційної роботи бакалавра є спрощення процесу діагностування шляхом автоматизованого визначення діагнозу за текстовим описом симптомів. Для досягнення мети слід виконати аналіз інформаційних моделей області діагностування, огляд теоретичних підходів та обрати підхід для розв'язку задачі визначення діагнозу за текстовим описом NLP-засобами, провести аналіз існуючих програмних рішень, створити метод визначення діагнозу за текстовим описом симптомів NLP-засобами та його програмну реалізацію, виконати тестування та дослідження ефективності запропонованого методу. Вихідними даними є діагноз користувача та текстові рекомендації щодо можливого лікування класифікованого захворювання.

7. Календарний план виконання кваліфікаційної роботи бакалавра:

№	Назва етапів (розділів) кваліфікаційної роботи бакалавра	Термін виконання	Примітка
1	*Вибір напряму дослідження та узгодження тематики кваліфікаційної роботи бакалавра з керівником, складання календарного графіка виконання роботи*	*січень 2024*	виконано
2	*Ознайомлення з предметною областю, формулювання мети та задач дослідження, визначення об'єкта та предмета дослідження*	*лютий 2024*	виконано
3	*Проектування та розробка загальної архітектури програмного забезпечення, інтерфейсу користувача, вибір засобів реалізації програмного забезпечення*	*березень 2024*	виконано
4	*Створення та тестування програмного забезпечення*	*квітень 2024*	виконано
5	*Написання пояснювальної записки, урахування зауважень керівника, оформлення згідно вимог*	*травень 2024*	виконано
6	*Розробка презентаційних матеріалів та попередній захист кваліфікаційної роботи*	*травень 2024*	виконано
7	*Отримання відгуку керівника, рецензії, перевірка на плагіат, нормоконтроль*	*червень 2024*	виконано
8	*Підготовка до захисту та захист кваліфікаційної роботи бакалавра*	*червень 2024*	виконано

Виконавець: *студентка групи КН-20-2* (Курс, група виконавця) _____ (Підпис) *Юлія КОЗЕНКО* (Ім'я, ПРІЗВИЩЕ)

Керівник: *викладач каф. КН* (Науковий ступінь, посада) _____ (Підпис) *Валерія КЛІМЕНКО* (Ім'я, ПРІЗВИЩЕ)

Анотація

Тема кваліфікаційної роботи бакалавра: *«Метод визначення діагнозу за текстовим описом симптомів NLP-засобами»*

Виконавець кваліфікаційної роботи бакалавра: *студентка групи КН-20-2 Юлія КОЗЕНКО*

Керівник кваліфікаційної роботи бакалавра: *викладач кафедри КН Валерія КЛІМЕНКО*

Кваліфікаційна робота бакалавра містить:

Пояснювальна записка				Кількість додатків
Сторінок	Рисунків	Таблиць	Джерел інформації	
65	24	10	41	4

Метою кваліфікаційної роботи бакалавра є спрощення процесу діагностування шляхом автоматизованого визначення діагнозу за текстовим описом симптомів, для чого виконувалась розробка методу визначення діагнозу за текстовим описом симптомів NLP-засобами, та відповідного програмного забезпечення у вигляді десктопного застосування та бази даних.

Для розробки інформаційної системи та тренування моделі машинного навчання було використано мову програмування Python, а також систему керування базами даних SQLite.

Програмну реалізацію на основі створеного методу визначення діагнозу, можна використовувати для раннього виявлення потенційних захворювань, дозволяючи пацієнтам негайно звертатися за медичною допомогою та лікуванням.

Ключові слова: NLP, діагностування хвороби за описом, KNN

Виконавець: *студентка групи КН-20-2* _____ *Юлія КОЗЕНКО*

Курс, група виконавця | Підпис | Ім'я, ПРІЗВИЩЕ

Зміст

Перелік скорочень

Скорочення, термін, позначення	Пояснення
NLP	Обробка природної мови
ВІЛ	Вірус імунодефіциту людини
СНІД	Синдром набутого імунодефіциту
КТ	Комп'ютерна томографія
МРТ	Магнітно-резонансна томографія
ШІ	Штучний інтелект
KNN	K-Nearest Neighbors
ПЗ	Програмне забезпечення
КРБ	Кваліфікаційна робота бакалавра
ІС	Інформаційна система
ІТ	Інформаційні технології
КН	Комп’ютерні науки
БД	База даних
СКБД	Системи керування базами даних
IDE	Інтегроване середовище розробки

Вступ

Кваліфікаційна робота бакалавра присвячена задачі спрощення процесу діагностування шляхом автоматизованого визначення діагнозу за текстовим описом симптомів розробка, для чого виконувалась розробка методу визначення діагнозу за текстовим описом симптомів NLP-засобами, та відповідного програмного забезпечення у вигляді десктопного застосування та бази даних.

Актуальність. Визначення діагнозу за текстовим описом симптомів з використанням методів обробки природної мови полягає у потенційній можливості створення ефективних інструментів для автоматизованої або підтримуючої медичної діагностики.

Програми основані на штучному інтелекті, які можуть аналізувати текстові описи симптомів, можуть прискорити процес діагностики, особливо у випадках, коли швидке реагування може врятувати життя пацієнта, а використання алгоритмів машинного навчання на основі NLP може допомогти відсіяти і враховувати широкий спектр можливих діагнозів, забезпечуючи більш точні результати, особливо в складних клінічних випадках.

На ряду з цим, зростання кількості медичних даних створює потребу в ефективних інструментах для їх аналізу. Використання NLP дозволяє автоматизувати процеси аналізу тексту, що допомагає відокремлювати важливі дані від шуму та забезпечувати більш комплексний аналіз.

Програмну реалізацію з визначення діагнозу за текстовим описом засобами NLP можна використовувати для раннього виявлення потенційних захворювань, дозволяючи пацієнтам негайно звертатися за медичною допомогою та лікуванням. Крім того, у ситуаціях, коли особисті консультації неможливі або бажані, таку програмну реалізацію можна використовувати для надання дистанційної діагностики та рекомендацій щодо лікування на основі симптомів користувача.

Об’єкт дослідження – процес визначення діагнозу за текстовим описом симптомів NLP-засобами.

Предмет дослідження – методи машинного навчання для роботи з текстовою інформацією.

Мета кваліфікаційної роботи бакалавра – спрощення процесу діагностування шляхом автоматизованого визначення діагнозу за текстовим описом симптомів.

Завдання кваліфікаційної роботи бакалавра – провести аналіз інформаційних моделей області діагностування; виконати огляд теоретичних підходів та обрати підхід для розв’язку задачі визначення діагнозу за текстовим описом засобами NLP; провести аналіз існуючих програмних рішень; створити метод визначення діагнозу за текстовим описом симптомів NLP-засобами; описати інформаційну структуру системи діагностування пацієнтів за текстовим описом засобами NLP; обрати набір даних для навчання класифікатора; створити відповідну програмну реалізацію на основі створеного методу визначення діагнозу; виконати тестування створеного ПЗ; виконати дослідження ефективності запропонованого методу визначення діагнозу з використанням розробленої інформаційної системи діагностування пацієнтів за текстовим описом NLP-засобами.

Розділ 1 Характеристика предметної області діагностики захворювань NLP-засобами

1.1 Аналіз інформаційних моделей визначення діагнозу за симптомами

Стан медицини у сучасному світі можна охарактеризувати як прогресивний та інноваційний, адже за останні десятиліття було досягнуто значних успіхів у галузі медицини, які призвели до зниження рівня смертності та поліпшення якості життя людей у всьому світі [1].

Одним із найважливіших досягнень сучасної медицини є розробка нових ліків та вакцин. Нові ліки дозволяють ефективно лікувати раніше невиліковні захворювання, такі як ВІЛ/СНІД, рак та серцево-судинні захворювання, а вакцини захищають людей від інфекційних захворювань, таких як поліомієліт, туберкульоз та дифтерія.

Іншим важливим досягненням сучасної медицини є розвиток нових технологій діагностики та лікування. Ці технології дозволяють лікарям більш точно діагностувати захворювання та надавати ефективніше лікування. Наприклад, КТ та МРТ дозволяють лікарям отримувати детальні зображення внутрішніх органів людини, що допомагає їм поставити правильний діагноз.

Розвиток медицини також сприяв покращенню доступу до медичних послуг у всьому світі. Завдяки міжнародній співпраці та розвитку приватного сектору більшість людей у світі мають доступ до базових медичних послуг, таких як вакцинація, профілактика, діагностування та лікування інфекційних захворювань [2].

Для того, щоб записатися до лікаря в Україні, також можна зателефонувати до лікарні або клініки. У рамках медичної реформи в Україні створено єдину систему запису до лікаря . Ця система доступна в Інтернеті та на мобільних пристроях. Щоб записатися на прийом до лікаря за допомогою цієї системи, потрібно зареєструватися за номером телефона, також використовуючи паспортні дані.

В Україні існує кілька систем, які дозволяють записатися до лікаря. Найпопулярнішою з них є система Helsi.me [3]. Ця система доступна в Інтернеті та на мобільних пристроях. Щоб записатися на прийом до лікаря за допомогою Helsi.me, потрібно створити обліковий запис і вказати свої контактні дані. Після цього можна вибрати лікаря та дату та час [4].

Також варто зазначити, що однією із ключових трансформацій в медичній сфері було впровадження телемедицини. Телемедицина використовує інформаційні та комунікаційні технології для забезпечення віддаленого доступу до медичних послуг. Це дозволяє лікарям проводити консультації, визначати діагнози та призначати лікування, необхідне для пацієнта, який знаходиться на відстані [5].

Застосування технологій сприяє ефективнішій обміну медичною інформацією між лікарями та пацієнтами. Електронні медичні записи дозволяють швидше та точніше документувати інформацію про пацієнта, а також полегшують обмін цією інформацією між різними медичними установами. Це підвищує координацію медичної допомоги та допомагає уникнути помилок через невірне тлумачення медичної інформації [6].

Зокрема, віддалена консультація та дистанційне спостереження за станом пацієнтів стали можливими завдяки технологічним засобам. Це особливо важливо в ситуаціях екстреної медичної допомоги та управління хронічними захворюваннями. Лікарі можуть віддалено моніторити показники здоров'я пацієнтів, надавати поради щодо лікування та вчасно втручатися у випадках загрозливих змін [7].

Під час пандемії COVID-19 віддалені консультації стали надзвичайно популярними через кілька ключових чинників, які вплинули на медичну систему та підходи до надання медичної допомоги. Віддалені консультації дозволили забезпечити безпечну взаємодію між лікарем та пацієнтом, не вимагаючи фізичного присутності в офісі чи лікарнях. Умови карантину та обмежень мобільності змусили пацієнтів шукати альтернативні способи отримання медичної допомоги, тому віддалені консультації забезпечили можливість

отримання медичної консультації без необхідності виходити з дому. Такого роду форма зв’язку з лікарями дозволила пацієнтам звертатися до лікарів зручним для них способом, уникати довгих черг та зменшити час очікування, що виявилось ефективним для розпізнавання та вирішення питань, які не потребують фізичного обстеження [8].

Комунікація між лікарем і пацієнтом через месенджери стала значущим аспектом віддаленої медичної консультації. Месенджери дозволяють лікарям отримати швидкий доступ до медичних питань та іншої інформації від пацієнтів, що корисно в ситуаціях, коли необхідно швидко реагувати на питання або симптоми пацієнта. Лікарі можуть використовувати месенджери для віддаленого спостереження за пацієнтами, особливо в умовах хронічних захворювань. Збір регулярної інформації від пацієнтів через чат може допомагати в ранньому виявленні змін у стані здоров'я [9].

1.2 Аналіз використання чат-ботів для консультації пацієнтів та діагностування хвороб

Використання чат-ботів для консультації пацієнтів та діагностування хвороб стало популярним напрямком в сучасній медицині, в особливості в контексті телемедицини та цифрового здоров'я. Чат-боти – це програми штучного інтелекту, які взаємодіють із користувачами через текстові чи голосові повідомлення, намагаючись вирішити їхні питання та надати необхідну інформацію [10].

Чат-боти можуть взаємодіяти з пацієнтами, надаючи їм інформацію про симптоми, захворювання, методи лікування та профілактику. Вони можуть відповідати на загальні медичні питання та надавати поради з основних аспектів здоров'я. Вони можуть запитувати пацієнтів про їхні симптоми та аналізувати цю інформацію для надання попередньої діагностики. Вони можуть слугувати як інструмент для попереднього визначення ступеня терміновості медичної консультації. Також чат-боти можуть служити засобом віддаленого моніторингу

хронічних захворювань, надсилаючи пацієнтам регулярні запитання та збираючи дані про їхній стан здоров'я [11].

Важливо зауважити, що чат-боти не можуть замінити повноцінну медичну консультацію та професійний медичний огляд. Вони можуть виступати як додатковий інструмент для надання інформації та попередньої діагностики, але остаточні рішення щодо лікування повинні приймати лікарі.

Для отримання лікування, лікарю необхідно здійснити діагностику пацієнта на прийомі. Діагностика хвороби – це процес встановлення причини захворювання на основі симптомів, анамнезу та результатів діагностичних тестів. Діагностика є важливою частиною медичного лікування, оскільки вона дозволяє лікарю розробити план лікування, який є найбільш ефективним для конкретного пацієнта [12]. Діагностика хвороби є складним процесом та має декілька етапів.

– збір анамнезу;
– фізичне обстеження;
– діагностичні тести.

Спочатку лікар запитує пацієнта про його симптоми, історію хвороби та інші фактори, які можуть бути пов'язані з захворюванням.

Симптом хвороби – це будь-яка відчутна зміна в організмі або його функціях, що виявляється на підставі скарг хворого (суб'єктивний симптом) або під час дослідження лікарем (об'єктивний симптом) [13].

Суб'єктивні симптоми – це те, що відчуває людина, наприклад біль, запаморочення, нудота, задишка. Об'єктивні симптоми – це те, що може виявити лікар під час обстеження, наприклад підвищення температури тіла, прискорене серцебиття, задишка, кашель [14].

Симптоми можуть бути загальними, тобто характерними для багатьох захворювань, або специфічними, тобто характерними для певного захворювання. Наприклад, біль у горлі є загальним симптомом, який може бути викликаний різними захворюваннями, такими як застуда, грип, тонзиліт, фарингіт. А ось

задишка, яка виникає при фізичному навантаженні, є специфічним симптомом, який може бути викликаний захворюваннями серця або легень [13].

Після збору анамнезу лікар проводить фізичне обстеження пацієнта. Фізичне обстеження включає в себе огляд пацієнта, а також пальпацію, перкусію та аускультацію. Під час огляду лікар звертає увагу на загальний стан пацієнта, його зовнішній вигляд, поставу, рухливість. Під час пальпації лікар обмацує тіло пацієнта, щоб виявити будь-які зміни у формі, розмірі або текстурі тканин.

Якщо лікар вважає це необхідним, він може призначити різні діагностичні тести. Діагностичні тести можуть бути лабораторними, інструментальними або генетичними. Лабораторні тести включають аналізи (наприклад загальний аналіз крові), які можуть допомогти лікарю визначити наявність або відсутність захворювання. Інструментальні тести включають рентгенографію, КТ, МРТ, ультразвукове дослідження та інші дослідження, які можуть допомогти лікарю отримати зображення внутрішніх органів або тканин пацієнта. Генетичні тести можуть допомогти лікарю визначити наявність генетичних захворювань [15].

Результатом діагностики є формування діагнозу. Діагноз – це медичний висновок про стан здоров'я обстежуваного, про наявне захворювання (травму) чи про причину смерті, виражений у термінах, передбачених прийнятими класифікаціями і номенклатурою хвороб [16].

Після проведення обстеження і постановки діагнозу лікар обговорює з пацієнтом результати обстеження та розробляє план лікування. План лікування може включати в себе медикаментозне лікування, фізіотерапію, дієту або інші заходи.

Отже, діагностика хвороби є важливою частиною медичного лікування, оскільки вона дозволяє лікарю розробити план лікування, який є найбільш ефективним для конкретного пацієнта. Традиційна діагностика включає в себе збір анамнезу, фізичне обстеження та проведення діагностичних тестів. Однак ці методи можуть бути трудомісткими і дорогими, а також можуть бути недоступними в віддалених районах. Завдяки розвитку телемедицини та

використання чат-ботів або месенджерів у якості засобу для віддаленої консультації, медицина стає все більш доступнішою для кожного. Тому розробка методу визначення діагнозу за текстовим описом симптомів NLP-засобами має потенціал для покращення доступності та ефективності медичної допомоги.

1.3 Огляд теоретичних підходів до розв’язку задачі визначення діагнозу засобами NLP

Штучний інтелект вже сьогодні має значний вплив на життя людини. Штучний інтелект – це галузь комп'ютерних наук, яка займається створенням інтелектуальних машин, здатних виконувати завдання, які зазвичай вимагають людського інтелекту [17].

Сьогодні ШІ має широке використання:

– у галузі фінансів для аналізу даних, прогнозування цін та управління активами;

– у галузі виробництва для автоматизації завдань, підвищення ефективності та якості продукції;

– у галузі роздрібної торгівлі для персоналізації пропозицій, оптимізації запасів та покращення обслуговування клієнтів;

– в медицині для діагностики захворювань, розробки нових ліків та створення персоналізованих планів лікування.

Засоби штучного інтелекту можна використовувати і під час віддаленої консультації та діагностики пацієнта лікарем у різних цілях. Засоби ШІ використовуються для автоматизації збору та аналізу інформації про пацієнта, що звільняє час лікаря для інших задач. Наприклад, для автоматичного заповнення медичних карт, аналізу лабораторних результатів та збору інформації про історію хвороби пацієнта. ШІ можна використовувати для аналізу медичних зображень, таких як рентгенівські знімки та МРТ, або для виявлення патологічних змін у крові. Також ШІ можна використовувати для аналізу даних про пацієнта, таких як його вік, стан здоров'я та історія хвороби,

для розробки плану лікування, який є найбільш ефективним для конкретного пацієнта [18].

ШІ можна розділити на кілька галузей, зокрема машинне навчання, розпізнавання образів, обробка природної мови.

Машинне навчання – це підгалузь ШІ, яка займається створенням алгоритмів, які можуть навчатися на даних і поліпшувати свої результати з часом. Машинне навчання використовується в різних областях, включаючи розпізнавання образів, розпізнавання мови, машинний переклад та рекомендаційні системи [19].

Комп’ютерний зір – це підгалузь ШІ, яка займається розробкою алгоритмів для розпізнавання об'єктів на зображеннях. Розпізнавання образів використовується в різних областях, включаючи охорону здоров'я, безпеку та транспорт [20].

Обробка природної мови – це підгалузь ШІ, яка займається розробкою алгоритмів для обробки та розуміння людської мови. Обробка природної мови використовується в різних областях, включаючи машинний переклад, розпізнавання мови та створення чат-ботів [21].

Для вирішення різних задач використовуються різного роду алгоритми, методи та моделі ШІ, залежно від конкретної задачі. Якщо розглядати методи ШІ в розрізі визначення діагнозу за текстовим описом симптомів, то одним із простих, але не менш популярних є алгоритм K-найближчих сусідів.

K-найближчі сусіди, K-Nearest Neighbors (KNN) – це простий, але ефективний алгоритм машинного навчання, який використовується для класифікації та регресії. Він використовує відстань між точками даних для визначення їх класу або значення (рисунок 1.5) [22].

KNN працює, використовуючи відстань між точками даних. Для класифікації KNN визначає клас точки даних, використовуючи її k найближчих сусідів. K – це параметр, який користувач повинен задати. Чим більше значення k, тим більшу роль грає більша відстань.

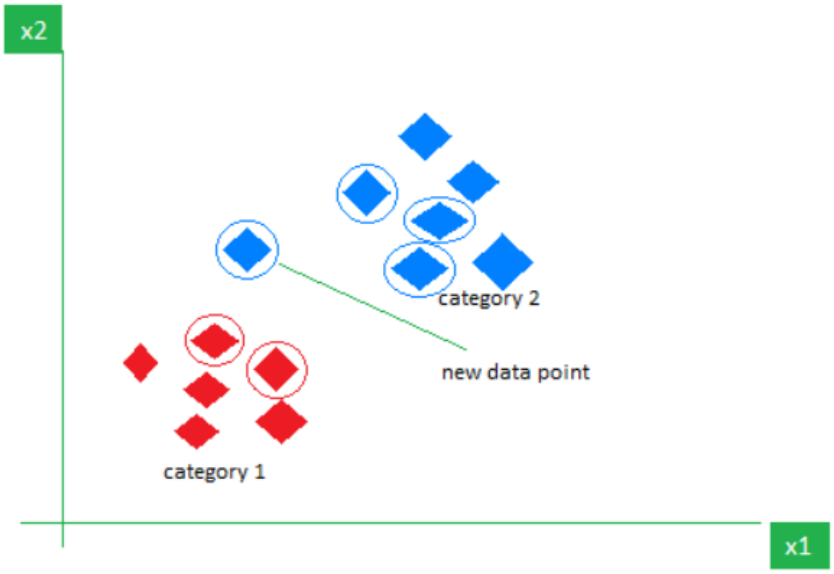


Рисунок 1.5 – Візуалізація методу KNN [23]

Наприклад, розглянемо набір даних з точками даних, що представляють людей. Кожна точка даних має два атрибути: зріст і вагу. Можна використовувати KNN для класифікації людей як «високих» або «низьких». Для цього можна вибрати значення K, наприклад, 5. Потім визначити клас кожної точки даних, використовуючи відстань до її п'яти найближчих сусідів. Якщо п'ять найближчих сусідів точки даних є «високими», то ми класифікуємо точку даних як «високу».

Значення K є важливим параметром KNN. Велике значення K може призвести до того, що KNN буде більш стійким до шуму, але також може призвести до того, що KNN буде менш точним. Мале значення K може призвести до того, що KNN буде більш точним, але також може призвести до того, що KNN буде більш чутливим до шуму [24].

KNN використовується в різних областях, включаючи класифікацію зображень, текстів, медичних даних, прогнозування погоди та для реалізації рекомендаційних систем. Також можна використовувати KNN для швидкого встановлення попереднього діагнозу за текстовим описом симптомів пацієнта.

Цей алгоритм має ряд переваг, як от [25]:

– він є непараметричним, що означає, що він не робить жодних припущень про розподіл даних;

– він є простим у використанні та реалізації;

– він може бути ефективним для класифікації даних, коли класифікатор має лише доступ до обмеженої кількості даних.

Отже, можна зробити висновок про те, що застосування алгоритму KNN для визначення діагнозу за текстовим описом симптомів є перспективним напрямком дослідження. Цей алгоритм може бути використаний для автоматизації процесу діагностики та надання лікарям додаткової інформації для прийняття рішень.

1.4 Аналіз існуючих програмних рішень

У мережі Інтернет є доволі багато різноманітних чат-ботів для консультації з лікарем. Вони представлені як у вигляді веб-застосунків, так і у вигляді мобільних додатків, далі наведено деякі з них.

EMed – це цифровий постачальник медичних послуг, який поєднує платформу на основі штучного інтелекту з віртуальними клінічними операціями для пацієнтів (рисунок 1.1). Пацієнти підключаються до медичних працівників через їх веб- та мобільний додаток [26].

Мета EMed – зробити медичну допомогу більш доступною та ефективною для людей у всьому світі. Компанія прагне зробити це, використовуючи технології для надання пацієнтам доступу до медичних працівників 24/7, незалежно від їхнього місця розташування чи фінансових можливостей.

На сьогоднішній день є застосунки, які за текстовим описом визначають захворювання людини. Вони використовують технологію машинного навчання для аналізу симптомів, які вводить користувач. Застосунки можуть бути корисними для швидкого виявлення потенційних проблем зі здоров'ям, але важливо пам'ятати, що вони не можуть замінити професійну медичну допомогу.

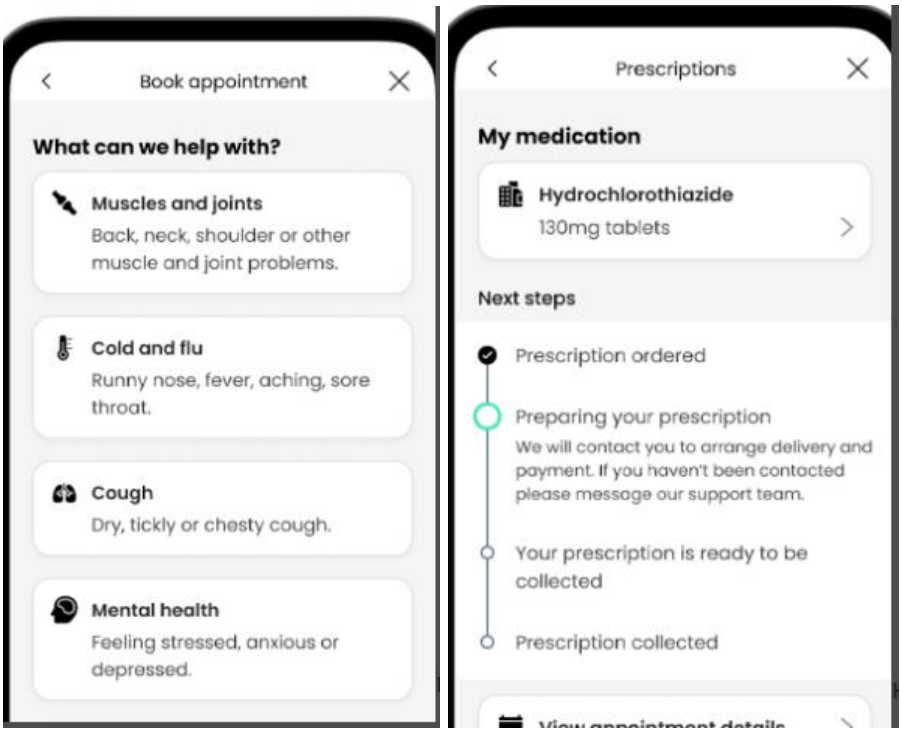


Рисунок 1.1 – Зовнішній вигляд застосунку EMed [26]

Ada – це застосунок, який використовує машинне навчання для аналізу симптомів, які вводить користувач (рисунок 1.2). Застосунок може визначити понад 100 різних захворювань, включаючи застуду, грип, інфекції сечовивідних шляхів та інші. Щоб використовувати застосунок Ada необхідно завантажити застосунок Ada з App Store або Google Play, запустити застосунок і створити обліковий запис. Після чого можна відповісти на запитання про свої симптоми, а застосунок Ada надасть попередній діагноз [27].

Застосунок Ada використовує технологію машинного навчання, яка навчається на величезному наборі даних симптомів та захворювань. Це дозволяє йому виявляти потенційні проблеми зі здоров'ям з високою точністю. Він простий у використанні.

WebMD – це один із найбільших і найпопулярніших медичних сайтів у світі (рисунок 1.3). Він пропонує широкий спектр інформації про здоров'я, включаючи статті, новини, симптоми, діагностику та лікування [28].

WebMD пропонує широкий спектр статей і новин про різні аспекти здоров'я, включаючи загальне здоров'я, захворювання, медицину, харчування, фізичні вправи та багато іншого. Сайт пропонує інформацію про симптоми

різних захворювань, включаючи опис симптомів, причини, діагностику та лікування. Також має корисні інструменти для самодіагностики та рекомендації щодо лікування різних захворювань.

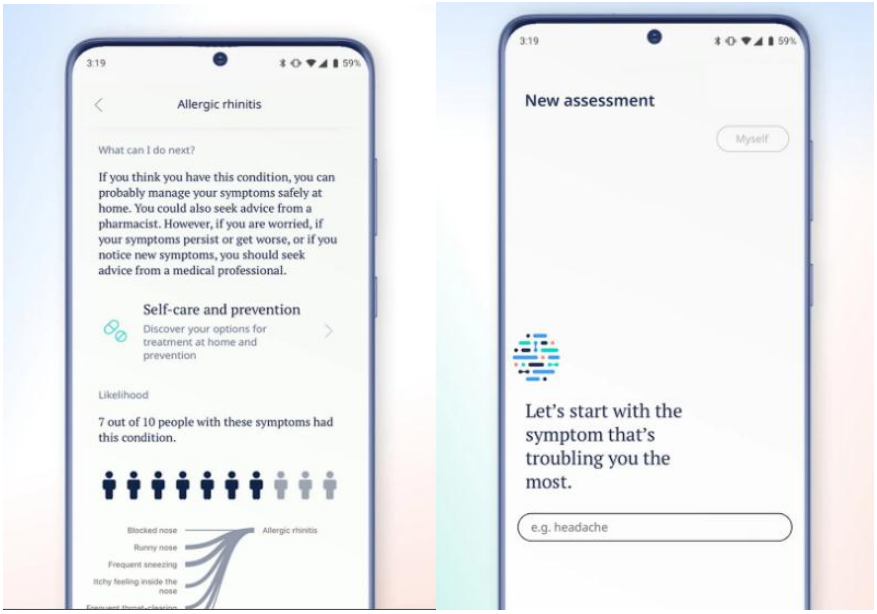


Рисунок 1.2 – Зовнішній вигляд застосунку Ada [27]

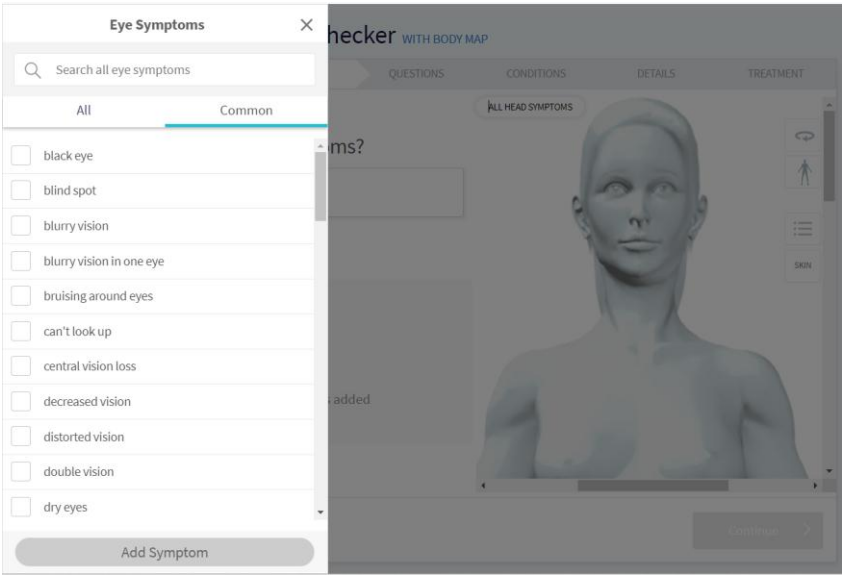


Рисунок 1.3 – Зовнішній вигляд застосунку WebMD [28]

Symptomate – це застосунок для самодіагностики, який використовує технологію машинного навчання для виявлення потенційних проблем зі здоров'ям (рисунок 1.4). Застосунок запитує користувачів про їхні симптоми, а

потім використовує ці дані для створення прогнозу щодо того, яке захворювання може бути у користувача. Symptomate також надає рекомендації щодо подальших дій, таких як відвідування лікаря або самолікування [29].

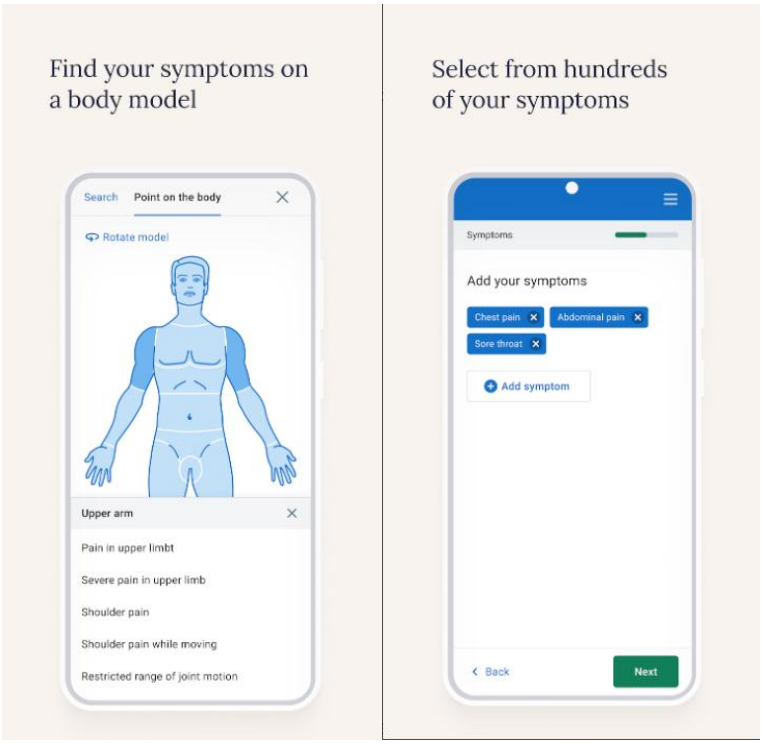


Рисунок 1.4 – Зовнішній вигляд застосунку Symptomate [29]

Symptomate доступний у 15 мовних версіях, включаючи англійську, іспанську, китайську, німецьку, французьку, португальську, арабську, голландську, чеську, турецьку, російську, українську, польську та словацьку.

Загалом Symptomate – це корисний інструмент для людей, які хочуть дізнатися більше про свої симптоми та отримати рекомендації щодо подальших дій. Застосунок простий у використанні і дає точну інформацію. Однак важливо пам'ятати, що Symptomate не може замінити професійну медичну допомогу.

Перспективи для розробки нового застосунку в галузі надання медичної допомоги та швидкої постановки попереднього діагнозу є значними. Ця галузь розвивається швидкими темпами, і постійно з'являються нові технології, які можна використовувати для поліпшення якості застосунків. Застосунки для онлайн-консультацій все ще є відносно новими, і їхня точність може бути не такою високою, як у лікаря. Хоча розробники застосунків працюють над

покращенням точності своїх продуктів, використовуючи такі методи, як машинне навчання та штучний інтелект, однак, вони все ще не є досконалими, тому подальша розробка застосування в даному напрямі є актуальним завданням.

Отже, діагностика хвороб є важливою частиною медичного лікування, оскільки вона дозволяє лікарю розробити план лікування, який є найбільш ефективним для конкретного пацієнта. Традиційна діагностика включає в себе збір анамнезу, фізичне обстеження та проведення діагностичних тестів. Однак ці методи можуть бути трудомісткими і дорогими, а також можуть бути недоступними в віддалених районах. Завдяки розвитку телемедицини та використання чат-ботів або месенджерів у якості засобу для віддаленої консультації, медицина стає все більш доступнішою для кожного. Тому визначення діагнозу за текстовим описом симптомів NLP-засобами має потенціал для покращення доступності та ефективності медичної допомоги та потребує автоматизації. Хоча розробники застосунків працюють над покращенням точності своїх продуктів, використовуючи такі методи, як машинне навчання та штучний інтелект, однак, вони все ще не є досконалими, тому подальша розробка застосування в даному напрямі є актуальним завданням.

1.5 Мета, задачі та вимоги до реалізації інформаційної системи

Метою кваліфікаційної роботи бакалавра є спрощення процесу діагностування шляхом автоматизованого визначення діагнозу за текстовим описом симптомів, для слід виконати розробку методу визначення діагнозу за текстовим описом симптомів NLP-засобами та відповідного програмного забезпечення у вигляді десктопного застосування.

Для досягнення поставленої мети слід вирішити такі завдання:

– виконати аналіз інформаційних моделей області діагностування;

– виконати огляд теоретичних підходів та обрати підхід для розв’язку задачі визначення діагнозу за текстовим описом NLP-засобами;

– провести аналіз існуючих програмних рішень;

– створити метод визначення діагнозу за текстовим описом симптомів NLP-засобами;

– описати функціональну структуру інформаційної системи діагностування пацієнтів за текстовим описом NLP-засобами;

– обрати набір даних для навчання класифікатора;

– створити відповідну програмну реалізацію на основі створеного методу визначення діагнозу;

– виконати тестування створеного ПЗ;

– виконати дослідження ефективності запропонованого методу визначення діагнозу з використанням розробленої інформаційної системи діагностування пацієнтів за текстовим описом NLP-засобами.

Розділ 2 Розробка методу визначення діагнозу за текстовим описом симптомів NLP-засобами

2.1 Схема методу визначення діагнозу за текстовим описом симптомів

Метод визначення діагнозу за текстовим описом симптомів NLP-засобами призначений для діагностування хвороб людини за текстовим описом симптомів користувачів. Схема методу наведена на рисунку 2.1.

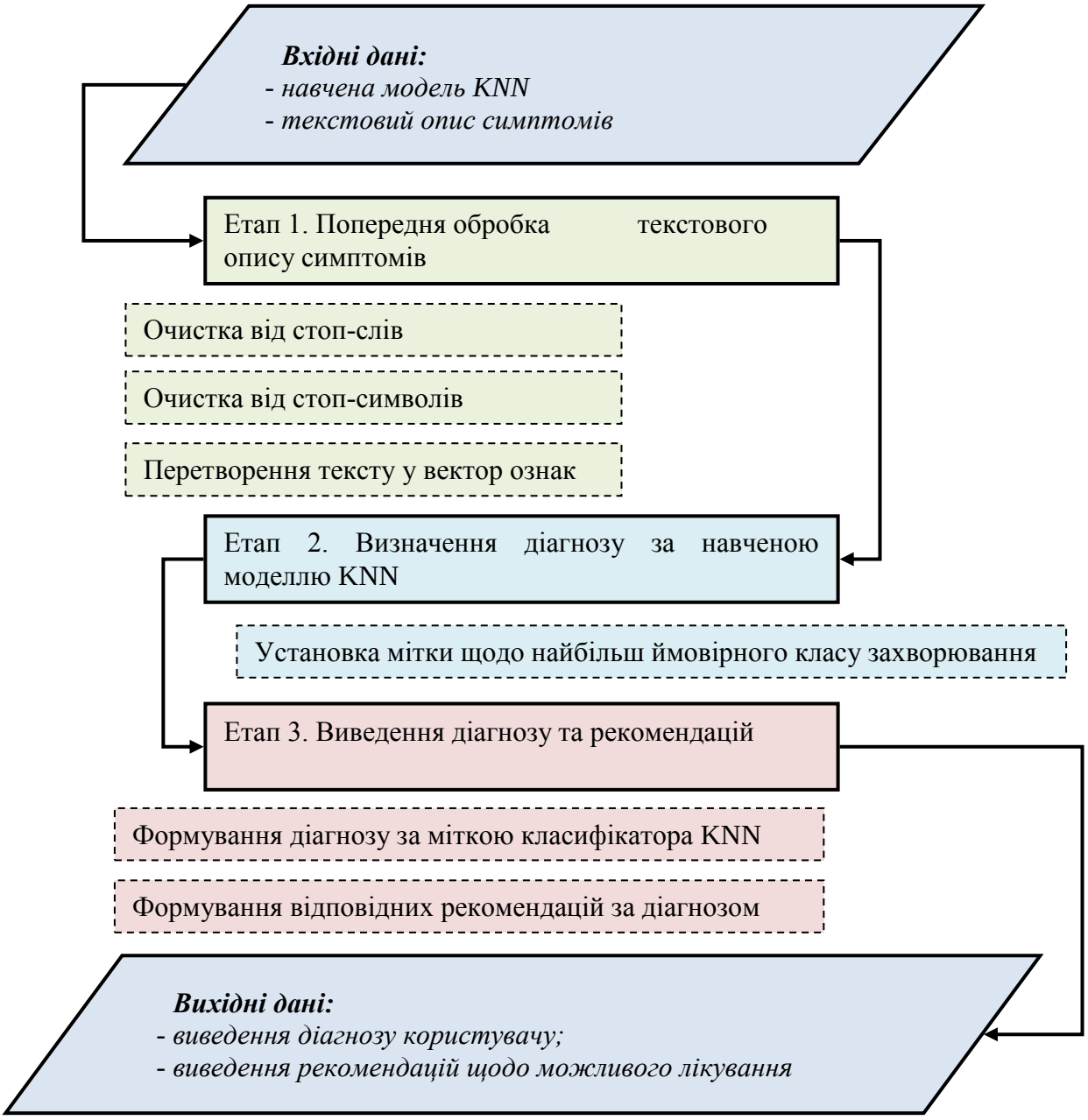


Рисунок 2.1 – Схема методу визначення діагнозу за текстовим описом симптомів

Метод перетворює вхідні дані у вигляді навченої моделі KNN та текстового опису симптомів у вихідні дані у вигляді виведення діагнозу користувачу та виведення рекомендацій щодо можливого лікування класифікованого захворювання.

Вхідними даними методу є навчена модель KNN та текстовий опис симптомів.

Перший етап роботи методу присвячений попередній обробці текстових даних. У попередню обробку входить очистка тексту опису симптомів від стоп-слів та стоп-символів та перетворення тексту у вектор ознак.

Наступним етапом є визначення діагнозу за попередньо навченою моделлю машинного навчання KNN, визначається його k найближчих сусідів у тренувальному наборі даних. Діагноз (клас) цього прикладу визначається на підставі більшості класів серед його найближчих сусідів, та присвоюється мітка, яка відповідає класу хвороби.

Третім етапом буде виведення діагнозу та відповідних рекомендацій, що включає в себе формування діагнозу за міткою класифікатора KNN та формування відповідних рекомендацій щодо лікування за діагнозом.

Вихідними даними роботи методу є діагноз користувача поставлений за текстовим описом та рекомендації щодо можливого лікування.

Отже, наведено схему та основні етапи методу визначення діагнозу за текстовим описом симптомів, що призначений для перетворення вхідних даних у вигляді навченої моделі KNN та текстового опису симптомів у вихідні дані у вигляді виведення діагнозу користувачу та виведення рекомендацій щодо можливого лікування класифікованого захворювання.

2.2 Функціональна структура інформаційної системи

Інформаційна система діагностування за текстовим описом складається із трьох підсистем та бази даних. Структура інформаційної системи діагностування захворювань та основні функції її складових наведені на рисунку 2.2.

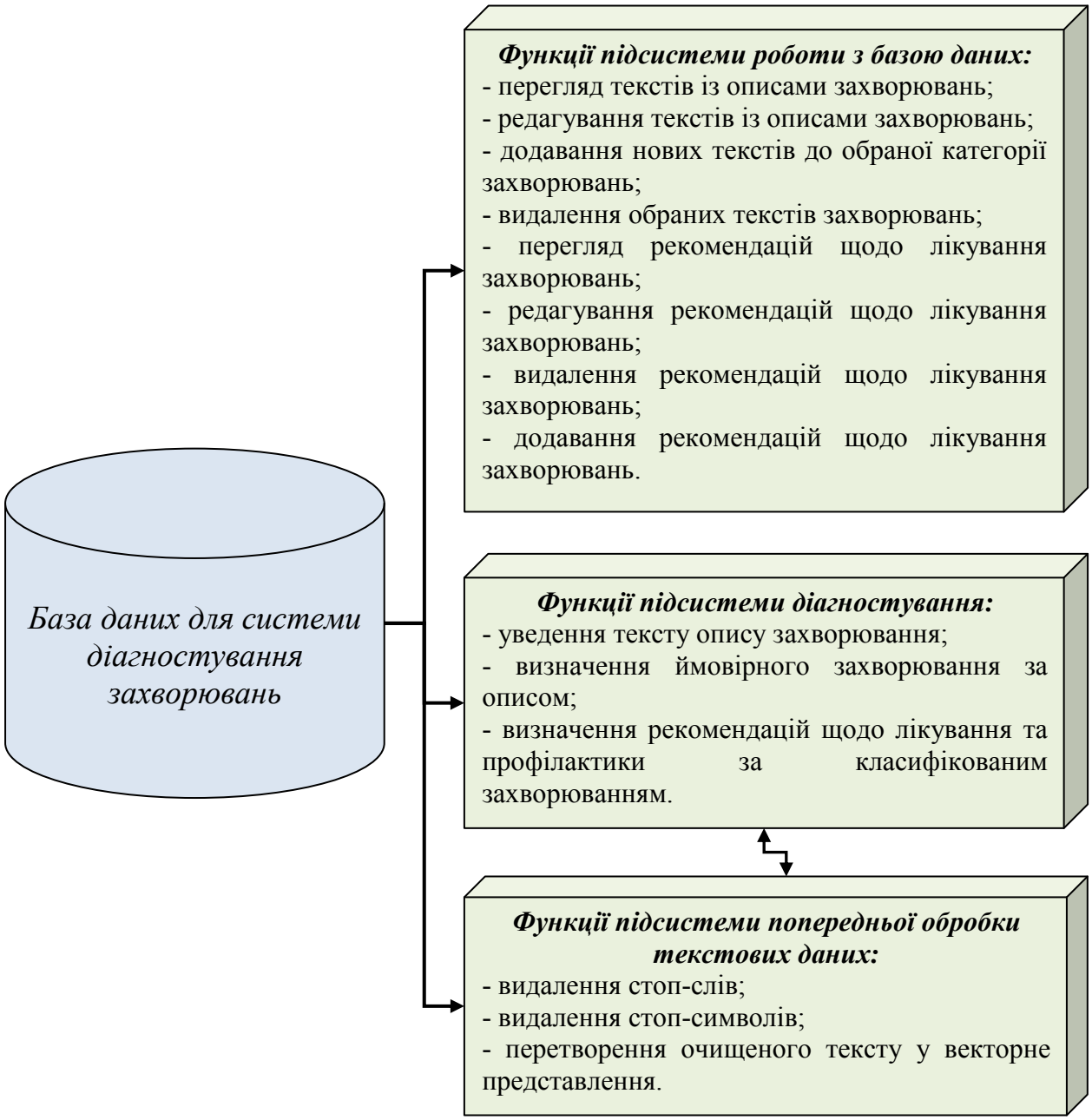


Рисунок 2.2 – Функціональна структура інформаційної системи діагностування захворювань

База даних для системи діагностування захворювань призначена для збереження даних захворювань та рекомендацій щодо їх лікування.

Підсистема роботи з базою даних основним призначенням має взаємодію з базою даних. Дана підсистема своїм функціями має перегляд, редагування, видалення та додавання нових текстів з описами захворювань, а також перегляд,

редагування, видалення та додавання нових рекомендацій щодо лікування захворювань.

Підсистеми діагностування є головною підсистемою, та призначена для безпосереднього діагностування захворювання за користувацьким текстовим описом. Своїми функціями має уведення тексту опису захворювання, визначення ймовірного захворювання за описом, а також визначення рекомендацій щодо лікування та профілактики за класифікованим захворюванням.

Підсистеми діагностування взаємодіє з підсистемою попередньої обробки текстових даних, яка призначена для очищення вхідного користувацького тексту від стоп-слів та стоп-символів та перетворення тексту-опису стану пацієнта на векторне представлення.

Отже, таким чином описано функціональну структуру інформаційної системи діагностування захворювань за текстовим описом, яка складається із трьох підсистем: роботи з базою даних, діагностування, попередньої обробки текстових даних і відповідної бази даних.

2.3 Формування пайплайну моделі KNN для визначення діагнозу за текстовим описом симптомів

У якості класифікатора буде використано модель машинного навчання KNN. Пайплайн для процесу установки діагнозу за текстовим описом симптомів наведено на рисунку 2.3.

Навчальна множина розмічених текстів за діагнозами проходить попередню обробку, де здійснюється очищення текстів від стоп-символів та стоп-слів. Очищені набори текстів перетворюються у векторний формат, де кожному тексту відповідає свій числовий вектор-представлення. Паралельно з множиною навчальних текстів формується множина міток, що відповідають кожному тексту. Мітка говорить про приналежність тексту визначеному класу захворювання. Мітки разом з очищеними векторизованими представленнями

текстів-описів хвороб йдуть вхідними даними, на яких буде навчатись класифікатор KNN.

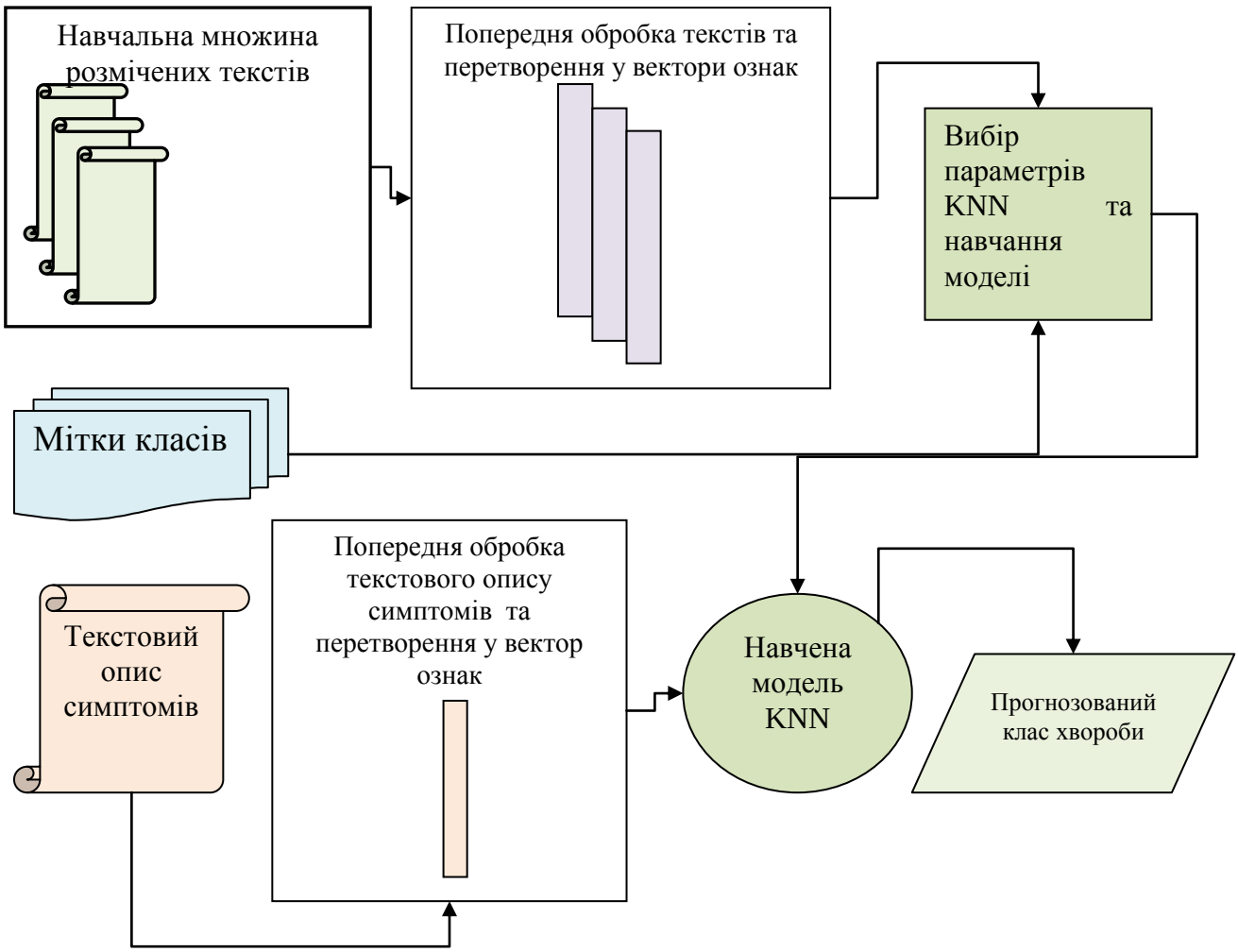


Рисунок 2.3 – Пайплайн KNN для установки діагнозу

Перед самим навчанням класифікатора задаються параметри ваг (*Weights*), кількості сусідів (*k*) та розміру листя (*Leaf Size*). Ваги використовуються для врахування впливу кожного симптома на кінцевий діагноз. Розмір листя відображає кількість симптомів, які вважаються для визначення діагнозу на кожному кроці алгоритму KNN. Це допомагає в оптимізації швидкості обчислень та уникненні перенавчання. Кількість сусідів визначає кількість найближчих сусідів, які будуть використовуватися для визначення діагнозу. За допомогою голосування або вагованого голосування (в залежності від параметра ваг) визначається кінцевий діагноз на основі класів цих сусідів. Після навчання модель готова до класифікації нових даних.

Текстовий опис симптомів аналогічно всьому текстовому набору також проходить попередню обробку даних та перетворення в вектор ознак. Далі перетворений вектор ознак подається навченій моделі KNN, яка видасть прогнозовану мітку хвороби, яка відповідає за клас захворювання.

Отже, таким чином наведено пайплайн моделі KNN для визначення діагнозу за текстовим описом симптомів.

2.4 Проєктування бази даних програмної системи

Ключовою складовою будь-якої інформаційної системи є належно структурована база даних, що відображає та зберігає необхідну інформацію для його оптимальної роботи. На рисунку 2.4 зображено модель бази даних інформаційної системи, яка базується на методі визначення діагнозу за текстовим описом симптомів за допомогою NLP-засобів.

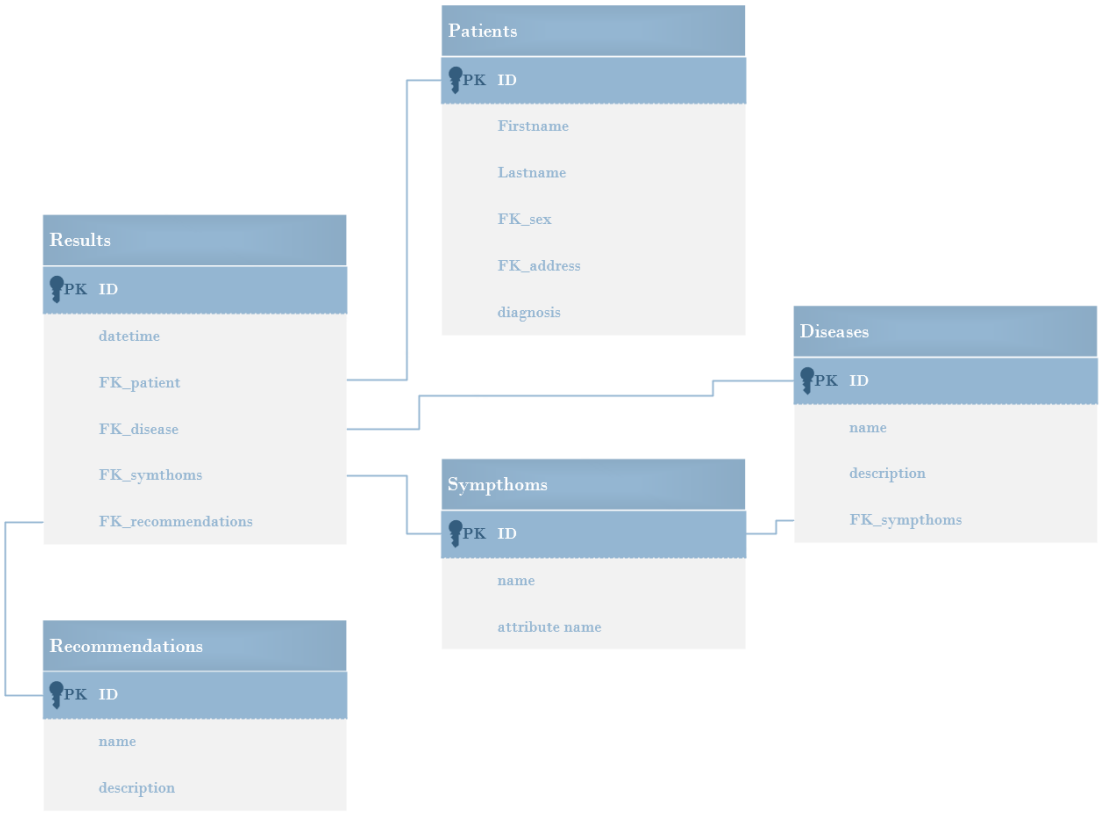


Рисунок 2.4 – Даталогічна модель бази даних на базі методу визначення діагнозу за текстовим описом симптомів

Створення бази даних, що ґрунтується на методі визначення діагнозу за текстовим описом симптомів за допомогою NLP-засобів, є важливим етапом у процесі проєктування Розробка структури бази даних відповідно до вимог інформаційної системи гарантує ефективне зберігання, організацію та швидкий доступ до даних, що є вирішальним для правильної роботи системи.

Таблиця «Patients» (таблиця 2.1) зберігатиме дані пацієнтів, а саме міститиме поля для прізвища та ім’я, статі, адреси проживання та визначений діагноз пацієнта.

Таблиця 2.1 – Атрибути таблиці «Patients»

№ п/п	**Назва**	**Тип даних**	**Опис**
1.	ID	int	Первинний ключ. Унікальний ідентифікатор пацієнта
2.	FirstName	varchar(255)	Ім’я пацієнта
3.	LastName	varchar(255)	Прізвище пацієнта
4.	FK_sex	int	Вторинний ключ, посилання на запис таблиці «Sex» для співставленням із відповідним записом про стать пацієнта
5.	FK_adress	int	Вторинний ключ, посилання на запис таблиці «Adress» для співставленням із відповідним записом про місце проживання пацієнта
6.	diagnosis	Varchar(255)	Назва діагнозу пацієнта

Таблиця «Results» (таблиця 2.2) призначена для збереження дати та часу запису результатів дослідження, інформації про пацієнтів, перелік існуючих діагнозів, симптомів, а також рекомендації з лікування.

Таблиця «Diseases» (таблиця 2.3) призначена для збереження інформації щодо захворювань, що можуть бути діагностовані при дослідженні. Таблиця

містить поля для запису назви захворювання, причин цього захворювання та симптомів.

Таблиця 2.2 – Атрибути таблиці «Results»

№ п/п	Назва	Тип даних	Опис
1.	ID	int	Первинний ключ. Унікальний ідентифікатор результату дослідження
2.	datetime	datetime	Дата запису результатів дослідження
3.	FK_patient	int	Вторинний ключ, посилання на запис таблиці «Patients» для співставленням із відповідним записом про пацієнта
4.	FK_disease	int	Вторинний ключ, посилання на запис таблиці «Diseases» для співставленням із відповідним записом про діагноз пацієнта
5.	FK_symptoms	int	Вторинний ключ, посилання на запис таблиці «Symptoms» для співставленням із відповідним записом про симптоми пацієнта
6.	FK_recommendations	int	Вторинний ключ, посилання на запис таблиці «Recommendations» для співставленням із відповідним записом про рекомендації щодо лікування пацієнта

Таблиця 2.3 – Атрибути таблиці «Diseases»

№ п/п	Назва	Тип даних	Опис
1.	ID	int	Первинний ключ. Унікальний ідентифікатор хвороби
2.	name	varchar(255)	Назва хвороби
3.	description	varchar(255)	Опис хвороби
4.	FK_symptoms	int	Вторинний ключ, посилання на запис таблиці «Symptoms» для співставленням із відповідним записом про симптоми, які притаманні хворобі

Таблиця «Symptoms» (таблиця 2.4) призначена для збереження назв симптомів, які є притаманними для тієї чи іншої хвороби.

Таким чином, була створена база даних для інформаційної системи діагностування на базі методу визначення діагнозу за текстовим описом симптомів за допомогою NLP-засобів. Це дозволить користувачеві ефективно та швидко отримувати доступ до необхідної інформації під час роботи з інформаційною системою.

Таблиця 2.4 – Атрибути таблиці «Symptoms»

№ п/п	Назва	Тип даних	Опис
1.	ID	int	Первинний ключ. Унікальний ідентифікатор назви симптому
2.	name	varchar(255)	Назва симптому

Таблиця «Recommendations» (таблиця 2.5) містить інформацію про рекомендації щодо лікування визначеної хвороби пацієнта.

Таблиця «Sexes» (таблиця 2.6) містить інформацію про назви статей.

Таблиця 2.5 – Атрибути таблиці «Recommendations».

№ п/п	Назва	Тип даних	Опис
1.	ID	int	Первинний ключ. Унікальний ідентифікатор назви симптому
2.	name	varchar(255)	Назва рекомендації
3.	description	varchar(255)	Опис рекомендації

Таблиця 2.6 – Атрибути таблиці «Recommendations».

№ п/п	Назва	Тип даних	Опис
1.	ID	int	Первинний ключ. Унікальний ідентифікатор назви симптому
2.	name	varchar(255)	Назва статі

Крім того, завдяки використанню бази даних SQLite, можливо забезпечити надійне зберігання та організацію медичної інформації. Це дозволить забезпечити високу швидкість доступу до даних та ефективну обробку запитів користувачів. Все це разом утворює надійну основу для створення корисного застосунку, який може значно полегшити процес визначення діагнозів у медичній практиці та покращити якість надання медичних послуг.

2.5 Особливості використання спеціалізованих програмних компонентів

Для полегшення процесу навчання класифікатора KNN можна використати ряд спеціалізованих програмних компонентів.

Pandas – це потужна бібліотека для обробки та аналізу даних, яка надає інструменти для роботи з табличними даними. Вона широко використовується в

області аналізу даних та наукових досліджень, а також в інших сферах програмування, таких як фінанси, економіка, соціальні науки, біологія і т.д [30]. Логотип бібліотеки Pandas наведено на рисунку 2.5.

Рисунок 2.5 – Логотип бібліотеки Pandas

Pandas дозволяє завантажувати, очищувати та підготовлювати дані для використання в моделі KNN. Це включає зчитування даних з файлів, обробку пропущених значень, перетворення категоріальних змінних в числові, масштабування функцій тощо.

Створення функцій відстані. Для роботи з алгоритмом KNN потрібно визначити міру відстані між об'єктами у просторі ознак. Зазвичай використовуються такі метри, як евклідова відстань або косинусна схожість. pandas може використовуватися для обчислення цих відстаней між векторами ознак.

Валідація моделі. Дана бібліотека допомагає здійснювати різні види валідації моделі, такі як перехресна перевірка, шляхом розділення даних на тренувальний та тестовий набори. Це дозволяє оцінювати ефективність моделі KNN на незалежних даних.

Також Pandas використовується для візуалізації результатів моделі KNN, таких як розподіл класів, матриці плутанини та інші важливі метрики ефективності.

Matplotlib.pyplot – це частина бібліотеки matplotlib, яка надає інтерфейс для створення графіків та візуалізації даних у середовищі Python. Ця бібліотека є потужним інструментом для створення різноманітних типів графіків, діаграм та

візуалізацій, що допомагає аналізувати дані, виявляти залежності та показувати результати досліджень чи аналізу у зрозумілій формі [31]. Логотип Matplotlib наведено на рисунку 2.6.

Рисунок 2.6 – Логотип Matplotlib

Основними функціями matplotlib.pyplot є можливість створення різних типів графіків, таких як лінійні графіки, гістограми, точкові графіки, графіки розподілу, контурні графіки та багато інших. Ці графіки можуть бути налаштовані з використанням різних параметрів для досягнення бажаного вигляду та стилю. Крім того, pyplot дозволяє додавати до графіків маркери, підписи, легенди, заголовки та інші елементи, які поліпшують їх інтерпретацію.

Matplotlib.pyplot є дуже гнучким інструментом, який легко інтегрується з іншими бібліотеками Python, такими як NumPy і pandas, що дозволяє використовувати його в різних сферах, включаючи аналіз даних, машинне навчання, наукові дослідження, інженерію та візуалізацію результатів. Багато функцій matplotlib.pyplot також можна налаштувати для автоматичної генерації великої кількості графіків та їх візуалізації у зручному форматі, що робить його незамінним інструментом для аналітики даних та досліджень.

NLTK (Natural Language Toolkit) – це бібліотека для обробки природної мови у Python. Вона є однією з найпопулярніших бібліотек для роботи з текстом та мовою завдяки своїм різноманітним функціям та можливостям. NLTK надає інструменти для аналізу тексту, токенізації, частиномовного розбору, визначення синтаксичних структур, роботи з корпусами текстів, класифікації тексту, роботи з морфологією та багато іншого [32]. Логотип бібліотеки NLTK наведено на рисунку 2.7.

Рисунок 2.7 – Логотип бібліотеки NLTK

Однією з ключових функцій NLTK є токенізація, яка включає розбиття тексту на окремі слова або фрази (токени). Це дозволяє легше обробляти текст та виконувати аналіз його структури. NLTK також містить багато різноманітних корпусів текстів для вивчення та експериментів, включаючи зразки текстів різних мов, літературні твори, статистичні дані та інше.

Додатково, NLTK має модуль для виконання частиномовного розбору, що дозволяє визначати частину мови кожного слова у тексті. Це корисна функція для багатьох завдань NLP, таких як побудова синтаксичних структур, виокремлення ключових слів та аналіз семантики тексту.

Крім того, NLTK включає модуль для роботи з машинним навчанням, який дозволяє тренувати моделі класифікації тексту на основі навчальних даних. Це дозволяє автоматично класифікувати текст на категорії або визначати відповідність тексту певним критеріям.

Scikit-learn, також відома як sklearn, є однією з найпопулярніших бібліотек для машинного навчання у Python. Вона надає широкий спектр алгоритмів машинного навчання, таких як класифікація, регресія, кластеризація, зменшення розмірності та багато інших, що дозволяє використовувати їх для різних завдань в області аналізу даних та інтелектуального аналізу [33].

Однією з ключових особливостей scikit-learn є простота використання та консистентний інтерфейс для різних алгоритмів. Вона має чітку та легко зрозумілу документацію, яка допомагає користувачам швидко розуміти особливості кожного алгоритму та правильно їх використовувати. Логотип бібліотеки scikit-learn наведено на рисунку 2.8.

Рисунок 2.8 – Логотип бібліотеки scikit-learn

Крім того, scikit-learn надає інструменти для підготовки даних, включаючи кодування категоріальних ознак, нормалізацію даних та видалення пропущених значень. Ці інструменти допомагають виконувати передобробку даних перед їх використанням у моделях машинного навчання, що забезпечує кращу якість та надійність моделі.

Крім того, scikit-learn має інструменти для оцінки та валідації моделей, включаючи різні метрики якості, розбиття навчального набору на навчальний та тестовий набори, а також крос-валідацію. Ці інструменти допомагають визначити ефективність моделі та уникнути перенавчання.

Узагальнюючи, scikit-learn є потужним інструментом для машинного навчання, який надає зручний та ефективний інтерфейс для використання різноманітних алгоритмів машинного навчання та їх застосування в практичних завданнях. Вона є незамінною бібліотекою для аналізу даних, класифікації, регресії та інших завдань машинного навчання у середовищі Python.

Враховуючи, що класифікатор KNN буде працювати із текстовими даними, то будуть використані такі програмні засоби: NLTK для попередньої обробки тестових на навчальних даних (видалення стоп-слів, стоп-символів, перетворення у вектор ознак), Pandas для завантаження навчального датасету та установки співвідношення між мітками та векторними представленнями текстів, matplotlib.pyplot для графічного представлення результатів, Scikit-learn для

навчання класифікатора інформаційної системи діагностування захворювань за текстовим описом симптомів.

2.6 Набір даних дослідження

Для навчання класифікатора KNN було використано набір даних «Symptom2Disease» [34].

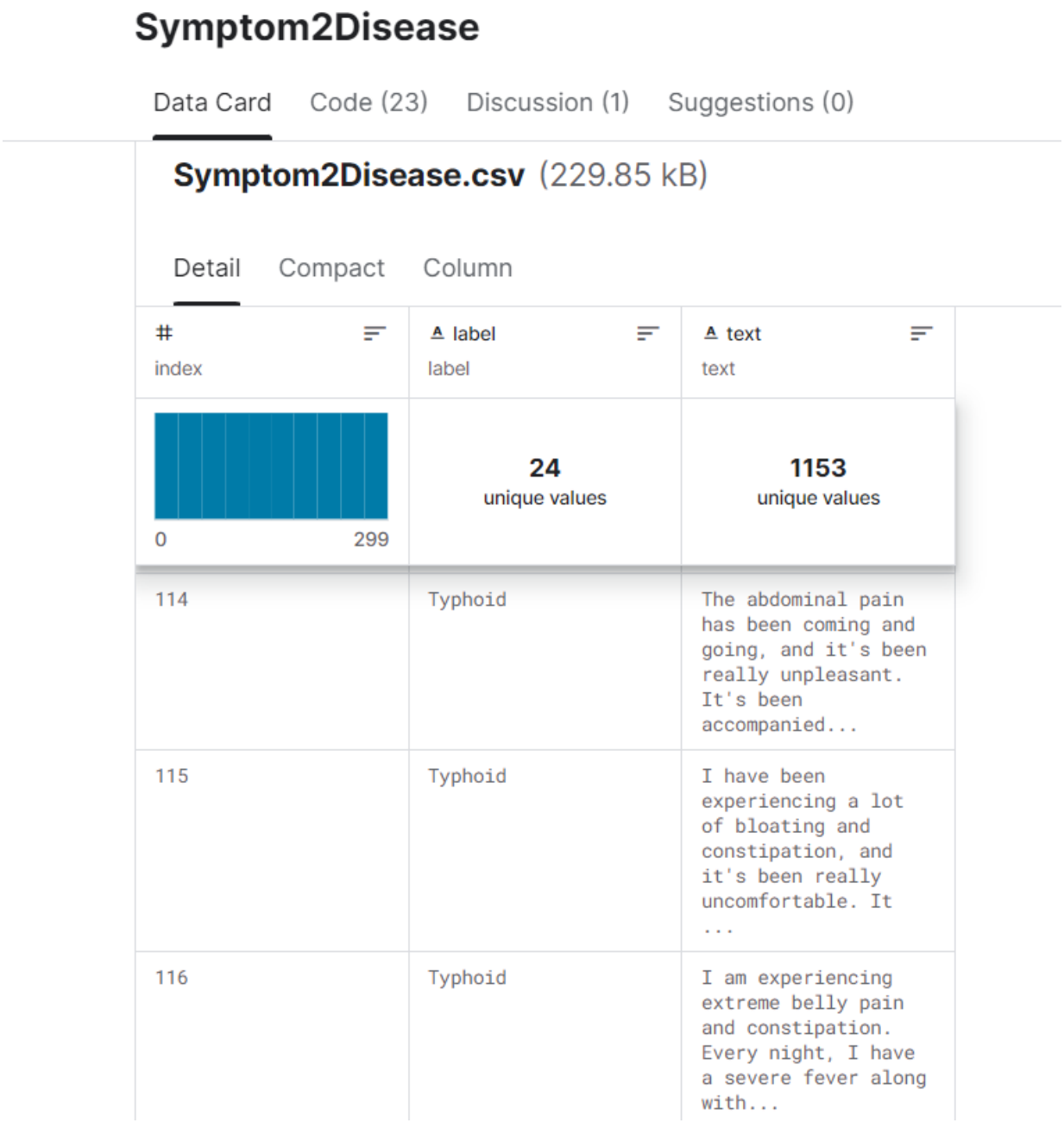


Рисунок 2.5 – Приклад даних набору для навчання класифікатора

Набір даних складається із 1200 записів даних і має два стовпці: «мітка» та «текст»: label містить мітки хвороби, а text містить описи симптомів природною мовою.

Набір даних англомовний, містить 24 різні захворювання, і кожна хвороба має 50 описів симптомів, що в результаті становить 1200 записів даних.

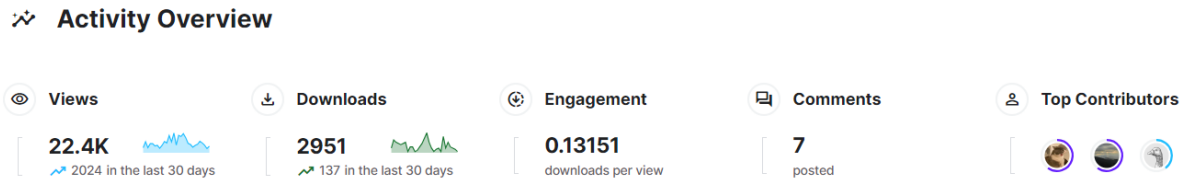
Наступні 24 хвороби були охоплені набором даних: псоріаз, варикозне розширення вен, тиф, вітряна віспа, імпетиго, денге, грибкова інфекція, застуда, пневмонія, диморфний геморой, артрит, акне, бронхіальна астма, гіпертонія, мігрень, шийний спондильоз, жовтяниця, малярія, інфекція сечовивідних шляхів, алергія, гастроезофагеальна рефлюксна хвороба, реакція на ліки, виразкова хвороба, цукровий діабет.

Приклад описів хвороб у наборі даних наведено на рисунку 2.5.

Також набір даних цілком збалансовано по довжині текстів. На рисунку 2.6 наведено розподіл середньої довжини текстів за діагнозами в наборі даних.

Рисунок 2.6 – Розподіл середньої довжини текстів за діагнозами

Набір даних є популярним серед науковців, про що свідчать перегляди та завантаження (рисунок 2.7)



Риснуок 2.7 – Статистика взаємодії з набором даних дослідження

Для подальшого використання набір даних було автоматично перекладено на українську мову.

2.7 Висновки до розділу 2

В рамках виконання другого розділу КРБ було створено метод визначення діагнозу за текстовим описом симптомів, а також наведено його схему та основні етапи. Метод призначений для перетворення вхідних даних у вигляді навченої моделі KNN та текстового опису симптомів у вихідні дані у вигляді виведення діагнозу користувачу та виведення рекомендацій щодо можливого лікування класифікованого захворювання.

Описано функціональну структуру інформаційної системи діагностування захворювань за текстовим описом, яка складається із трьох підсистем: роботи з базою даних, діагностування, попередньої обробки текстових даних і відповідної бази даних.

Наведено пайплайн моделі KNN для визначення діагнозу за текстовим описом симптомів, описано основні етапи його формування.

Створена база даних для інформаційної системи діагностування на базі методу визначення діагнозу за текстовим описом симптомів за допомогою NLP-засобів. Це дозволить користувачеві ефективно та швидко отримувати доступ до необхідної інформації під час роботи з інформаційною системою.

Оскільки класифікатор KNN буде працювати із текстовими даними, то будуть використані такі програмні засоби: NLTK для попередньої обробки тестових на навчальних даних (видалення стоп-слів, стоп-символів, перетворення у вектор ознак), Pandas для завантаження навчального датасету та установки співвідношення між мітками та векторними представленнями текстів, matplotlib.pyplot для графічного представлення результатів, Scikit-learn для навчання класифікатора інформаційної системи діагностування захворювань за текстовим описом симптомів.

Для навчання класифікатора KNN було використано набір даних «Symptom2Disease», що складається із 1200 записів даних і має два стовпці: «мітка» та «текст»: label містить мітки хвороби, а text містить описи симптомів природною мовою. Набір даних містить дані про 24 хвороби.

За методом визначення діагнозу за текстовим описом симптомів NLP-засобами потрібно в подальшому розробити застосунок, за допомогою якого провести дослідження ефективності методу. Також для доведення коректності результатів його треба окремо функціонально дослідити й протестувати.

Розділ 3 Експериментальне дослідження методу та програмна реалізація інформаційної системи

3.1 Визначення шляхів дослідження та засобів створення програмного забезпечення

Для створення програмного віконного застосунку, який визначатиме діагноз за текстовим описом симптомів за допомогою NLP-засобів, необхідно інтегрувати навчену модель класифікації KNN з базою даних та інтерфейсом користувача.

Одним із завдань застосунку буде зчитування інформації з бази даних про симптоми та діагнози, внесення нових записів про симптоми та їх діагнози, видалення існуючих записів, вибір текстового опису симптомів для аналізу, виведення результатів аналізу та відображення переліку ймовірних причин захворювання та рекомендацій щодо лікування.

Також потрібно провести тестування функціоналу та оптимізувати параметри навчання NLP моделі для досягнення найкращих результатів.

3.2 Вибір засобів розробки інформаційної системи визначення діагнозу за текстовим описом симптомів NLP-засобами

Для створення програмного застосунку на базі методу визначення діагнозу за текстовим описом симптомів за допомогою NLP-засобів необхідно обрати засоби, платформу та мову програмування, що найкраще підходить до обраної теми.

Для реалізації застосунків на базі ШІ сьогодні найбільшою популярністю користується Python.

Python – це високорівнева, інтерпретована, мультипарадигмальна програмна мова, яка була розроблена у 1991 році Гвідо ван Россумом. Мова Python має простий синтаксис, що дозволяє розробникам писати код швидше і ефективніше. Основною філософією Python є принцип «читабельності коду», що

робить його ідеальним вибором для початківців та досвідчених програмістів. Python підтримує об'єктно-орієнтований, процедурний, функціональний та аспектно-орієнтований підходи до програмування, що дає розробникам велику гнучкість у розробці програмного забезпечення [35].

Однією з найбільших переваг Python є велика та активна спільнота користувачів та розробників. Ця спільнота підтримує безліч бібліотек та фреймворків, що робить Python відмінним вибором для різноманітних проектів, від веб-розробки до аналізу даних та штучного інтелекту. Багато з цих бібліотек, таких як NumPy, pandas, TensorFlow, і PyTorch, роблять Python незамінним інструментом для роботи з науковими даними [36].

Ще однією перевагою Python є його платформенна незалежність. Це означає, що є можливість писати код на Python і запускати його на різних операційних системах, таких як Windows, macOS та різних дистрибутивах Linux. Це робить Python універсальною мовою програмування, яка підходить для різних середовищ розробки та різних задач.

Одна із найкращих IDE для цієї мови програмування на сьогодні – PyCharm. PyCharm – це інтегроване середовище розробки для мови програмування Python, що розроблене компанією JetBrains. Воно надає розробникам різні зручні інструменти для створення, редагування, тестування та відлагодження Python-коду. PyCharm має інтуїтивний інтерфейс, який дозволяє зосередитися на розробці без відволікання на налаштування середовища [37].

Однією з ключових особливостей PyCharm є його розширені можливості автодоповнення, що допомагають розробникам швидше писати код. Воно також включає в себе інтегровану систему керування версіями, що дозволяє зручно працювати з репозиторіями, такими як Git, безпосередньо з інтерфейсу IDE.

PyCharm підтримує розробку різноманітних типів проектів, включаючи веб-розробку, наукову розробку, робототехніку та машинне навчання. Воно також інтегрується з популярними фреймворками Python, такими як Django, Flask, та іншими, що спрощує розробку веб-застосунків.

Крім того, PyCharm має розширену систему відлагодження, яка дозволяє розробникам ефективно відлагоджувати свій код, виявляти та виправляти помилки. Воно підтримує різні типи точок зупинки, перегляд значень змінних, відстеження стеку викликів та багато іншого [38].

Загалом, PyCharm є потужним інструментом для розробки Python-програм, який допомагає розробникам зосередитися на розвитку своїх проектів та підвищити продуктивність.

Також для роботи застосунку передбачено використання БД. Для реалізації бази даних було обрано СКБД SQLite. Це компактна, вбудована база даних, яка зберігається у вигляді одного файлу, що робить її ідеальним вибором для вбудованих систем, мобільних додатків та невеликих веб-проектів. Вона реалізує легковаговий, серверний, самостійний, транзакційний SQL-рушій, який має високий ступінь стандартизації та сумісності з SQL-92.

Однією з ключових переваг SQLite є його простота використання. База даних SQLite не вимагає жодних серверів або конфігураційних файлів, що робить її дуже легкою встановити та використовувати. Також, SQLite підтримує багатопоточність, що дозволяє кільком процесам одночасно отримувати доступ до бази даних [39].

Ще однією перевагою SQLite є його кросплатформенність. Файли баз даних SQLite можна використовувати на різних операційних системах, таких як Windows, macOS та різних дистрибутивах Linux, що робить його відмінним вибором для розробки кросплатформенних додатків.

SQLite також підтримує багато функцій, які зазвичай асоціюються з реляційними базами даних, включаючи транзакції, індекси, обмеження цілісності даних та підзапити. Це робить SQLite потужним інструментом для роботи з даними та дозволяє йому конкурувати з іншими реляційними СКБД [40].

Загалом, SQLite є надійним, ефективним та легким використанням базою даних, яка підходить для різних типів проектів та додатків.

Таким чином, поєднання цих компонентів, дає міцну основу для розробки застосунку на базі методу визначення діагнозу за текстовим описом симптомів за допомогою NLP-засобів. Використання мови програмування Python та інтегрованого середовища розробки PyCharm дозволяє швидко та ефективно створювати та тестувати код. Бібліотеки та фреймворки Python, надають доступ до потужних інструментів для роботи з нейронними мережами та обробки природної мови. Використання бази даних SQLite дозволить зберігати та організовувати дані про симптоми та діагнози в зручному форматі. Це поєднання дозволить створити ефективний та корисний інструмент для автоматизованого визначення діагнозів на основі текстових описів симптомів, що може знайти застосування у сферах медицини та охорони здоров'я.

3.3 Структура та функціональне призначення програмних складових системи

Відповідно по поставленого завдання, необхідно реалізувати програмне забезпечення на базі методу визначення діагнозу за текстовим описом симптомів за допомогою NLP-засобів, що полягає у реалізації системи обробки природної мови, яка аналізує введені медичні дані та видає можливий діагноз на основі описаних симптомів. На рисунку 3.1 проілюстровано діаграму класів, створену для подальшої реалізації застосунку.

Клас textPreprocessing містить такі основні методи: tokenization, «stopwords та lemmatization. Метод tokenization призначений для розбиття тексту на токени. Токени – це окремі слова або фрази, які можуть бути використані для подальшої обробки тексту. Наприклад, якщо на вхід подається речення «Це класний день», то метод tokenization розбиває його на два токени: «Це» та «класний день».

Метод stopwords видаляє зайві слова з тексту, такі як «а», «в», «у» тощо. Ці слова не містять значення і не допомагають у розумінні тексту. Видалення

зайвих слів допомагає зменшити розмір тексту та покращити ефективність алгоритмів обробки тексту.

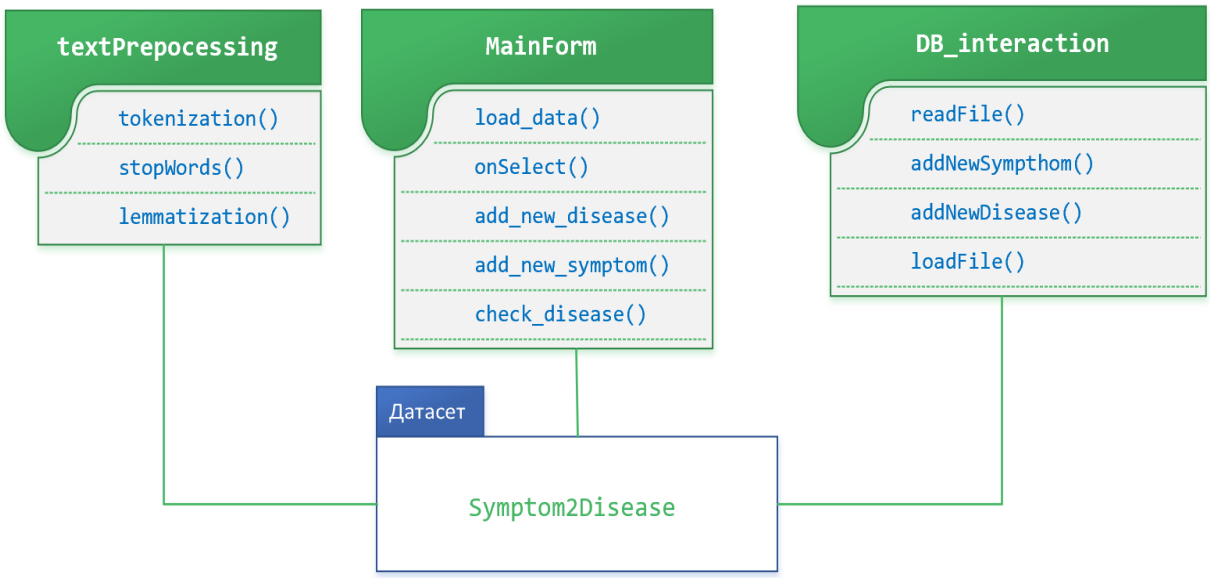


Рисунок 3.1 – Діаграма класів програмного застосунку

Метод lemmatization приводить слова до їх базової форми. Наприклад, слова “біжить”, “біг” та “бігти” будуть приведені до базової форми “бігти”. Це допомагає зменшити кількість унікальних слів у тексті та поліпшити ефективність алгоритмів обробки тексту.

Клас MainForm містить функції, що необхідні при роботі із визначенням діагнозу за описаними користувачем симптомами. Клас містить load_data, що використовується для завантаження даних датасету в програму.

Метод onSelect викликається при виборі елемента програмі, а саме запису користувача із скаргою. Окрім того, користувач може вручну ввести симптоми та отримати висновок щодо ймовірного захворювання.

add_new_disease – метод, що призначений для додавання нових захворювань до бази даних. Схожим методом є add_new_symptom, що використовується для додавання нових симптомів, пов’язаних з захворюваннями в базу даних.

check_disease – метод, що використовується для виведення ймовірного захворювання на основі заданих симптомів або критеріїв.

Таким чином, було створено діаграму класів програмного застосунку на базі методу визначення діагнозу за текстовим описом симптомів NLP-засобами.

3.4 Особливості реалізації програмних складових інформаційної системи визначення діагнозу за текстовим описом симптомів

Програмний застосунок на базі методу визначення діагнозу за текстовим описом симптомів за допомогою NLP-засобів, інформаційна система визначення діагнозу за текстовим описом симптомів NLP-засобами, передбачає ряд кроків обробки та аналізу тексту.

Першим кроком в роботі застосунку є нормалізація тексту, а саме очищення від розділових знаків, видалення стоп-слів та приведення тексту до нижнього регістру.

На етапі препроцесингу тексту буде використовуватись функція clean_text. На цьому кроці видаляються пунктуаційні знаки з речення, після цього використовується метод strip, щоб видалити можливі пробіли на початку та в кінці рядка.

Наступним етапом є видалення слів, які не несуть інформації про зміст тексту. Завантажуються стоп-слова з бібліотеки NLTK, розбиваються речення на слова за допомогою word_tokenize, та залишаються лише ті слова, які не є стоп-словами.

Для візуалізації результати роботи методів було створено хмару слів із частовживаними словами, що зустрічаються в датасеті (рисунок 3.2).

На останньому етапі очищений текст об’єднується назад у рядок за допомогою функції join, і весь текст перетворюється до нижнього регістру за допомогою функції lower.

Рисунок 3.2 – Хмара частовживаних слів

Для визначення діагнозу було обрано метод k-найближчих сусідів (KNN) є простим та ефективним алгоритмом машинного навчання для класифікації та регресії. Цей алгоритм використовує ідею того, що об’єкти, які близькі один до одного в просторі ознак, мають схожі мітки. Для перевірки результати роботи методу, що визначатиме ймовірний діагноз за описом симптомів, було створено функцію report. Це функція, яка використовується для створення звіту про результати класифікації. В звіті виводяться основні метрики оцінки класифікації, такі як точність (accuracy) та звіт про класифікацію (classification report). Використовуючи функцію accuracy_score з бібліотеки scikit-learn, обчислюється точність, а функція classification_report з бібліотеки scikit-learn, генерує та виводить детальний звіт про здійснену класифікацію, а саме точність, функція втрат та F1-оцінка для кожного класу, а також узагальнені значення по всім класам.

В таблиці 3.1 наведено результати, що повертає функція report при аналізі реалізованого методу визначення ймовірного діагнозу. Отримані значення точності в результаті перевірки методу на тестовій вибірці даних – 0,97

Таблиця 3.1 – Фрагмент результатів роботи функції report

Назва захворювання	**Accuracy**	**Recall**	**F1-score**
Акне	1,00	1,00	1,00
Артрит	1,00	1,00	1,00
Бронхіальна астма	0,92	1,00	0,96
Шийний спондильоз	1,00	1,00	1,00
Вітряна віспа	0.85	0.92	0.88
Лихоманка денге	0.79	0.92	0.85
Грибкова інфекція	1,00	1,00	1,00

Окрім того, користувачеві інформаційної системи визначення діагнозу за текстовим описом симптомів NLP-засобами забезпечується можливість взаємодіяти із набором даних, а саме додавати нові записи симптомів та захворювань. Для цього було реалізовано окреме вікно інтерфейсу та відповідні функції (рисунок 3.3).

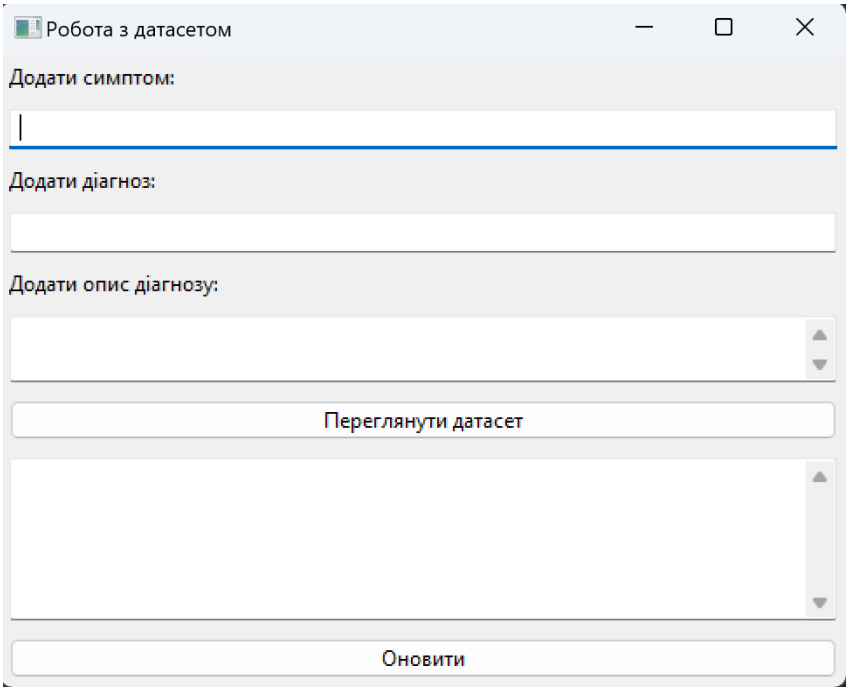


Рисунок 3.3 – Вікно застосунку для роботи з датасетом

На цьому вікні користувачеві реалізовано функції для внесення нового симптому, діагнозу та його опису. Окрім того, можна переглянути датасет, результат виведення датасету на екран наведено на рисунку 3.4.

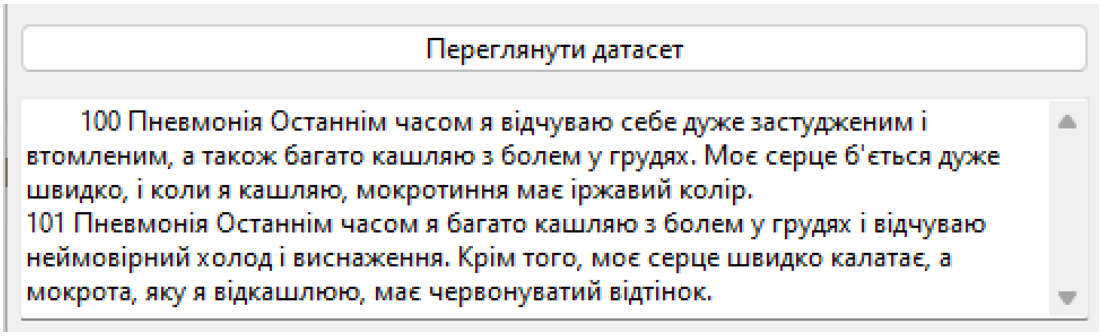


Рисунок 3.4 – Результат відображення датасету

Для отримання результатів у програмному застосунку відносно ймовірного діагнозу користувачеві надано наступний інтерфейс (рисунок 3.5).

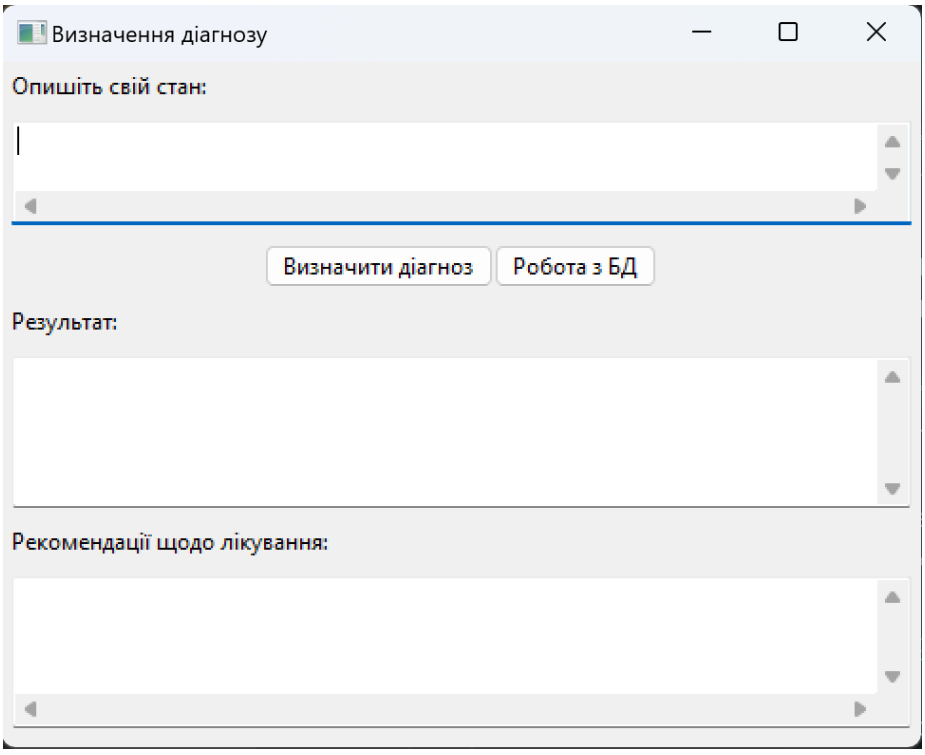


Рисунок 3.5 – Інтерфейс інформаційної системи для визначення ймовірного діагнозу за текстовим описом симптомів NLP-засобами

Для отримання результату користувачеві необхідно ввести текстовий опис свого стану, у перше текстове поле, наприклад «*Я сильно схуднув за останній тиждень, тому що не міг багато їсти через нудоту. До цього додалася висока температура, головний біль і біль у шлунку*». Після натиснення на кнопку «Визначити діагноз» результат виводиться в друге текстове поле, «Результат». Також окрім імовірного діагнозу в текстовому полі нижче виводитиметься інформація щодо рекомендацій з лікування. Результат роботи програмного коду наведено нижче (рисунок 3.6)

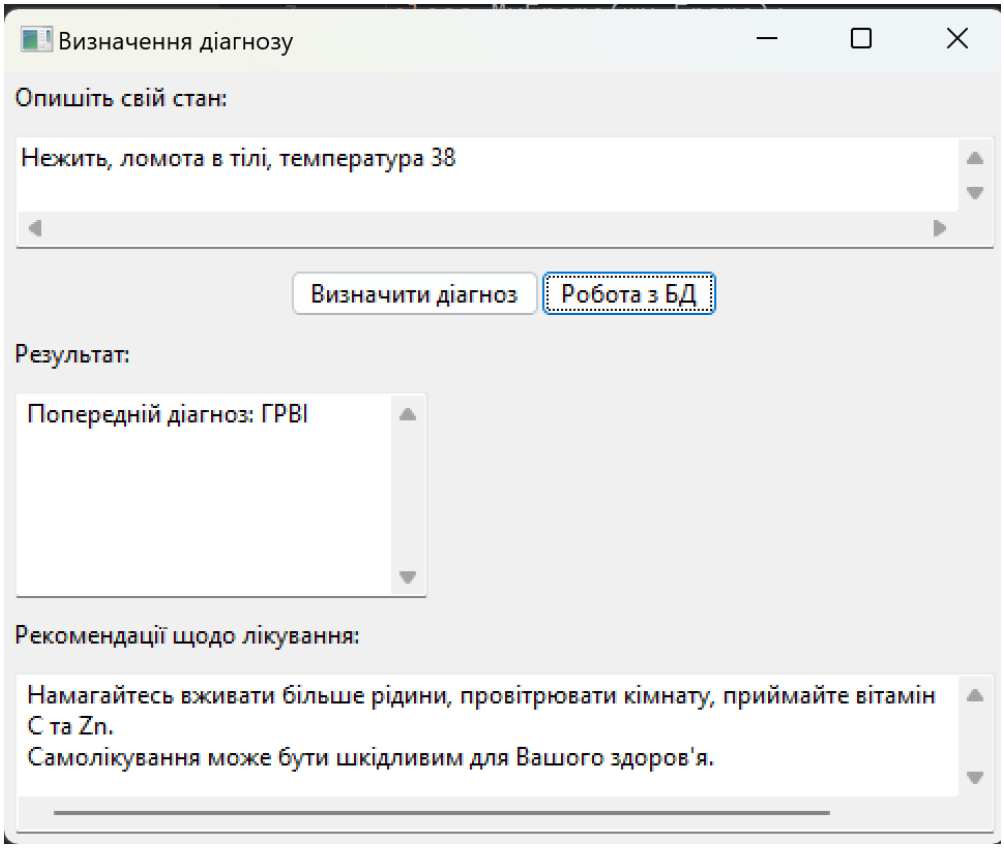


Рисунок 3.6 – Результат роботи програмного коду інформаційної системи визначення діагнозу за текстовим описом симптомів NLP-засобами

Таким чином, було реалізовано інформаційну систему визначення діагнозу за текстовим описом симптомів NLP-засобами. Створений програмний продукт містить функції для визначення ймовірного діагнозу та текстовим описом, внесення нової інформації та роботи з датасетом.

3.5 Тестування інформаційної системи визначення діагнозу за текстовим описом симптомів та вимоги до розгортання

Для валідації отриманої інформаційної системи визначення діагнозу за текстовим описом симптомів NLP-засобами необхідно провести належне тестування. Для цього було виконано ряд тест-кейсів та юніт-тестування.

Перший тест-кейс (таблиця 3.2) створено для перевірки методів визначення діагнозу за описом симптомів. Користувачеві необхідно запустити програмний застосунок, у відповідне поле для введення симптомів помістити потрібний текст та переглянути результати.

Таблиця 3.2 – Тест-кейс AA-0001

<table>
<tr><td>Тест-кейс ID: AI0001</td><td>Пріоритет: 1</td><td>Створено:25.03.2024</td></tr>
<tr><td colspan="3">Назва: Тест-кейс для перевірки методів визначення діагнозу за описом симптомів.</td></tr>
<tr><td colspan="2">Кроки</td><td>Очікуваний результат</td></tr>
<tr><td colspan="2">1. Запустити програмний застосунок;
2. Ввести необхідний опис стану хворого у відповідне текстове поле;
3. Натиснути кнопку «Визначити діагноз»;
4. Порівняти отриманий результат з очікувним.</td><td>У текстовому полі для виведення «Результат» діагноз «Пневмонія»</td></tr>
<tr><td colspan="3">Результат виконання тест-кейсу: пройдено успішно</td></tr>
</table>

Для виконання тесту було обрано текст-скаргу від пацієнта сімейному лікарю: «*У мене розвивається інтенсивний кашель, який посилюється вночі. Кашель супроводжується великою кількістю мокроти. Відчуваю дискомфорт та біль в грудях, особливо при глибокому вдиханні або кашлю. Моє дихання стало швидшим і менше ефективним. Я часто відчуваю нестачу повітря, навіть при невеликому фізичному навантаженні. У мене піднята температура, та я відчуваю загальну слабкість та втомленість.*»

Результат виконання тест-кейсу наведено на рисунку 3.7. Програмний продукт визначає, що ймовірне захворювання за описом скарг пацієнта – пневмонія, а також в рекомендаціях зазначено візит до сімейного лікаря.

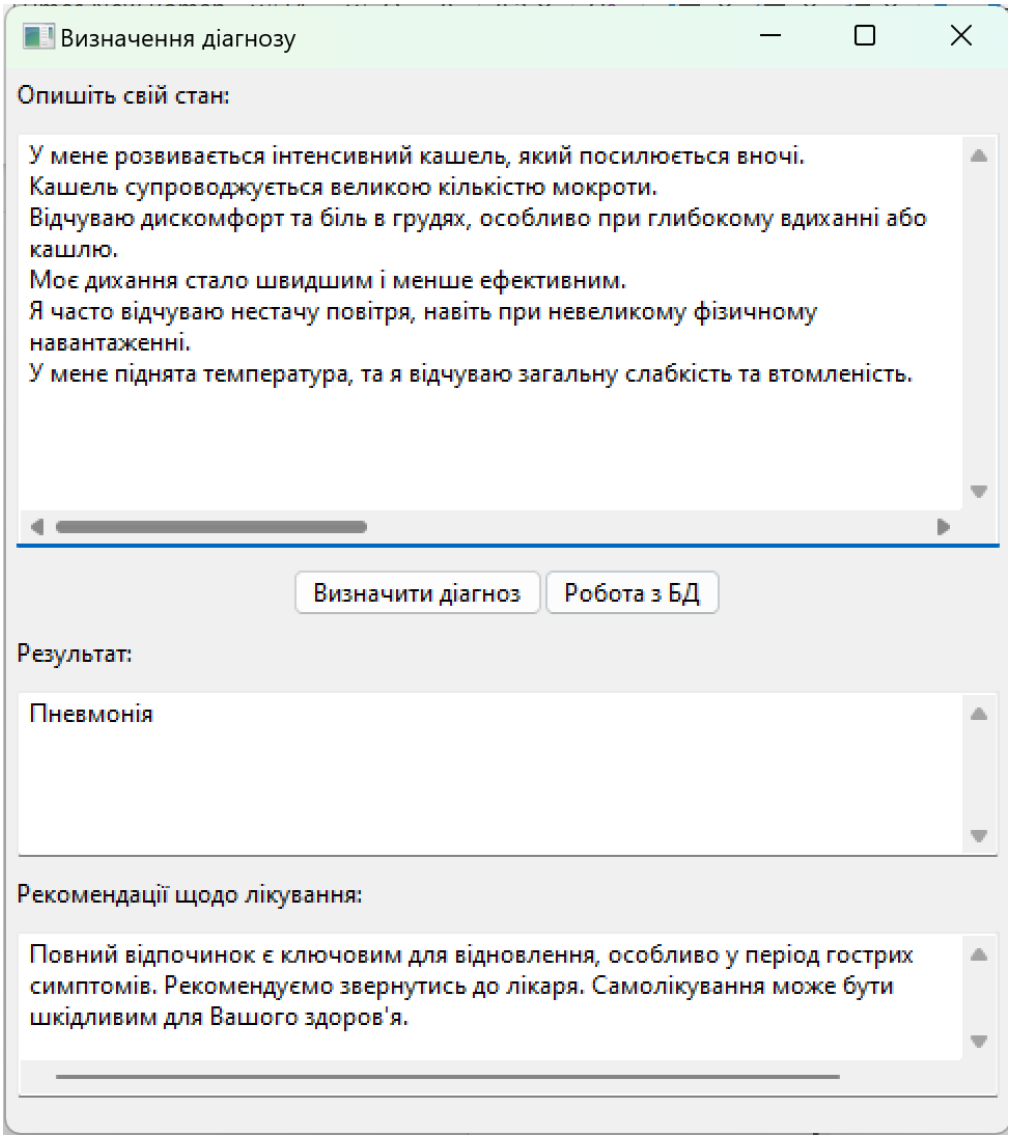


Рисунок 3.7 – результат виконання тест-кейсу АА-0001

Наступний тест-кейс (таблиця 3.3) призначений для перевірки роботи методів для взаємодії з датасетом. Тест-кейс передбачає перевірку методу для запису нових даних. Для цього необхідно запустити застосунок, на головному вікні натиснути кнопку «Робота з БД». У вікні, що з’явилось на екрані необхідно ввести дані у зазначені текстові поля.

Таблиця 3.3 – Тест-кейс AA-0002

Тест-кейс ID: AI0002	**Пріоритет: 2**	**Створено:30.03.2024**
Назва: Тест-кейс для перевірки методів роботи із датасетом		
Кроки		**Очікуваний результат**
1. Запустити програмний застосунок; 2. Натиснути кнопку «Робота з БД»; 3. У вікні, що з’явилось ввести необхідну інформацію; 4. Натиснути кнопку «Оновити»; 5. Порівняти отриманий результат з очікуваним.		Виведення повідомлення «Дані оновлено!»
Результат виконання тест-кейсу: пройдено успішно		

Якщо користувач коректно заповнив поля, застосунок повертає повідомлення «Дані оновлено!».

Результат виконання тест-кейсу наведено на рисунку 3.8. На зображенні можна переглянути, що дані було додано й виведено відповідне повідомлення.

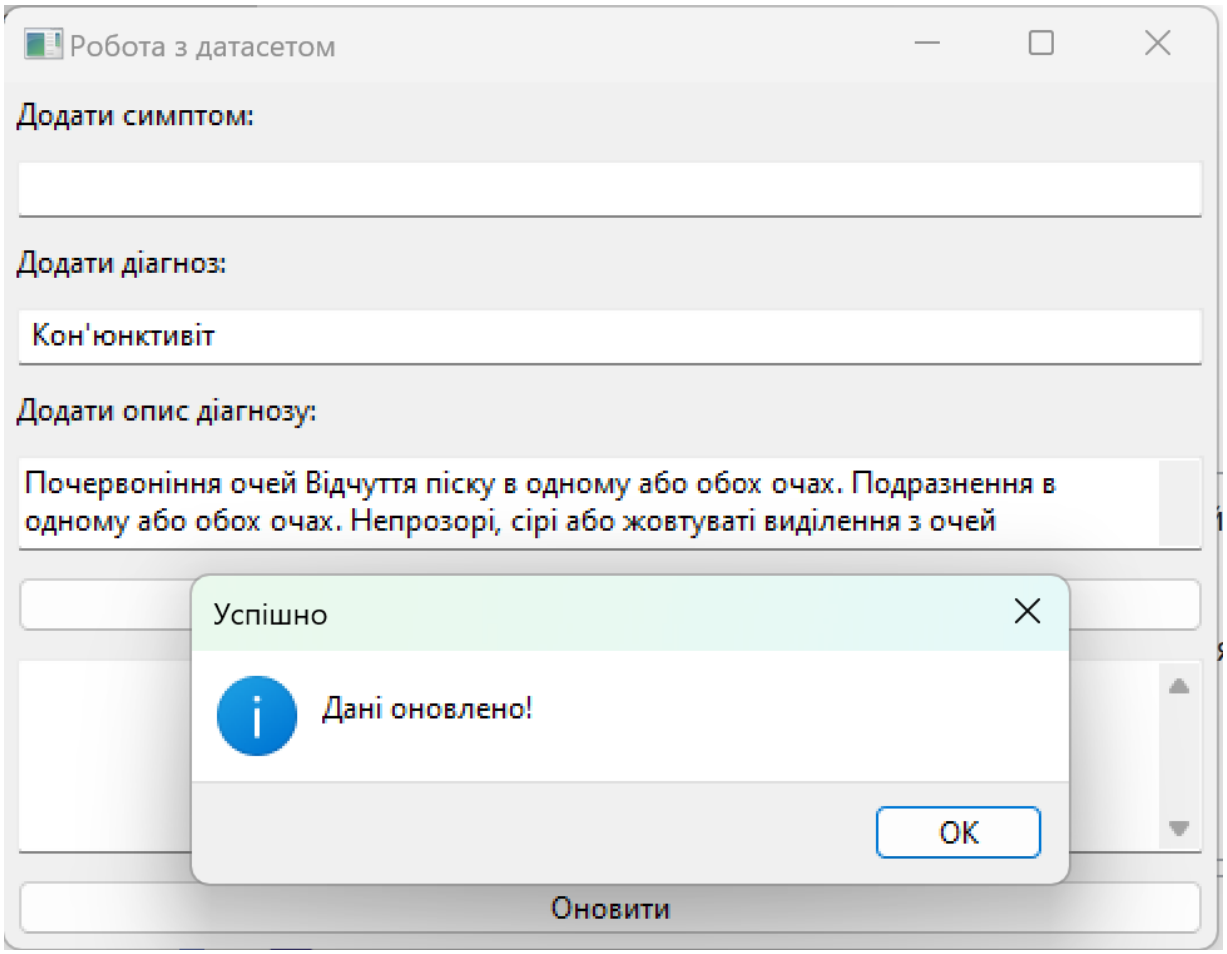


Рисунок 3.8 – Результат виконання тест-кейсу AA-0002

Також було реалізовано unit-тест для перевірки головної функції застосунку – визначення ймовірного діагнозу за описом симптомів користувача. Результат успішного виконання тесту наведено на рисунку 3.9.

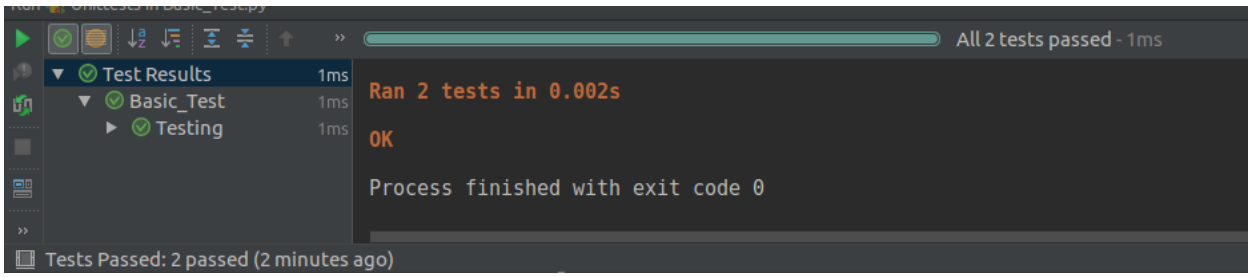


Рисунок 3.9 – Результат виконання те unit-тесту

Таким чином, було виконано ряд тест-кейсів та ю unit-тестування для валідації отриманого програмного продукту на базі методу визначення діагнозу за текстовим описом симптомів за допомогою NLP-засобів необхідно провести

належне тестування. Отримані результати свідчать про належну реалізацію програмного коду, інформаційна система визначення діагнозу за текстовим описом симптомів NLP-засобами виконує функції, поставлені при написанні мети роботи.

3.6 Аналіз функціональності інформаційної системи визначення діагнозу за текстовим описом симптомів NLP-засобами

Написання доступної та зрозумілої інструкції для користувача є критично важливим етапом у розробці інформаційної системи визначення діагнозу за текстовим описом симптомів NLP-засобами.

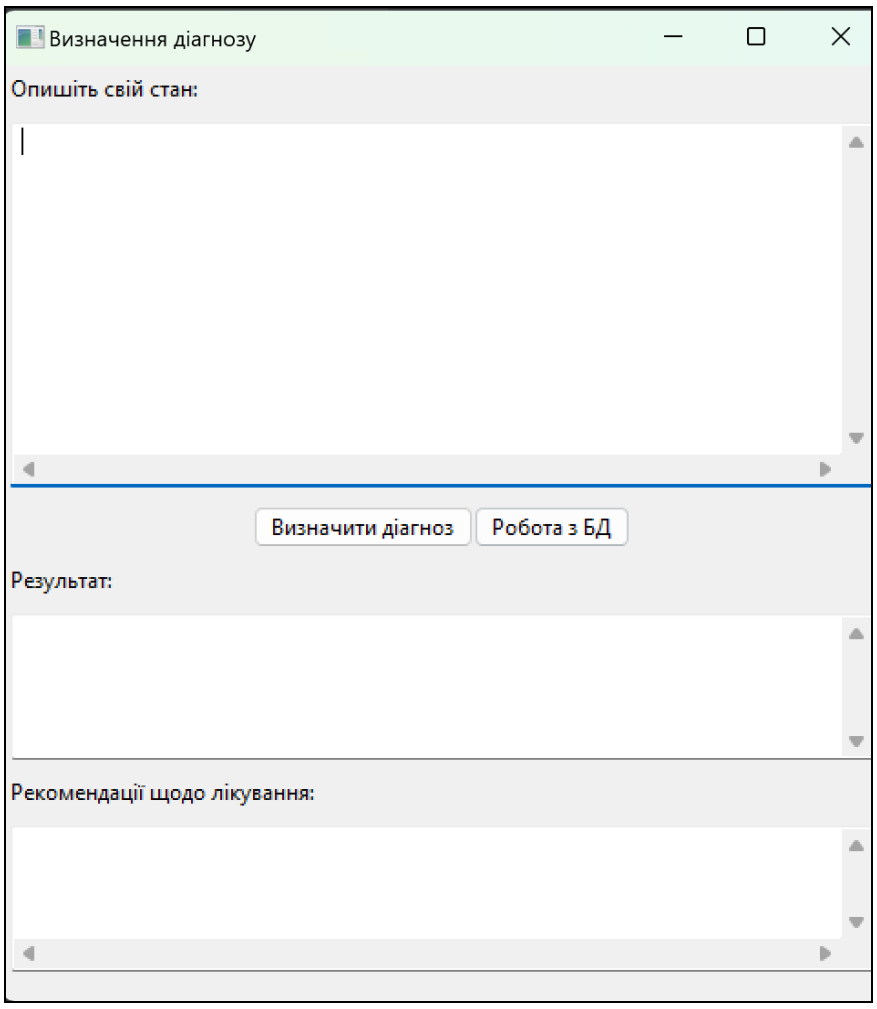


Рисунок 3.10 – Вигляд головного вікна інформаційної системи визначення діагнозу за текстовим описом симптомів NLP-засобами

Це допомагає забезпечити ефективне використання програми широким спектром користувачів. Для цього нижче представлено інструкцію з використання програного застосунку на базі методу визначення діагнозу за текстовим описом симптомів за допомогою NLP-засобів.

Запустивши інформаційну систему визначення діагнозу за текстовим описом симптомів NLP-засобами, користувач бачить головне вікно, на якому можна взаємодіяти із основним функціоналом програми чи перейти до розділу для роботи з датасетом (рисунок 3.10).

Натиснувши на кнопку «Робота з БД» користувач має змогу перейти до вкладки для здійснення взаємодії із датасетом, а саме: переглянути його, оновити, внести нові симптоми чи захворювання з відповідним описом (рисунок 3.11).

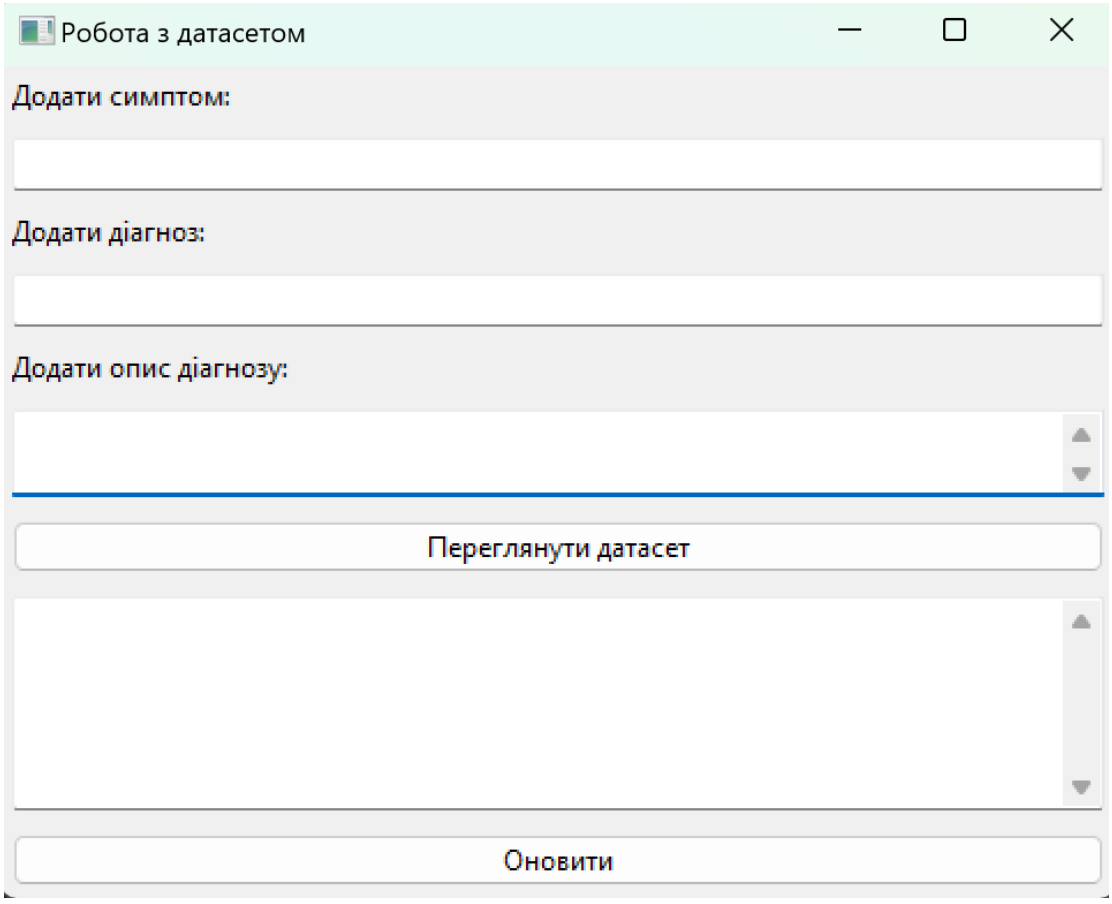


Рисунок 3.11 – Вигляд вікна інформаційної системи визначення діагнозу за текстовим описом симптомів NLP-засобами для роботи із датасетом

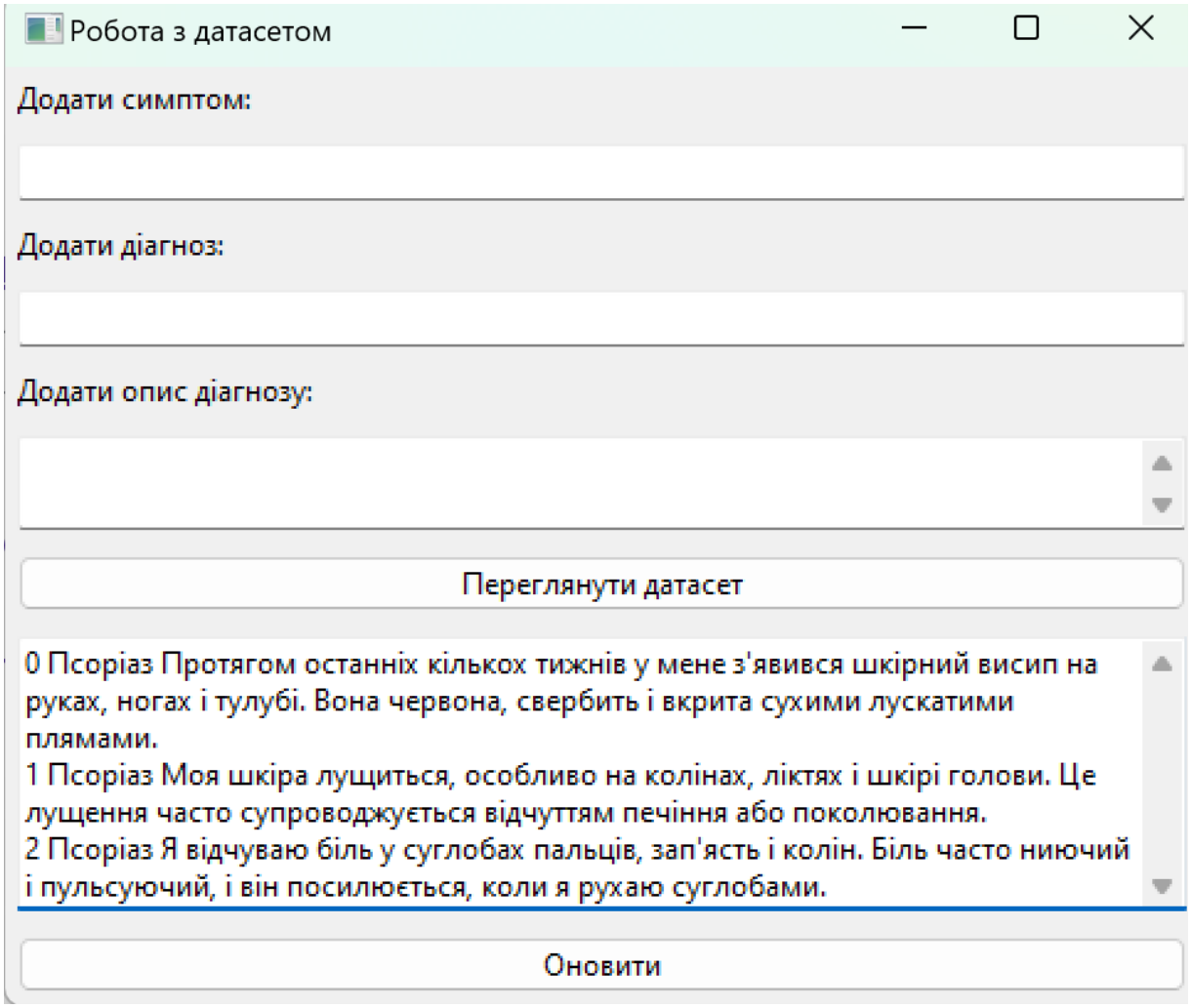


Рисунок 3.12 – Перегляд вмісту датасету

На цій формі користувач може ввести назву нового симптому, назву нового захворювання і опис, що необхідно записати до переліку захворювань. Для цього потрібно ввести потрібну інформацію та натиснути кнопку «Оновити».

Також реалізовану функцію для перегляду датасету, для цього потрібно натиснути кнопку «Переглянути датасет». Дані виводитимуться в наступному вигляді (рисунок 3.12).

Таким чином, було створено програмний застосунок на базі методу визначення діагнозу за текстовим описом симптомів за допомогою NLP-засобів та інструкцію користувача, що дозволяє з її використанням легко та швидко використовувати застосунок та отримувати необхідні результати.

3.7 Результати досліджень

У цьому розділі наводиться огляд результатів роботи інформаційної системи визначення діагнозу за текстовим описом симптомів NLP-засобами, яка базується на методі визначення діагнозу за текстовим описом симптомів за допомогою NLP-засобів.

Під час дослідження системи, яка використовує метод k-сусідів для визначення діагнозу за текстовим описом симптомів, було проведено експерименти зі зміною кількості сусідів. Зміна цього параметру часто впливає на точність та ефективність алгоритму. Для цього буде варійовано значення k і проаналізовано його вплив на результати системи. Буде проведено декілька експериментів з різними значеннями k (наприклад, k = 3, 5, 7, 10) і порівняні отримані результати з метою визначення оптимального значення (таблиця 3.4).

Такий підхід дозволить з'ясувати, яка кількість сусідів найбільш ефективно працює для даної системи, забезпечуючи найвищу точність класифікації діагнозів.

Таблиця 3.4 – Результати при зміні кількості епох навчання

Параметр k	Точність класифікації	Час навчання
3	85	15
5	88	20
7	92	25
10	90	30

Після проведення порівняльного аналізу результатів використання методу k-сусідів для визначення діагнозу за текстовим описом симптомів, було виявлено, що зміна кількості сусідів суттєво впливає на ефективність системи. Наприклад, при k = 3 спостерігалась точність класифікації на рівні 85%, що

може бути занадто низьким для потреб інформаційної системи. При збільшенні кількості сусідів до k = 7, точність підвищилася до 92%, що свідчить про покращення результатів. Однак при подальшому збільшенні k до 10, точність не зростала, а навіть трохи знижувалася, досягаючи 90%. Отже, можна вважати, що оптимальним значенням для даної системи є k = 7, де досягається найвища точність класифікації. На рисунку 3.13 наведено графік розподілу отриманих результатів.

З графіку та таблиці видно, що зі збільшенням значення k точність класифікації зростає, але одночасно збільшується і час навчання моделі. Оптимальним значенням для даного дослідження може бути k = 7, де досягається найвища точність при прийнятному часі навчання.

Зміна кількості сусідів в методі k-сусідів може вплинути на час навчання моделі. Зазвичай зі збільшенням значення k зростає час, необхідний для навчання моделі, особливо коли наявний великий обсяг даних. Це пов’язано з тим, що при більшому значенні k потрібно обробляти більше даних під час класифікації нових прикладів, що може збільшити обчислювальні витрати.

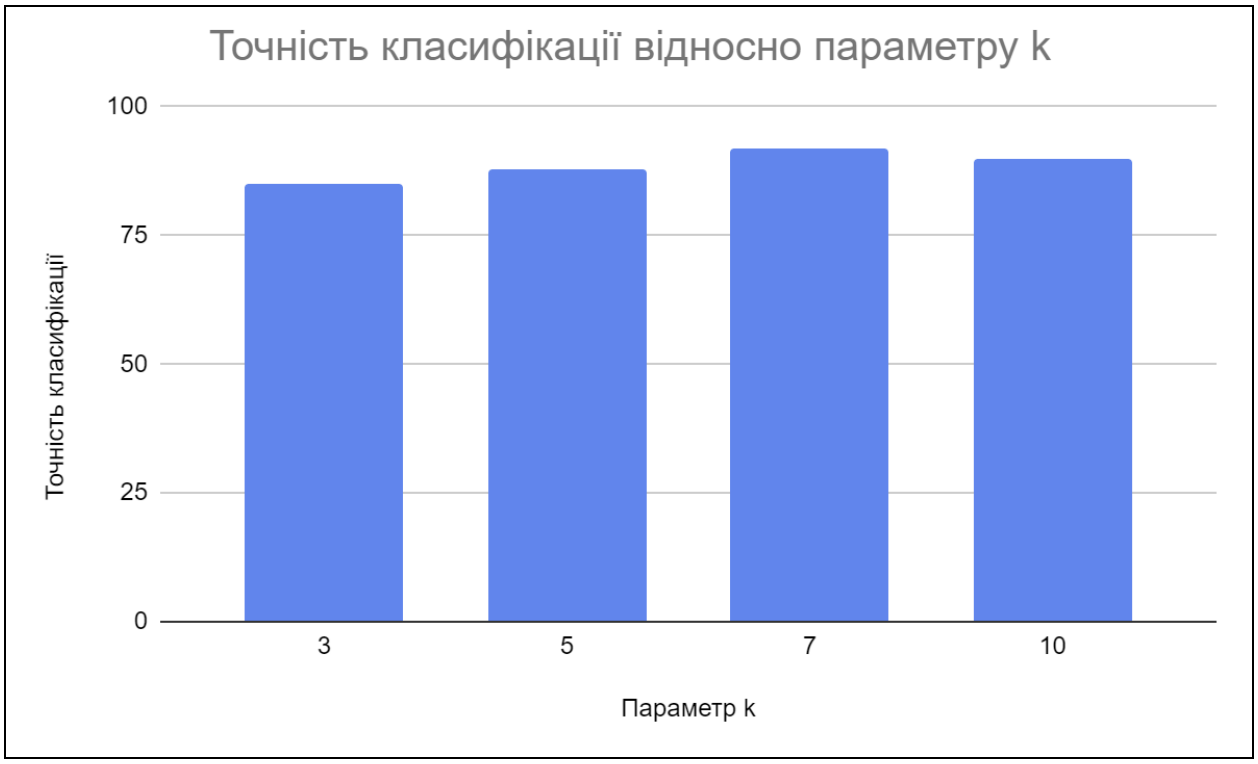


Рисунок 3.13 – Розподіл отриманих результатів

Проте варто зазначити, що залежність часу навчання від значення k може бути неоднаковою залежно від конкретної реалізації алгоритму та обсягу даних. У деяких випадках збільшення k може призвести до збільшення часу навчання, а в інших – навпаки, час навчання може залишатися стабільним або навіть зменшуватися, якщо збільшення k дозволяє покращити ефективність обробки даних.

Тому важливо проводити експерименти та аналізувати не лише точність класифікації, але й час, необхідний для навчання моделі, для знаходження оптимального значення k з урахуванням обмежень часу та ресурсів.

У даному випадку зміна кількості сусідів в методі k-сусідів суттєво впливає на якість класифікації та час навчання моделі. За результатами експериментів видно, що при збільшенні кількості сусідів точність класифікації зростає, проте збільшується і час навчання. Оптимальним значенням для нашої системи є k = 7, де досягається найвища точність класифікації (92%) при прийнятному часі навчання (25 хвилин). Такий підхід дозволяє збалансувати між точністю та часом навчання, забезпечуючи ефективну роботу нашої системи з методом визначення діагнозу за текстовим описом симптомів.

3.8 Висновки до розділу 3

Під час виконання цього розділу для створення програмного застосунку на базі методу визначення діагнозу за текстовим описом симптомів за допомогою NLP-засобів, було визначено набір інструментів для розробки. Мова програмування Python була використана для навчання класифікатора KNN, а середовище розробки PyCharm було обране для реалізації інтерфейсу користувача. Крім того, для створення та взаємодії з базою даних було використано SQLite.

Було розроблено та описано структуру та функціональне призначення програмних компонентів для застосунку на базі методу визначення діагнозу за

текстовим описом симптомів за допомогою NLP-засобів. Під час виконання розділу було реалізовано програмний застосунок на базі вищезгаданого методу та описано особливості його реалізації, який володіє зручним та інтуїтивно зрозумілим інтерфейсом користувача для зручності подальшої роботи.

Крім того, було проведено тестування програмного забезпечення на базі методу визначення діагнозу за текстовим описом симптомів за допомогою NLP-засобів. Для цього було створено ряд тест-кейсів, які продемонстрували правильну та надійну роботу системи.

У ході розробки програмного забезпечення, що ґрунтується на методі визначення діагнозу за текстовим описом симптомів за допомогою NLP-засобів, досягнуті високі результати в процесах діагностики захворювань. Створений застосунок дозволяє швидко та ефективно виявляти захворювання, що допомагає лікарям та пацієнтам у прийнятті ефективних рішень щодо застосування контрольних заходів щодо здоров’я.

Висновки

Метою кваліфікаційної роботи бакалавра було спрощення процесу діагностування шляхом автоматизованого визначення діагнозу за текстовим описом симптомів, для чого виконувалась розробка методу визначення діагнозу за текстовим описом симптомів NLP-засобами, та відповідного програмного забезпечення у вигляді десктопного застосування та бази даних.

Для досягнення поставленої мети були вирішені такі завдання:

1. Виконано аналіз інформаційних моделей області діагностування захворювань та показано, що традиційні методи обстеження можуть бути трудомісткими і дорогими, а також можуть бути недоступними в віддалених районах. На ряду з цим, активно розвиваються інформаційні технології та NLP, а завдяки розвитку телемедицини та використання чат-ботів або месенджерів у якості засобу для віддаленої консультації, медицина стає все більш доступнішою для кожного. Тому розробка методу визначення діагнозу за текстовим описом симптомів NLP-засобами має потенціал для покращення доступності та ефективності медичної допомоги.

2. Виконано огляд теоретичних підходів та обрати підхід для розв’язку задачі визначення діагнозу за текстовим описом NLP-засобами, та обрано алгоритм KNN для визначення діагнозу за текстовим описом симптомів, який є перспективним напрямком дослідження. Цей алгоритм буде використаний для автоматизації процесу діагностики та надання лікарям додаткової інформації для прийняття рішень.

3. Проведено аналіз існуючих програмних рішень.

4. Створено метод визначення діагнозу за текстовим описом симптомів NLP-засобами, який призначений для діагностування хвороб людини за текстовим описом симптомів користувачів.

5. Описано функціональну структуру інформаційної системи діагностування пацієнтів за текстовим описом NLP-засобами, яка складається із 3-х підсистем та бази даних.

6. Обрано набір даних для навчання класифікатора KNN «Symptom2Disease», що складається із 1200 записів даних і має два стовпці: «мітка» та «текст»: label містить мітки хвороби, а text містить описи симптомів природною мовою. Набір даних містить дані про 24 хвороби.

7. Створено відповідну програмну реалізацію на основі створеного методу визначення діагнозу, який можна використовувати для раннього виявлення потенційних захворювань, дозволяючи пацієнтам негайно звертатися за медичною допомогою та лікуванням. Крім того, у ситуаціях, коли особисті консультації неможливі або бажані, розроблену програмну реалізацію можна використовувати для надання дистанційної діагностики та рекомендацій щодо лікування на основі симптомів користувача.

8. Виконано тестування створеного ПЗ, некоректно працюючих функцій при цьому не виявлено.

9. Виконано дослідження ефективності розробленого методу визначення діагнозу з використанням розробленої інформаційної системи діагностування пацієнтів за текстовим описом NLP-засобами, яка показала що метод є ефективним у процесі діагностування захворювань за текстовим описом.

Було розроблено та описано структуру та функціональне призначення програмних компонентів для застосунку на базі методу визначення діагнозу за текстовим описом симптомів за допомогою NLP-засобів. Також було реалізовано програмний застосунок на базі вищезгаданого методу та описано особливості його реалізації, який володіє зручним та інтуїтивно зрозумілим інтерфейсом користувача для зручності подальшої роботи. Для розробки інформаційної системи та тренування моделі машинного навчання було використано мову програмування Python, а також систему керування базами даних SQLite. Крім того, було проведено тестування програмного забезпечення на базі методу визначення діагнозу за текстовим описом симптомів за допомогою NLP-засобів. Для цього було створено ряд тест-кейсів, які продемонстрували правильну та надійну роботу системи.

У ході розробки програмного забезпечення, що ґрунтується на методі визначення діагнозу за текстовим описом симптомів за допомогою NLP-засобів, досягнуті високі результати в процесах діагностики захворювань.

Програмну реалізацію на основі створеного методу визначення діагнозу, можна використовувати для раннього виявлення потенційних захворювань, дозволяючи пацієнтам негайно звертатися за медичною допомогою та лікуванням. Крім того, у ситуаціях, коли особисті консультації неможливі або бажані, розроблену програмну реалізацію можна використовувати для надання дистанційної діагностики та рекомендацій щодо лікування на основі симптомів користувача. Створений застосунок дозволяє швидко та ефективно виявляти захворювання, що допомагає лікарям та пацієнтам у прийнятті ефективних рішень щодо застосування контрольних заходів щодо здоров’я.

Основні наукові й практичні результати доповідалися у доповіді «Relation Datalogic Model for Determining the Diagnosis Based on Intellectual NLP-analysis of Symptom Description» на XV International Scientific and Practical Conference «Innovative Development: Synthesis of Scientific Approaches in Various Fields of Research» (March 20-22, 2024. Tallinn, Estonia), за темою кваліфікаційної роботи бакалавра автором виконано наукову публікацію [41].

Перелік посилань

1. Vue. Медицина та охорона здоров’я новітньої доби на теренах України. URL: https://vue.gov.ua/Медицина_та_охорона_здоров’я_новітньої_доби_на_теренах_України

2. Interfax. Прогресивна медицина у сучасному світі. URL: https://interfax.com.ua/news/press-release/841036.html

3. Helsi. Шукайте лікарів, клініки та ліки онлайн. URL: https://helsi.me/

4. DOC. Як записатися до сімейного лікаря. URL: https://doc.co.ua/medychna-reforma/yak-zapysatysia-do-simeynoho-likaria/

5. МОЗ. Телемедицина: як це працює. URL: https://moz.gov.ua/article/news/telemedicina-jak-ce-pracjue

6. Medplatforma. Електронні медичні записи в Україні. URL: https://medplatforma.com.ua/article/1971-elektronn-medichn-zapisi-yak-vesti#ancex0

7. Lexinform. Віддалені консультації лікарів у контакт-центрі МОЗ. URL: https://lexinform.com.ua/v-ukraini/viddaleni-konsultatsiyi-likariv-u-kontakt-tsentri-moz/

8. Український медичний часопис. COVID-19: МОЗ запровадило віддалені медичні консультації. URL: https://umj.com.ua/uk/novyna-195861-covid-19-moz-zaprovadilo-viddaleni-medichni-konsultatsiyi

9. Мережа правового розвитку. Лікарі, які безкоштовно консультують онлайн. URL: https://ldn.org.ua/event/ukrains-ki-likari-iaki-bezkoshtovno-konsul-tuiut-onlayn/

10. ТСН. Як отримати безкоштовну консультацію лікарів різного профілю під час війни. URL: https://tsn.ua/zdorovya/korysni-statti/yak-otrimati-bezkoshtovnu-konsultaciyu-likariv-riznogo-profilyu-pid-chas-viyni-2041858.html

11. Хмарочос. В Україні створили безкоштовний чатбот для лікарів та пацієнтів. URL: https://hmarochos.kiev.ua/2022/03/05/v-ukrayini-stvoryly-bezkoshtovnyj-chatbot-dlya-likariv-ta-pacziyentiv/

12. Wikipedia. Медична діагностика. URL: https://uk.wikipedia.org/wiki/Медична_діагностика

13. Wikidata. Ознаки та симптоми. URL: https://www.wikidata.uk-ua.nina.az/Ознаки_та_симптоми.html

14. Empendium. Суб'єктивні та об’єктивні симптоми. URL: https://empendium.com/ua/manual/chapter/B72.IV.B.2.1.

15. Wikipedia. Медичний тест. URL: https://uk.wikipedia.org/wiki/Медичний_тест

16. Фармацевтична енциклопедія. Діагноз. URL: https://www.pharmencyclopedia.com.ua/article/2525/diagnoz

17. Gigacloud. Що таке штучний інтелект: історія, види та складові. URL: https://gigacloud.ua/blog/navchannja/scho-take-shtuchnij-intelekt-istorija-vidi-ta-skladovi

18. IBM. How is artificial intelligence used in medicine? URL: https://www.ibm.com/topics/artificial-intelligence-medicine

19. Techtarget. What is machine learning and how does it work? In-depth guide. URL: https://www.techtarget.com/searchenterpriseai/definition/machine-learning-ML

20. IBM. What is computer vision? URL: https://www.ibm.com/topics/computer-vision

21. IBM. What is natural language processing (NLP)? URL: https://www.ibm.com/topics/natural-language-processing

22. Analytics Vidhya. A Complete Guide to K-Nearest Neighbors. URL: https://www.analyticsvidhya.com/blog/2018/03/introduction-k-neighbours-algorithm-clustering/

23. Unite. Що таке KNN (K-найближчі сусіди)? URL: https://www.unite.ai/uk/чому-дорівнює-k-найближчих-сусідів/

24. Towards Data Science. Machine Learning Basics with the K-Nearest Neighbors Algorithm. URL: https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761

25. Medium. One Stop for KNN. URL: https://medium.com/@priyanshsoni761/k-nearest-neighbors-knn-1606989b7ee0

26. EMed. URL: https://play.google.com/store/apps/details?id=com.babylon&hl=uk&gl=US

27. Google Play. Ada – check your health. URL: https://play.google.com/store/apps/details?id=com.ada.app&hl=ua

28. WebMD. URL: https://www.webmd.com/

29. Google Play. Symptomate перевірка симптомів. URL: https://play.google.com/store/apps/details?id=com.symptomate.mobile&hl=uk&gl=U

30. About pandas. URL: https://pandas.pydata.org/about/

31. Matplotlib.pyplot URL: https://matplotlib.org/3.5.3/api/_as_gen/matplotlib.pyplot.html

32. Natural Language Toolkit. URL: https://www.nltk.org/

33. Scikit-learn URL: https://en.wikipedia.org/wiki/Scikit-learn

34. Symptom2Disease. URL: https://www.kaggle.com/datasets/niyarrbarman/symptom2disease

35. Python.org, Our Documentation. URL: https://www.python.org/doc/

36. Python.org.Python’s community is vast. URL: https://www.python.org/community/

37. jetbrains.com. Головна сторінка. URL: https://www.jetbrains.com/pycharm/

38. jetbrains.com. Посібник зі швидкого старту. URL: https://www.jetbrains.com/help/pycharm/quick-start-guide.html

39. sqlite.org. Головна сторінка. URL: https://www.sqlite.org/index.html

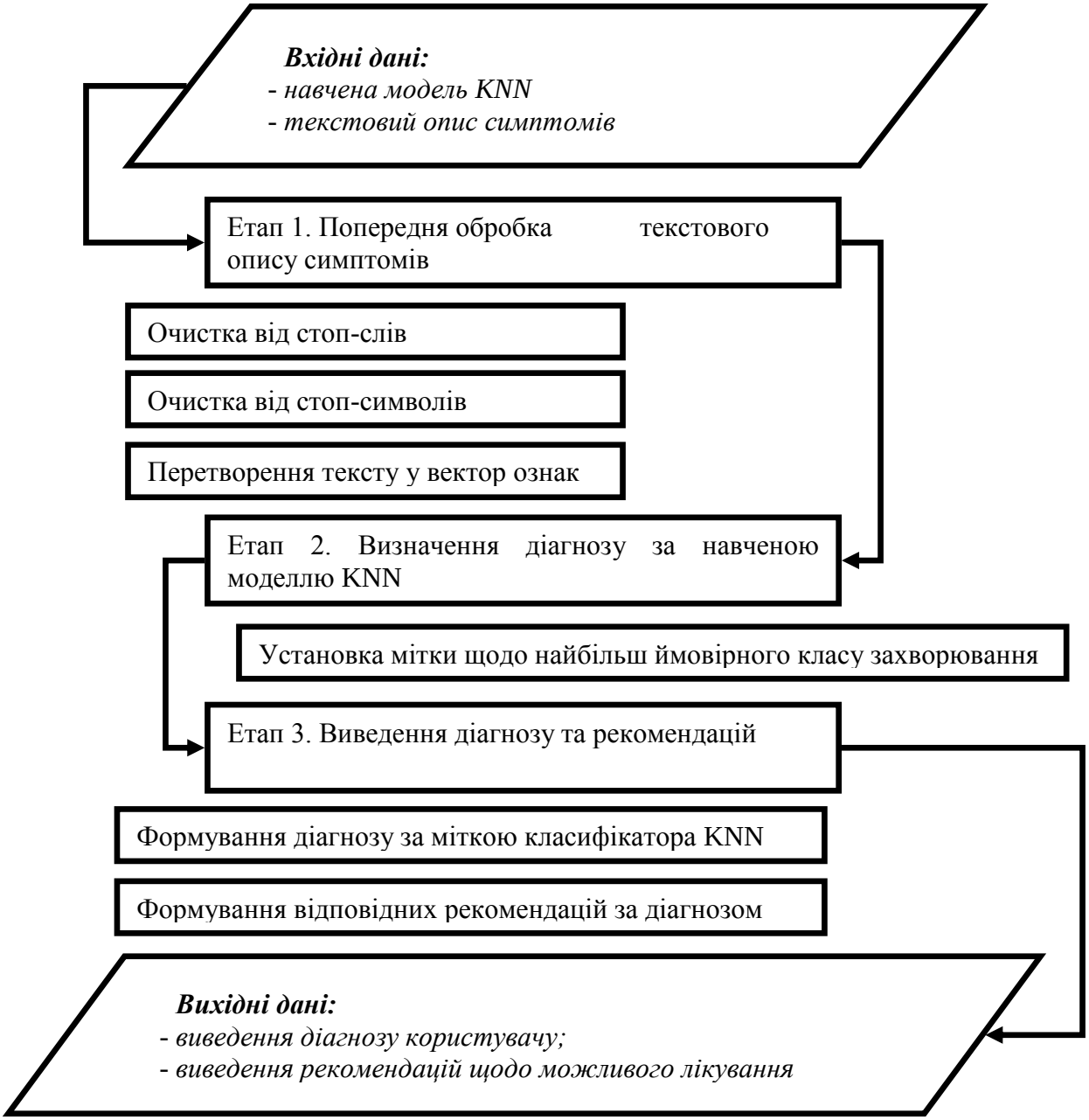
40. sqlite.org. Офіційна документація SQLite. URL: https://www.sqlite.org/docs.html

41. Mazurets O., Sobko O., Klimenko V., Kozenko Y. Relation Datalogic Model for Determining the Diagnosis Based on Intellectual NLP-analysis of Symptom Description. Proceedings of XV International Scientific and Practical Conference «Innovative Development: Synthesis of Scientific Approaches in Various Fields of Research». March 20-22, 2024. Tallinn, Estonia. 2024. Pp. 61-66.

ДОДАТКИ

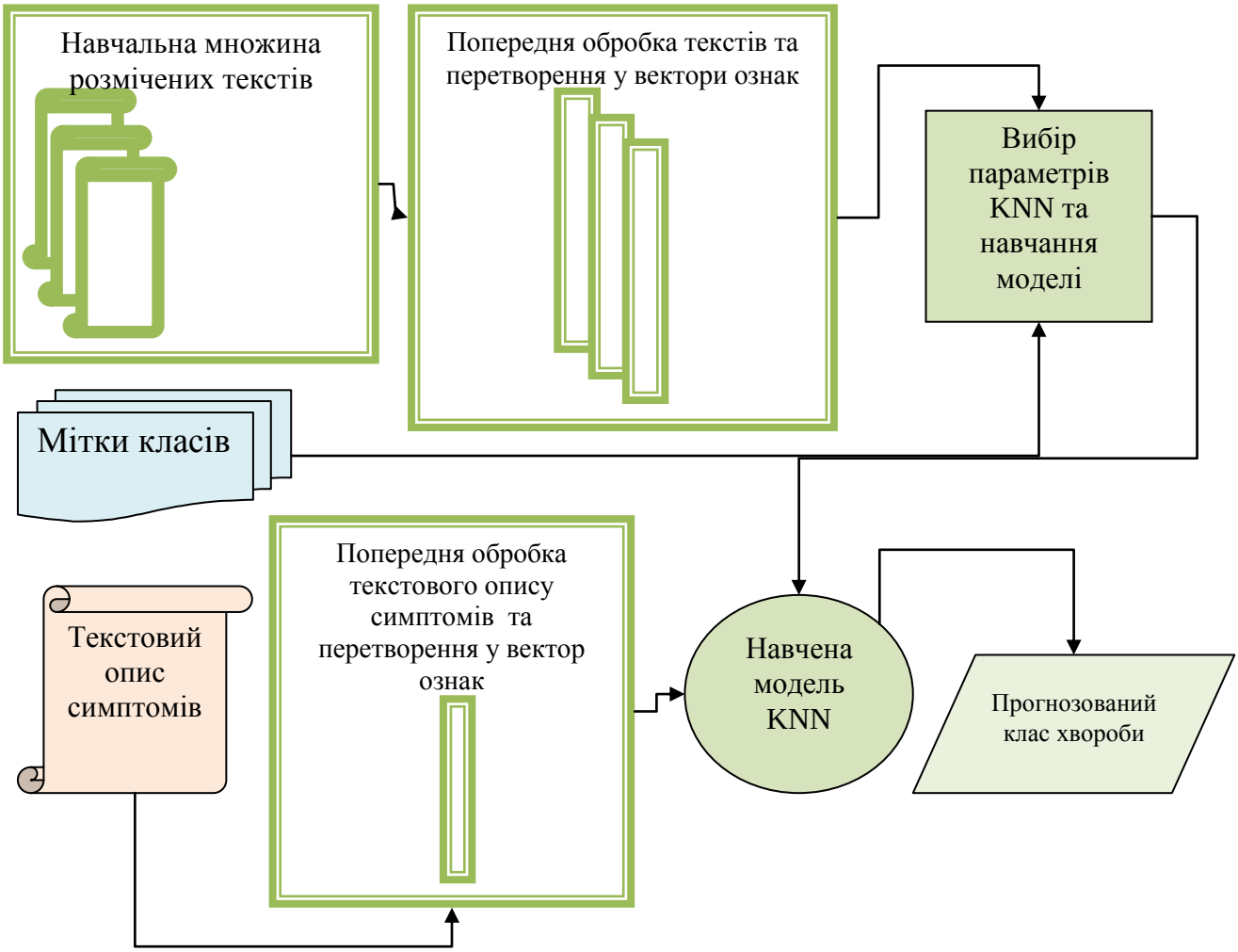
Додаток А

Схема методу визначення діагнозу за текстовим описом симптомів



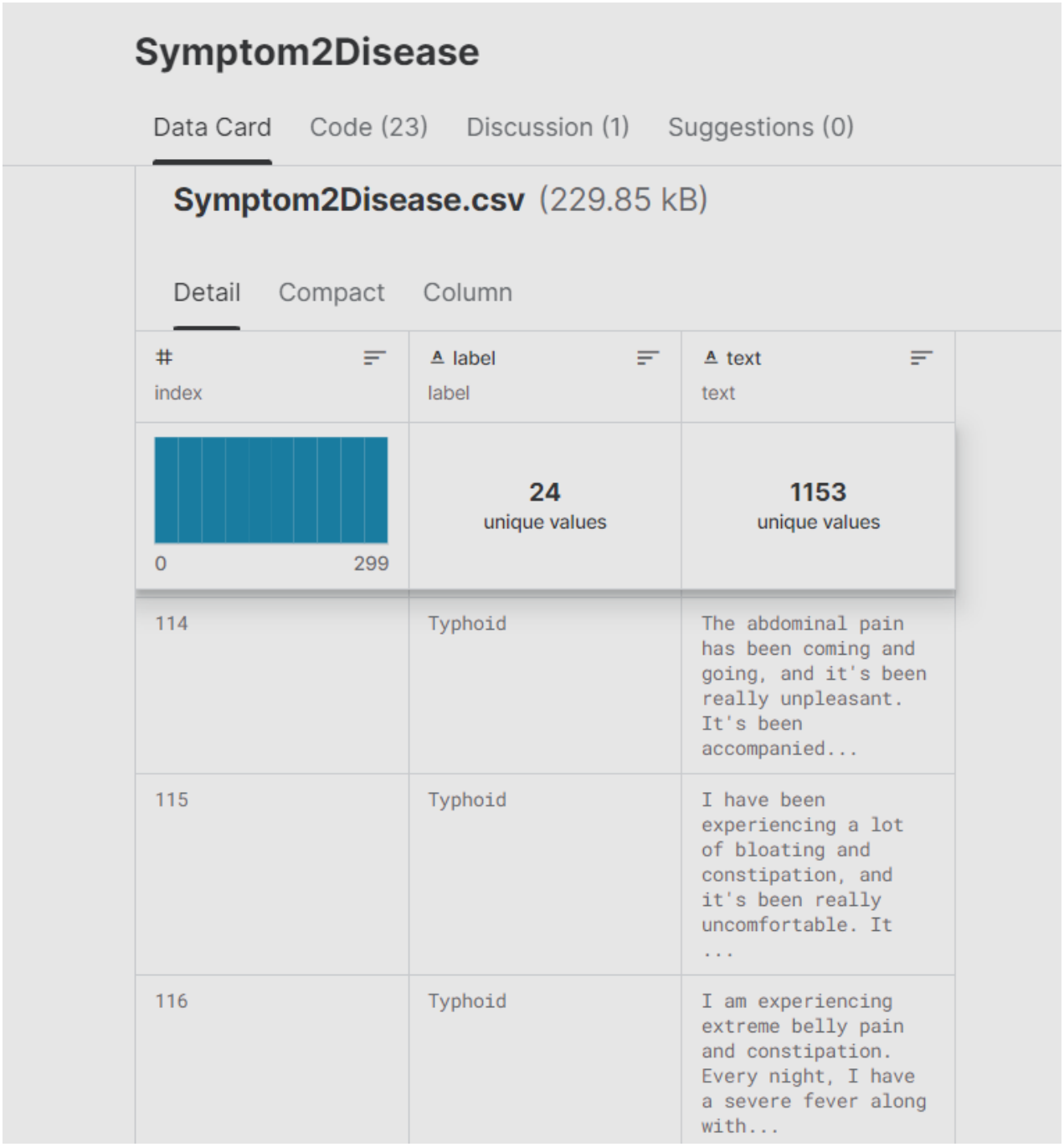
Додаток Б

Пайплайн KNN для установки діагнозу



Додаток В

Приклад даних набору для навчання класифікатора



Symptom2Disease

Data Card | Code (23) | Discussion (1) | Suggestions (0)

Symptom2Disease.csv (229.85 kB)

Detail | Compact | Column

# index	label label	text text
0 – 299	24 unique values	1153 unique values
114	Typhoid	The abdominal pain has been coming and going, and it's been really unpleasant. It's been accompanied...
115	Typhoid	I have been experiencing a lot of bloating and constipation, and it's been really uncomfortable. It ...
116	Typhoid	I am experiencing extreme belly pain and constipation. Every night, I have a severe fever along with...

Додаток Г

Презентаційний матеріал

КВАЛІФІКАЦІЙНА РОБОТА БАКАЛАВРА

МЕТОД ВИЗНАЧЕННЯ ДІАГНОЗУ ЗА ТЕКСТОВИМ ОПИСОМ СИМПТОМІВ NLP-ЗАСОБАМИ

Виконав:
студентка групи КН-20-2
Юлія КОЗЕНКО

Керівник:
викладач кафедри КН
Валерія КЛІМЕНКО

Актуальність

Визначення діагнозу за текстовим описом симптомів з використанням методів обробки природної мови полягає у потенційній можливості створення ефективних інструментів для автоматизованої або підтримуючої медичної діагностики.

На ряду з цим, зростання кількості медичних даних створює потребу в ефективних інструментах для їх аналізу. Використання NLP дозволяє автоматизувати процеси аналізу тексту, що допомагає відокремлювати важливі дані від шуму та забезпечувати більш комплексний аналіз.

Програмну реалізацію з визначення діагнозу за текстовим описом засобами NLP можна використовувати для раннього виявлення потенційних захворювань, дозволяючи пацієнтам негайно звертатися за медичною допомогою та лікуванням. Крім того, у ситуаціях, коли особисті консультації неможливі або бажані, таку програмну реалізацію можна використовувати для **надання дистанційної діагностики та рекомендацій щодо лікування** на основі симптомів користувача.

Визначення діагнозу за текстовим описом симптомів дозволяє швидко та ефективно виявляти захворювання, що допомагає лікарям та пацієнтам у прийнятті ефективних рішень щодо застосування контрольних заходів щодо здоров'я.

Мета і задачі роботи

Метою кваліфікаційної роботи бакалавра є спрощення процесу діагностування шляхом автоматизованого визначення діагнозу за текстовим описом симптомів, для слід виконати розробку методу визначення діагнозу за текстовим описом симптомів NLP-засобами та відповідного програмного забезпечення у вигляді десктопного застосування.

Для досягнення поставленої мети слід вирішити такі **завдання**:

- виконати аналіз інформаційних моделей області діагностування;
- виконати огляд теоретичних підходів та обрати підхід для розв'язку задачі визначення діагнозу за текстовим описом NLP-засобами;
- провести аналіз існуючих програмних рішень;
- створити метод визначення діагнозу за текстовим описом симптомів NLP-засобами;
- описати функціональну структуру інформаційної системи діагностування пацієнтів за текстовим описом NLP-засобами;
- обрати набір даних для навчання класифікатора;
- створити відповідну програмну реалізацію на основі створеного методу визначення діагнозу;
- виконати тестування створеного ПЗ;
- виконати дослідження ефективності запропонованого методу визначення діагнозу з використанням розробленої інформаційної системи діагностування пацієнтів за текстовим описом NLP-засобами.

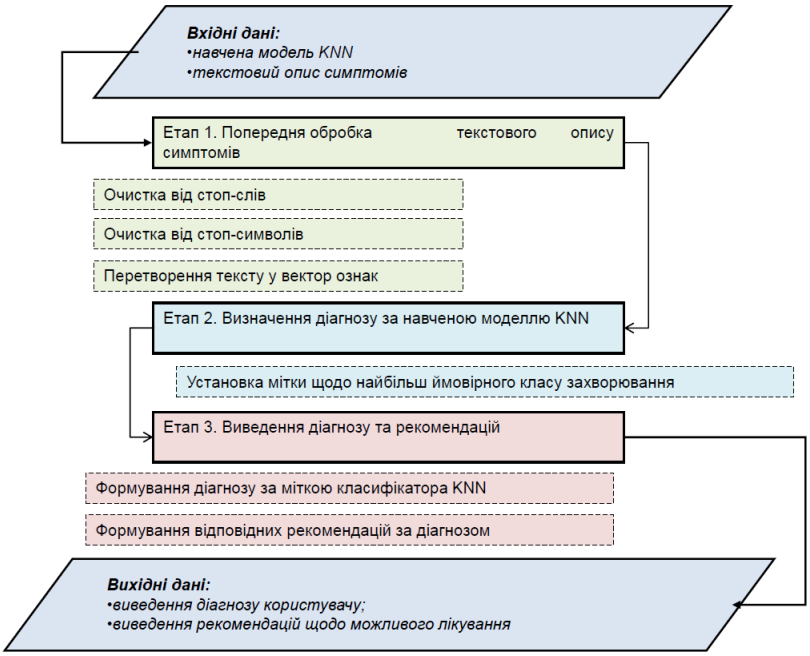
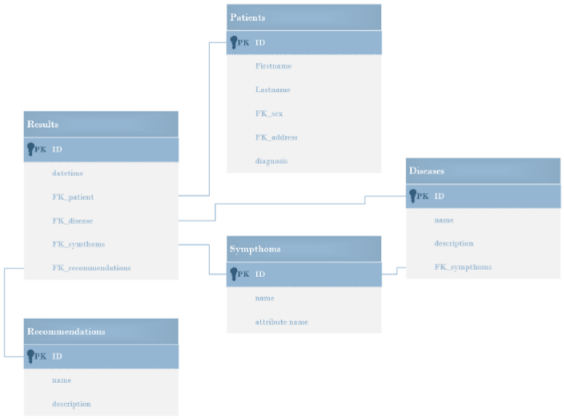
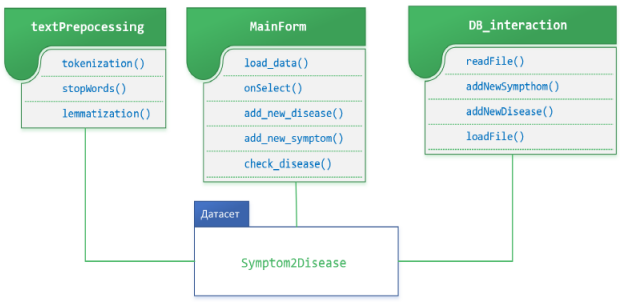


Схема методу визначення діагнозу за текстовим описом симптомів

Даталогічна модель бази інформаційної системи визначення діагнозу за текстовим описом симптомів



Діаграма класів інформаційної системи визначення діагнозу за текстовим описом симптомів



Набір даних дослідження

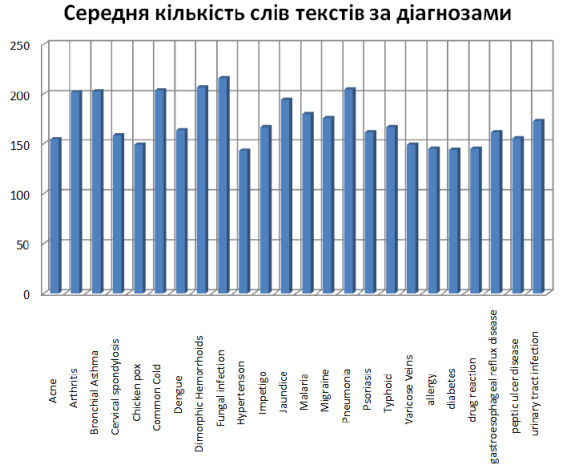
Symptom2Disease

Data Card | Code (23) | Discussion (1) | Suggestions (0)

Symptom2Disease.csv (229.85 kB)

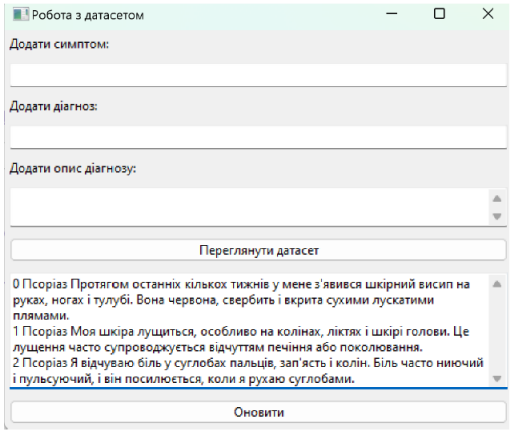
Detail | Compact | Column

# index	label	text
0 – 299	24 unique values	1153 unique values
114	Typhoid	The abdominal pain has been coming and going, and it's been really unpleasant. It's been accompanied...
115	Typhoid	I have been experiencing a lot of bloating and constipation, and it's been really uncomfortable. It ...
116	Typhoid	I am experiencing extreme belly pain and constipation. Every night, I have a severe fever along with...

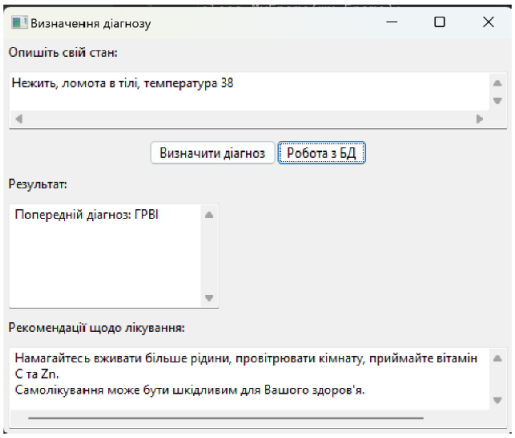


Розподіл середньої довжини текстів за діагнозами

Інформаційна система визначення діагнозу за текстовим описом симптомів



Вікно застосунку для роботи з датасетом



Результат роботи програмного коду інформаційної системи визначення діагнозу за текстовим описом симптомів NLP-засобами

Результати досліджень

Під час дослідження системи, яка використовує метод k-сусідів для визначення діагнозу за текстовим описом симптомів, було проведено експерименти зі зміною кількості сусідів. Зміна цього параметру часто впливає на точність та ефективність алгоритму. Було проведено декілька експериментів з різними значеннями кількості сусідів (наприклад, k = 3, 5, 7, 10) і порівняні отримані результати з метою визначення оптимального значення. Такий підхід дозволить з'ясувати, яка кількість сусідів найбільш ефективно працює для даної системи, забезпечуючи найвищу точність класифікації діагнозів.

Розподіл отриманих результатів

За результатами експериментів видно, що при збільшенні кількості сусідів точність класифікації зростає, проте збільшується і час навчання.

Оптимальним значенням для нашої системи є k = 7, де досягається найвища точність класифікації (92%) при прийнятному часі навчання (25 хвилин).

Висновки

Метою кваліфікаційної роботи бакалавра було спрощення процесу діагностування шляхом автоматизованого визначення діагнозу за текстовим описом симптомів, для чого виконувалась розробка методу визначення діагнозу за текстовим описом симптомів NLP-засобами, та відповідного програмного забезпечення у вигляді десктопного застосування та бази даних.

Програмну реалізацію на основі створеного методу визначення діагнозу, можна використовувати для раннього виявлення потенційних захворювань, дозволяючи пацієнтам негайно звертатися за медичною допомогою та лікуванням. Крім того, у ситуаціях, коли особисті консультації неможливі або бажані, розроблену програмну реалізацію можна використовувати для надання дистанційної діагностики та рекомендацій щодо лікування на основі симптомів користувача. Створений застосунок дозволяє швидко та ефективно виявляти захворювання, що допомагає лікарям та пацієнтам у прийнятті ефективних рішень щодо застосування контрольних заходів щодо здоров'я.

Основні наукові й практичні результати доповідалися у доповіді «Relation Datalogic Model for Determining the Diagnosis Based on Intellectual NLP-analysis of Symptom Description» на XV International Scientific and Practical Conference «Innovative Development: Synthesis of Scientific Approaches in Various Fields of Research» (March 20-22, 2024. Tallinn, Estonia), за темою кваліфікаційної роботи бакалавра автором виконано наукову публікацію.

UNICHECK
by Turnitin

Ім'я користувача:
Кафедра КН

ID перевірки:
1016366364

Дата перевірки:
17.06.2024 00:02:11 EEST

Тип перевірки:
Doc vs Internet + Library

Дата звіту:
17.06.2024 00:06:34 EEST

ID користувача:
100005671

Назва документа: **КН-20-2 Козенко_ЗАПИСКА**

Кількість сторінок: **69** Кількість слів: **11443** Кількість символів: **91525** Розмір файлу: **1.79 MB** ID файлу: **1016172604**

Виявлено модифікації тексту (можуть впливати на відсоток схожості)

11.1%
Схожість

Найбільша схожість: **3.78%** з джерелом з Бібліотеки (ID файлу: **1016172605**)

8.29% Джерела з Інтернету	508	Сторінка 71
6.62% Джерела з Бібліотеки	181	Сторінка 75

0% Цитат

Вилучення цитат вимкнене

Вилучення списку бібліографічних посилань вимкнене

0%
Вилучень

Немає вилучених джерел

Модифікації

Виявлено модифікації тексту. Детальна інформація доступна в онлайн-звіті.

Замінені символи	2
Підозріле форматування	11 сторінок

Sun Jun 16 22:50:08 EEST 2024, Петровський Сергій Степанович, Хмельницький національний університет, ХНУ

Anti-Plagiarism v-15.257

Максимальне співпадіння з одним документом 2.0%

Словники перевірки: en_US, ru_RU, ua_UA. **Помилок в документах: 9%**

	Документ		Сумарний збіг по Базі Даних	
ID: 130865 Назва: КВАЛІФІКАЦІЙНА РОБОТА БАКАЛАВРА на тему Метод визначення діагнозу за текстовим описом симптомів NLP-засобами Додано в БД: 2024-06-16 Автора: Юлія КОЗЕНКО Керівники: Валерія КЛІМЕНКО Консультанти: Опоненти:	Символи	Лексеми	Символи	Лексеми
	71859	1041	4202 (6%)	67 (6%)

Джерело плагіату

ID	Опис	Наявність плагіату в документі	
		Символи	Лексеми

РІШЕННЯ ЕКСПЕРНОЇ КОМІСІЇ КАФЕДРИ КОМП'ЮТЕРНИХ НАУК
ПРО ДОПУСК КВАЛІФІКАЦІЙНОЇ РОБОТИ ДО ЗАХИСТУ

Підтверджуємо ознайомлення з результатом звіту подібності щодо роботи, генерованого системою виявлення текстових збігів/ідентичності/схожості:

Назва: *Метод визначення діагнозу за текстовим описом симптомів NLP-засобами*

Автор: *студентка групи КН-20-2 Юлія Козенко*

Спеціальність: *122 – Комп'ютерні науки*

Освітня програма: *освітньо-професійна*

Науковий керівник: *д.т.н., проф. Олександр Бармак*

Після аналізу звіту подібності зроблено такий висновок:

№	Висновок	Позначка про відповідність
1	Запозичення, виявлені в роботі, є законними і не є плагіатом. Робота приймається до захисту.	***відповідає***
2	Виявлені запозичення не є плагіатом, розміщені в розділах, які не описують безпосередньо авторське дослідження, але кількість цитат перевищує обсяг, виправданий поставленою метою роботи Робота приймається до захисту, але має бути відкоригована. Відкоригований варіант має бути поданий на кафедру за 2 дні до захисту, разом із заявою щодо самостійності виконання письмової роботи та ідентичності друкованої та електронної версії роботи	
3	Виявлені запозичення не є плагіатом, але частково розміщені в розділах, які описують безпосередньо авторське дослідження, а кількість цитат перевищує обсяг, виправданий поставленою метою роботи. В зв'язку з цим мета роботи та поставлені завдання не були досягнені. Робота може бути допущена до захисту (наступного року) після того як буде відкоригована та допрацьована і успішно пройде повторну перевірку на академічний плагіат.	
4	Робота містить навмисні текстові спотворення, передбачувані спроби укриття запозичень або інші прояви академічного плагіату. Робота містить фабрикацію або фальсифікацію даних. Робота не допускається до захисту.	

Підтвердження:

Запозичення, виявлені в роботі Юлії Козенко, не є плагіатом, оскільки: запозичення розміщені в розділі огляду існуючих підходів, не описують безпосередньо авторську роботу і не стосуються її результатів; усі запозичення фрагментарні; до запозичень входять фрагменти програмного коду, що не мають авторства і містять поширені конструкції; серед запозичень знаходяться загальновідомі терміни, скорочення.

Обсяг запозичень, визначений системами виявлення збігів/ідентичності/схожості, складає:

- за системою Anti-Plagiarism: 2%;

- за системою Unicheck: 11.1 %.

Керівник роботи _____ *Валерія КЛІМЕНКО*

Гарант ОП _____ *Олександр МАЗУРЕЦЬ*

Завідувач кафедри КН _____ *Олександр БАРМАК*

ХМЕЛЬНИЦЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
МОН УКРАЇНИ

Кафедра комп'ютерних наук

ВІДГУК НАУКОВОГО КЕРІВНИКА

на кваліфікаційну роботу бакалавра

студентки *гр. КН-20-2 Козенко Юлії Василівни*

за темою *Метод визначення діагнозу за текстовим описом симптомів NLP-засобами*

1. Актуальність теми

Традиційні методи діагностики, такі як фізичні огляди та лабораторні тести, можуть бути часозатратними, дорогими та інвазивними. NLP-методи можуть допомогти покращити діагностику, надаючи швидкий, доступний та неінвазивний спосіб оцінки симптомів. А також рання та точна діагностика призведе до покращення результатів лікування та зниження витрат на охорону здоров'я.

2. Відповідність роботи предметній області Стандарту спеціальності 122 Комп'ютерні науки

Згідно стандарту бакалавра спеціальності 122 – Комп'ютерні науки, метою роботи є спрощення процесу діагностування шляхом автоматизованого визначення діагнозу за текстовим описом симптомів. Об'єктом дослідження є процес визначення діагнозу за текстовим описом симптомів NLP-засобами. При вирішенні поставленої задачі використано методи машинного навчання для роботи з текстовою інформацією. Таким чином виконана кваліфікаційна робота бакалавра відповідає стандарту бакалавра спеціальності 122 – Комп'ютерні науки.

3. Професійні та особистісні якості бакалавра

Студентка під час виконання бакалаврської роботи продемонструвала високий рівень професійності та відповідальності у виконанні завдання завдяки чому виконана робота є повною та структурованою, а також чіткою у викладенні матеріалу. Студентка уміє впевнено використовувати наукові джерела та методи, що підкріплюють наведене дослідження, тому це свідчить про глибоке розуміння мети та поставлених завдань.

4. Ступінь самостійності під час виконання кваліфікаційної роботи

Усі завдання кваліфікаційної роботи бакалавра виконанні студенткою самостійно, тому усі отримані результати роботи є наслідком індивідуальної діяльності авторки.

5. Ступінь оволодіння методами дослідження

При виконанні кваліфікаційної роботи студентка показала хороший рівень компетентностей та володіє необхідними засобами, методами та технологіями передбачених стандартом бакалавра спеціальності 122 – Комп'ютерні науки.

6. Повнота та якість розкриття теми роботи

Обрана тема роботи ретельно обґрунтована та всебічно розкрита. В роботі здійснено аналіз актуальності та сучасних досліджень у вибраній області, що свідчить про високий рівень підготовки студентки. Поставлені завдання були успішно виконані, а розроблена інформаційна система виявилась ефективним інструментом для валідації та верифікації запропонованого в роботі методу.

7. Логічність, послідовність, аргументованість, літературна грамотність викладення матеріалу

Структура роботи і її логічна послідовність відповідають поставленим завданням і демонструють високий рівень організації дослідження. Матеріал викладено послідовно, з обґрунтованими аргументами і відповідно до вимог наукової літератури.

8. Можливість практичного застосування кваліфікаційної роботи бакалавра, окремих її частин

Програмна реалізація для визначення діагнозу за текстовим описом з використанням NLP може служити ефективним інструментом для раннього виявлення потенційних захворювань, що дозволяє пацієнтам швидше отримувати медичну допомогу та лікування, забезпечуючи своєчасну реакцію на медичні проблеми і покращуючи результати терапії.

9. Висновок про можливість допуску кваліфікаційної роботи бакалавра до захисту, на яку оцінку заслуговує робота

*Враховуючи високий рівень виконання та забезпечення усіх необхідних вимог, робота може бути допущена до захисту. Рекомендована оцінка «**добре**» .*

Керівник _____ *викладач кафедри КН Валерія КЛІМЕНКО*

ХМЕЛЬНИЦЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
МОН УКРАЇНИ

Кафедра комп'ютерних наук

РЕЦЕНЗІЯ

на кваліфікаційну роботу бакалавра

студентки *гр. КН-20-2 Козенко Юлії Василівни*

за темою: *Метод визначення діагнозу за текстовим описом симптомів NLP-засобами*

1. Актуальність обраної теми

Розробка такого роду систем дозволяє автоматизувати процеси діагностики, зменшуючи навантаження на медичний персонал і забезпечуючи швидше реагування, що особливо важливо в критичних випадках. Крім того, такі системи можуть надавати точніші діагнози шляхом аналізу великої кількості можливих варіантів, що допомагає в складних клінічних ситуаціях.

2. Повнота розкриття мети та завдань роботи

Авторка кваліфікаційної роботи бакалавра повністю розкрила мету та завдання даної роботи. Було детально розглянуто предметну область, можливі шляхи вирішення завдання, а також послідовно викладено кроки розробленого методу визначення діагнозу за текстовим описом симптомів NLP-засобами. А також детально описано розроблену інформаційну систему та проведено її дослідження, що показало високу ефективність розробленого методу.

3. Зміст кожного розділу роботи

Наведені в роботі розділи повністю відповідають змісту. У першому розділі кваліфікаційної роботи бакалавра було наведено характеристику предметної області діагностики захворювань NLP-засобами, у другому розділі розроблено метод визначення діагнозу за текстовим описом симптомів NLP-засобами та детально наведено його кроки. У третьому розділі проведено експериментальне дослідження методу та наведено особливості програмної реалізації інформаційної системи.

4. Оцінка розробленої інформаційної системи, її практична цінність

Створена інформаційна система з методом визначення діагнозу має великий потенціал використання для раннього виявлення можливих захворювань, що сприяє швидкому доступу пацієнтів до медичної допомоги та лікування. Наведений в роботі підхід забезпечує ефективне реагування на медичні проблеми та здатен покращувати результати лікування завдяки своєчасній і точній діагностиці.

5. Якість оформлення кваліфікаційної роботи бакалавра

Кваліфікаційна робота бакалавра має чітку структуру та послідовно викладений матеріал. Текст роботи супроводжується рисунками, графіками, також табличним поданням інформації, що полегшує сприйняття наведеної інформації. Також робота має великий перелік посилань на джерела, що вказує на всебічне дослідження матеріалів за темою роботи.

6. Недоліки кваліфікаційної роботи бакалавра

Деякі рисунки наприклад рис. 3.12, розміщений раніше ніж згадується у тексті пояснювальної записки. Проведене дослідження ефективності є обмеженим з точки зору досліджених параметрів. Невідповідно велика увага в ілюстраціях приділяється логотипам програмних засобів.

7. Загальний висновок (допускається чи не допускається до захисту), та оцінка на яку заслуговує кваліфікаційна робота.

Враховуючи рівень виконання та забезпечення усіх необхідних вимог, робота може бути допущена до захисту. Рекомендована оцінка ***«добре»****.*

Рецензент _____ Говорущенко Т. О.