Міністерство освіти і науки України
Харківський національний університет радіоелектроніки

Факультет Комп'ютерних наук
(повна назва)

Кафедра Штучного інтелекту
(повна назва)

АТЕСТАЦІЙНА РОБОТА
Пояснювальна записка

рівень вищої освіти другий (магістерський)

Дослідження ансамблю алгоритмів класифікації у підсистемі симптоматичної діагностики захворювань
(тема)

Виконав:
студент 2 курсу, групи СШІм-19-1
Дудар В.В.
(прізвище, ініціали)

Спеціальність 122 Комп'ютерні науки
(код і повна назва спеціальності)

Тип програми освітньо-професійна
(освітньо-професійна або освітньо-наукова)

Освітня програма Системи штучного інтелекту
(повна назва спеціалізації)

Керівник к.т.н., доц. Магдаліна І.В.
(посада, прізвище, ініціали)

Допускається до захисту

Зав. кафедри _____ В.О. Філатов
(підпис) (прізвище, ініціали)

2020 р.

Харківський національний університет радіоелектроніки

Факультет Комп'ютерних наук
(повна назва)

Кафедра Штучного інтелекту
(повна назва)

Рівень вищої освіти другий (магістерський)

Спеціальність 122 Комп'ютерні науки
(код і повна назва)

Тип програми освітньо-професійна
(освітньо-професійна або освітньо-наукова)

Освітня програма Системи штучного інтелекту (СШІ)
(повна назва)

ЗАТВЕРДЖУЮ:
Зав. кафедри _____
(підпис)
«_____» _____ 20 ___ р.

ЗАВДАННЯ
НА АТЕСТАЦІЙНУ РОБОТУ

студентові Дудару Владиславу Вадимовичу
(прізвище, ім'я, по батькові)

1. Тема роботи Дослідження ансамблю алгоритмів класифікації у підсистемі симптоматичної діагностики захворювань
затверджена наказом університету від __ _____ 20 ___ р. № _____
2. Термін подання студентом роботи до екзаменаційної комісії _____ _____ 20 ___ р.
3. Вихідні дані до роботи Науково-технічні публікації, Інтернет-ресурси за темою, дані статей, публікації в наукових журналах

4. Перелік питань, що потрібно опрацювати в роботі
Аналіз предметної галузі та постановка задачі, аналіз існуючих систем, збір даних для моделі діагностики, дослідження методів визначення діагнозу на основі відомих симптомів, розробка системи діагностики як сервісу, розробка мобільного клієнту

5. Перелік графічного матеріалу із зазначенням креслеників, схем, плакатів, комп'ютерних ілюстрацій (п.5 включається до завдання за рішенням випускової кафедри) Рисунок 1 - Досліджень джерел збору даних високого рівня для машинного навчання, Рисунок 2 - Етапи процесу збору даних, Рисунок 3 - Ілюстрація автоматизованого протоколу для отримання бібліографічних даних пов'язаних із захворюваннями та симптомами, Рисунок 4 - Приклад спільного виникнення (co-occurrence) захворювання – симптому, Рисунок 5 - Список зібраних хвороб та кількість їх згадок у PubMed, фрагмент, Рисунок 6 - Список зібраних симптомів та кількість їх згадок у PubMed, фрагмент, Рисунок 7 - Приклад байєсовської мережі для діагностики захворювань, Рисунок 8-9 - Гістограма точності класифікації за допомогою коефіцієнта Жаккара, коректні/ помилкові тестові дані, Рисунок 10-11 - Гістограма точності класифікації за допомогою байєсовської мережі, коректні/ помилкові тестові дані, Рисунок 12-13 - Гістограма точності класифікації за допомогою показника TF-IDF, коректні/ помилкові тестові дані, Рисунок 14-15 - Гістограма точності класифікації за допомогою алгоритма з ансамбля, коректні/помилкові тестові дані, Рисунок 16 - Гістограма середньої точності класифікації розробленого алгоритму з існуючими системами.

6. Консультанти розділів роботи (п.6 включається до завдання за наявності консультантів згідно з наказом, зазначеним у п.1)

Найменування розділу	Консультант (посада, прізвище, ім'я, по батькові)	Позначка консультанта про виконання розділу	
		підпис	дата
Основна частина	к.т.н., доц. Магдаліна І.В.		

КАЛЕНДАРНИЙ ПЛАН

№	Назва етапів роботи	Терміни виконання етапів роботи	Примітка
1	Отримання завдання на дипломну роботу	01.11.2020	виконано
2	Аналіз предметної галузі та постановка задачі	02.11.2020 – 07.11.2020	виконано
3	Дослідження методів розпізнавання	07.11.2020 – 12.11.2020	виконано
4	Аналіз існуючих проблем даної предметної області	12.11.2020 – 17.11.2020	виконано
5	Збір даних для моделі діагностики	17.11.2020 – 21.11.2020	виконано
6	Модель системи діагностики хвороб на підставі симптомів	21.11.2020 – 26.11.2020	виконано
7	Обробка і оформлення результатів	26.11.2020 – 28.11.2020	виконано
8	Оформлення пояснювальної записки	28.11.2020 – 29.11.2020	виконано
9	Оформлення графічних матеріалів	30.11.2020 – 01.12.2020	виконано
10	Попередній захист	14.12.2020	виконано
11	Захист перед ЕК	16.12.2020	

Дата видачі завдання 01 11 20 20 р.

Студент _____
(підпис)

Керівник роботи _____ _____
(підпис) (посада, прізвище, ініціали)

РЕФЕРАТ

Записка пояснювальна: 69 с., 16 рис., 1 табл., 2 дод., 25 джерел.

БАЗА ЗНАНЬ, ЕКСПЕРТНА СИСТЕМА, КЛАССИФІКАЦІЯ, МАШИННЕ НАВЧАННЯ, МЕДИЧНА ДІАГНОСТИКА, СИМПТОМИ, СИСТЕМА ПРИЙНЯТТЯ РІШЕНЬ, ХВОРОБИ, ШТУЧНИЙ ІНТЕЛЛЕКТ

Об'єкт дослідження – це процес медичної діагностики хвороб за наявними симптомів за допомогою використання штучного інтелекту і методів машинного навчання.

Предмет дослідження – це методи отримання даних для створення бази знань з бази даних медичних досліджень за остання століття PubMed та застосування індекса Жаккара і статистичного показника tf-idf для класифікації хвороб за наявними симптомами, як характеристиками.

Мета роботи – створення підсистеми первинної медичної діагностики хвороб людей за наявними симптомами.

Методи дослідження – пошук та збір даних для бази знань системи діагностики хвороб, підготовка та перетворення даних у базу знань, розробка методів класифікації хвороб за наявних симптомі, дослідження їх точності при умовах близьких до робочих умов розробленої системи, інтегрування методу збору даних у форматі інтервью, створення программного інтерфейсу та мобільного клієнту.

РЕФЕРАТ

Пояснительная записка: 69 с., 16 рис., 1 табл., 2 прил., 25 источников.

БАЗА ЗНАНИЙ, БОЛЕЗНИ, ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ, КЛАССИФИКАЦИЯ, МАШИННОЕ ОБУЧЕНИЕ, МЕДИЦИНСКАЯ ДИАГНОСТИКА, СИМПТОМЫ, СИСТЕМА ПРИНЯТИЯ РЕШЕНИЙ, ЭКСПЕРТНАЯ СИСТЕМА

Объект исследования – это процесс медицинской диагностики болезней по имеющимся симптомам посредством использования искусственного интеллекта и методов машинного обучения.

Предмет исследования – это методы получения данных для создания базы знаний из базы данных медицинских исследований с последняя века PubMed и применения индекса Жаккара и статистического показателя tf-idf для классификации болезней по имеющимся симптомам, как характеристиками.

Цель работы – создание подсистемы первичной медицинской диагностики болезней людей по имеющимся симптомам.

Методы исследования – поиск и сбор данных для базы знаний системы диагностики болезней, подготовка и преобразования данных в базу знаний, разработка методов классификации болезней при имеющихся симптоме, исследование их точности при условиях близких к рабочим условиям разработанной системы, интеграции метода сбора данных в формате интервью, создание программного интерфейса и мобильного клиента.

ABSTRACT

Explanatory note: 69 p., 16 fig., 1 tabl., 2 ann., 25 sources.

ARTIFICIAL INTELLIGENCE, BASE OF KNOWLEDGE, CLASSIFICATION, DECISION SYSTEM, DISEASE, EXPERT SYSTEM, MACHINE LEARNING, MEDICAL DIAGNOSIS, SYMPTOMS

The object of study is the process of medical diagnosis of diseases according to the symptoms through the use of artificial intelligence and machine learning methods.

The subject of the study is the methods of obtaining data for creating a knowledge base from the medical research database from the last century PubMed and using the Jaccard index and the tf-idf statistical indicator to classify diseases according to symptoms as characteristics.

The purpose of the work is to create a subsystem for the primary medical diagnosis of human diseases according to the symptoms.

Research methods – search and collection of data for the knowledge base of the system for diagnosing diseases, preparing and converting data into a knowledge base, developing methods for classifying diseases with existing symptoms, studying their accuracy under conditions close to the working conditions of the developed system, integrating the method of collecting data in an interview, creating a application programing interface service and a mobile client.

ЗМІСТ

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ, СИМВОЛІВ, ОДИНИЦЬ, СКОРОЧЕНЬ ТА ТЕРМІНІВ

AI – Artificial Intelligence – штучний інтелект;

API – Application Programming Interface – програмний інтерфейс;

CBR – Case-Based Reasoning – обгрунтування на основі випадків;

CDSS – Clinical Decision Support System – клінічна система підтримки прийняття рішень;

Co-Occurence – міра кількості сумісних згадок симптома і хвороби;

DDSS – Diagnosis Decision Support System – діагностична система підтримки прийняття рішень;

ES – Expert System – експертна система;

KBS – Knowledge Based System – системи зосновані на базі знань;

MeSH – Medical Subject Headings – медичні предметні рубрики;

Medline – MEDlars onLINE;

RANSAC – RANdom SAmple Consensus – випадковий вибір консенсусу.

ВСТУП

Машинне навчання – це дисципліна штучного інтелекту, яка займається навчанням систем штучного інтелекту за допомогою великої кількості даних. Це поєднання інформатики та статистики. Комп'ютерні науки в основному орієнтовані на вирішення проблем і визначення того, чи проблеми вирішуються на всіх етапах. Ідея статистики – це моделювання даних, гіпотези та вимірювання надійності. Машинне навчання – це парадигма, яка вивчає досвід минулого для поліпшення майбутньої діяльності. Основною метою цієї галузі є автоматичні методи навчання. Навчання означає модифікацію або вдосконалення алгоритму на основі попереднього досвіду без участі людини. Машинне навчання в основному зосереджено на розробці програм, які використовують дані для самонавчання. Машинне навчання забезпечує алгоритми та інструменти, які роблять систему розумною. В основному воно використовується для вирішення проблем без детермінованого рішення, де немає конкретних моделей проблеми. Алгоритми розроблені на основі різноманітних дисциплін і використовуються в основному для точності, швидкості та настроюваності.

Машинне навчання також сприяє медичному діагнозу для прогнозування захворювань, аналізу даних, планування терапії тощо. Машинне навчання інтегрує комп'ютерну систему з медичною сферою для ефективної діагностики та якісного лікування медичними експертами. Медичний діагноз є важливим завданням різних інтелектуальних систем. Будь-яке медичне лікування починається з скринінгу, діагностики, лікування та частого моніторингу. У теперешні дні медична сфера все більше покладається на комп'ютерні технології, а машинне навчання використовується для діагностування у великій кількості проектів. У медичній діагностиці точний діагноз дуже важливий для вибору правильної

терапії на ранній стадії. Але в багатьох випадках для експерта дуже важко визначити стан пацієнта. За допомогою клінічних записів можна використовувати методи машинного навчання для описового аналізу клінічних ознак. Алгоритми машинного навчання широко використовуються в діагностиці різних захворювань, таких як діабет, проблеми з серцем, рак.

Серед різних алгоритмів, найчастіше використовувані – це метод опорних векторів і дерева рішень. В основному існує багато типів алгоритмів ML:

- навчання з вчителем;
- навчання без вчителя;
- гібридне навчання;
- навчання з підкріпленням;
- глибоке навчання.

Алгоритми навчання з вчителем навчаються з використанням навчального набору даних, на основі якого прогнозуються результати. Метою є прогнозування вихідного значення для даного вхідного вектора. Вихід може бути безперервним значенням або дискретним значенням. Безперервне значення використовується для задачі регресії, а дискретне значення – для задач класифікації. Навчальний набір даних містить вхідні та вихідні значення вибірки. Популярними алгоритмами навчання є класифікація та регресія. Виходячи з набору даних навчання, прогнозується вихід для нового вхідного значення [1]. Як правило, навчання з вчителем має два типи: параметричні моделі і непараметричні моделі. У параметричних моделях функція прогнозування є комбінацією фіксованого числа параметричних. Перший етап є етапом навчання з використанням навчального набору даних. Після цього етапу тренувальні дані можуть бути відкинуті, оскільки передбачення для нового входу базується на вивчених параметрах. Лінійна регресія і класифікація є деякими з параметричних

моделей. Найбільш успішною параметричною моделлю є нейронні мережі. У непараметричних моделях кількість параметрів залежить від навчального набору даних. Тренувальний набір даних підтримується для прогнозування. Найбільш часто використовуваними непараметричними моделями є векторні машини підтримки і алгоритм найближчого сусіда. Ці алгоритми можуть бути використані як для регресії, так і для класифікації.

Алгоритми навчання без вчителя прогнозують результати на основі подібності між вхідними даними [2]. Кластеризація – це часто використовуваний метод навчання без вчителя. Кластеризація – це метод групування подібних даних на основі їх відстані. Загальна властивість кластерів полягає в тому, що подібність внутрішнього кластера повинна бути високою, а схожість між кластерами повинна бути низькою [3]. Кластери залежать від вхідних значень даних. Аналіз соціальних мереж, генетична кластеризація, аналіз ринку – це деякі з поширених застосувань безнавчаного навчання.

Гібридне навчання є поєднанням як навчання з вчителем, так і без вчителя. Гібридне навчання використовується в основному для навчальних додатків під наглядом, де дані з позначкою не доступні або недорогі для отримання. Він знаходиться між позначеними та немеченими даними. Найдавнішою формою цього алгоритму є модель самонавчання [4]. Це повторюваний процес, де спочатку позначені значення даних використовуються з навчанням з вчителем, а далі нелейблені значення даних позначаються попередніми відомими значення і, нарешті, всі значення даних використовуються для прогнозування. В основному, моделі гібридного навчання можна класифікувати за чотирма класами:

– генеративної моделі;

– модель, де межа рішення знаходиться в області з низькою щільністю;

– на основі графіків;

– двоступенева модель з навчанням без вчителя, за яким йде навчання з вчителем.

Навчання з підкріпленням досліджує тестові дані і знаходить правильний висновок за допомогою оціночного зворотного зв'язку. Основними ознаками є затримка винагороди та проби-і-помилки. Цей алгоритм слідує за процесами прийняття рішень Маркова. Він нагадується, коли результат неправильний, і тоді алгоритм досліджує можливості знайти правильний результат. Цей алгоритм використовується в основному в галузі фінансів, робототехніки, управління запасами.

Глибоке навчання – це ще одна форма машинного навчання, де існує абстракція високого рівня. Він використовує різні шари обробки з лінійними та нелінійними перетвореннями.

Існує велика кількість даних для клінічної оцінки захворювань і симптомів. Ці великі дані повинні підтримуватися і розглядатися в процесі діагностики. Існує потреба у належному управлінні для ефективного вилучення та обробки даних [5]. Це можна зробити за допомогою алгоритмів машинного навчання. Дані поділяються і аналізуються в класифікаторах машинного навчання [6].

В ході цієї роботи буде спроектована та розроблена система медичної діагностики захворювань по наявним у користувача симптомам. Ця система буде розроблена за допомогою алгоритмів машинного навчання та класифікації, з ціллю утилізувати зібрані дані для підвищення точності класифікації хвороби користувача.

1 АНАЛІЗ ПРЕДМЕТНОЇ ГАЛУЗІ І ПОСТАНОВКА ЗАДАЧІ

1.1 Аналіз предметної галузі

Експерті системи (ES) є галуззю штучного інтелекту (AI) і були розроблені спільнотою штучного інтелекту в середині 1960-х років. Експертна система може бути визначена як «інтелектуальна комп'ютерна програма, яка використовує знання і процедури виводу для вирішення проблем, які є досить складними, щоб вимагати значної людської експертизи для їх вирішення».

З цього визначення ми можемо зробити висновок, що експертиза може бути передана від людини до комп'ютера, а потім збережена на комп'ютері у відповідній формі, яку користувачі можуть використовувати на комп'ютері для отримання конкретних рекомендацій. Тоді система може обробити інформацію і прийти до конкретного висновку, щоб надати поради і, при необхідності, пояснити логіку, що стоїть за порадою. Експертні системи надають потужні та гнучкі засоби для вирішення різноманітних проблем, які часто не можуть бути вирішені іншими, більш традиційними та ортодоксальними методами. Терміни експертна система та система знань (KBS) часто використовуються як синоніми. Чотирма основними компонентами таких систем є: база знань, механізм виводу, інструмент інженерії знань і певний інтерфейс користувача. Деякі з важливих сфер використання експертних систем та систем знань включають: медичне лікування, аналіз збоїв в техніці, підтримка прийняття рішень, представлення знань, прогнозування клімату, прийняття рішень і навчання, і керування хімічними процесами.

Експертні системи мають застосування в багатьох областях. Вони в основному підходять у ситуаціях, коли експерт не доступний. Для розробки експертної системи знання повинні бути отримані із експерта до домену.

Потім ці знання використовуються у комп'ютерній програмі. Інженер з знань виконує завдання вилучення знань з експерта домену. Експертні системи, що базуються на правилах, є найбільш відомим типом систем на основі знань. Зачасти знання представлено у формі правил IF-THEN, але зустрічаються і більш складні моделі.

Знання – це теоретичне або практичне розуміння предмета або домену. Іншими словами, знання є сумою того, що зараз відомо. Система діагностики – це система, яка може діагностувати хвороби шляхом перевірки симптомів. Системи медичної діагностики засновані на знаннях та розроблені для діагностики захворювань на основі знань, наданих лікарями.

Медичний діагноз (скорочено МД) – це процес визначення того, яке захворювання або стан пояснює симптоми та ознаки людини. Найчастіше його називають діагнозом, а медичний контекст – неявним. Інформація, необхідна для діагностики, зазвичай збирається з історії хвороби та фізичного огляду особи, яка звертається за медичною допомогою. Часто під час процесу проводять одну або більше діагностичних процедур, таких як медичні тести. Іноді посмертний діагноз вважається своєрідним медичним діагнозом.

Діагностика часто є складною, оскільки багато ознак і симптомів неспецифічні. Наприклад, почервоніння шкіри (еритема) само по собі є ознакою багатьох розладів і, таким чином, не говорить лікареві про те, що це неправильно. Таким чином, необхідно виконати диференційну діагностику, в якій порівнюються і контрастують кілька можливих пояснень. Це передбачає співвідношення різних частин інформації з подальшим розпізнаванням і диференціацією закономірностей. Іноді цей процес полегшується ознакою або симптомом (або групою з декількох), які є патогномонічними.

Діагностика є основним компонентом процедури візиту до лікаря. З

точки зору статистики, діагностична процедура передбачає класифікаційні тести.

Діагноз, у сенсі діагностичної процедури, можна розглядати як спробу класифікації стану людини на окремі категорії, які дозволяють приймати медичні рішення щодо лікування та прогнозу. Згодом діагностична думка часто описується в термінах захворювання або іншого стану, але у випадку неправильного діагнозу, фактичне захворювання або стан людини не є таким, як діагноз людини.

Діагностична процедура може виконуватися різними медичними працівниками, такими як лікар, фізіотерапевт, окуліст, медичний вчений, мануальний терапевт, стоматолог, ортолог, медсестра або лікар-асистент.

Діагностична процедура (а також отриманий результат) не обов'язково передбачає з'ясування етіології захворювань, тобто, що викликало захворювання або стан. Таке з'ясування може бути корисним для оптимізації лікування, подальшого уточнення прогнозу або запобігання рецидиву захворювання або стану в майбутньому.

Початковим завданням є виявлення медичної індикації для виконання діагностичної процедури. Показання включають:

– виявлення будь-якого відхилення від того, що, як відомо, є нормальним, таке як, наприклад, анатомія (структура людського організму), фізіологія (як працює організм), патологія (що може бути неправильним з анатомія та фізіологією), психологія (мислення і поведінка) і людський гомеостаз (механізми збереження систем тіла в рівновазі). Знання того, що є нормальним і вимірювання поточного стану пацієнта проти цих норм, може допомогти у визначенні конкретного відходу пацієнта від гомеостазу і ступеня відходу, що, в свою чергу, може допомогти у визначенні показників для подальшої діагностичної обробки;

– скарги, виражені пацієнтом;

– той факт, що пацієнт звернувся за діагностикою, може сам бути

вказівкою на проведення діагностичної процедури. Наприклад, при візиті лікаря лікар може вже розпочати виконання діагностичної процедури, спостерігаючи за ходою пацієнта з залу очікування до кабінету лікаря ще до того, як він почав подавати скарги.

Навіть під час вже проведеної діагностичної процедури може бути проведена інша, окрема діагностична процедура для іншої, потенційно супутньої, хвороби або стану. Це може статися внаслідок випадкового виявлення ознаки, що не має відношення до параметра, але може бути потенційною ознакою іншої хвороби.

Клінічна система підтримки прийняття рішень (CDSS) – це система інформаційних технологій охорони здоров'я, яка призначена для надання лікарям та іншим медичним працівникам клінічної підтримки прийняття рішень (CDS), тобто допомогу у вирішенні клінічних завдань прийняття рішень. Роберт Хейворд з Центру Доказів Здоров'я запропонував робоче визначення: «Клінічні системи підтримки прийняття рішень пов'язують спостереження за станом здоров'я зі знаннями про здоров'я, щоб вплинути на вибір здоров'я лікарів для поліпшення медичного обслуговування».

Клінічна система підтримки прийняття рішень являють собою важливу галузь застосування штучного інтелекту в медицині.

Система клінічної підтримки прийняття рішень була визначена як «активна система знань, яка використовує два або більше елементів даних пацієнтів для створення рекомендацій, що стосуються конкретних випадків». Це означає що CDSS це просто система прийняття рішень яка фокусується на використанні обробки та управління знань таким чином, щоб досягти клінічних рекомендацій щодо догляду за пацієнтами на основі декількох елементів даних пацієнтів.

Головною метою сучасного CDSS є надання допомоги медичним спеціалістам на етапі медичної діагностики. Це означає, що спеціалісти взаємодіють з CDSS щоб допомогти проаналізувати і досягти фінального

діагнозу на основі даних пацієнтів.

У перші дні, CDSS були задумані як системи, які буквально приймають рішення за спеціаліста. Спеціаліст вводив інформацію і чекав, коли CDSS виведе «правильний» вибір, і спеціаліста просто буде засновуватись на цьому виході. Тим не менш, сучасна методологія використання CDSS для надання допомоги означає, що клініцист взаємодіє з CDSS, використовуючи як свої власні знання, так і CDSS, щоб зробити кращий аналіз даних пацієнта, ніж будь-яка людина або CDSS може зробити самостійно. Як правило, CDSS робить рекомендації для спеціаліста для перегляду, і спеціаліст, як очікується, вибере корисну інформацію з представлених результатів і відсіє помилкові пропозицій CDSS.

Існує два основних типи CDSS, описані нижче:

– knowledge-based;

– non-knowledge-based.

Прикладом того, як медична допомога може бути використана медичною системою підтримки прийняття рішень, є певний тип CDSS – DDSS (системи підтримки рішень для діагностики). DDSS запитує деякі дані пацієнтів і у відповідь пропонує набір відповідних діагнозів. Потім лікар приймає вихідні дані DDSS і визначає, які діагнози можуть бути релевантними і які ні і, при необхідності, наказує подальші тести для конкретизації діагнозу.

Іншим прикладом CDSS була б система обґрунтування на основі випадків – case-based reasoning system (CBR). Система CBR може використовувати дані попереднього випадку, щоб допомогти визначити відповідну кількість променів і оптимальні кути променя для використання в радіотерапії для пацієнтів з раком мозку; Медичні фізики та онкологи потім переглянуть рекомендований план лікування, щоб визначити його життєздатність.

Інша важлива класифікація CDSS ґрунтується на часу її використання.

Лікарі використовують ці системи під час догляду, коли вони мають справу з пацієнтом, при цьому термін використання – предіагностика, діагностика, або постдіагностика. Предіагностичні CDSS системи використовуються для допомоги лікарю визначити діагноз. CDSS, що використовується під час діагностики, допомагають переглянути та відфільтрувати попередні діагностичні рішення лікаря, щоб поліпшити їхні кінцеві результати. Постдіагностичні системи CDSS використовуються для аналізу даних для встановлення зв'язків між пацієнтами та їх минулою історією хвороби та клінічними дослідженнями для прогнозування майбутніх хвороб.

Knowledge-based CDSS складаються з трьох частин: бази знань, механізму виводу і механізму спілкування. База знань містить правила та асоціації складених даних, які найчастіше приймають форму правил IF-THEN. Якщо це була система для визначення взаємодії лікарських засобів, то правило може полягати в тому, що IF був прийнятий препарат X та IF був прийнятий препарат Y THEN необхідно попередити користувача. Використовуючи інший інтерфейс, користувач може редагувати базу знань для оновлення нових лікарських засобів. Механізм виведення об'єднує правила з бази знань з даними пацієнта. Механізм зв'язку дозволяє системі показувати результати користувачеві, а також вводити дані в систему.

Non-knowledge-based CDSS, які не використовують базу знань, використовують форму штучного інтелекту, що називається машинним навчанням, що дозволяє комп'ютерам навчитися з минулого досвіду та/або знаходити моделі в клінічних даних. Це виключає необхідність написання правил та внесення експертів. Однак, оскільки системи, засновані на машинному навчанні, не можуть пояснити причини своїх висновків (це так звані «чорні ящики», оскільки жодна осмислена інформація про те, як вони працюють, може бути розпізнана людським оглядом), більшість лікарів не використовують їх безпосередньо для діагностики, з причини надійності та підзвітності. Тим не менш, вони можуть бути корисними як

постдіагностичні системи, щоб запропонувати спеціалістам більш глибоко вивчити закономірності.

Три типи систем, не заснованих на знаннях, – це векторні системи підтримки, штучні нейронні мережі та генетичні алгоритми.

Штучні нейронні мережі використовують вузли і зважені зв'язки між ними, щоб проаналізувати закономірності, знайдені в даних пацієнтів, для отримання зв'язку між симптомами і діагнозом.

Генетичні алгоритми засновані на спрощених еволюційних процесах з використанням спрямованої селекції для досягнення оптимальних результатів CDSS. Алгоритми вибору оцінюють компоненти випадкових множин рішень задачі. Рішення, які демонструють найліпші результати, потім рекомбінують і мутують та знову проходять процес. Це відбувається знову і знову, поки не буде виявлено належне рішення. Вони функціонально схожі з нейронними мережами в тому, що вони є також «чорними ящиками», які намагаються отримати знання з даних пацієнтів.

Мережі, що не базуються на знаннях, часто зосереджуються на вузькому списку симптомів, таких як симптоми для однієї хвороби, на відміну від підходу, що базується на знаннях, що охоплює діагностику багатьох різних захворювань.

1.2 Аналіз існуючих систем

До початку робіт з дослідження, проектування та розробки мобільної підсистеми прийняття рішень для ранньої діагностики захворювань було прийнято рішення знайти і проаналізувати можливі аналогічні існуючі системи, які повністю, частково або модульно повторюють функціонал системи, яка була поставлена завданням розробки.

Існує вже неймовірна кількість технологій та автоматизації в медицині, незалежно від того, розуміємо ми це чи ні – медичні записи

оцифруються, призначення до лікаря можуть бути заплановані в режимі онлайн, пацієнти можуть пройти в медичні центри або клініки, використовуючи свої телефони або комп'ютери. Оскільки використання технологій зросло у всіх сферах життя, вона спокійно змінила способи, якими ми звертаємося за медичною допомогою.

В ході дослідження були проаналізовані наступні існуючі системи, розроблені з використанням штучного інтелекту і/або машинного навчання, які так чи інакше зачіпають сферу медичної діагностики в тому чи іншому вигляді, будь це повноцінні системи, сервіси або вбудовані підсистеми:

– DXplain;
– MYCIN;
– Babylon Health;
– Infermedica Symptomate;
– Your.MD;
– Ada.

Далі, кожна з вищезазначених медичних систем, що використовують штучний інтелект для медичної діагностики захворювань, будуть розібрані більш докладно.

DXplain – це система підтримки прийняття рішень, була розроблена в лабораторії комп'ютерних наук в Массачусетській загальній лікарні, має характеристики як електронного медичного підручника, так і медичної довідкової системи.

У своєму режимі аналізу або аналізу ситуацій DXplain приймає набір клінічних даних (ознак, симптомів, лабораторних даних), щоб створити список діагнозів, які могли б пояснити (або бути пов'язані) з клінічними проявами. DXplain дає обґрунтування того, чому кожне з цих захворювань може бути розглянуте, припускає, яка подальша клінічна інформація може бути корисною для кожного захворювання, і перераховує, які клінічні прояви, якщо такі є, були б незвичайними або нетиповими для кожного

конкретного захворювання.

У ролі медичного підручника DXplain може надати опис більш ніж 2400 різних захворювань, підкреслюючи ознаки і симптоми, які виникають у кожній хворобі, етіологію, патологію і прогноз. DXplain також надає до 10 посилань для кожної хвороби, вибраних для підкреслення клінічних оглядів, де вони були доступні. Крім того, DXplain може надати список захворювань, які слід враховувати для будь-якого з більш ніж 5000 різних клінічних проявів (ознаки, симптоми та лабораторні дослідження).

DXplain широко використовувався вже більше 25 років, і за цей час вона зросла і розвивалася. Розробка почалася в 1984 році, і перша версія, з інформацією про приблизно 500 захворювань, була випущена в 1986 році. Національне розповсюдження DXplain з базою приблизно 2000 захворювань почалося в 1987 році над набором AMANET. Після припинення експлуатації AMANET у 1990 році DXplain продовжувала розповсюджуватися по комутованих мережах до 1995 року. У період між 1991 і 1996 роками DXplain також розповсюджувалася як окрема версія, яка може бути завантажена на окремий ПК. З 1996 року доступ до Інтернету до веб-версії DXplain замінив всі попередні методи розподілу.

Поточна база знань DXplain включає понад 2400 захворювань та більше 5000 клінічних даних (симптоми, ознаки, епідеміологічні дані та лабораторні, ендоскопічні та рентгенологічні дані). Середній опис захворювання включає 53 висновки, діапазон від 10 до понад 100. Кожна хвороба / знахідна пара має два атрибути, що описують зв'язок: один представляє частоту, з якою відбувається виявлення хвороби, а інша – ступінь присутності. знахідки припускає розгляд захворювання. У базі знань представлені понад 230 000 індивідуальних точок даних, які представляють відносини захворювання / знаходження. Крім того, кожен висновок має пов'язане з хворобою незалежне термін, що вказує на важливість пояснення наявності знахідки. Кожна хвороба також має два асоційованих ознаки: той,

який є грубим наближенням її поширеності (дуже поширеним, поширеним, рідкісним або дуже рідкісним) та іншим його значенням, призначеним для відображення впливу не розглядають захворювання, якщо воно присутнє.

DXplain використовує інтерактивний формат для збору клінічної інформації і використовує модифіковану форму байєсівської логіки для виведення клінічних інтерпретацій. Система використовувалася десятками тисяч лікарів та студентів-медиків з моменту її випуску, як самостійної версії (більше не підтримується), так і через Інтернет. База даних і система постійно вдосконалюються і адаптуються в результаті коментарів користувачів. DXplain також використовується в ряді лікарень і медичних шкіл для клінічної освіти і як освітня допомога у вирішенні клінічних проблем.

MYCIN була експертною системою, яка використовувала штучний інтелект для виявлення бактерій, що викликають важкі інфекції, такі як бактеріємія і менінгіт, і рекомендувала антибіотики, причому дозування коригувалося за вагою тіла пацієнта – назва, отримана з самих антибіотиків, так само багато антибіотиків мають суфікс «-mycin». Систему Mycin також використовували для діагностики захворювань згортання крові. MYCIN був розроблений протягом п'яти або шести років на початку 1970-х років в Стенфордському університеті. Він був написаний в Лісп як докторська дисертація Едуарда Шаркліффа під керівництвом Брюса Г. Бьюкенена, Стенлі Н. Коена та інших.

MYCIN працював з використанням досить простого механізму виведення і бази знань з ~ 600 правил. Було б запитати лікаря, який керує програмою, через довгий ряд простих так / ні або текстових питань. Зрештою, він надав список можливих винуватців бактерій, віднесених від високих до низьких на основі ймовірності кожного діагнозу, його впевненості у вірогідності кожного діагнозу, міркування кожного діагнозу (тобто MYCIN також перелічує питання та правила що призвело його до

ранжирування діагнозу особливим способом), і його рекомендований курс медикаментозного лікування.

MYCIN викликала дебати про використання своєї спеціальної, але принципової, невизначеності, відомої як «фактори визначеності». Розробники провели дослідження, що показують, що продуктивність MYCIN мінімально впливала на збурення метрик невизначеності, пов'язаних з індивідуальними правилами, що свідчить про те, що потужність системи більше пов'язана з його представленням знань і схемою міркувань, ніж до деталей його чисельної моделі невизначеності. Деякі спостерігачі вважали, що необхідно було використовувати класичну байєсовську статистику. Розробники MYCIN стверджували, що для цього потрібні або нереалістичні припущення про імовірнісну незалежність, або вимагають від експертів надавати оцінки для невиправдано великої кількості умовних ймовірностей.

Наступні дослідження пізніше показали, що модель коефіцієнта визначеності дійсно може бути інтерпретована в імовірнісному сенсі, і висвітлено проблеми з передбачуваними припущеннями такої моделі. Однак модульна структура системи виявилася б дуже успішною, що призвело б до розробки таких графічних моделей, як байєсівські мережі.

У MYCIN було можливим, що два або більше правил можуть зробити висновки про параметр з різними вагами доказів. Наприклад, одне правило може зробити висновок, що цей організм є E. Coli з визначеністю 0,8, а інший робить висновок, що це E. Coli з певністю 0,5 або навіть -0,8. У випадку, якщо визначеність є меншою за нуль, докази насправді проти гіпотези. Для обчислення коефіцієнта визначеності MYCIN об'єднали ці ваги, щоб отримати один фактор достовірності (формула 1.1).

$$CF(x,y) = \begin{cases} X + Y - XY, & \text{якщо X,Y} > 0 \\ X + Y + XY, & \text{якщо X,Y} < 0 \\ \dfrac{X + Y}{1 - \min(|X|,|Y|)}, & \text{інакше} \end{cases} \quad (1.1)$$

де X і Y є факторами визначеності. Ця формула може застосовуватися більш ніж один раз, якщо більше двох правил роблять висновки про один і той же параметр. Вона комутативно, тому не має значення, в якому порядку ваги були об'єднані.

Дослідження, проведені в Стенфордській медичній школі, показали, що MYCIN отримала рейтинг прийнятності 65% на плані лікування від групи з восьми незалежних фахівців, що було порівняно з 42,5% до 62,5% для п'яти викладачів. Це дослідження часто цитується як показ потенціалу для розбіжностей щодо терапевтичних рішень, навіть серед експертів, коли немає «золотого стандарту» для правильного лікування.

MYCIN ніколи не використовувався на практиці. Це було не через слабкість у його виконанні. Деякі спостерігачі піднімали етичні та правові питання, пов'язані з використанням комп'ютерів у медицині. Однак найбільшою проблемою, а також тим, що MYCIN не використовувалася в рутинній практиці, був стан технологій системної інтеграції, особливо на час її розробки. MYCIN була автономною системою, яка вимагала від користувача вводити всю релевантну інформацію про пацієнта, ввівши відповіді на поставлені запитання MYCIN. Програма працювала на великій системі, доступній у ранній мережі Інтернет (ARPANet), до того як були розроблені персональні комп'ютери.

Найбільший вплив MYCIN були, таким чином, демонстрація потужності її представництва та підхід міркування. Системи, засновані на правилах, в багатьох немедичних областях були розроблені в роки, що слідували за впровадженням підходу MYCIN. У 1980-х роках були введені «снаряди» експертної системи (в тому числі на основі MYCIN, відомої як E-

MYCIN (з подальшою інженерною середовищем знань – KEE)) і підтримали розробку експертних систем у різних сферах застосування. Труднощі, що виникли під час розробки MYCIN та наступних складних експертних систем, полягали у вилученні необхідних знань для використання механізму висновку від експертів з людських ресурсів у відповідних галузях в основу правил.

Заснований у Великобританії стартап Babylon Health – це зоснована на підписках служба для пілкуванні о стані здоров'я, яка розробила chatbot для профілактики та діагностики захворювань.

Використовуючи розпізнавання мови, чат-бот, як повідомляється, порівнює симптоми, які він отримує від користувача, з базою даних хвороб. У відповідь він рекомендує відповідний хід дій, заснований на поєднанні повідомлених симптомів, історії хвороби та обставин пацієнта.

Наприклад, відповідь програми на когось, хто описує грипоподібні симптоми, може бути рекомендацією відвідати аптеку для безрецептурного лікування. На противагу цьому, якщо користувач повідомляє про більш серйозні симптоми, програма може рекомендувати набрати гарячу лінію або вийти безпосередньо до лікарні.

На додаток до своєї функції діагностики, програма також призначена для інтеграції даних пацієнтів з пристроїв, які можна носити, для моніторингу життєво важливих факторів, таких як частота серцевих скорочень і рівень холестерину.

Infermedica розробила сервіс, який використовує штучний інтелект і машинне навчання для управління чатботом для перевірки симптомів, Symptomate.

Алгоритми, що пройшли навчання у великій базі даних медичної літератури та випадків пацієнтів компанії Infermedica, навчаються розпізнавати загальні симптоми за допомогою обробки природних мов.

У центрі всіх систем медичної діагностики Infermedica лежить їх

сервіс – модель, яка використовує як свою основу велику базу знань хвороб і їх симптомів, а також показання пацієнтів, на основі якої будується байєсовські мережі класифікації.

«Мозок» Symptomate – складний алгоритм AI, який контролює послідовність питань і обчислює остаточні рекомендації. Вона вивчає відому медичну літературу та набори даних, зібрані з їхніми діловими партнерами. Весь процес ретельно опікується медичною командою з 15 осіб. Symptomate може легко розрізняти 600 хвороб і 1500 симптомів.

Your.MD стверджує, що використовує AI і машинне навчання для надання персоналізованої медичної інформації та відповідних продуктів і послуг.

Алгоритми, підготовлені на «перевіреній медичній літературі, що охоплюють більше 1000 медичних умов», дозволяють чат-боту вивчати загальні симптоми та надавати рекомендації щодо відповідних ресурсів.

Алгоритми, підготовлені на «перевіреній медичній літературі, що охоплюють більше 1000 медичних умов», дозволяють чат-боту вивчати загальні симптоми та надавати рекомендації щодо відповідних ресурсів.

Наприклад, користувачі вводять свої симптоми через чат, можуть переглядати список відповідних умов і через ряд підказок віртуальний помічник може визначити потенційний стан пацієнта. Chatbot доступний через 6 платформ, включаючи Facebook Messenger і Skype.

Ada Health GmbH є компанією, що базується в Берліні і виробляє програму для перевірки симптомів Ada.

Її заснували Клер Новорол, британський педіатр, Мартін Хірш і Даніель Натрат. Натрат – випускник юридичного центру університету Х'юстона. Компанія залучила $ 69,3 млн. З моменту свого заснування в 2011 році.

Додаток почав працювати як платформа для лікарів і був адаптований у 2016 році, щоб зосередитися на бітах, які пацієнти могли зрозуміти.

Додаток приймає повідомлені симптоми, порівнює їх з симптомами пацієнтів подібного віку та статі і повідомляє про статистичну ймовірність того, що пацієнт має певний стан. Детальний звіт, складений Ada, може бути надісланий лікареві як PDF.

Програма доступна англійською, німецькою, іспанською та португальською мовами. У вересні 2018 р. Його було завантажено близько п'яти мільйонів разів. Він безкоштовний і має найвищі споживчі ряди серед подібних програм. Вона буде доступна на суахілі та румунській мові завдяки фінансуванню Фонду Білла і Мелінди Гейтс та Фонду Ботнар.

Ada порівнювали з WebMD, GP Babylon's Hand and Your.MD. У жовтні 2017 року, коли було протестовано три програми з симптомами астми, черепиці, захворювань печінки, пов'язаних з алкоголем, та інфекції сечовивідних шляхів, Ада показала дуже добрі результати – сервіс запитав про найважливіші симптоми і надав найкращі діагнози. Він давав діаграми, що показують, які симптоми для кожної хвороби були присутні, а також силу зв'язку і діаграму відсотка людей, які, ймовірно, мають такий діагноз.

1.3 Постановка задачі

Перед початком проектування системи необхідно проаналізувати існуючі аналоги реалізованих в даній сфері систем і сформулювати функціонал майбутньої системи. Далі слід розбити функціонал системи на окремі компоненти та визначити план майбутніх робіт і список реалізації необхідного функціоналу в системі.

Після проведення теоретичних досліджень була сформульована концепція завдання – реалізація підсистеми прийняття рішень для ранньої діагностики захворювань. Дана система дозволить працювати з користувачем в форматі інтерв'ю, на зразок того як користувача опитував б реальний експерт. На початку система дізнається у користувача причину

його звернення до даної програми – основні скарги користувача в форматі 2-3 основних симптомів, наприклад головний біль і жар. Далі система, грунтуючись на даних основних симптомах, запропонує користувачеві на вибір кілька симптомів, які були часто помічені разом з симптомами, які користувач ввів перед цим. Після того як користувач дасть відповідь почнеться основна частина інтерв'ю з системою, а саме система буде по черзі цікавитися у користувача є у нього той чи інший симптом. Але, щоб скоротити час інтерв'ю, система буде підбирати нові питання грунтуючись на переданої користувачем інформації про наявність і відсутність тих чи інших симптомів раніше в інтерв'ю. Питання про присутність симптомів будуть підібрані таким чином, щоб скоротити час інтерв'ю до мінімуму із збереженням високої точності правильності діагнозу, навіть при наявності невірно переданих користувачем даних щодо кількох симптомів. Після певної кількості питань, система виведе користувачеві результат, що показує список його можливих діагнозів і ймовірність їх присутності у користувача. Також система повинна попереджати користувача що дана інформація є лише ознайомчою, не є достовірно вірною і якщо у користувача є скарги, то йому слід звернутися до фахівця, а не займатися самолікуванням. Після цього система надаст користувачеві можливість зберегти дані щодо опитування, а також поділитися ними або відправити лікарю.

В результаті аналізу поставленого завдання, існуючих систем та захисту інформації відноситься до розробки подібних підсистем був розроблений план реалізації даного завдання. План проектування і розробки даної системи є ітеративним і містить наступні етапи:

- визначити і конкретизувати проблему;
- зібрати і опрацювати інформацію що відноситься до даної сфери;
- знайти та зібрати дані;
- обробити дані, скласти з них базу знань;

– визначитися з механізмом підбору релевантних питань (симптомів), інформація про присутність яких необхідна для більш точного результату;

– визначитися з алгоритмом класифікації введеної користувачем інформації, яка є свідомо неповною і ймовірно частково помилковою.

Пункти a і b необхідні для початку роботи і будуть виконуватися за кадром даної роботи, але пункти c-d відносяться безпосередньо до реалізації даної системи, так що як теоретичні так і практичні дослідження, що пов'язані із зазначеними пунктами, а також їх підсумкова практична реалізація будуть детально описані далі в даній роботі у пунктах відносящихся к теоретичним та практичним частинам дослідження.

2 ЗБІР ДАНИХ ДЛЯ МОДЕЛІ ДІАГНОСТИКИ

2.1 Аналіз процесу збору даних

Збір даних є основною перепоною у машинному навчанні та активною темою досліджень у багатьох спільнотах. Існує в основному дві причини за якими збір даних став критично важливим питанням. По-перше, оскільки машинне навчання стає все більш широко використовуваним, ми бачимо нові програми, які не обов'язково мають достатньо маркірованих даних. По-друге, на відміну від традиційного машинного навчання, де інженерна характеристика є основною перепоною, глибокі методи навчання автоматично генерують характеристики, але замість цього вимагають великих обсягів маркірованих даних.

Ми живемо в захоплюючі часи, коли машинне навчання має величезний вплив на широке коло додатків, від розуміння тексту, розпізнавання зображень і мови, до медичної допомоги та геноміки. Яскравим прикладом, як відомо, є методики глибокого навчання точність виявлення діабетичних захворювань очей на зображеннях яких стоїть на рівні з офтальмологами [7]. Значна частина недавнього успіху пояснюється кращою обчислювальною інфраструктурою та великими обсягами навчальних даних. Серед численних проблем, що виникають у машинному навчанні, збір даних стає однією з критичних проблем. Відомо, що більша частина часу для впровадження алгоритмів машинного навчання у систему витрачається на підготовку даних, що включає в себе збір, очищення, розуміння та інженерну характеристику. Хоча всі ці кроки вимагають багато часу, збір даних нещодавно став проблемою через наступні причини.

По-перше, оскільки машинне навчання використовується в нових системах, то, як правило, не вистачає навчальних даних. Традиційні системи, такі як машинний переклад або виявлення об'єктів, користуються

величезною кількістю навчальних даних, накопичених десятиліттями. З іншого боку, нові програми мають мало або взагалі не мають даних для навчання. Як приклад, інтелектуальні заводи все частіше стають автоматизованими, де контроль якості продукції здійснюється за допомогою машинного навчання. Кожного разу, коли з'являється новий продукт або новий дефект для виявлення, на початку існує мало або зовсім відсутні навчальні дані. Наївний підхід до ручного маркування може бути неможливим з обмеженим бюджетом. Ця проблема стосується будь-якої нової та конкретної програми.

Більш того, оскільки глибоке навчання [8] стає популярним, додається ще більше необхідності в навчальних даних. У традиційному машинному навчанні інженерна характеристика є одним з найбільш складних етапів, коли користувачеві необхідно зрозуміти програму та забезпечити функції, що використовуються для моделей навчання. Глибоке навчання, з іншого боку, може автоматично генерувати особливості, але замість цього вимагає великих обсягів навчальних даних, щоб видавати добрий результат [9].

Цікаве спостереження полягає в тому, що методи збору даних надходять не тільки від спільноти машинного навчання (включаючи обробку природних мов і комп'ютерного зору, які традиційно використовують машинне навчання), але також із спільноти менеджменту даними (рисунок 2.1). На рисунку нижче показаний огляд дослідницького ландшафту, де теми, які мають внески спільноти управління даними, виділені курсивом. Традиційно дані маркування були природним фокусом досліджень для задач машинного навчання. Наприклад, напів-контрольоване навчання є класичною проблемою, коли навчання моделі проводиться на невеликій кількості мічених даних і більшій кількості немечених даних. Однак, оскільки машинне навчання необхідно проводити на великих обсягах навчальних даних, питання управління даними, включаючи способи отримання великих наборів даних, як виконувати

маркування даних у масштабі, і як покращити якість великих обсягів існуючих даних, стають більш актуальними.

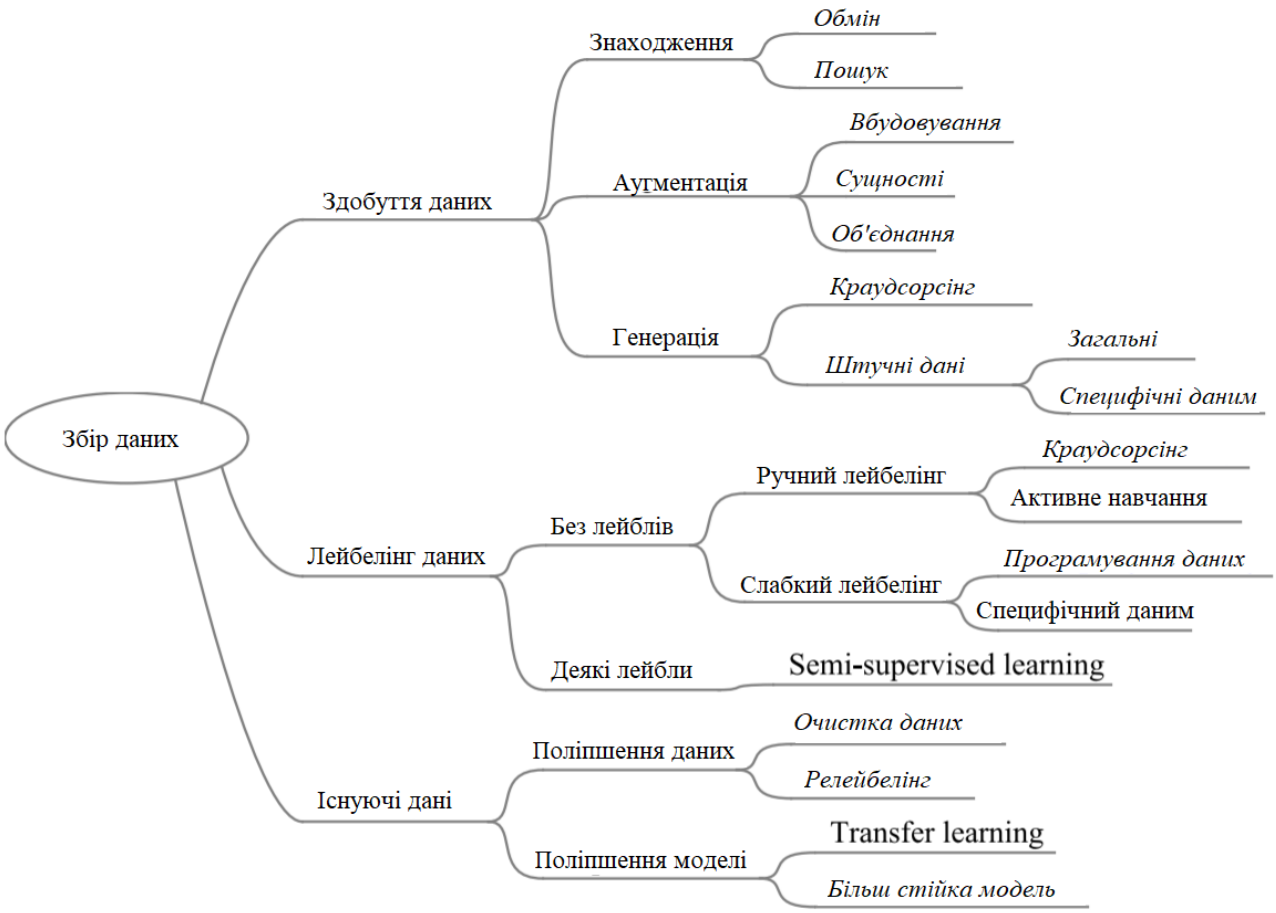

Рисунок 2.1 – Досліджень джерел збору даних високого рівня для машинного навчання. Теми, які, принаймні частково, надаються спільнотою з менеджменту даними, виділяються за допомогою курсиву.

Є багато способів отримати набір даних, наприклад, налаштування API, інтернет, бази даних і т.д. Щоб перетворити зібрані дані в корисні дані, нам необхідно виконати певні процеси, які включають розпакування файлів, очистку, маркування тощо. Попередня обробка є найважливішою частиною обробки даних, коли дані є неповними або відсутні деякі значення.

Необхідно заповнити деякі значення і обробити ці дані так, щоб уникнути помилок (рисунок 2.2).

Рисунок 2.2 – Етапи процесу збору даних

Проаналізувавши методи збору даних для моделей які використовують алгоритми машинного навчання, було визначено етапи збору набору даних у форматі відносин симптом-хвороба для розробляємої моделі діагностики:

– пошук джерел даних;

– збір структурованих і неструктурованих даних;

– обробка і перетворення в необхідний формат відносин симптом-хвороба.

2.2 Аналіз можливих джерел даних

Медичні предметні рубрики (Medical Subject Headings, скорочено MeSH) – всеосяжний контрольований словник, що індексує журнальні статті та книги з природничих наук; може також служити в якості тезауруса, що полегшує пошук інформації. Створено і оновлюється Національної медичної бібліотекою США, використовується в базах статей Medline і PubMed.

MeSH може бути переглянутий і безкоштовно завантажений по Інтернету через Medline. Щорічне друковане видання припинило

випускатися в 2007 році, і в даний час MeSH доступний тільки через Інтернет. Спочатку створений англійською мовою, MESH був переведений на багато інших мов.

Версія MeSH 2009 року містить у в цілому 25186 предметних рубрик (дескрипторів). Більшість з них супроводжуються коротким описом або визначенням, посиланнями на інші дескриптори, а також списком синонімів або схожих термінів. Завдяки спискам синонімів, MeSH може використовуватися як тезаурус.

Medline (MEDlars onLINE) – найбільша бібліографічна база статей з медичних наук, створена Національною медичною бібліотекою США (U.S. National Library of Medicine, NLM). Охоплює близько 75% світових медичних видань. Використовує словник MeSH. MEDLINE є ключовою складової PubMed.

База містить понад 18,8 млн записів про публікації з 1950 року і до теперішнього часу. Спочатку база містила статті, написані після 1965 року народження, але потім були додані і більш ранні публікації. В даний час публікації з 1966 року входять в MEDLINE, а раннє цього року – в OLDMEDLINE. Для записів, доданих в 1995-2003 році: близько 48% опубліковано в США, близько 88% англійською мовою і для 76% є абстракти англійською мовою, надані авторами роботи.

Існує кілька інтерфейсів, за допомогою яких здійснюється доступ до бази даних Medline. Серед них є інтерфейси у відкритому доступі, такі як PubMed і HubMed, і комерційні, наприклад, Ovid Technologies, SwetsWise і деякі інші.

PubMed – англомовна текстова база даних медичних та біологічних публікацій, створена Національним центром біотехнологічної інформації (NCBI) на основі розділу «біотехнологія» Національної медичної бібліотеки США (NLM). Ключовою складовою PubMed є MEDLINE.

PubMed документує медичні та біологічні статті зі спеціальної літератури, а також дає посилання на повнотекстові статті. PubMed включає в себе дані з наступних областей: медицина, стоматологія, ветеринарія, загальну охорону здоров'я, психологія, біологія, генетика, біохімія, цитологія, біотехнологія, біомедицина і т.д. Документувано близько 3800 біомедичних видань. Щорічно база даних PubMed збільшується на 500 000 документів. Пошук відбувається за принципом Medical Subject Headings (MeSH). Кожній статті присвоюється унікальний ідентифікаційний номер PMID (англ. PubMed Identifier – ідентифікатор PubMed).

SNOMED-CT (Систематизована медична номенклатура – Клінічні терміни) – це систематизована машинно-оброблювана медична номенклатура. До складу SNOMED-CT входить сукупність елементів: медичні терміни (terms), коди термінів (codes) і визначники кодів (definitions). SNOMED-CT застосовується в медичній документації і звітах для підвищення ефективності роботи з клінічними даними. Систематизація клінічної інформації сприяє загальному підвищенню якості послуг, що надаються з лікування. SNOMED-CT є базовою термінологією, яка використовується для ведення електронних медичних записів. Терміни номенклатури SNOMED-CT відображають поняття багатьох категорій медицини і охорони здоров'я.

Міжнародна статистична класифікація хвороб і проблем, пов'язаних зі здоров'ям (англ. International Statistical Classification of Diseases and Related Health Problems) – документ, який використовується як провідна статистична та класифікаційна основа в охороні здоров'я. Раз в десять років переглядається під керівництвом Всесвітньої організації охорони здоров'я (ВООЗ). Міжнародна класифікація хвороб (англ. ICD) є нормативним документом, що забезпечує єдність методичних підходів і міжнародну порівнянність матеріалів.

2.3 Теоретичні дослідження збору бібліографічних даних, пов'язаних з симптомами та хворобами

Побудова необхідного набору даних на основі симптомів вимагає базової таксономії захворювань і симптомів, а також корпусу даних, з яких можна витягти їхні відносини. Після оцінки декількох можливих варіантів ми вибрали комбінацію лексики MeSH і бази даних PubMed. Класифікація MeSH визначається фахівцями і пропонує комплексний словник для всіх категорій хвороб (на відміну від, наприклад, OMIM, який фокусується на моногенних захворюваннях), систематично організований в ієрархічному дереві (на відміну від, наприклад, ICD, який має лише два рівні). Найважливішою перевагою для наших цілей є те, що MeSH використовується безпосередньо для індексування всіх статей у масовій базі даних PubMed. Індексація проводиться вручну кваліфікованими фахівцями і відповідно до стандартизованих процедур, що забезпечує високу точність виконання завдань. Крім того, цей процес полегшує основний виклик у видобутку медичних текстів – неоднозначності та різні синоніми одного терміна у номенклатурі, оскільки номенклатура MeSH включає синонімічні псевдоніми для будь-якого терміну.

Основні дані, що використовуються в нашому дослідженні, також мають певні обмеження. Словник MeSH відносно старий і негнучкий, з лише щорічними оновленнями. Це може обмежити ступінь, до якого визначені асоціації охоплюють останні результати досліджень, що швидко розвиваються, у галузі медицини. З іншого боку, стабільні та усталені умови можуть також призвести до більш надійних асоціацій для наших цілей. Іншими важливими недоліками є те, що MeSH має відносно небагато термінів захворювань (порівняно з, наприклад, ICD) і що наші асоціації не походять безпосередньо з клінічної діагностики, а з наукових статей. У майбутньому було б дуже бажано розробити методи, які дозволять

автоматично отримувати інформацію з клінічних записів. Наявні в даний час методи для цієї дуже складної проблеми автоматизованого повнотекстового аналізу у великомасштабних даних не дають результатів з порівнянною точністю. Проблемою, властивою всім таксономіям хвороби, є те, що відмінність між симптомами і захворюваннями не завжди зрозуміла. Наприклад, ожиріння. Згідно з експертною класифікацією MeSH, ожиріння належить до чотирьох різних широких категорій, а саме: «Харчові та метаболічні захворювання», «Діагноз», «Фізіологічні явища» і «Патологічні стани, ознаки та симптоми». Враховуючи його визначення MeSH як «стан з вагою тіла, що значно перевищує прийнятну або бажану вагу, зазвичай через накопичення надлишкових жирів в тілі [...], очевидно, що точна і унікальна класифікація в одну категорію тяжка, і ожиріння може розглядатися як хвороба, симптом, діагноз і фізіологічне явище одночасно.

2.4 Практична реалізація збору та обробки бібліографічних даних, пов'язаних з симптомами та хворобами

Ми використовуємо термінологію Medical Subject Headings (MeSH) [10] для формування зв'язків захворювання-симптоми з метаданих, витягнутих з бібліографічних записів PubMed [11]. PubMed – це найповніша база даних літератури з біомедичних наук. Вона включає в себе MEDLINE [12] і використовує MeSH для кожного цитування для полегшення пошуку інформації. MeSH є контрольованим тезаурусом, який використовується для анотації опублікованих статей, що призводить до високоякісного представлення їх основних тем і внесків. Терміни MeSH призначаються вручну підготовленими індексаторами і використовуються в численних біомедичних текстах та літературних дослідженнях [13], [14], [15], [16].

Ми завантажили версію ASCII 2011 року MeSH [17], яка містить

26,142 різні терміни та їх уніфіковані ідентифікатори. Словник MeSH структурований як ієрархічне дерево з 16 верхніми вузлами, що представляють загальні категорії, такі як «Анатомія», «Хвороби» і «Явища і процеси». Широка категорія «Хвороби» містить підкатегорію «Симптоми і ознаки». (Код MeSH дерева C23.888), який включає терміни, пов'язані з клінічними проявами, що спостерігаються лікарями або сприймаються пацієнтами. Ми використовували всі терміни, що містяться в категорії «Хвороби» (таблиця 2.1), за винятком «Хвороб тварин», а також двадцять термінів, які являють собою лише неспецифічну інформацію про захворювання, таку як «Хвороби», «Синдроми», «Хронічні захворювання». Загалом ми отримали 4 422 чіткі терміни хвороб MeSH і 327 виразних термінів симптомів MeSH для запиту PubMed. Щоб переконатися, що були отримані тільки записи з відповідними індексованими термінами хвороби як основними темами, проводиться пошук по MEDLINE з обмеженням "[Majr:NoExp]", який фільтрує бібліографічні записи з вивченням конкретного захворювання як основною темою.

Таблиця 2.1 – Головні категорії хвороб MeSH

Категорії хвороб	Коди корня дерева MeSH
Bacterial Infections and Mycoses	C01
Virus Diseases	C02
Parasitic Diseases	C03
Neoplasms	C04
Musculoskeletal Diseases	C05
Digestive System Diseases	C06

Продовження таблиці 2.1

Stomatognathic Diseases	C07
Respiratory Tract Diseases	C08
Otorhinolaryngologic Diseases	C09
Nervous System Diseases	C10
Eye Diseases	C11
Male Urogenital Diseases	C12
Female Urogenital Diseases and Pregnancy Complications	C13
Cardiovascular Diseases	C14
Hemic and Lymphatic Diseases	C15
Congenital, Hereditary, and Neonatal Diseases and Abnormalities	C16
Skin and Connective Tissue Diseases	C17
Nutritional and Metabolic Diseases	C18
Endocrine System Diseases	C19
Immune System Diseases	C20
Pathological Conditions, Signs	C23(Виключая C23.888)
Occupational Diseases	C24
Substance-Related Disorders	C25
Wounds and Injuries	C26

Використовуючи інтерфейс веб-сервісу E-Utility API Національного центру біотехнологічної інформації, була розроблена програма JAVA для автоматичного пошуку всіх бібліографічних записів MEDLINE, що публікуються з 1966 року (рисунок 2.3).

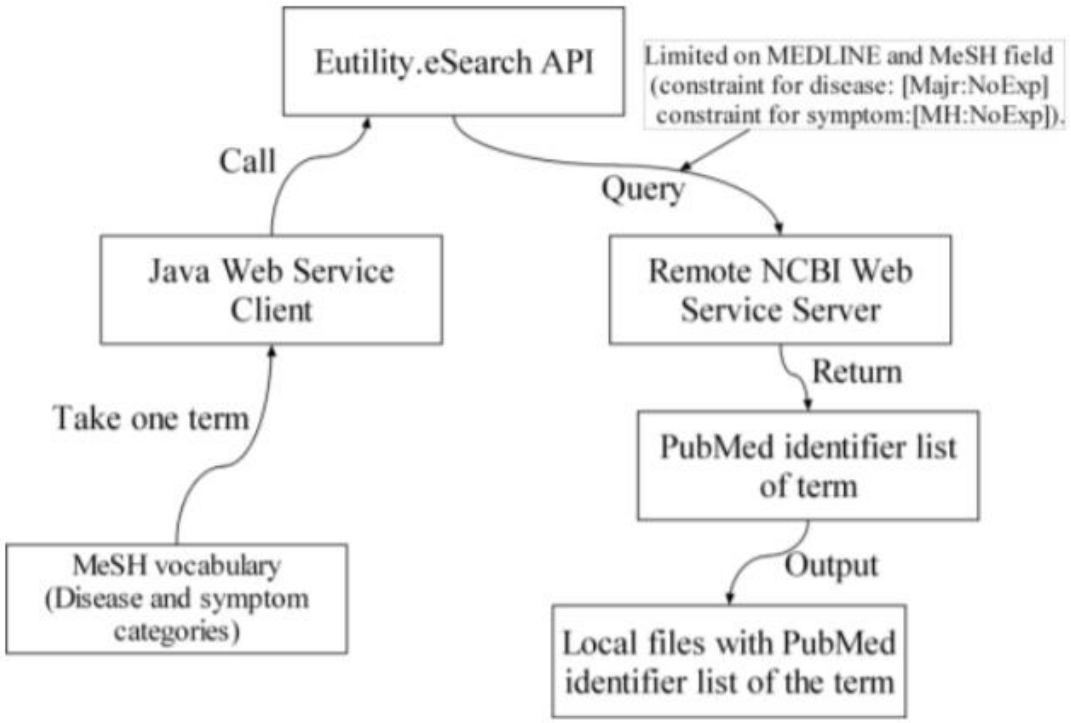

Рисунок 2.3 – Ілюстрація автоматизованого протоколу для отримання бібліографічних даних пов'язаних із захворюваннями та симптомами

Зауважу, що не був виконан повнотекстовий пошук статей або їх рефератів, а лише пошук по їх метаданим. Потім були визначені кількісні асоціації між симптомами і захворюваннями за допомогою терміну co-occurence (кількість ідентифікаторів PubMed, в яких два терміни з'являються разом) (рисунок 2.4). Подібні методи широко використовуються як надійний підхід до виявлення асоціацій між різними медичними сутностями [18]. Зауважимо, що цей підхід не враховує можливих взаємодій між симптомами, але вважає, що різні симптоми даного захворювання є

незалежними один від одного.

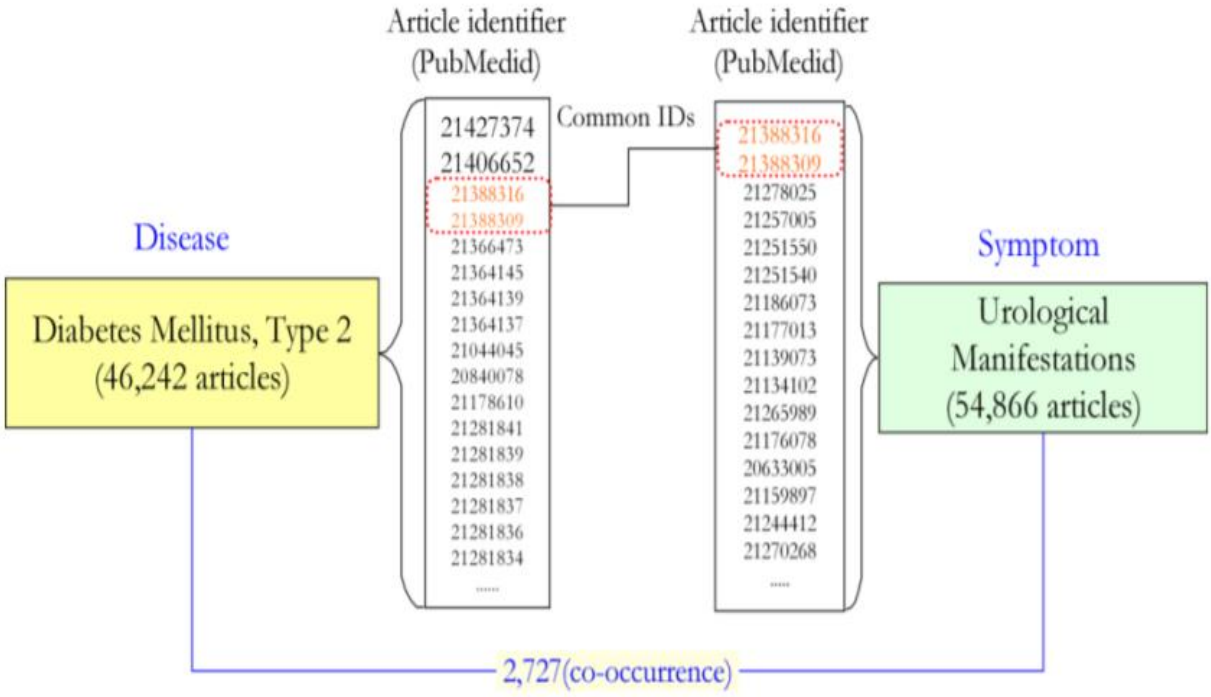

Рисунок 2.4 – Приклад спільного виникнення (co-occurrence) захворювання – симптому

Загальна кількість зібраних записів PubMed становила 7 109 929, з яких 6 553 494 включали захворювання, а 1 405 038 – терміни симптомів. Кількість записів, які містять як хворобу, так і термін симптомів, становила 849,103. Вони включали всі 4442 терміни хвороби MeSH (рисунок 2.5) і майже всі (322, тобто 98%) терміни симптомів (рисунок 2.6).

```
MeSH Disease Term       PubMed occurrence
Breast Neoplasms        122226
Hypertension    107294
Coronary Artery Disease 82819
Lung Neoplasms  78009
Myocardial Infarction   75945
HIV Infections  66601
Coronary Disease        64339
Asthma  63669
Adenocarcinoma  60855
Dementia        55782
Prostatic Neoplasms     54606
Skin Neoplasms  53590
Obesity 50477
Carcinoma, Squamous Cell        50241
Pain    49599
Liver Neoplasms 49005
Arthritis, Rheumatoid   48854
Schizophrenia   46537
Diabetes Mellitus, Type 2       46242
Brain Neoplasms 45792
```

Рисунок 2.5 – Список зібраних хвороб та кількість їх згадок у PubMed, фрагмент

```
MeSH Symptom Term       PubMed occurrence
Body Weight     147857
Pain    103168
Obesity 100301
Anoxia  47351
Mental Retardation      43883
Seizures        36959
Diarrhea        34850
Angina Pectoris 29473
Edema   29446
Birth Weight    29364
Fever   27734
Pain, Postoperative     22793
Deafness        21470
Headache        20212
Vision Disorders        19403
```

Рисунок 2.6 – Список зібраних симптомів та кількість їх згадок у PubMed, фрагмент

2.5 Порівняння побудованого набору даних з медичними термінологічними системами

Детальне знання про зв'язок симптом-хвороба є основою для будь-якого клінічного діагнозу, особливо для клінічних синдромів і традиційних класифікацій захворювань. Для систематичного запису та обробки таких клінічних даних було розроблено декілька медичних термінологій. Найбільш помітним прикладом є SNOMED-CT (клінічні терміни) [19], що містить комплексний набір понять з концепцій клінічної практики та їх взаємозв'язків. Основним застосуванням SNOMED-CT є управління клінічними даними, але він також може бути використаний для біоінформатики [20]. Оскільки відносини в SNOMED-CT безпосередньо випливають з клінічної практики, вони забезпечують ідеальну основу для вивчення зв'язків із симптомами захворювання. Тому цей варіант був проаналізовані, використовуючи дані SNOMED-CT, включені в систему UMLS (UMLS 2012 AA). Оскільки семантичні типи «Знак і Симптом» і «Хвороба або Синдром» в UMLS кодуються відповідно T184 і T047 відповідно, були отфільтрувані записи відносин з обмеженнями цих двох семантичних типів. Всього було отримано 2340 зв'язків між 1623 захворюваннями та 817 симптомами. Було виявлено, що 1250 (77,0%) захворювань мають лише одну споріднену симптоматику і 236 (14,5%) захворювань мають два пов'язаних симптоми. Розлад мігрені, наприклад, має уніфікований ідентифікатор концепції UMLS (CUI) C0149931 і ідентифікатор SNOMED 37796009. У цих даних є лише один запис відносини, в якому симптом є «головний біль» (CUI: C0018681). Хоча головний біль є найбільш домінуючим симптомом мігрені, він, звичайно, не єдиний, ця хвороба має різні супутні симптоми, такі як нудота, запаморочення [21], блювота, світлобоязнь і втома [22]. Далі ми виявляємо, що підтипи, такі як «Мігрень з аурою» (CUI: C0154723), «Мігрень без

аури» (CUI: C1827190) та «Абдомінальна мігрень» (CUI: C0270858), також мають «головний біль», як їх єдиний симптом, тому вони не можуть бути диференційовані за ознаками, пов'язаними з ними в SNOMED-CT. Був зроблений висновок, що відносини захворювання-симптом, що містяться в SNOMED-CT є занадто обмеженими, щоб бути корисними при розробці системи.

Існує кілька інших медичних термінологічних систем, зосноваих на симптомах, такі як ICD 9CM [23], Disease Ontology [24] і Symptom Ontology [25]. Проте жодна з них не містить значного числа зв'язків із симптомами захворювання, які є настільки ж всеосяжними, як ті, що були вилучені з записів PubMed.

3 МОДЕЛЬ СИСТЕМИ ДІАГНОСТИКИ ХВОРОБ НА ПІДСТАВІ СИМПТОМІВ

3.1 Загальний опис модулів системи

Медична діагностика – це процес визначення того, яке захворювання або стан пояснює симптоми та ознаки людини. Найчастіше його називають просто діагнозом мая на увазі медичний контекст. Інформація, необхідна для діагностики, зазвичай збирається експертом під час інтерв’ю та медичного огляду особи, яка звертається за медичною допомогою.

Хоча така частина діагностики як фізичне обстеження пацієнта є значною частиною обстеження і підвищує точність діагнозу, але являється набагато складніше в реалізації, ніж напрямок розглядаємий у цій роботі. Також у багатьох випадках для виконання фізичного обстеження потрібна інтеграція медичного обладнання та результатів їх показників в процес діагностики. Але різноманітність даного обладнання і велике розмаїття хвороб і їх проявів роблять процес впровадження штучного інтелекту в цю процедуру дуже складним і матеріально затратним.

Тому в рамках даної роботи розглядається перша складова діагностики, а саме – інтерв'ю. Формат розроблюваної системи реалізує цей процес у вигляді самодіагностики, де експертом, який опитуватиме пацієнта є додаток, що використовує розробляєму в даній роботі модель.

Наведемо опис функціоналу системи: «На початку система дізнається у користувача причину його звернення до даної програми – основні скарги користувача в форматі 2-3 основних симптомів, наприклад головний біль і жар. Далі система, грунтуючись на даних основних симптомах, запропонує користувачеві на вибір кілька симптомів, які були часто помічені разом з симптомами, які користувач ввів перед цим. Після того як користувач дасть відповідь почнеться основна частина інтерв'ю з системою, а саме система

буде по черзі цікавитися у користувача є у нього той чи інший симптом. Але, щоб скоротити час інтерв'ю, система буде підбирати нові питання грунтуючись на переданої користувачем інформації про наявність і відсутність тих чи інших симптомів раніше в інтерв'ю. Питання про присутність симптомів будуть підібрані таким чином, щоб скоротити час інтерв'ю до мінімуму із збереженням високої точності правильності діагнозу, навіть при наявності невірно переданих користувачем даних щодо кількох симптомів. Після певної кількості питань, система виведе користувачеві результат, що показує список його можливих діагнозів і ймовірність їх присутності у користувача».

Проаналізувавши вимоги до функціоналу системи можна виділити її два основних модуля:

– модуль класифікації;

– модуль ведення інтерв'ю (отримання даних).

Модуль класифікації – це частина системи яка приймає на вхід від користувача список введених їм симптомів, а на виході видає список можливих діагнозів і їх вірогідність. За описом можна зрозуміти що це ніщо інше як процес класифікації, де характеристиками виступають дані про присутність симптомів, представлені у вигляді вектора бінарних значень. Використовуючи ці дані, можна за допомогою навченої на зібраних раніше даних моделі або використовуючи інших алгоритмів класифікації, що не вимагають навчання, отримати необхідний вихід – найбільш ймовірний клас (хворобу) до якої відноситься введені користувачем дані користувача (список симптомів). Також алгоритм реалізований в даному модулі повинен бути стійкий до помилкового введення частини даних від користувача.

Модуль ведення інтерв'ю – це частина системи яка відповідає за збір даних від користувача – списку симптомів, які допоможуть правильно класифікувати наявний стан користувача і скоротять тривалість інтерв'ю до

мінімуму без втрати точності класифікації. Відштовхуючись від основних симптомів, введених користувачем користувачем на початку інтерв’ю, даний модуль проаналізує наявні дані відносин хвороба-симптом і задасть користувачеві ті питання, відповідь на які наблизить інтерв'ю до фіналу – визначення діагнозу. Також даний модуль повинен аналізувати введені користувачем дані на наявність в них помилкових даних – даних які вибиваються із загального потоку спостережень – і мінімізувати їх вплив на хід інтерв'ю і підсумковий результат.

Далі будуть будуть досліджені можливі методи реалізації обох цих модулів і проаналізовані, а також налаштовані під конкретну задачу, існуючі алгоритми і математичні моделі.

3.2 Аналіз існуючих методів класифікації задовольняючих умови поставленої задачі

Існують різні алгоритми машинного навчання які підходять під виконання поставленого завдання, з огляду на наявні для роботи дані.

Дерева рішень.

Дерево рішень є одним з найважливіших і найбільш часто використовуваних алгоритмів класифікації. Цей алгоритм використовує стратегію поділу та завоювання для створення дерева. Існує набір випадків, які пов'язані з набором атрибутів. Дерево рішень складається з вузлів і листя, в яких вузли є тестами на оцінки характеристик або атрибутів, а листя є класами прикладу, що відповідає заданим умовам. Результат може бути «true» або «false». Правила можуть бути отримані з шляху, який починається від кореневого вузла і закінчується на листовому вузлі, і, крім того, використовує вузли в транзиті як передумови для прийнятого правила, щоб передбачити клас на аркуші. Необхідно зробити обрізку дерев, щоб евакуювати безглузді передумови та дублювання. Дерева рішень зазвичай

використовуються в дослідженні операцій, зокрема в аналізі рішень, щоб допомогти визначити стратегію, яка найбільш ймовірно досягає мети, але також є популярним інструментом у машинному навчанні.

K-Means.

Метод k-середніх (англ. K-means) – найбільш популярний метод кластеризації. Алгоритм є версією EM-алгоритму, що застосовується також для поділу суміші Гауссіан. Він розбиває безліч елементів векторного простору на заздалегідь відоме число кластерів k. Основна ідея полягає в тому, що на кожній ітерації переобчислюється центр мас для кожного кластера, отриманого на попередньому кроці, а потім вектори розбиваються на кластери знову відповідно до того, який з нових центрів виявився ближчим за обраною метриці. Алгоритм завершується, коли на якийсь ітерації не відбувається зміни внутрікластерного стану. Це відбувається за кінцеве число ітерацій, так як кількість можливих розбиттів кінцевого безлічі звичайно, а на кожному кроці сумарна квадратичне відхилення V зменшується, тому зациклення неможливо.

K-NN.

Метод k-найближчих сусідів (англ. K-nearest neighbors algorithm, K-NN) – метричний алгоритм для автоматичної класифікації об'єктів або регресії. У разі використання методу для класифікації об'єкт присвоюється того класу, який є найбільш поширеним серед k сусідів даного елемента, класи яких вже відомі. У разі використання методу для регресії, об'єкту присвоюється середнє значення по k найближчим до нього об'єктів, значення яких вже відомі. Алгоритм може бути застосований до вибірок з великою кількістю атрибутів (багатовимірним). Для цього перед застосуванням потрібно визначити функцію відстані; класичний варіант такої функції – евклидова метрика.

SVM.

Метод опорних векторів (англ. SVM, support vector machine) – набір

схожих алгоритмів навчання з учителем, що використовуються для задач класифікації та регресійного аналізу. Основна ідея методу – переклад вихідних векторів в простір більш високої розмірності і пошук розділяє гиперплощини з максимальним зазором в цьому просторі. Дві паралельні гіперплощини будуються по обидва боки гіперплощини, що розділяє класи. Розділяємою гиперплощиною буде гиперплощина, що максимізує відстань до двох паралельних гиперплощин. Алгоритм працює в припущенні, що чим більша різниця або відстань між цими паралельними гиперплощинами, тим менше буде середня помилка класифікатора.

Naive bayes.

Наївний байесовский класифікатор – простий імовірнісний класифікатор, заснований на застосуванні теореми Байеса зі строгими (наївними) припущеннями про незалежність. Залежно від точної природи ймовірнісної моделі, наївні байєсовські класифікатори можуть навчатися дуже ефективно. У багатьох практичних додатках для оцінки параметрів для наївних Байесови моделей використовують метод максимальної правдоподібності; іншими словами, можна працювати з наївною байєсівської моделлю, не вірячи в Байєсова ймовірність і не використовуючи байєсовські методи. Незважаючи на наївний вигляд і, безсумнівно, дуже спрощені умови, наївні байєсовські класифікатори часто працюють набагато краще в багатьох складних життєвих ситуаціях. Перевагою наївного байєсівського класифікатора є мала кількість даних необхідних для навчання, оцінки параметрів і класифікації. Даний алгоритм класифікації, а іноді і його різновид – Байєсова мережу, часто використовують в медичної діагностики через його особливості і переваги.

3.3 Теоретичні дослідження методів визначення діагнозу на основі відомих симптомів

Природа системи ставить такі умови, що у нас не буде спочатку необхідної кількості даних для класифікації діагнозу, тому просте впровадження класифікатора не є вирішенням проблеми. За цією причиною в системі присутній модуль збору даних для класифікації діагнозу у форматі інтерв'ю – тобто процесу задавання питань про наявність симптомів у певній черговості.

На перший погляд здається що для вирішення даного завдання підходить алгоритм побудови дерева рішень на основі початкових симптомів, за яким і буде визначатися хід інтерв'ю і класифікуватися діагноз. Вузлами б в даних деревах були симптоми, переходи далі по гілках здійснювалися б на підставі відповідей про наявність у користувача цих симптомів, а кінцями гілок – діагнози.

Але недоліками дерев рішень є те що вони чутливі до помилкового введення (наприклад користувач помилився в тому що у нього є певний симптом, або він просто в нього не проявився). Хоча є певні підходи до вирішення даної проблеми (наприклад RANSAC – стабільний метод оцінки параметрів моделі на основі випадкових вибірок). Але проаналізувавши плюси і мінуси даного підходу, а також провівши додаткові практичні дослідження було прийнято рішення відмовитися від використання даного алгоритму.

Наступним кандидатами стали алгоритми використовуючи байєсову логіку, наприклад байєсовські мережі. Хоча даний підхід і показує хороші результати в сфері даної задачі, а також використовується в більшості комерційних аналогічних системах, саме з цієї причини було вирішено відмовитися від даного підходу. Надалі, при розробці і проведенні досліджень над обраним методом класифікації, порівняння проводилося

саме з системами на основі даного алгоритму. При побудові класифікатора у вигляді байєсівської мережі ми оперуємо зв'язками у вигляді Parent -> Child, а вся мережа має вигляд спрямованого зваженого графа. Для поточної предметної області у вигляді Parent виступають симптоми (наприклад кашель або головний біль), а у вигляді Child – хвороби або діагнози (наприклад застуда або COVID-19) (рисунок 3.1).

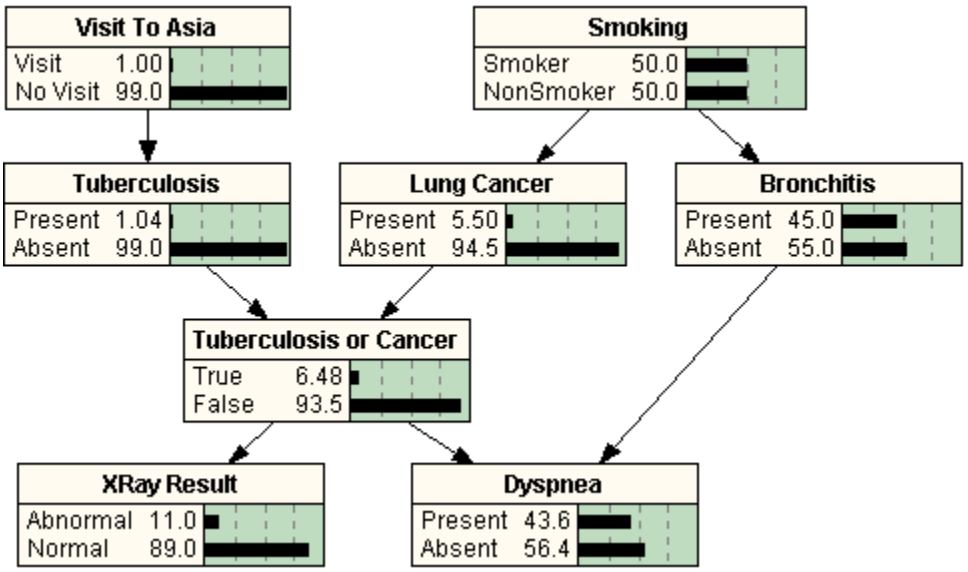

Рисунок 3.1 – Приклад байєсовської мережі для діагностики захворювань

Класифікація при використанні байесовских мереж відбувається за формулою підрахунку ймовірностей (формула 3.1):

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}, \qquad 0 \leq P(A|B) \leq 1, \tag{3.1}$$

Якщо говорити байєсовської логікою, то питання виду "Яка ймовірність того що у мене хвороба D при наявності симптомів A, B, C?" буде виглядати наступним чином (формула 3.2):

$$P(D|A,B,C) = \frac{P(A,B,C|D)P(D)}{P(A,B,C)}, \quad (3.2)$$

За ітогом було прийнято рішення розробити алгоритм класифікації спеціально для даного завдання, а потім використати ансамбль із нового алгоритма та байєсовскої мережі. В результаті як новий алгоритм був використаний алгоритм з сімейства nearest-neighbor, розроблений в двух варіантах за використанням двох мір подібності – на основі коефіцієнта Жаккара і підході, що використовується в текстовому пошуку – term frequency - inverse document frequency.

Індекс Жаккара, також відомий як перетин над союзом і коефіцієнт подібності Жаккар, є статистичною мірою, що використовується для оцінки подібності та різноманітності наборів прикладів. Коефіцієнт Жаккара вимірює подібність між множинами кінцевих прикладів і визначається як розмір перетину, поділений на розмір об'єднання наборів вибірок (формула 3.3):

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}, \qquad 0 \leq J(A,B) \leq 1, \quad (3.3)$$

В даному випадку набором є хвороба, а саме асоційований з нею набір симптомів. Аналізуючи раніше зібрані дані, було виявлено що якщо уявити значення co-occurence у зв'язку хвороба-симптом як bag of words або просто бінарне значення 1 – присутній, а відсутність зв'язку хвороба-симптом як бінарне значення 0 – відсутня, і висловити у вигляді двовимірної матриці, то на виході вийде матриці бінарних значень, де у рядів, відповідних за хвороби буде мала кількість значень 1, і велика кількість значень 0. Якщо по такому набору розраховувати стандартну міру подібності (наприклад cosine similarity), то вийде ситуація де у хвороб, у яких практично немає перетинів присутніх симптомів, буде високе значення міри подібності, так

як у них велика кількість співпадаючих симптомів. Проблема полягає в тому що у ситуацій рівності позитивних бінарних значень вага така ж, як і ситуацій рівності негативних значень. Але якщо мати на увазі предметну область, то можно побачити, що якщо у списку введених користувачем симптомів і у розглянутій хвороби є перетини присутніх симптомів, то для діагностики вага цього спостереження вище ніж збіги відсутності симптомів. Для даних ситуацій відмінно підходить міра подібності Жаккара. Тоді формулу заходи подібності між даними, введеними користувачем I і перевіряється в наборі даних хвороби Di можна виразити в наступному вигляді (формула 3.4):

$$sim_{(I,Di)} = \frac{S}{A + B - S}, \tag{3.4}$$

де S – кількість співпадаючих присутніх симптомів в I та Di;

A – кількість присутніх симптомів в I;

B – кількість присутніх симптомів в Di.

Хоча даний підхід добре підходить для задач, де дані представлені в такому вигляді, він ставить умовою що всі позитивні зв'язки хвороба-симптом мають однакову вагу і вплив на підсумкове значення міри подібності, що не є добре для такої сфери як діагностика хвороб. Це пов'язано з тим що у різних симптомів однієї хвороби може бути різна ступінь і частота прояву, отже присутність у хворого симптомів, які є більш показовими для певної хвороби, має більше впливати на підсумкову ймовірність присутності саме даної хвороби.

Хорошим показником сили зв'язку хвороба-симптом (який також може буди представлений як вірогідність для байєсовської мережі) може бути зібрані в наборі даних показники їх co-occurence Wi, j, адже логічно припустити що більш часте згадування симптому разом з хворобою, показує те що цей симптом більш виразний для даної хвороби. Однак, так як дані

були зібрані з медичних публікацій і не є рівноцінною для всіх хвороб і симптомів вибіркою, а також присутня така ситуація що певні симптоми зустрічаються у багатьох хвороб, з'являється проблема переважання певних симптомів і хвороб, а то й враховувати яку буде помічена втрата в точності діагностики. Поширеність різних симптомів і захворювань дуже відрізняється, наприклад, існують поширені симптоми, такі як біль, і упередженість публікацій щодо певних захворювань, таких як рак молочної залози. Для врахування цієї неоднорідності ми не використовуємо абсолютну co-occurrence $W_{i,j}$ щоб виміряти силу зв’язку між симптомом *i* та хворбою *j*, а частоту слова - обернену частоту документа (англ. term frequency-inverse document frequency) *wi,j* (формула 3.5):

$$w_{i,j} = \frac{w_{i,j}}{w_j} log \frac{N}{n_i}, \tag{3.5}$$

де W_j – позначає сумму co-occurence всіх симптомів у хвороби *j*;

N – позначає кількість всіх захворювань у наборі даних;

n_i – кількість захворювань, де з'являється симптом *i*.

Оскільки всі симптоми в зібраних даних мають принаймні одне асоційоване захворювання, потенційна проблема ділення на нуль не виникає.

У інформаційному пошуку, tf – idf або TFIDF, коротко для частоти слова - оберненої частоти документа, є числовою статистикою, яка має на меті відобразити, наскільки важливим є слово для документа в збірці або корпусі. Він часто використовується як ваговий коефіцієнт у пошуках інформаційного пошуку, видобування тексту та моделювання користувачів. Значення tf – idf збільшується пропорційно кількості разів, коли слово з'являється в документі, і компенсується кількістю документів у корпусі, які містять слово, що допомагає пристосуватися до того, що деякі слова

з'являються частіше. tf – idf є однією з найпопулярніших схем термінового зважування сьогодні; 83% систем на основі текстових рекомендацій у цифрових бібліотеках використовують tf – idf.

Варіації схеми зважування tf – idf часто використовуються пошуковими системами як центральний інструмент у підрахунку та ранжуванні релевантності документа з урахуванням запитів користувача. Tf – idf може бути успішно використано для фільтрації стоп-слів у різних тематичних полях, включаючи узагальнення та класифікацію тексту.

Одна з найпростіших функцій ранжирування обчислюється шляхом підсумовування tf – idf для кожного терміну запиту; багато більш складних функцій ранжування є варіантами цієї простої моделі.

Відштовхуючись від процесу текстового пошуку, якщо представити симптом як слово / термін, список симптомів, який ввів користувач як пошуковий запит, а набір даних, в якому кожна хвороба представлена як вектор зі значень co-occurence симптомів, – корпусом документів, то можна впровадити в модель механізм пошукових систем на основі tf-idf, що призведе до значного підвищення точності класифікації, навіть при наявності помилкового введення користувача. Реалізувати це можна, якщо розрахувавши ці дані і представивши їх як репрезентативні вектори характеристик, а потім використавши їх для розрахунку широко застосовуваємої міри подібності як у видобутку тексту, так і в біомедичній літературі для кількісної оцінки подібності двох понять – косинусної подібності відповідних векторів. Подібність між векторами *dx* і *dy* двох захворювань *x* і *y* обчислюється наступним чином (формула 3.6):

$$P(D) = \cos(d_x, d_y) = \frac{\sum_i d_{x,i}, d_{y,i}}{\sqrt{\sum_i d_{x,i}^2}\sqrt{\sum_i d_{y,i}^2}}, \quad (3.6)$$

де значення міри подібності коливається від 0 (без спільних симптомів)

до 1 (ідентичні симптоми).

Ансамбль методів в статистиці і навчанні машин використовує кілька навчальних алгоритмів з метою отримання кращої ефективності класифікації, ніж могли б отримати від кожного навчального алгоритму окремо. Ансамбль був створений з комбінування двох кращих методів для класифікації хвороб при наявності неповних і помилкових даних – байєсовской мережі і підрахунок подібності на основі tf-idf. При побудові ансамблю був використаний принцип бутстреп-агрегування, часто скорочується до беггінг, який дає кожній моделі в ансамблі однакову вагу (формула 3.7).

$$\mathrm{P(D)} = 0.5 * P_{bayes}(D) + 0.5 * P_{TF-IDF}(D), \quad (3.7)$$

3.4 Практичні дослідження методів визначення діагнозу на основі відомих симптомів

У цьому розділу були проведені практичні дослідження розроблених під час теоретичних досліджень алгоритмів класифікації. Так як модель побудована не основі великої кількості навчальних даних, а на основі сформованої бази знань, а також для того щоб при дослідженні максимально приблизити дані для тестування до даних, які можуть вводити користувачі, було прийнято рішення сформувати ці тестові дані на основі бази знань, але взяти меншу випадкову кількість характеристик і сформувати з них тестові вектори, репрезентуючі хвороби.

Для кожної хвороби у базі знань є відповідна їй інформація о наявності 322 симптомів. Очевидно що під час інтерв’ю користувач не буде відповідати на 322 питання о наявності в нього цих симптомів. Завдяки зарактеристикам системи і методам визначення важливих для ходу інтерв’ю симптомів середня кількість питань в одному інтерв’ю знаходиться у

діапазоні 10-20 питань. Тому у проведених тестах використовується саме такий розмір векторів, представляючих хвороби, з випадково обраними симптомами. Також для більшого реализму у тестові дані були добавлені помилки, як будто їх джерелом була людина.

Перший тест проводився для класифікації за мірою подібності Жаккара, за якою для визначення результата хвороби сортуються. Так як система не повинна видавати один результат, а видає декілька найімовірніших діагнозів, точність була розрахована для 1-3 найвищих діагнозів у списку результатів. Перший тест проводився для даних з відсутніми шумами (помилками) (рисунок 3.2).

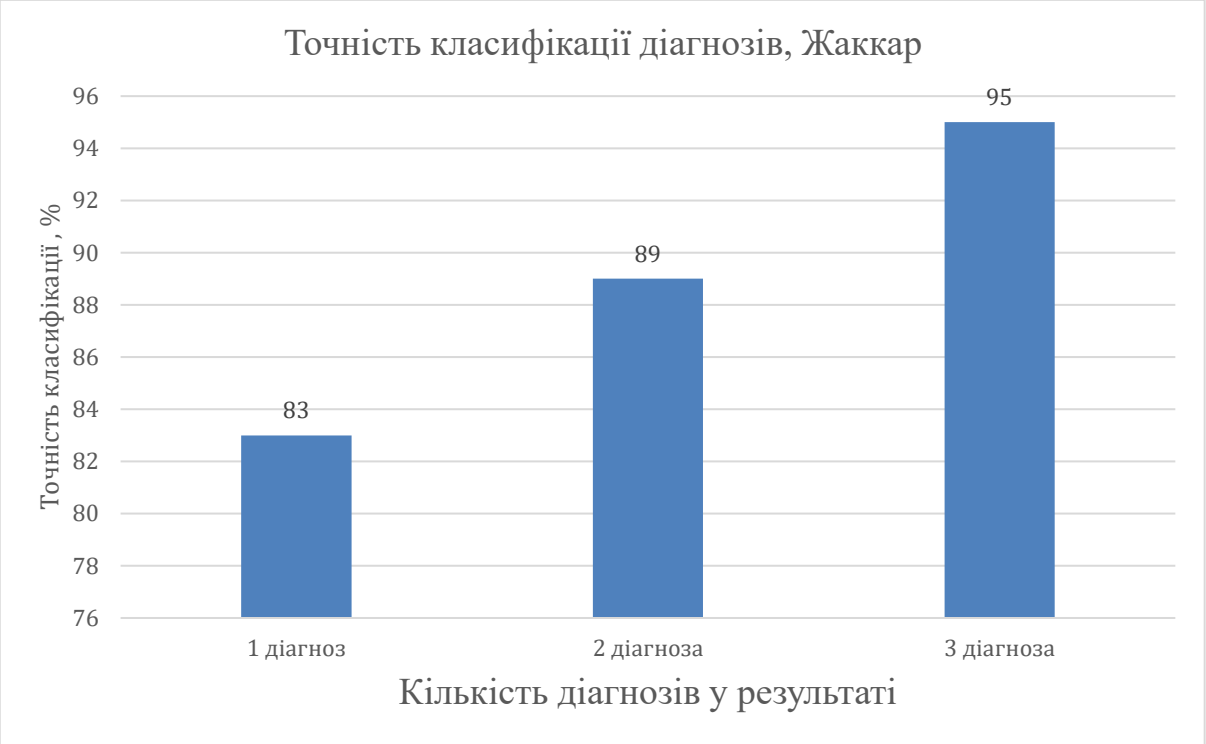

Рисунок 3.2 – Гістограма точності класифікації за допомогою коефіцієнта Жаккара, коректні тестові дані

Як можна побачити використовування міри подібності Жаккара показує добрі результаті, навіть тільки при класифікації за 15

характеристиками (симптомами) при наявних 322 можливих характеристики. Треба помітити що при реальному інтерв'ю точність повинна бути більше, так як там буде застосовуватися система визначення важливих характеристик, а не простий випадковий вибір як при практичних дослідженнях.

Далі були проведені дослідження про результати класифікації за допомогою коефіцієнта Жаккара при наявності помилок у тестових даних (рисунок 3.3).

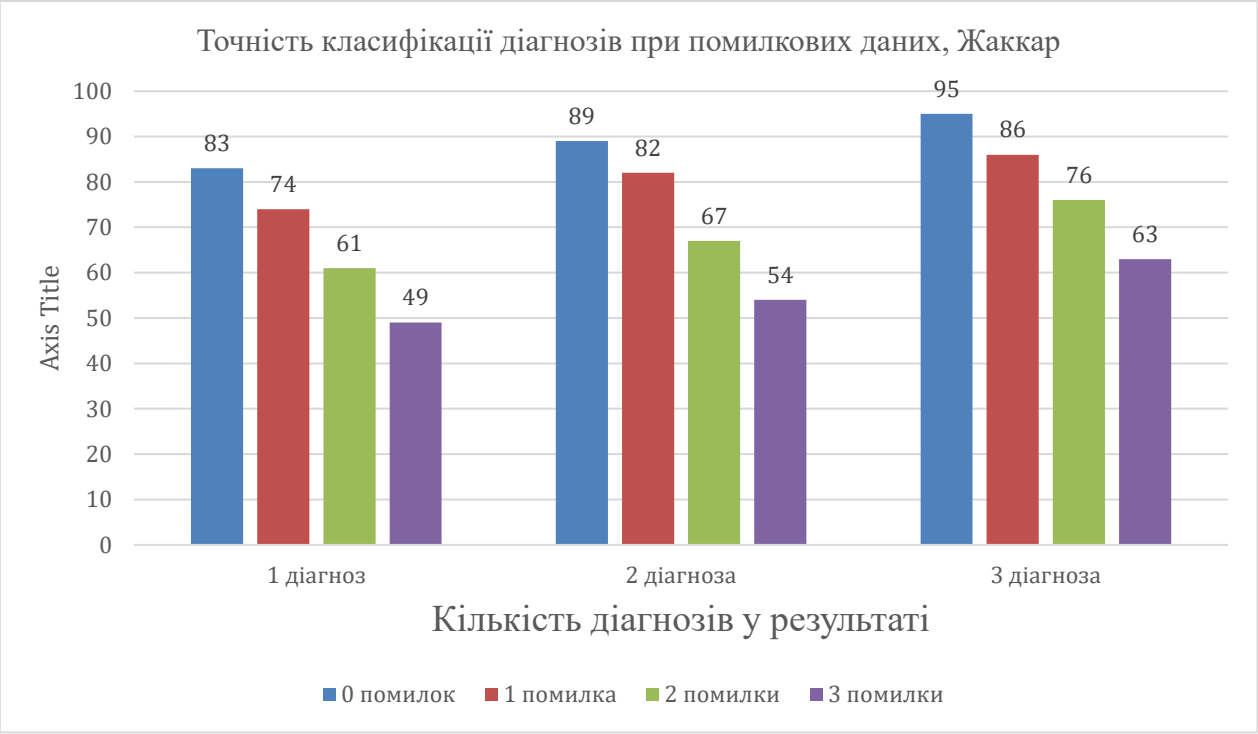

Рисунок 3.3 – Гістограма точності класифікації за допомогою коефіцієнта Жаккара, помилкові тестові дані

Як можна побачити за результатами дослідження, точність класифікації за цим методом швидко падає при збільшенні кількості помилок у введених(тестових) даних. Так як така ситуація буде часто повторюватись у реальній системі, був зроблений висновок що

використання міри подібності Жаккара, принаймі у такому вигляді не є підходящим варіантом для розроблюваної системи.

Далі були проведені практичні дослідження результатів роботи алгоритму класифікації на основі байэсовської мережі побудованої із зв’язків Симптом -Хвороба при відсутності шумів (рисунок 3.4).

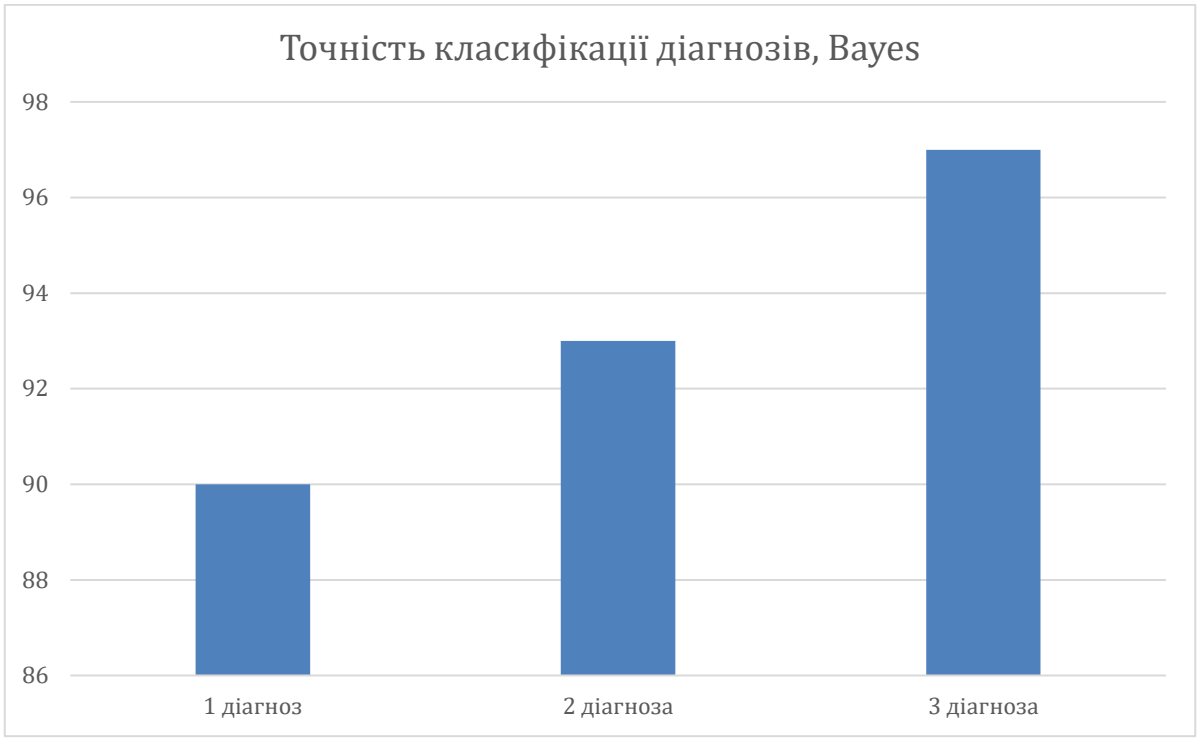

Рисунок 3.4 – Гістограма точності класифікації за допомогою байєсовської мережі, коректні тестові дані

Як можна побачити показники даного алгоритму для тестових даних без помилок знаходяться на рівні з попереднім варіантом використання коефіцієнта Жаккара, але в варіантах з меншою кількістю діагнозів в результаті показники декілька вище, що говорить о том, що даний алгоритм краще діференціює між хворобами з подібним набором симптомів.

Далі були проведені дослідження про результати класифікації за наявності помилок у тестових даних (рисунок 3.5).

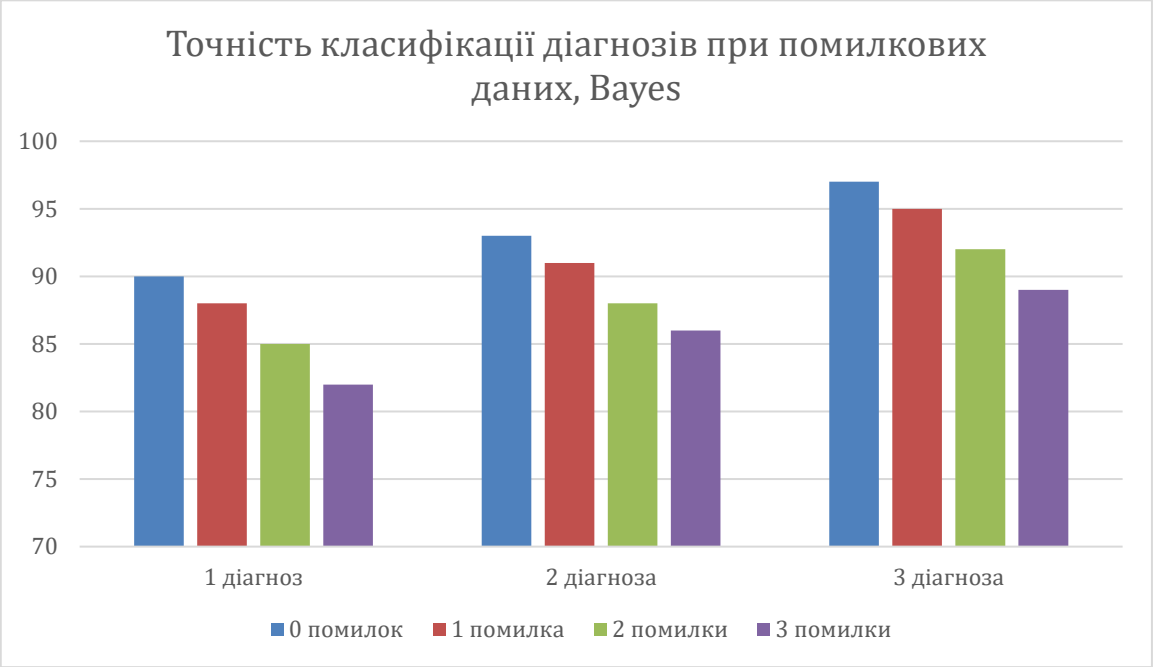

Рисунок 3.5 – Гістограма точності класифікації за допомогою байєсовської мережі, помилкові тестові дані

Як можна побачити із результатів даний підхід не демонструє значне зниження точності при збільшенні кількості помилок у введених даних. При роботі реальній системи ці показники будуть ще більш вищими, так як буде використовуватися система підбору важливих характеристик, для яких значення будуть вищими і в яких у користувача менше вірогідність звершити помилку.

Далі були проведені практичні дослідження результатів роботи алгоритму класифікації на основі косинусної міри подібності, що використовує вектори значень tf-idf для симптомів (рисунок 3.6).

Як можна побачити показники даного алгоритму для тестових даних без помилок знаходяться на рівні з попереднім варіантом використання коефіцієнта Жаккара, але в варіантах з меншою кількістю діагнозів в результаті показники декілька вище, як і у байєсовської мережі, що також говорить о том, що даний алгоритм краще діференціює між хворобами з подібним набором симптомів.

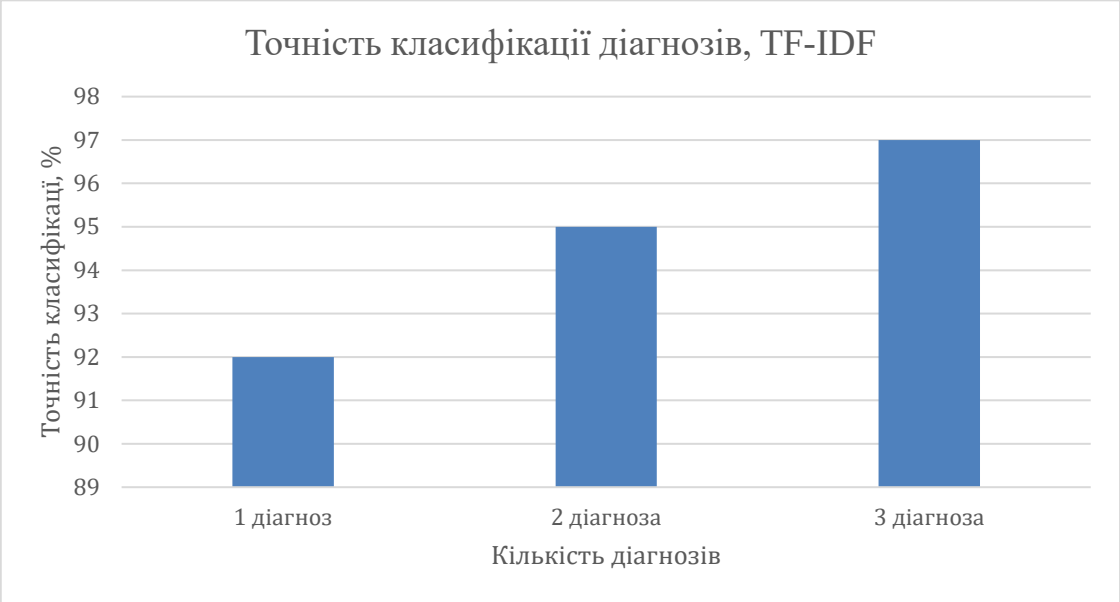

Рисунок 3.6 – Гістограма точності класифікації за допомогою показника TF-IDF, коректні тестові дані

Далі були проведені дослідження про результати класифікації за наявності помилок у тестових даних (рисунок 3.7).

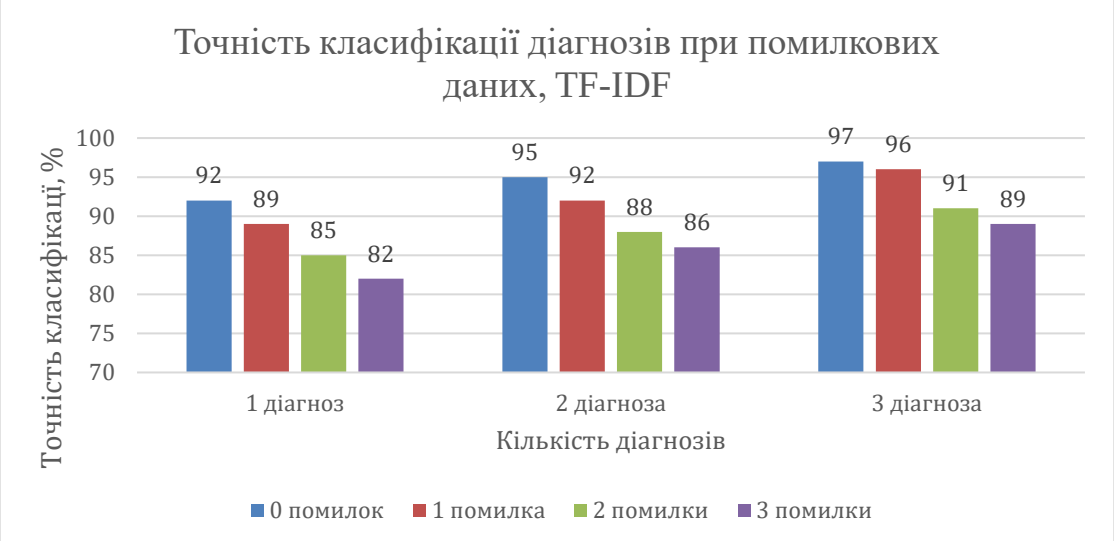

Рисунок 3.7 – Гістограма точності класифікації за допомогою показника TF-IDF, помилкові тестові дані

Як можна побачити із результатів даний підхід не демонструє значне зниження точності при збільшенні кількості помилок у введених даних. При роботі реальній системи ці показники будуть ще більш вищими, так як буде використовуватися система підбору важливих характеристик, для яких значення будуть вищими і в яких у користувача менше вірогідність звершити помилку.

Далі були проведені практичні дослідження результатів роботи алгоритму класифікації побудованому на основі ансамбля двух методів – байесу та TF-IDF (рисунок 3.8).

Завдяки комбінації двух методів класифікації хвороб у представлених умовах, які показали покращення середньої точності класифікації.

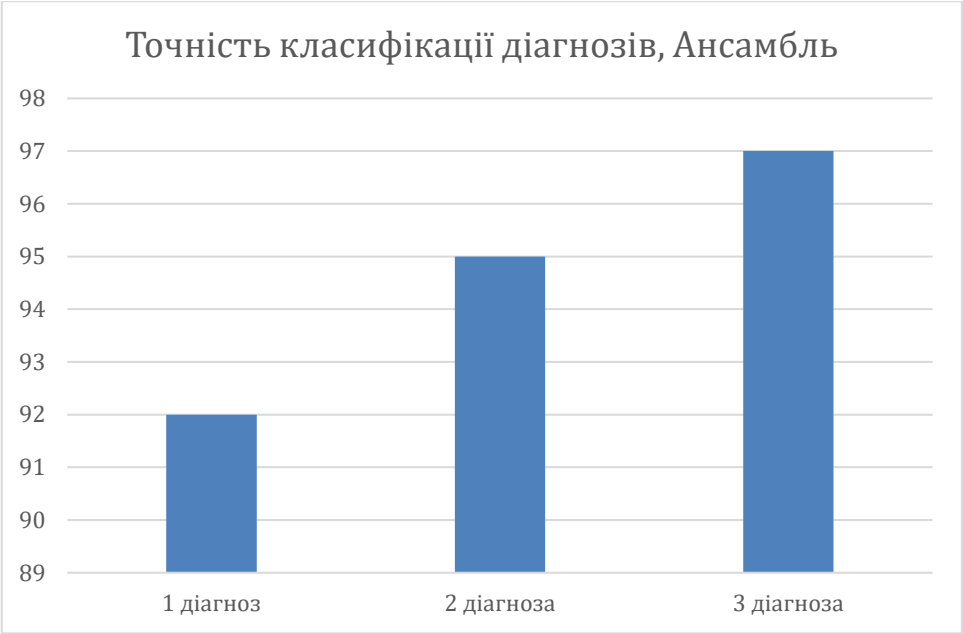

Рисунок 3.8 – Гістограма точності класифікації за допомогою алгоритма з ансамбля, коректні тестові дані

Далі були проведені дослідження про результати класифікації за наявності помилок у тестових даних (рисунок 3.9).

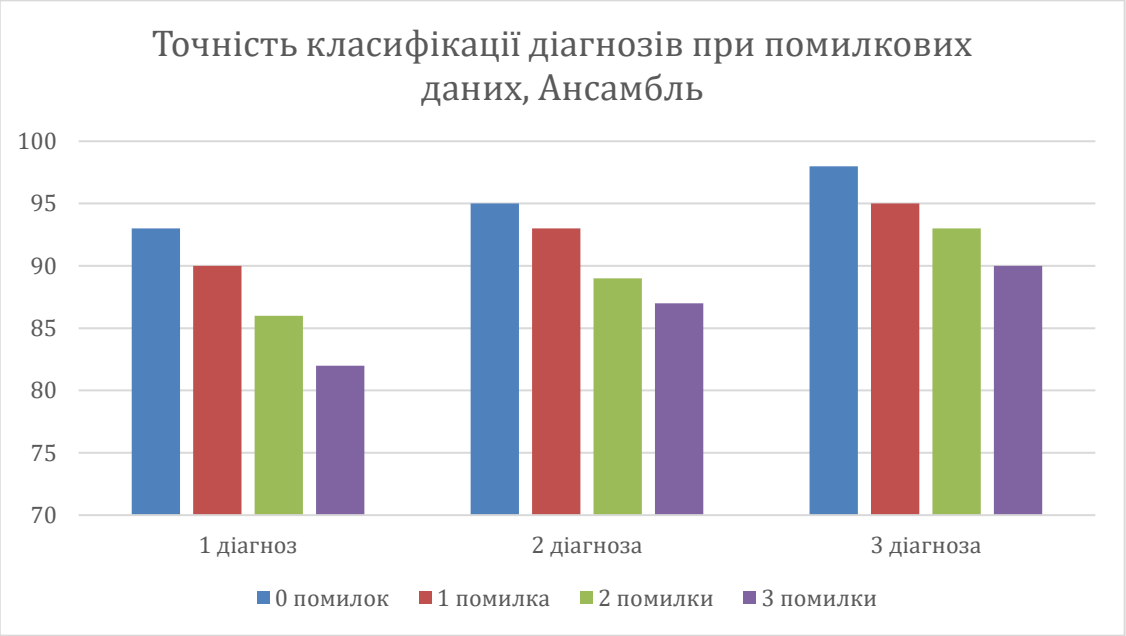

Рисунок 3.9 – Гістограма точності класифікації за допомогою показника TF-IDF, помилкові тестові дані

Як можна побачити із результатів даний підхід завдяки комбінації двух найкращих методів класифікації хвороб у представлених умовах призвів до покращення середньої точності класифікації порівняно з оригінальними методами путем покращення стабільності класифікації та зменшення кількісті вибросів.

Також, порівнюючи заявлену точність аналогічних систем, можно побачити що розроблений алгоритм може конкурувати з комерційними системами, які використовують байесову логіку и дані зібрані з медичних карток хворих (рисунок 3.10). Був зроблений висновок що даний алгоритм удовлетворяє вимоги поставленої задачі.

Рисунок 3.10 – Гістограма середньої точності класифікації розробленого алгоритму з існуючими системами

3.5 Дослідження методів визначення важливих для ходу інтерв'ю симптомів

При отриманні даних від користувача, експертна система повинна задати якомога менше запитань і уникати питань, відповіді яких вимагатимуть важкого або тривалого дослідження користувача. Існує два способи встановлення стратегії, яка буде використовуватися експертною системою для вибору запитань: програміст, який розробляє систему, може запрограмувати стратегію явно або стратегія може автоматично визначатися системою виконання формалізму в якій експертна система реалізована. Явно запрограмований контроль над введенням / виведенням вимагає імперативного формалізму – наприклад, імперативних систем на основі правил. У чисто декларативних формалізмах, таких як декларативні системи, засновані на правилах, генерація питання повинна виконуватися

автоматично системою під час виконання формалізму. Обмеження є декларативними конструкціями і, отже, не можуть використовуватися програмістом експертної системи для вказівки вводу / виводу. Замість цього, в експертних системах, заснованих на обмеженнях, генерація питання повинна виконуватися автоматично за допомогою спеціальних алгоритмів.

Для експертної системи, що базується на декларативних правилах, досить легко вирішити, які питання задати користувачеві. Гіпотези системи містять коріння набору переплетених дерев, листя яких являють собою можливі точки даних; при розгляді гіпотези, система розгортає ланцюжок від кореня до її листя, задаючи питання про них.

Вирішення, які питання задавати, є більш складним для експертної системи на основі обмежень. Для поточної розроблюваної системи був розроблений спеціальний підхід, який генерує питання про наявність симптомів у користувача грунтуючись на попередніх відповідях.

Інтерв'ю складається з трьох стадій:

– введення початкових симптомів. На початку користувач просто шукає симптоми, на які він скаржиться, в списку, що надається системою і відзначає їх;

– пропозиція часто зустрічаються симптомів. На цьому етапі система пропонує користувачу список з декількох симптомів, які найчастіше зустрічаються з тими, які він ввів на початку;

– опитування по черзі. Це основний етап інтерв'ю в якому користувачеві по черзі задається питання про наявність симптомів. Вибір цих симптомів здійснюється на основі попереднього введення за допомогою спеціально розробленого евристичного алгоритму.

Суть алгоритму основної частини інтерв'ю полягає в наступному. На вхід системи передається список симптомів. Використовуючи метод випадкових підмножин система робить з цього списку симптомів кілька

пересічних підмножин. Далі проводиться пошук найбільш часто зустрічаються симптомів разом з цими підмножинами симптомів. Це проводиться для кожної підмножини. На виході з цього виходить список симптомів. Далі для кожного з цих симптомів розраховується коефіцієнт вигоди за формулою (формула 3.7):

$$Q_i = \frac{N^+}{N^-}, \tag{3.7}$$

де N^+ – це кількість хвороб у яких спостерігається цей симптом;

N^- – то кількість хвороб у яких відсутній цей симптом.

Далі симптоми сортуються відповідно отриманими значеннями, і симптом з найменшим значенням коефіцієнта вибирається для наступного питання.

Даний підхід показує себе набагато краще ніж підхід у систем на основі правил, де питання йдуть по заздалегідь побудованому дереву. На відміну від цього підходу, підхід розроблений в даній системі є стійким до помилкового введення симптомів користувачем.

ВИСНОВКИ

Величезна кількість експертних систем – медична. Головною метою будь-якої медичної експертної системи є виявлення та лікування захворювань. Медична експертна система складається з програм і бази медичних знань. Інформація, отримана з медичної експертної системи, подібна до інформації, наданої фахівцями в цій галузі.

Було розроблено програмне забезпечення діагностики та моніторингу стану здоров'я для поширених захворювань у всіх частинах людського тіла. Програмне забезпечення буде дуже хорошим керівництвом для осіб та медичного персоналу у віддалених районах, де медичний персонал є неадекватним. Це також зменшить навантаження на лікарні, лікарів і медсестер, що особливо важливо в період оверзавантаження медичних закладів та робітників сфері лікування. Розроблене програмне забезпечення пропонує первинну медичну діагностику хвороб практично без вартості пацієнтам. Програмне забезпечення ще більше зменшить зловживання наркотиками та покращить якість надання медичних послуг та зменшить смертність у деяких ситуаціях.

Розроблена система дозволяє дуже легко і доступно визначати можливі хвороби користувачів, просто поставивши кілька запитань щодо присутності певних симптомів у користувача. Завдяки використанню технологій машинного навчання та впровадження штучного інтелекту в механізм прийняття рішень вийшло домогтися високої швидкості і точності системи навіть при наявності помилкових відповідей користувача. Розроблена система була виконана у вигляді сервісу, що дозволяє реалізовувати клієнтське програмне забезпечення на будь-якій платформі, з якої можливий доступ в інтернет, не обмежуючись лише однією. Створена API дозволяє створювати індивідуальні діагностичні рішення з нуля. Як демонстрація цього функціоналу був розроблений мобільний клієнт

Medicare, що надає доступ до цього сервісу через простий і зручний інтерфейс. Використовуючи штучний інтелект для перевірки симптомів, цей клієнт потім направляє пацієнтів до відповідних медичних служб. Це дозволить користувачам легко і просто перевірити стан їх здоров'я, без необхідності записуватися на прийом до лікаря з будь-якого приводу.

Також, основною перевагою розробленої системи є те, що для її роботи не потрібен збір великої кількості медичних даних хворих ті їх історії хвороб, що є дуже затратною задачею. Ця система показує високу точність при використанні для роботи даних, зібраних з відкритого для всіх історії хвороб, що є дуже затратною задачею. Ця система показує високу точність при використанні для роботи даних, зібраних з відкритого для всіх джерелу – медичних публікацій PubMed, що нараховує більш семи мільйонів публікацій за останні 60 років. Саме це дозволяє системі демонструвати такі високі показники точності.

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ

1. Vikas T. T. Design and implementation of an efficient relative model in cancer disease recognition. Toronto: IJARCSSE, 2013. 10 p.

2. Peleg S. T. Decision Support, Knowledge Representation and Management. Toronto: IMIA, 2006. 15 p.

3. Khaleel M. A. A Survey of Data Mining Techniques on Medical Data for Finding frequent diseases. Toronto: IJARCSSE, 2013. 20 p.

4. Vembandasamy S. D. Heart Diseases Detection Using Naive Bayes Algorithm. Deli: IJISET, 2015. 20 p.

5. Chaurasia.V. P. Data Mining Approach to Detect Heart Disease. New York: IJACSIT, 2013. 15 p.

6. Parthiban G. S. Applying Machine Learning Methods in Diagnosing Heart Disease for Diabetic Patients. Ontario: IJAIS, 2015. 10 p.

7. Deep learning for detection of diabetic eye disease. URL: https://research.googleblog.com/2016/11/deep-learningfor-detection-of-diabetic.html (дата звернення: 17.09.2020).

8. Goodfellow I., Bengio Y., Courville A. Deep Learning. New York: The MIT Press, 2016. 10 p.

9. Bach S. H., He B. D., Ratner A., Re C. Learning the structure of generative models without labeled data. Deli: ICML, 2017. 273 p.

10. Lowe H.J., Barnett G.O. Understanding the medical subject headings (MeSH) vocabulary to perform literature searches. Deli: JAMA, 1994. 20 p.

11. Wheeler D.L. Database resources of the National Center for Biotechnology Information. New York: Nucleic Acids Res 35, 2007. 30 p.

12. Dee C.R. The development of the Medical Literature Analysis and Retrieval System (MEDLARS). Beijing: J Med Libr Assoc, 2007. 25 p.

13. Perez-Iratxeta C., Bork P., Andrade M.A. Association of genes to genetically inherited diseases using data mining. Deli: Nat Genet, 2002. 20 p.

14. Bhattacharya S., Ha-Thuc V., Srinivasan P. MeSH: a window into full text for document summarization. London: Bioinformatics, 2011. 32 p.

15. Swanson D.R., Smalheiser N.R., Torvik V.I. Ranking indirect connections in literature-based discovery: The role of medical subject headings. *Journal of the American Society for Information Science and Technology*. 2006. No. 57. P. 42-49.

16. Jensen L.J., Saric J., Bork P. Literature mining for the biologist: from information retrieval to biological discovery. New York: Nat Rev Genet, 2006. 15c.

17. NLM. The download page of 2011 MeSH. URL: http://www.nlm.nih.gov/mesh/2011/download/termscon.html (дата звернення: 17.09.2020).

18. Jensen, L. J., Saric, J., Bork P. Information retrieval for biological discovery. New York: Nat Rev Genet, 2006. 20 p.

19. Cote R.A., Robboy S. Progress in medical information management. Systematized nomenclature of medicine (SNOMED). Deli: JAMA, 1980. 22 p.

20. Cimino J.J. High-quality, standard, controlled healthcare terminologies come of age. Toronto: Methods Inf Med, 2011. 65 p.

21. Bisdorff A. Migraine and dizziness. *Curr Opin Neurol*. 2014. No. 27, P. 105-109.

22. Roger P. A., David A. S., Michael J. G. Clinical neurology, New York: Lange Medical Books/McGraw-Hill, 2009. 73 p.

23. NCHS. Vol. 2014 ICD-9-CM home page. URL: http://www.cdc.gov/nchs/icd/icd9cm.htm (дата звернення: 17.09.2020).

24. Schriml L.M. Disease Ontology: a backbone for disease semantic integration. New York: Nucleic Acids Res, 2012. 44 p.

25. Baclawski K., Matheus C.J., Kokar M.M., Letkowski J., Kogut P.A. Towards a Symptom Ontology for Semantic Web Applications. New York: McGraw-Hill, 2009. 73 p.