Хмельницький національний університет
Факультет інформаційних технологій
Кафедра комп'ютерних наук

КВАЛІФІКАЦІЙНА РОБОТА

на тему *Метод діагностики захворювань за описом симптомів з використанням обробки природної мови*

Рівень вищої освіти *другий (магістерський)*

Галузь знань *12 – Інформаційні технології*
Шифр і найменування

Спеціальність *122 – Комп'ютерні науки*
Код і найменування

Освітня програма *Комп'ютерні науки*
Назва

Виконав: *студент 2 курсу, група КНм-24-1* — Курс, група виконавця; Підпис; *Олександр БОНДАР* — Ім'я, ПРІЗВИЩЕ

Керівник: *к.т.н., доцент кафедри КН* — Науковий ступінь, посада; Підпис; *Олександр ПАСІЧНИК* — Ім'я, ПРІЗВИЩЕ

Нормоконтроль: *к.т.н., доцент кафедри КН* — Науковий ступінь, посада; Підпис; *Руслан БАГРІЙ* — Ім'я, ПРІЗВИЩЕ

До захисту допускаю:

Зав. кафедри КН, д.т.н., професор — Підпис; *Олександр БАРМАК* — Ім'я, ПРІЗВИЩЕ

15 грудня 2025 р.

Хмельницький 2025

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

ЗАТВЕРДЖУЮ
Завідувач кафедри *комп'ютерних наук*
_____
(підпис)
д.т.н., професор Олександр БАРМАК
«*28*» *08* *2025* року

ЗАВДАННЯ
НА КВАЛІФІКАЦІЙНУ РОБОТУ

1. Тема кваліфікаційної роботи: *«Метод діагностики захворювань за описом симптомів з використанням обробки природної мови»*

2. Завдання видано студенту *Олександру БОНДАРУ*
(Ім'я, ПРІЗВИЩЕ)

3. Керівник роботи *доцент кафедри КН Олександр ПАСІЧНИК*
(Ім'я, ПРІЗВИЩЕ)

4. Затверджені наказом університету від «*25*» *08* *2025* р. № *65*

5. Дата видачі завдання студенту: «*28*» *08* *2025* р.

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

Мета роботи полягає у підвищенні точності та швидкості діагностики захворювань на основі текстових описів симптомів шляхом розробки методу з використанням рекурентних нейронних мереж та механізмів уваги для автоматизованого аналізу описових медичних текстів. Задачі: проаналізувати сучасні технології та підходи до діагностування захворювань на основі текстової інформації; спроектувати метод класифікації симптомів із використанням рекурентних нейронних мереж типу LSTM та механізмів уваги; реалізувати модуль попередньої обробки текстових даних з урахуванням специфіки медичної термінології; виконати експериментальну перевірку запропонованого рішення.

7. Календарний план виконання кваліфікаційної роботи:

№	Назва етапів (розділів) кваліфікаційної роботи магістра	Термін виконання	Примітка
1	*Вибір напряму дослідження та узгодження теми кваліфікаційної роботи з керівником, складання календарного графіка виконання роботи*	*вересень 2025*	*Виконано*
2	*Ознайомлення з предметною областю, аналіз існуючих методів і моделей, формулювання мети та завдань дослідження, визначення об'єкта й предмета дослідження*	*вересень 2025*	*Виконано*
3	*Розробка методу чи моделі для вирішення обраного завдання, опис архітектури рішення*	*жовтень 2025*	*Виконано*
4	*Програмна реалізація методу чи моделі*	*жовтень 2025*	*Виконано*
5	*Дослідження ефективності та експериментальна перевірка результатів, порівняння з відомими підходами*	*листопад 2025*	*Виконано*
6	*Написання пояснювальної записки, оформлення відповідно до вимог, врахування зауважень керівника*	*листопад 2025*	*Виконано*
7	*Підготовка презентаційних матеріалів та попередній захист*	*листопад 2025*	*Виконано*
8	*Перевірка пояснювальної записки на відповідність вимогам оформлення (нормоконтроль) та перевірка на академічну доброчесність. Отримання відгуку керівника та рецензії.*	*грудень 2025*	*Виконано*
9	*Публічний захист кваліфікаційної роботи*	*грудень 2025*	*Виконано*

Виконавець: *студент групи КНм-24-1* _____ *Олександр БОНДАР*
Група виконавця | Підпис | Ім'я, ПРІЗВИЩЕ

Керівник: *к.т.н., доцент каф. КН* _____ *Олександр ПАСІЧНИК*
Науковий ступінь, посада | Підпис | Ім'я, ПРІЗВИЩЕ

Реферат

Кваліфікаційна робота присвячена розробці методу автоматизованої діагностики захворювань на основі аналізу текстових описів симптомів із використанням сучасних технологій обробки природної мови.

Актуальність теми. Своєчасна та точна діагностика захворювань залишається однією з найважливіших проблем сучасної медицини. Традиційні підходи до первинної діагностики потребують значних часових витрат та високої кваліфікації медичного персоналу. Водночас спостерігається зростання навантаження на систему охорони здоров'я, що ускладнює доступ пацієнтів до своєчасної медичної допомоги. Розвиток технологій штучного інтелекту, а саме методів глибокого навчання в тому числі обробки природної мови, відкриває нові можливості для створення інтелектуальних систем підтримки прийняття медичних рішень. Пацієнти часто описують свій стан неструктурованою природною мовою, використовуючи різноманітні формулювання, медичну та розмовну термінологію. Автоматизований аналіз таких описів дозволить прискорити процес первинної діагностики, зменшити ймовірність людських помилок та забезпечити доступ до медичних консультацій у віддалених регіонах через телемедичні платформи. Актуальність дослідження підтверджується необхідністю створення надійних методів інтерпретації текстової медичної інформації, які можуть стати інструментом підтримки лікарів у клінічній практиці.

Мета роботи полягає у підвищенні точності та швидкості діагностики захворювань на основі текстових описів симптомів шляхом розробки методу з використанням рекурентних нейронних мереж та механізмів уваги для автоматизованого аналізу описових медичних текстів.

Задачі:

– провести аналіз існуючих методів та підходів до автоматизованої діагностики захворювань на основі текстових даних з використанням технологій обробки природної мови та машинного навчання;

– розробити метод класифікації захворювань за описом симптомів на основі архітектури рекурентних нейронних мереж LSTM з механізмом уваги, що забезпечує виділення найбільш інформативних фрагментів у медичних текстах;

– спроєктувати програмну реалізацію метода попередньої обробки текстових даних та обробку заперечень для коректного врахування відсутності симптомів;

– провести експериментальне дослідження запропонованого методу шляхом порівняння базової та модифікованої архітектури нейронної мережі.

Об'єкт дослідження – процес автоматизованої діагностики захворювань на основі текстового опису клінічних симптомів.

Предмет дослідження – методи, моделі та засоби обробки природної мови для виявлення закономірностей у текстових описах симптомів та їх класифікації за типами захворювань.

Методи дослідження. У роботі застосовано методи глибокого навчання, рекурентні нейронні мережі LSTM, механізми уваги для виділення найбільш інформативних фрагментів тексту, техніки попередньої обробки текстових даних, методи регуляризації для запобігання перенавчанню.

Наукова новизна одержаних результатів. Удосконалено метод діагностики захворювань за текстовими описами симптомів, який відрізняється від наявних комбінованим використанням двонаправленої LSTM-архітектури з механізмом уваги та додатковим шаром максимального пулінгу, що дозволяє одночасно враховувати як контекстуальну важливість окремих симптомів, так і найбільш виражені прояви захворювання, забезпечуючи підвищення точності класифікації до 6% порівняно з базовою архітектурою.

Апробація результатів кваліфікаційної роботи магістра та публікації.

Бондар О.А., Пасічник О.А., Скрипник Т.К. Метод діагностики захворювань за описом симптомів на основі рекурентних нейронних мереж. Збірник наукових праць за матеріалами XVII Всеукраїнської науково-практичної конференції «Актуальні проблеми комп’ютерних наук АПКН-2025». - Хмельницький, 2025. - С.39 – 41.

Структура та обсяг роботи. Робота складається зі вступу, чотирьох розділів, висновків, переліку використаних джерел та додатків. Загальний обсяг основного тексту становить 77 сторінок, включаючи 19 рисунків, 1 таблицю та список літератури з 47 джерел.

Ключові слова: діагностика захворювань, обробка природної мови, LSTM, механізм уваги, класифікація симптомів, глибоке навчання, медичні текстові дані, нейронні мережі, машинне навчання.

Зміст

Перелік скорочень

Скорочення, термін, позначення	**Пояснення**
NLP	Natural Language Processing (Обробка природної мови)
LSTM	Long Short-Term Memory (Довга короткочасна пам'ять)
CNN	Convolutional Neural Network (Згорткова нейронна мережа)
RNN	Recurrent Neural Network (Рекурентна нейронна мережа)
ML	Machine Learning (Машинне навчання)
DL	Deep Learning (Глибоке навчання)
TP	True Positive (Істинно позитивний результат)
TN	True Negative (Істинно негативний результат)
FP	False Positive (Хибно позитивний результат)
FN	False Negative (Хибно негативний результат)
F1	F1-score (F1-міра)

Вступ

Кваліфікаційна робота присвячена розробці методу автоматизованої діагностики захворювань на основі аналізу текстових описів симптомів із використанням сучасних технологій обробки природної мови.

Актуальність теми. Своєчасна та точна діагностика захворювань залишається однією з найважливіших проблем сучасної медицини. Традиційні підходи до первинної діагностики потребують значних часових витрат та високої кваліфікації медичного персоналу. Водночас спостерігається зростання навантаження на систему охорони здоров'я, що ускладнює доступ пацієнтів до своєчасної медичної допомоги. Розвиток технологій штучного інтелекту, зокрема методів глибокого навчання в тому числі обробки природної мови, відкриває нові можливості для створення інтелектуальних систем підтримки прийняття медичних рішень. Пацієнти часто описують свій стан неструктурованою природною мовою, використовуючи різноманітні формулювання, медичну та розмовну термінологію. Автоматизований аналіз таких описів дозволить прискорити процес первинної діагностики, зменшити ймовірність людських помилок та забезпечити доступ до медичних консультацій у віддалених регіонах через телемедичні платформи. Актуальність дослідження підтверджується необхідністю створення надійних методів інтерпретації текстової медичної інформації, які можуть стати інструментом підтримки лікарів у клінічній практиці.

Мета роботи полягає у підвищенні точності та швидкості діагностики захворювань на основі текстових описів симптомів шляхом розробки методу з використанням рекурентних нейронних мереж та механізмів уваги для автоматизованого аналізу описових медичних текстів.

Задачі:

– провести аналіз існуючих методів та підходів до автоматизованої діагностики захворювань на основі текстових даних з використанням технологій обробки природної мови та машинного навчання;

– розробити метод класифікації захворювань за описом симптомів на основі архітектури рекурентних нейронних мереж LSTM з механізмом уваги, що забезпечує виділення найбільш інформативних фрагментів у медичних текстах;

– спроєктувати та реалізувати програмну реалізацію метода попередньої обробки текстових даних та обробку заперечень для коректного врахування відсутності симптомів;

– провести експериментальне дослідження запропонованого методу шляхом порівняння базової та модифікованої архітектури нейронної мережі.

Об'єкт дослідження – процес автоматизованої діагностики захворювань на основі текстового опису клінічних симптомів.

Предмет дослідження – методи, моделі та засоби обробки природної мови для виявлення закономірностей у текстових описах симптомів та їх класифікації за типами захворювань.

Методи дослідження. У роботі застосовано методи глибокого навчання, зокрема рекурентні нейронні мережі LSTM, механізми уваги для виділення найбільш інформативних фрагментів тексту, техніки попередньої обробки текстових даних, методи регуляризації для запобігання перенавчанню.

Наукова новизна одержаних результатів. Удосконалено метод діагностики захворювань за текстовими описами симптомів, який відрізняється від наявних комбінованим використанням двонаправленої LSTM-архітектури з механізмом уваги та додатковим шаром максимального пулінгу, що дозволяє одночасно враховувати як контекстуальну важливість окремих симптомів, так і найбільш виражені прояви захворювання, забезпечуючи підвищення точності класифікації до 6% порівняно з базовою архітектурою.

Апробація результатів кваліфікаційної роботи магістра та публікації.

Бондар О.А., Пасічник О.А., Скрипник Т.К. Метод діагностики захворювань за описом симптомів на основі рекурентних нейронних мереж. Збірник наукових праць за матеріалами XVII Всеукраїнської науково-практичної конференції

«Актуальні проблеми комп’ютерних наук АПКН-2025». - Хмельницький, 2025. - С.39 – 41.

Структура та обсяг роботи. Робота складається зі вступу, чотирьох розділів, висновків, переліку використаних джерел та додатків. Загальний обсяг основного тексту становить 77 сторінок, включаючи 19 рисунків, 1 таблицю та список літератури з 47 джерел.

Розділ 1 Аналіз та огляд методів діагностики захворювань за описом симптомів

1.1 Характеристика задачі діагностики захворювань за описом симптомів

Задача діагностики захворювань за описом симптомів з використанням обробки природної мови представляє собою важливий напрямок у сфері медичної інформатики. Основна мета такого підходу полягає в автоматизованому тлумаченні текстових описів симптомів, які надають пацієнти, з подальшим встановленням попереднього діагнозу [1]. Цей процес охоплює аналіз неструктурованих даних, таких як вільні текстові описи, які можуть надходити через різноманітні канали комунікації. Серед таких каналів виділяються інтерфейси чат-ботів, мобільні застосунки для охорони здоров'я, а також електронні медичні карти пацієнтів [2, 3].

Автоматизація первинного аналізу симптомів дозволяє підвищити точність попередньої діагностики та суттєво скоротити час, необхідний для встановлення діагнозу. Особливо актуальним це є для регіонів з обмеженим доступом до кваліфікованої медичної допомоги, де системи автоматичної діагностики можуть частково компенсувати дефіцит спеціалістів [4]. Крім того, такі системи здатні знизити навантаження на медичний персонал, дозволяючи лікарям зосередитися на більш складних випадках, які потребують глибокого професійного аналізу.

Складність реалізації цієї задачі обумовлена значною лінгвістичною варіативністю, з якою стикаються розробники систем. Пацієнти часто описують свій стан, використовуючи синоніми медичних термінів, діалектні вирази, розмовний сленг або неповні формулювання симптомів [5, 6]. Ця варіативність вимагає застосування технік нормалізації тексту, які включають токенізацію для розділення тексту на складові, стемінг для виділення основи слів, лематизацію для того, щоб слова привести до базової форми, а також видалення шумових елементів, які не несуть смислового навантаження.

Розглянемо конкретний приклад для ілюстрації складності задачі. Опис симптомів "сильний головний біль з нудотою та запамороченням" може свідчити про

різні захворювання залежно від контексту. Такі симптоми характерні для мігрені, але можуть також вказувати на гіпертензивний криз або навіть неврологічні ускладнення більш серйозного характеру. Система повинна враховувати семантичний контекст опису, супутні симптоми, тривалість їх прояву та інші фактори для уникнення помилкових висновків.

У контексті предметної області медичної діагностики ця задача тісно інтегрується з системами підтримки прийняття клінічних рішень. Такі системи класифікують симптоми за різними категоріями, від гострих станів, які потребують невідкладної допомоги, до хронічних захворювань з тривалим перебігом або рідкісних патологій [7]. При цьому враховуються додаткові фактори, такі як демографічні дані пацієнта, включаючи вік і стать, тривалість проявів симптомів, інтенсивність їх вираження, а також супутні умови та захворювання в анамнезі.

Для забезпечення узгодженості діагнозів системи часто використовують стандартизовані медичні онтології. Серед найпоширеніших онтологій виділяються SNOMED CT, яка містить детальну систематичну номенклатуру медицини, та міжнародну класифікацію хвороб ICD-10 [8–10].

Серед ключових викликів, які постають перед розробниками таких систем, особливо виділяється проблема полісемії медичних термінів. Одне й те саме слово може мати різні значення залежно від контексту використання. Наприклад, слово "біль" може стосуватися різних локалізацій в організмі, мати різну інтенсивність та характер прояву [11, 12]. Для адекватної інтерпретації таких термінів необхідний глибокий контекстуальний аналіз за допомогою алгоритмів семантичного парсингу, які здатні врахувати всі нюанси опису.

Суттєвим обмеженням систем діагностики є їхня залежність від якості вхідних даних. Суб'єктивні або неточні описи симптомів, які надають пацієнти, можуть призводити до хибних висновків системи [13]. Методи обробки природної мови в цій задачі акцентують увагу на витягуванні медичних сутностей з тексту. Розпізнавання іменованих сутностей дозволяє ідентифікувати в тексті назви симптомів, таких як "лихоманка" або "кашель", а також встановлювати відносини між ними [14]. Для

цього застосовуються різноманітні алгоритми, від простіших підходів на базі правил до складніших методів машинного навчання та глибоких нейронних мереж, які можуть добре навчатися на великих медичних даних.

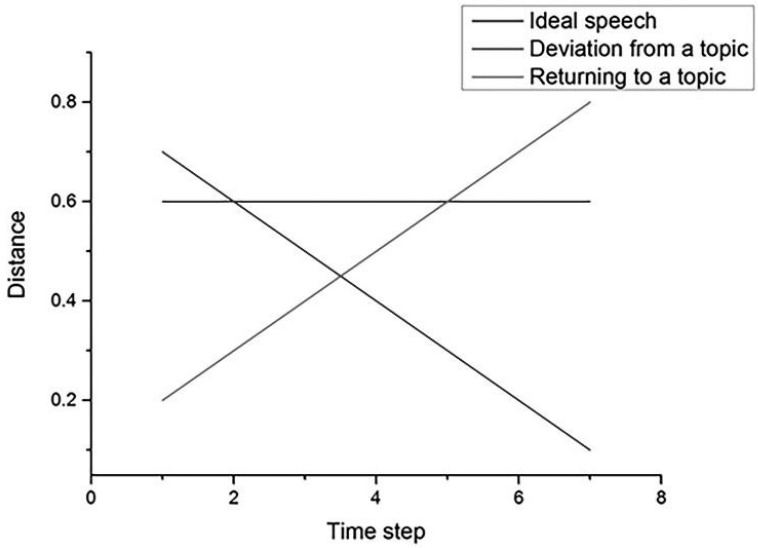


Рисунок 1.1 – Застосування для оцінки трьох типів мовних зразків: правильне мовлення, відхід від теми та часткове дотримання теми [14]

Параметри систем діагностики охоплюють широкий спектр характеристик. Розмір словникового запасу може варіюватися від десяти тисяч до одного мільйона токенів залежно від спеціалізації системи та обсягу медичної термінології, яку вона покриває [15]. Глибина архітектур нейронних мереж зазвичай становить від чотирьох до двадцяти чотирьох шарів для трансформерних моделей, що дозволяє захоплювати складні залежності в текстах. Розмір векторних представлень слів зазвичай коливається від 768 до 1024 розмірностей.

Оцінка ефективності систем діагностики здійснюється за допомогою типових метрик машинного навчання. Точність визначається як частка вірних передбачень серед усіх зроблених передбачень. Прецизійність важлива для мінімізації помилкових позитивних діагнозів, коли система помилково діагностує захворювання, якого у пацієнта немає [16]. Повнота характеризує здатність системи охопити всі релевантні випадки захворювань. F1-score, як гармонійне середнє між прецизійністю та

повнотою, є критичною метрикою для незбалансованих даних, де деякі захворювання зустрічаються значно частіше за інші.

У реалізаціях на сучасних платформах, таких як Hugging Face, моделі на основі трансформерів демонструють високу ефективність. Модель для витягування симптомів з клінічних нотаток може досягати F1-score більше 90%. Однак для досягнення таких результатів необхідне тонке налаштування моделі на великій кількості епох навчання, зазвичай від 500 до 2000, з відповідно підібраною швидкістю навчання та розміром партії даних. Обмеження існуючих підходів включають ризик перенавчання на специфічних доменах. Модель може демонструвати високу ефективність на тренувальних даних, але виявлятися слабкою на нових даних через недостатню різноманітність навчальної вибірки [17]. Крім того, системи часто стикаються з проблемами при обробці рідкісних або атипових симптомів, які недостатньо представлені в тренувальних даних.

На практиці розв'язання задачі структуровано за декількома послідовними етапами. Перший етап включає попередню обробку тексту, яка робиться за допомогою інструментів для токенізації та нормалізації. Наступний етап передбачає витягування ознак з тексту, де використовуються різні підходи від класичних статистичних методів до сучасних контекстних ембеддінгів [18, 19]. Заключний етап включає власне класифікацію симптомів та встановлення діагнозу за допомогою алгоритмів машинного навчання. Параметри алгоритмів класифікації потребують ретельного налаштування. Для випадкового лісу необхідно визначити максимальну глибину дерев, кількість оцінювачів та критерій розділення вузлів. Для нейронних мереж критично важливими є швидкість навчання, коефіцієнт відкидання для запобігання перенавчанню, а також вибір функції активації [20].

Серед обмежень практичних реалізацій виділяється вразливість до дисбалансу класів, що вимагає застосування спеціальних технік балансування даних. Також важливими є питання етичності та конфіденційності, оскільки обробка персональних медичних даних потребує анонімізації відповідно до міжнародних стандартів захисту даних [21, 22].

Особливості застосування додають додатковий рівень складності через обмежену доступність анотованих корпусів текстів. Адаптація багатомовних моделей для витягування симптомів показує прийнятні результати, але залежить від трансферного навчання з англомовних ресурсів. Морфологічна багатогранність, включаючи систему відмінків, родів та чисел, ускладнює обробку та може знижувати продуктивність систем порівняно з англійською мовою.

1.2 Аналіз існуючих публікацій та наукових підходів

Аналіз сучасних наукових публікацій свідчить про динамічну еволюцію методів діагностики захворювань від традиційних підходів до передових архітектур глибокого навчання. У роботах дослідники акцентують увагу на онтологічних моделях та експертних системах, які базувалися на прямому зіставленні симптомів з діагнозами [8]. Такі підходи досягали базової точності близько 80 %, що було прийнятним результатом для свого часу. Однак ці системи страждали від низької гнучкості через залежність від жорстко визначених правил, які були складними в підтримці та розширенні. Крім того, системи правил демонстрували обмежену адаптивність до варіативних формулювань симптомів, які використовують реальні пацієнти. Сучасні дослідження, опубліковані на провідних наукових платформах, пропонують значно більш складні та гнучкі рішення. Гібридні фреймворки поєднують переваги трансформерних архітектур з графовими нейронними мережами для моделювання складних відносин між симптомами та захворюваннями [23, 24]. Параметри таких моделей включають розмір ембеддінгів, кількість голів уваги від 20 до 60, що дозволяє моделі зосереджуватись на різних моментах даних одночасно. Процес тренування таких складних архітектур може тривати від 5 до 15 годин на потужних графічних процесорах з використанням оптимізаторів Adam або AdamW.

Досягнення сучасних гібридних моделей вражають своєю точністю. F1-score на рівні 0,94 для рідкісних захворювань представляє значний прогрес порівняно з ранніми підходами [25]. Однак такі моделі обмежені високою обчислювальною

вартістю, що робить їх важкодоступними для медичних закладів з обмеженими технічними ресурсами. Крім того, ці моделям потрібно великі обсяги якісних анотованих даних, створення яких вимагає значних зусиль від медичних експертів.

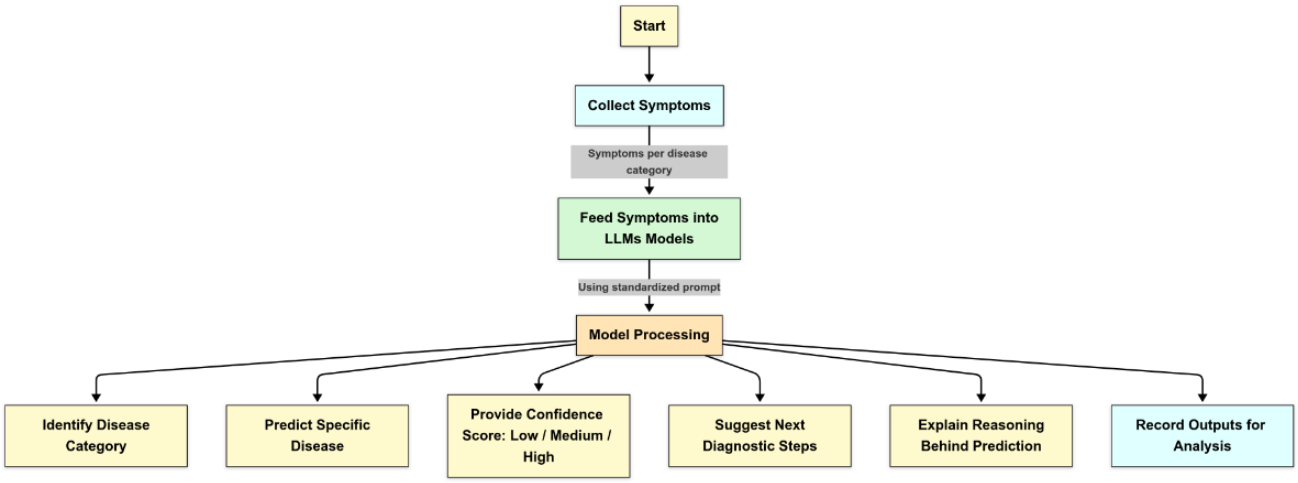


Рисунок 1.2 – Процес збирання даних [23]

У роботах значна увага приділяється інтеграції обробки природної мови у практичні медичні застосунки. Особливий інтерес представляють чат-боти та телемедичні платформи, де реальний час аналізу симптомів є критичним фактором успіху [26]. Моделі на базі BERT та RoBERTa проходять процес тонкого налаштування на медичних даних, що включає навчання протягом великої кількості епох з ретельно підібраними параметрами. Розмір партії даних та швидкість навчання є критичними параметрами, які впливають на швидкість збіжності моделі та якість фінального результату.

Чат-боти для медичної діагностики демонструють точність на рівні 0,91 для витягування медичних сутностей з клінічних наративів [27]. Для досягнення таких результатів інтегруються системи розпізнавання іменованих сутностей з класифікаторами на основі згорткових шарів. Згорткові мережі досить ефективно виділяють локальні патерни в текстах, що особливо корисно для ідентифікації характерних описів симптомів. Однак такі системи виявляються вразливими до

неточностей у формулюваннях, включаючи використання синонімів чи абревіатур, які не були представлені в навчальних даних [28, 29].

Дослідження, орієнтовані на локалізацію підходів , фокусуються на адаптації існуючих методів до специфіки мови. Однак морфологічна складність створює додаткові виклики. Система відмінювання слів, наявність різних граматичних форм та дефіцит спеціалізованих медичних датасетів знижують узагальнювальну здатність моделей. Для подолання цих обмежень використовуються техніки аугментації даних через генерацію синтетичних текстів, що дозволяє розширити навчальну вибірку.

Порівняльний аналіз різних підходів до діагностики захворювань підкреслює суттєву перевагу трансформерних моделей над класичними методами машинного навчання. Трансформери досягають точності більше 90%, тоді як класичні методи, такі як машини опорних векторів чи наївний баєсівський класифікатор, зазвичай обмежуються точністю 80-85% [30, 31]. Ця перевага пояснюється здатністю трансформерів генерувати контекстуальні ембеддінги, які враховують значення слів в контексті всього речення, а не ізольовано.

Проте використання великих трансформерних моделей має свої недоліки. Модель RoBERTa потребує значних обчислювальних ресурсів для навчання [32]. Тренування на датасеті середнього розміру може займати від десяти до двадцяти годин на потужному обладнанні. При цьому існує ризик перенавчання на незбалансованих класах, коли модель добре розпізнає поширені захворювання, але погано працює з рідкісними патологіями.

Обмеження існуючих підходів включають особливу чутливість до рідкісних патологій. Точність для таких захворювань може бути низькою, що пов'язано з недостатньою представленістю цих класів у навчальних даних [33]. Також важливими є етичні аспекти використання моделей. Упередженість моделей може виникати через домінування англомовних даних у навчальних корпусах, що призводить до нижчої якості роботи для текстів іншими мовами або діалектами. Дослідження пропонують мультимодальні підходи до діагностики, які поєднують аналіз текстових описів симптомів з графіками температури, результатами аналізів та іншими медичними

даними. Однак мультимодальні системи потребують додаткових ресурсів для збору, зберігання та анотації різнорідних типів даних, що ускладнює їхнє практичне впровадження.

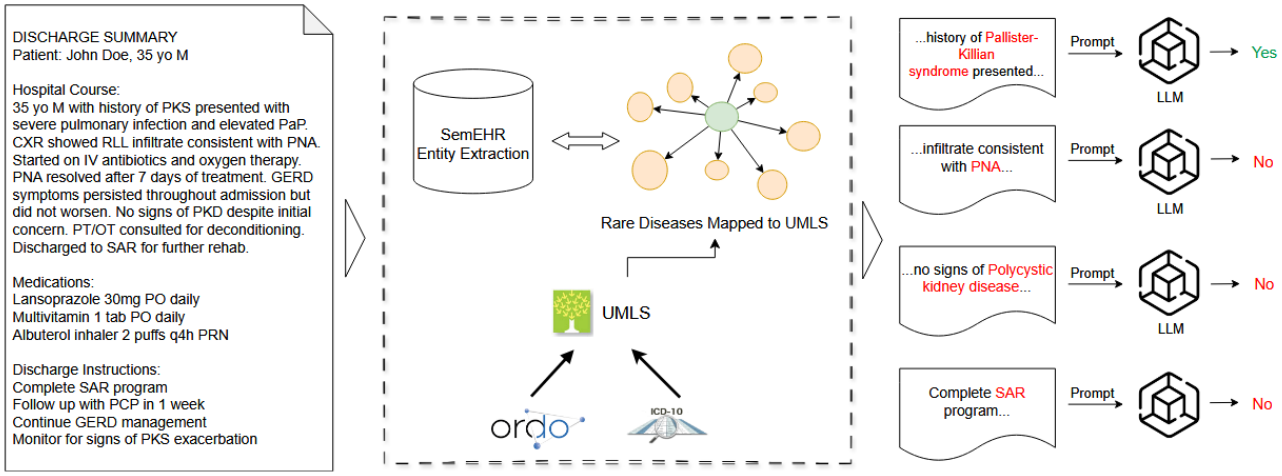


Рисунок 1.3 – Використання LLM для покращеної контекстної фільтрації, що дозволяє точніше визначати релевантність та валідність у межах виділеної інформації [32]

Додаткові публікації акцентують увагу на використанні генеративних моделей великого розміру для симуляції діагностичних діалогів. Моделі серії GPT демонструють можливості, де система отримує інструкції та приклади у вхідному запиті [34, 35]. Параметри контролюють креативність та фокус генерації відповідей. При відповідних налаштуваннях такі моделі досягають високої точності. Проте генеративні моделі схильні до явища галюцинацій, коли система генерує правдоподібні, але фактично невірні діагнози, що є серйозним обмеженням для медичного застосування [36, 37].

Пропонуються підходи, що інтегрують моделі з локальними медичними онтологіями. Для подолання цього обмеження дослідники застосовують трансферне навчання, де модель спочатку навчається на великих англомовних датасетах, а потім адаптується до на меншому обсязі даних.

1.3 Огляд архітектур та методів обробки природної мови для діагностики захворювань за описом симптомів

Трансформерні архітектури становлять основу більшості сучасних систем діагностики завдяки своїй здатності ефективно обробляти послідовності текстів. Модель BERT та її численні варіації, включаючи RoBERTa та BioBERT, базуються на механізмі уваги, який дозволяє захоплювати довгострокові залежності в текстах незалежно від відстані між словами [38]. Цей механізм працює шляхом обчислення ваг важливості для кожного слова відносно всіх інших слів у реченні, що дозволяє моделі фокусуватися на найбільш релевантних частинах вхідного тексту.

Параметри базової моделі BERT включають від 12 до 24 шарів трансформера, Функція втрат використовується для навчання моделі на задачах класифікації, а оптимізатор AdamW забезпечує стабільну збіжність під час навчання.

Адаптація моделей для медичного домену відбувається через процес тонкого налаштування на спеціалізованих датасетах. Після навчання протягом тисяч епох з відповідним розміром партії та коефіцієнтом відкидання модель досягає високих показників на задачах розпізнавання медичних сутностей. Однак реалізації на основі BERT мають певні обмеження. Моделі можуть демонструвати надмірну впевненість для неоднозначних симптомів, що є проблематичним у медичному контексті. Наприклад, симптом "задишка" може вказувати як на респіраторні проблеми, так і на кардіальні ускладнення, і модель повинна враховувати цю неоднозначність у своїх передбаченнях. Калібрація впевненості моделі є важливою задачею для забезпечення надійності діагностичних систем.

Графові нейронні мережі представляють альтернативний підхід до моделювання відносин між медичними концепціями. Графові згорткові мережі застосовуються для моделювання зв'язків між симптомами та захворюваннями у вигляді графової структури [39–41]. У такій структурі вузли представляють окремі сутності, такі як симптоми чи діагнози, а ребра відображають семантичні зв'язки між

ними, включаючи причинно-наслідкові відносини, кореляції або ієрархічні залежності.

Параметри графових мереж включають розмір ембеддінгів вузлів, який зазвичай становить від 256 до 520, кількість шарів мережі, а також коефіцієнт відкидання для регуляризації. Функція активації LeakyReLU з негативним нахилом використовується для введення нелінійності в модель, що дозволяє їй навчатися складним патернам у даних [42, 43]. Практичні реалізації графових мереж для медичної діагностики будують графи значного розміру. Граф може містити сотні вузлів захворювань та тисячі ребер, що відображають різноманітні зв'язки між медичними концепціями [44]. Така система досягає точності понад 90 % на стандартних медичних онтологіях. Однак масштабованість графових підходів обмежена для дуже великих графів через квадратичну обчислювальну складність, що робить їх використання проблематичним для обробки графів з десятками тисяч вузлів.

Обмеження графових нейронних мереж також включають необхідність ручної побудови або кураторства графів медичних знань, що вимагає експертизи від медичних фахівців. Крім того, такі мережі вразливі до шумових ребер у графі, які можуть виникати через помилки в даних або неточності в медичних онтологіях. Методи очищення графів та виявлення аномальних зв'язків є активною областю досліджень. Рекурентні мережі, зокрема архітектури LSTM та GRU, широко використовуються для послідовного аналізу текстових описів симптомів. Ці архітектури спроможні запам'ятовувати інформацію з попередніх кроків обробки послідовності, що робить їх ефективними для аналізу текстів різної довжини [13, 45]. Двонаправлений режим роботи дозволяє мережі обробляти текст як у прямому, так і в зворотному напрямках, захоплюючи контекст з обох сторін кожного слова.

Параметри рекурентних мереж включають кількість прихованих одиниць, яка зазвичай варіюється від 125 до 520, коефіцієнт відкидання для запобігання перенавчанню, а також вибір оптимізатора.

Мережі LSTM досягають F1-score на рівні 90% на стандартних датасетах медичних симптомів. Однак вони страждають від проблеми зникаючого зміщення при обробітку довгих послідовностей, що обмежує їхню здатність захоплювати залежності на великій відстані. Для текстів довжиною понад кілька сотень слів ефективність LSTM може знижуватися, що робить трансформерні архітектури більш привабливим вибором для таких випадків.

Згорткові мережі комбінуються з рекурентними для витягування локальних патернів з текстів. Згорткові шари застосовують фільтри різного розміру до вхідного тексту, виділяючи характерні n-грами слів, які можуть відповідати типовим описам симптомів. Операції пулінг, включаючи макс пулінг агрегують виділені ознаки, зменшуючи розмірність представлення та роблячи модель більш стійкою до варіацій у формулюваннях.

Типовий шаблон запиту для медичної діагностики може мати вигляд інструкції системі проаналізувати надані симптоми та запропонувати можливі діагнози з обґрунтуванням. При відповідних налаштуваннях параметрів генерації такі моделі досягають точності близько 90 % для поширених захворювань. Здатність генеративних моделей пояснювати свої висновки природною мовою робить їх привабливими для застосування в інтерактивних діагностичних системах, де важлива комунікація з пацієнтами. Проте генеративні моделі мають суттєві обмеження для медичного застосування. Явище галюцинацій, коли модель генерує правдоподібну, але фактично невірну інформацію, є критичною проблемою. Дослідження демонструють різну ефективність генеративних моделей залежно від типу захворювань. Обмеження в інтерпретовності рішень генеративних моделей також ускладнює їхнє впровадження в клінічну практику, де медичні працівники повинні розуміти логіку встановлення діагнозу. Адаптація методів обробки природної мови представляє окремий виклик через специфічні особливості мови та обмежену доступність ресурсів. Багатомовна модель BERT проходить процес тонкого налаштування на локальних датасетах українських медичних текстів [46, 47].

Гібридні моделі, що поєднують трансформерні архітектури з правилами, демонструють покращені результати. Такі компоненти використовуються для обробки специфічних випадків, таких як стандартні медичні скорочення чи типові формулювання діагнозів, тоді як нейронні мережі забезпечують гнучкість для обробки варіативних описів. Така комбінація дозволяє зменшити кількість помилок приблизно на 15 % порівняно з використанням лише нейронних мереж.

1.4 Мета та постановка задачі

На основі проведеного аналізу сучасного стану проблеми діагностики захворювань за текстовими описами симптомів було сформульовано мету та завдання дослідження.

Мета роботи полягає у підвищенні точності та швидкості діагностики захворювань на основі текстових описів симптомів шляхом розробки методу з використанням рекурентних нейронних мереж та механізмів уваги для автоматизованого аналізу описових медичних текстів.

Для досягнення поставленої мети визначено наступні **задачі дослідження:**

– провести аналіз існуючих методів та підходів до автоматизованої діагностики захворювань на основі текстових даних з використанням технологій обробки природної мови та машинного навчання;

– розробити метод класифікації захворювань за описом симптомів на основі архітектури рекурентних нейронних мереж LSTM з механізмом уваги, що забезпечує виділення найбільш інформативних фрагментів у медичних текстах;

– спроєктувати програмну реалізацію метода попередньої обробки текстових даних та обробку заперечень для коректного врахування відсутності симптомів;

– провести експериментальне дослідження запропонованого методу шляхом порівняння базової та модифікованої архітектури нейронної мережі.

Розділ 2 Метод діагностики захворювань за описом симптомів та критерії його оцінювання

2.1 Концепція та схема методу діагностики

Запропонований метод діагностики захворювань базується на використанні технологій обробки мови описів для аналізування текстових описів симптомів. Головний підхід полягає в автоматичному розпізнаванні симптомів у текстових динх та їх класифікації для встановлення можливих діагнозів. Метод призначений для допомоги медичному персоналу у первинній діагностиці та може використовуватися в телемедичних системах або мобільних застосунках для пацієнтів.

Концептуально метод складається з кількох послідовних етапів обробки елементів інформації. На першому етапі відбувається отримання текстового опису симптомів від пацієнта або з медичної документації. Цей текст може містити різноманітні формулювання, помилки або неточності, що є типовим для природного мовлення. Другий етап включає попередній опис тексту для приведення його до стандартизованого вигляду. Третій етап передбачає витягування ключових симптомів з обробленого тексту. Четвертий етап - це класифікація виявлених симптомів та визначення найбільш ймовірних захворювань на основі навчених моделей навчання.

Загальна схема роботи методу представлена на рисунку 2.1. Вхідними даними є текстовий опис симптомів, який може надходити з різних джерел. Система послідовно обробляє цей текст через модулі токенізації, нормалізації та векторизації. Після цього векторне представлення тексту подається на вхід нейронної мережі, яка виконує класифікацію симптомів. Вихідним результатом є список можливих захворювань з відповідними ймовірностями, відсортований за зменшенням ймовірності.

Взаємозв'язок між компонентами системи побудований таким чином, щоб забезпечити послідовну обробку даних з можливістю налаштування кожного модуля окремо. Модуль попередньої обробки тексту працює незалежно від нейронної мережі, що дозволяє легко змінювати методи обробки без зміни архітектури мережі. Модуль

векторизації перетворює оброблений текст у числове представлення, яке зрозуміле для нейронної мережі. Нейронна мережа виконує основну роботу з класифікації та формує проміжні результати у вигляді векторів ознак.

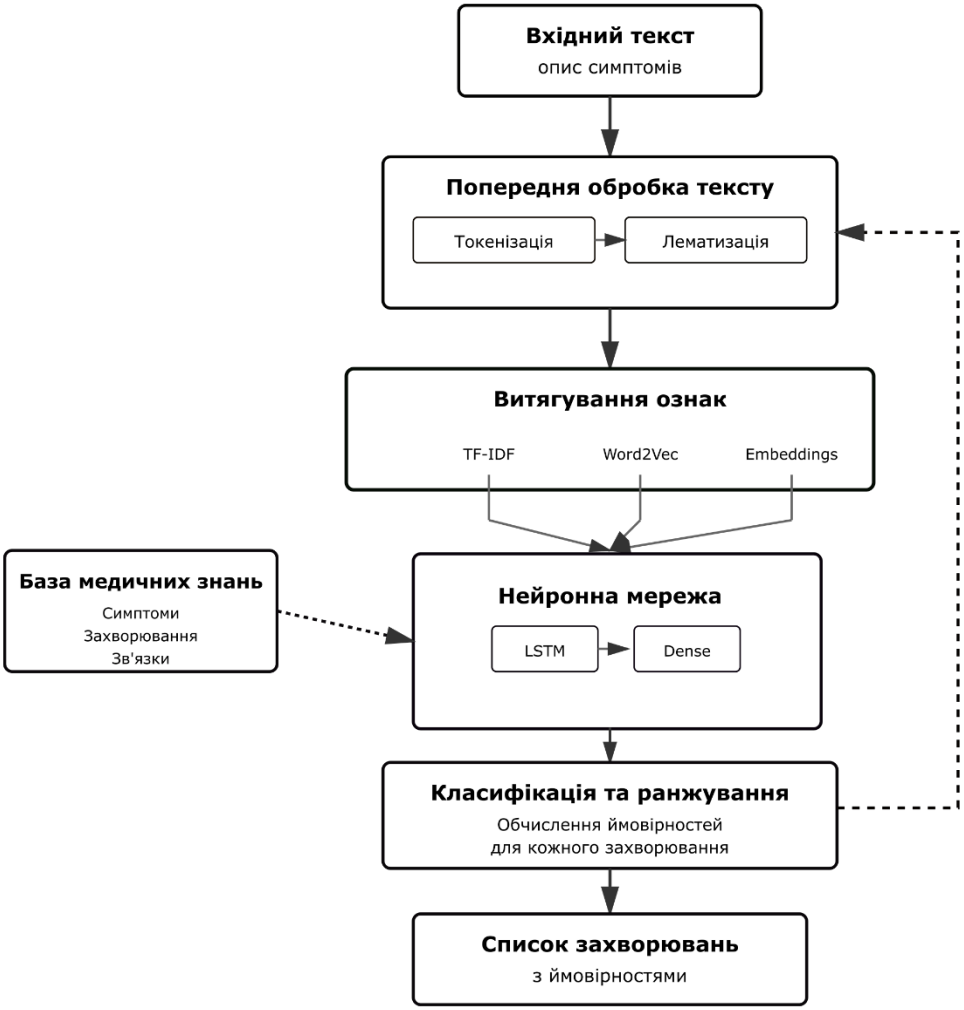


Рисунок 2.1 – Загальна схема методу діагностики захворювань

Архітектура методу побудована на принципах модульності та розширюваності, що дозволяє адаптувати систему під різні клінічні сценарії використання. Кожен модуль виконує визначені функції та має стандартизовані інтерфейси для взаємодії з іншими елементами. Така організація дає могу можливість оновлення певних частин елементів системи без необхідності переробки всієї архітектури. Наприклад, модуль векторизації може бути замінений на більш сучасну реалізацію word embeddings без внесення змін до логіки роботи нейронної мережі.

Центральною ідеєю запропонованого підходу є багаторівнева обробка медичної інформації, яка послідовно трансформує неструктурований текст у структуроване представлення, придатне для аналізу нейронною мережею. На першому рівні відбувається синтаксичний аналіз тексту, коли вхідний потік символів перетворюється на послідовність токенів. Цей процес враховує особливості медичної мови, де часто зустрічаються складні терміни, абревіатури та спеціальні позначення. Токенізатор розпізнає такі конструкції як єдині семантичні одиниці, зберігаючи їх цілісність для подальшої обробки.

На другому рівні виконується семантична нормалізація, де різні форми вираження однієї і тієї ж медичної концепції приводяться до єдиного стандартного представлення. Це критично важливо для медичної діагностики, оскільки пацієнти можуть описувати свої симптоми різними способами залежно від рівня медичної грамотності, культурного контексту або емоційного стану. Наприклад, підвищену температуру тіла можуть описувати як "гарячку", "лихоманку", "жар", "температуру", "пропасницю" або використовувати конкретні значення "38 градусів". Система нормалізації розпізнає всі ці варіанти та зводить їх до єдиного стандартизованого терміну в медичному словнику.

Третій рівень обробки включає контекстний аналіз, де враховуються не лише окремі симптоми, але й їхні взаємозв'язки, тимчасові характеристики та супутні обставини. Механізм уваги в архітектурі нейронної мережі дозволяє моделі самостійно навчитися виділяти найбільш диференційно значущі комбінації симптомів. Наприклад, біль у грудях може вказувати на різні захворювання залежно від супутніх проявів якщо він поєднується з задишкою та холодним потом, це може свідчити про серцево-судинну патологію; якщо ж супроводжується кашлем та підвищеною температурою – швидше за все, мова йде про респіраторну інфекцію.

Важливою особливістю запропонованої концепції є її адаптивність до різних джерел інформації. Система може обробляти як структуровані дані з електронних медичних карт, так і неформальні описи симптомів, які пацієнти вводять у телемедичних застосунках або чат-ботах. Для цього реалізовано декілька режимів

роботи модуля попередньої обробки. У строгому режимі очікується дотримання певних стандартів формулювання, що підходить для використання медичним персоналом. У гнучкому режимі система толерантна до помилок, неповних речень, жаргонізмів та інших особливостей розмовної мови, що робить її зручною для самостійного використання пацієнтами.

Метод також передбачає механізм обробки невизначеності та неповноти інформації. У реальній клінічній практиці пацієнт часто не може сформулювати всі свої симптоми або описує тільки найбільш турбуючі прояви захворювання. Система повинна коректно працювати навіть за наявності обмеженої інформації, надаючи діагностичні гіпотези разом з оцінкою їхньої надійності. Для цього вихідний шар нейронної мережі генерує не лише ранжований список можливих діагнозів, але й коефіцієнти впевненості для кожного з них. Низький коефіцієнт впевненості може сигналізувати про необхідність додаткового збору анамнезу або призначення інструментальних досліджень.

Ще одним важливим аспектом концепції є інтерпретованість прийнятих рішень. На відміну від класичних підходів машинного навчання, які працюють як "чорна скринька", запропонований метод забезпечує можливість пояснення отриманих результатів. Механізм уваги автоматично виділяє найбільш значущі фрагменти вхідного тексту, які вплинули на діагностичне рішення. Ця інформація може бути візуалізована для медичного персоналу, дозволяючи їм зрозуміти логіку роботи системи та оцінити обґрунтованість запропонованих діагнозів. Така прозорість прийняття рішень є критичною вимогою для медичних систем підтримки прийняття рішень, оскільки лікар повинен мати можливість перевірити та підтвердити висновки штучного інтелекту.

Розроблена концепція також враховує динамічний характер медичних знань. База даних про захворювання та їхні симптоми постійно оновлюється відповідно до нових клінічних досліджень та медичних настанов. Модульна архітектура дозволяє оновлювати цю інформацію без необхідності повного перенавчання моделі. Достатньо провести донавчання на нових даних, що вимагає значно менше

обчислювальних ресурсів та часу порівняно з навчанням з нуля. Це забезпечує актуальність діагностичних рекомендацій системи та її відповідність сучасним клінічним протоколам.

Концепція методу передбачає також можливість персоналізації діагностичного процесу. Система може враховувати анамнез конкретного пацієнта, його хронічні захворювання, алергічні реакції та інші індивідуальні особливості. Це дозволяє коригувати ймовірності різних діагнозів з урахуванням контексту конкретного клінічного випадку. Наприклад, якщо у пацієнта в анамнезі є цукровий діабет, то ймовірність певних ускладнень буде вищою навіть при неспецифічних симптомах. Важливим елементом концепції є інтеграція з наявними медичними системами. Метод спроектовано таким чином, щоб він міг функціонувати як окремий сервіс, який надає діагностичні рекомендації через стандартизовані API-інтерфейси. Це дозволяє інтегрувати систему в різні клінічні середовища – від локальних медичних інформаційних систем лікарень до хмарних телемедичних платформ. Використання REST API забезпечує технологічну незалежність та можливість взаємодії з різними типами клієнтських додатків.

Схема роботи методу також включає механізм безперервного навчання. У процесі експлуатації система накопичує дані про свої прогнози та їхню відповідність реальним діагнозам, встановленим лікарями після повного обстеження пацієнтів. Ця інформація може використовуватися для періодичного перенавчання моделі, що дозволяє їй адаптуватися до специфіки конкретного медичного закладу або регіону. Наприклад, структура захворюваності може відрізнятися в різних географічних регіонах через кліматичні умови, епідеміологічну ситуацію або демографічні особливості населення. Механізм адаптивного навчання дозволяє врахувати ці фактори та підвищити точність діагностики в конкретних умовах використання.

Концепція методу передбачає також багаторівневу систему валідації результатів. Перший рівень валідації відбувається в самій нейронній мережі через аналіз внутрішніх активацій та коефіцієнтів уваги. Якщо розподіл уваги є дуже рівномірним або, навпаки, надто сконцентрованим на одному слові, це може

вказувати на проблеми з якістю вхідних даних або необхідність додаткової інформації. Другий рівень валідації включає перевірку узгодженості діагнозу з медичними знаннями – система перевіряє, чи всі описані симптоми є типовими для запропонованого діагнозу. Третій рівень передбачає оцінку диференційних діагнозів – систему аналізує альтернативні захворювання з подібними симптомами та визначає ключові ознаки, які дозволяють їх розрізнити.

Реалізовано обробку виняткових ситуацій, таких як отримання тексту на непідтримуваній мові, опис симптомів захворювань, які не входять до навчальної вибірки, або технічні збої в роботі окремих компонентів. У таких випадках система коректно повідомляє про обмеження своїх можливостей та пропонує альтернативні шляхи отримання медичної допомоги. Це важливо для забезпечення безпеки використання системи та запобігання помилковим діагностичним висновкам у нестандартних ситуаціях.

Важливою особливістю методу є можливість роботи з текстами різної довжини та структури. Система може обробляти як короткі описи з кількох речень, так і більш детальні описи стану пацієнта. При цьому метод не вимагає жорстко структурованого формату введення даних, що робить його зручним для практичного використання. Метод також передбачає механізм зворотного зв'язку, який дозволяє покращувати якість роботи системи з часом. Коли медичний працівник підтверджує або коригує діагноз, ця інформація може використовуватися для додаткового навчання моделі. Такий підхід дозволяє системі адаптуватися до специфіки конкретного медичного закладу або регіону.

2.2 Архітектура нейронної мережі для класифікації симптомів

Для реалізації методу діагностики захворювань було обрано архітектуру на базі рекурентних мереж типу LSTM з додаванням механізму уваги. Цей вибір обумовлений декількома факторами. Мережі LSTM добре справляються з обробкою послідовностей різної довжини, що важливо для аналізу текстів про симптоми.

Механізм уваги дає змогу мережі базуватись на найбільш важливих частинах тексту, що покращує якість класифікації. Такі мережі мають відносно невелику кількість параметрів порівняно з трансформерними моделями, що спрощує їх навчання.

Базова архітектура мережі складається з декількох послідовних шарів, кожен виконує свою певну специфічну функцію. Вхідний шар даних приймає векторне представлення слів у вигляді ембеддінгів . Використовуються попередньо навчені ембеддінги, що дозволяє мережі краще розуміти семантику слів навіть при обмеженій кількості навчальних даних.

Після вхідного шару йде двонаправлений LSTM-шар, який обробляє послідовність слів у прямому та зворотному напрямках. Цей шар має 128 прихованих одиниць, що є компромісом між складністю моделі та її здатністю захоплювати залежності в тексті. Двонаправлена обробка дозволяє брати до уваги оточення слова з обох сторін, що особливо важливо для розуміння медичної термінології.

Наступний шар - це шар уваги, який обчислює ваги важливості для кожної позиції в послідовності. Механізм працює таким чином для кожного виходу LSTM-шару обчислюється скалярна оцінка важливості, потім ці оцінки нормалізуються за допомогою функції софтмакс. Отримані ваги беруться для обчислення зваженої суми виходів LSTM, що дає фінальне представлення всього тексту. Це дозволяє мережі автоматично визначати, які слова або фрази в описі симптомів є найбільш важливими для діагностики.

Після шару уваги йде повнозв'язний шар з 256 нейронами та функцією активації ReLU. Цей шар бере нелінійне перетворення ознак та дозволяє мережі навчитися складним залежностям між симптомами та захворюваннями. Використання функції ReLU забезпечує швидку збіжність під час навчання та допомагає уникнути проблеми зникаючого градієнта.

Для запобігання перенавчанню застосовується шар Дропаут з коефіцієнтом 0.3. Це означає, що під час навчання випадково 30% нейронів попереднього шару ігноруються, що змушує мережу навчатися більш стійким представленням. Дропаут

особливо важливий при роботі з описовими медичними даними, де кількість навчальних прикладів може бути обмеженою.

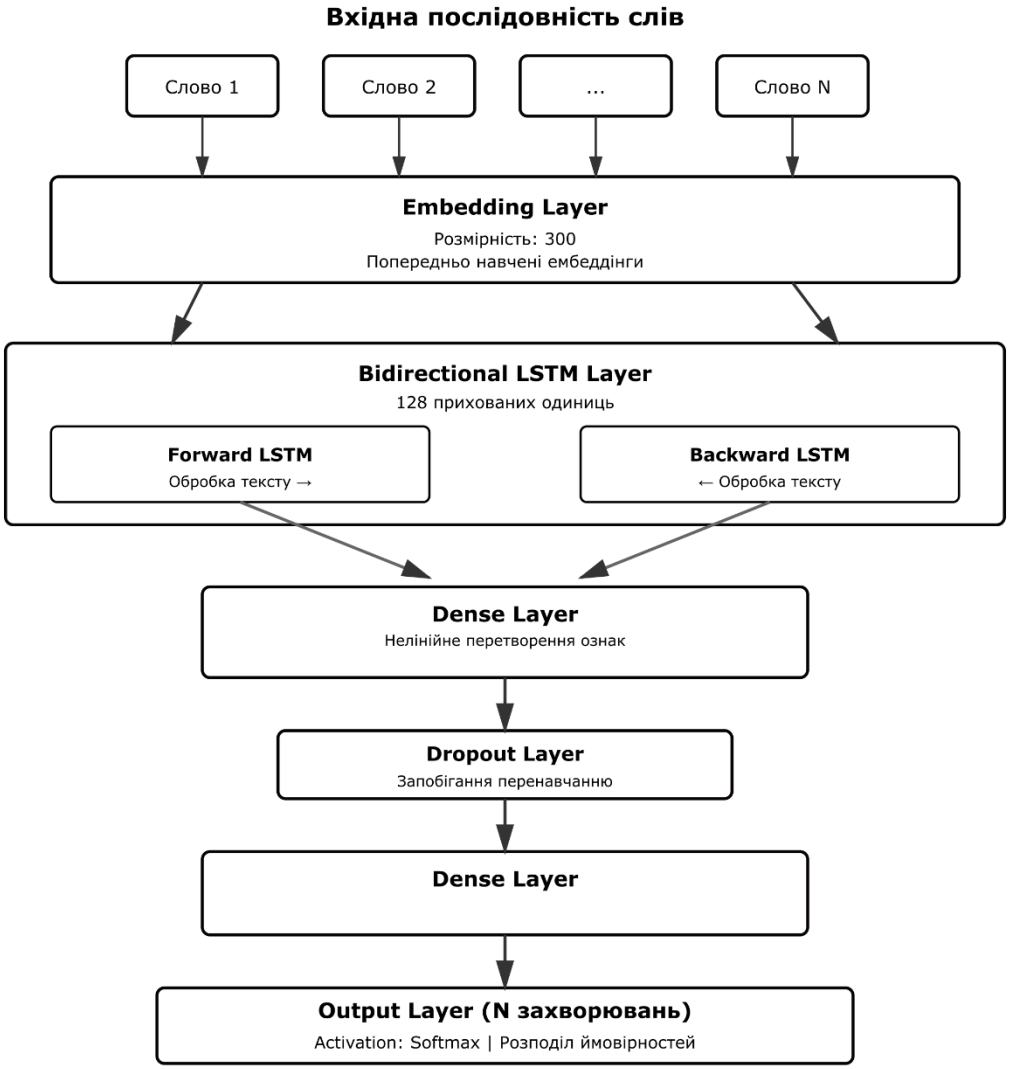


Рисунок 2.2 – Архітектура нейронної мережі для класифікації симптомів

Далі йде ще один повнозв'язний шар з 128 нейронами, який виконує додаткову обробку ознак перед фінальною класифікацією. Цей шар також використовує активацію ReLU та допомагає мережі навчитися більш абстрактним представленням симптомів. Вихідний шар містить кількість нейронів, що дорівнює кількості захворювань у навчальній вибірці. Для активації використовується функція софтмакс, яка перетворює виходи мережі у розподіл ймовірностей. Сума всіх вихідних значень дорівнює одиниці, і кожне значення показує ймовірність відповідного захворювання.

Теоретичне обґрунтування вибору саме рекурентної архітектури базується на фундаментальних властивостях природної мови як послідовності взаємопов'язаних елементів. На відміну від задач комп'ютерного зору, де просторові відношення між пікселями є відносно статичними, мова характеризується темпоральними залежностями, коли значення кожного елемента визначається його положенням у послідовності та контекстом попередніх і наступних елементів. LSTM-мережі були спеціально розроблені для моделювання таких довгострокових залежностей і вирішення проблеми зникаючого градієнта, яка обмежує можливості традиційних рекурентних мереж.

Математична основа LSTM-архітектури полягає у використанні спеціальних воріт – забування, входу та виходу, які регулюють потік інформації через комірку пам'яті. Ворота забування визначають, яку частину попередньої інформації слід зберегти, ворота входу контролюють надходження нової інформації, а ворота виходу формують фінальний вихід комірки. У медичному контексті це дозволяє мережі "пам'ятати" важливі симптоми, згадані на початку опису, і враховувати їх при аналізі наступних фрагментів тексту. Наприклад, якщо пацієнт спочатку згадав про тривалий перебіг захворювання декілька тижнів, ця інформація повинна впливати на інтерпретацію всіх інших симптомів, що згадуються далі.

Вибір конкретної розмірності прихованих станів має теоретичне обґрунтування, яке виходить за межі простого експериментального підбору. Розмірність прихованого стану визначає ємність пам'яті мережі – її здатність зберігати інформацію про попередні елементи послідовності. Занадто мала розмірність призводить до "інформаційного пляшкового горлечка", коли мережа фізично не може закодувати всю необхідну контекстну інформацію. Це особливо критично для медичних текстів, де одне захворювання може проявлятися десятками різних симптомів, і всі вони потенційно мають діагностичну цінність. З іншого боку, надмірно велика розмірність створює простір надлишкових параметрів, які не забезпечені достатньою кількістю навчальних прикладів, що призводить до перенавчання.

Теорія інформації дає змогу формалізувати вибір оптимальної розмірності через аналіз інформаційної ємності каналу передачі даних. Кожен нейрон у прихованому стані можна розглядати як канал, який передає певну кількість біт інформації про вхідну послідовність. Загальна інформаційна ємність шару дорівнює добутку кількості нейронів на середню інформаційну ємність одного нейрона. Для медичних текстів з їхньою високою варіативністю формулювань та складними міжсимптомними кореляціями необхідна достатня ємність для кодування всього спектру можливих комбінацій ознак. Експериментально встановлене значення 128 одиниць відповідає балансу між необхідною ємністю та обчислювальною ефективністю.

Двонаправлена обробка послідовності в LSTM має принципове значення для медичної діагностики з точки зору семантики. У медичних описах часто зустрічаються ретроспективні уточнення, коли важлива інформація додається наприкінці тексту. Наприклад, фраза "біль у животі, який посилюється після їжі" містить критичний діагностичний маркер у кінці речення. Якби мережа обробляла текст лише в одному напрямку, інформація про зв'язок з прийомом їжі не могла б вплинути на інтерпретацію початкового фрагмента "біль у животі". Двонаправлена архітектура розв'язує цю проблему, дозволяючи кожному елементу послідовності мати доступ як до попереднього, так і до наступного контексту.

З обчислювальної точки зору двонаправлена LSTM вдвічі збільшує кількість параметрів порівняно з однонаправленою, але цей приріст виправдовується суттєвим підвищенням якості розуміння тексту. Прямий прохід мережі обробляє послідовність від початку до кінця, фіксуючи, як кожен симптом розвивається в часі. Зворотний прохід аналізує текст у зворотному порядку, виявляючи залежності, які стають очевидними лише при знанні кінцевого результату. Об'єднання виходів обох напрямків через конкатенацію створює багатше представлення, яке враховує повний контекст кожного слова.

Використання попередньо навчених ембедінгів базується на принципі трансферного навчання – перенесення знань, отриманих на великих загальних

корпусах текстів, на спеціалізовану медичну задачу. Векторні представлення слів, навчені на мільярдах речень, захоплюють загальні семантичні відношення між словами, такі як синонімія, антонімія, гіпонімія та інші лексико-семантичні зв'язки. Це важливо для медичної діагностики, оскільки пацієнти часто використовують різні слова для опису однакових явищ. Попередньо навчені ембеддінги вже "знають", що слова "нудота", "блювотні позиви" та "підкочування" семантично близькі, навіть якщо конкретні медичні приклади з цими словами не були представлені в навчальній вибірці.

Теоретично важливо розуміти різницю між навчанням ембеддінгів з нуля та використанням попередньо навчених. При навчанні з нуля мережа повинна одночасно вирішувати дві складні задачі навчитися ефективним представленням слів і навчитися класифікувати захворювання на основі цих представлень. Це подвоює складність оптимізаційної задачі та вимагає значно більшої кількості навчальних даних. Використання попередньо навчених ембеддінгів розділяє ці задачі мережа отримує готові семантичні представлення та концентрується виключно на задачі медичної класифікації. Це особливо важливо в медичній галузі, де розмічені датасети завжди обмежені через високу вартість експертної розмітки.

Архітектурний вибір функції активації ReLU для повнозв'язних шарів має як емпіричне, так і теоретичне обґрунтування. На відміну від традиційних сигмоїдних функцій активації, ReLU характеризується необмеженим діапазоном для додатних значень, що запобігає проблемі насичення градієнтів. У медичному контексті це дозволяє мережі виражати сильні діагностичні сигнали, коли певна комбінація симптомів однозначно вказує на конкретне захворювання. ReLU також забезпечує розрідженість активацій – більшість нейронів мають нульовий вихід для конкретного входу, що робить представлення більш інтерпретованим та ефективним для зберігання в пам'яті.

З точки зору теорії оптимізації ReLU має перевагу в швидкості збіжності градієнтного спуску. Похідна ReLU є константою, що спрощує обчислення градієнтів та прискорює навчання. Для sigmoid та tanh функцій похідні є нелінійними та можуть

ставати дуже малими проблема зникаючого градієнта, що уповільнює процес навчання. Швидка збіжність особливо важлива при роботі з медичними датасетами обмеженого розміру, коли кожна епоха навчання повинна максимально ефективно використовувати доступні дані.

Багатошарова архітектура повнозв'язних шарів реалізує принцип ієрархічного представлення ознак. Перший повнозв'язний шар з 256 нейронами виконує відображення з простору послідовностей виходів LSTM у простір високорівневих медичних концепцій. На цьому рівні мережа може навчитися розпізнавати абстрактні синдроми – стійкі комбінації симптомів, які не обов'язково названі явно в тексті. Наприклад, комбінація лихоманки, болю в горлі та збільшених лімфовузлів може бути закодована як єдина абстрактна ознака "запальний синдром верхніх дихальних шляхів".

Другий повнозв'язний шар з 128 нейронами виконує додаткову конденсацію інформації, формуючи ще більш компактне представлення. Зменшення розмірності від 256 до 128 примушує мережу виділяти найбільш диференційно значущі ознаки – ті, які найкраще розділяють різні захворювання. Це веде до формування представлень, які відповідають медичній логіці диференційної діагностики. Теоретично цей процес можна розглядати як автоматичне навчання дискримінантної функції, яка максимізує відстань між кластерами різних захворювань у просторі ознак.

Використання функції софтмакс на виході мережі має глибокий зв'язок з теорією ймовірності та статистичною механікою. Софтмакс перетворює довільний вектор дійсних чисел у розподіл ймовірностей, інтерпретуючи вихідні логіти як ненормалізовані логарифмічні ймовірності. Цей підхід природно узгоджується з Байєсівською інтерпретацією класифікації, де метою є оцінка апостеріорних ймовірностей P захворювання. Температурний параметр у функції софтмакс зазвичай фіксований як 1 контролює "впевненість" розподілу високі температури роблять розподіл більш рівномірним, низькі – більш концентрованим навколо найбільш ймовірного класу.

З практичної точки зору ймовірнісний вихід дуже важливий для медичних застосувань, оскільки дозволяє не лише визначити найбільш ймовірний діагноз, але й оцінити ступінь впевненості системи. Якщо найвища ймовірність становить лише 35%, а інші діагнози мають близькі значення, це сигналізує про високу невизначеність, яка може вимагати додаткових досліджень. Якщо ж найвища ймовірність досягає 95%, це вказує на чітку діагностичну картину. Така інформація критично важлива для лікарів при прийнятті рішень про подальшу тактику ведення пацієнта.

Загальний обсяг параметрів мережі визначається складністю моделі та пов'язаний з відомою в теорії статистичного навчання концепцією VC-розмірності – міри ємності простору гіпотез. Більша кількість параметрів означає вищу ємність моделі та здатність апроксимувати більш складні функції. Однак існує фундаментальний компроміс між ємністю моделі та узагальнюючою здатністю надмірно складні моделі схильні до перенавчання, особливо при обмеженій кількості навчальних даних.

Теорія статистичного навчання дає нам принцип для надійного навчання моделі кількість навчальних прикладів повинна бути значно більшою за кількість параметрів. Емпірично встановлено, що для нейронних мереж це співвідношення може бути порядку 101 або навіть менше при використанні регуляризації. У випадку медичного датасету з 1200 прикладами це означає, що модель з 2 мільйонами параметрів знаходиться на межі того, що може бути надійно навчено без серйозного ризику перенавчання. Саме тому критично важливі регуляризаційні техніки дропаут, рання зупинка, які ефективно зменшують активну ємність моделі під час навчання.

Архітектурне рішення використовувати відносно невелику мережу відображає фундаментальний принцип в машинному навчанні з множини моделей, які однаково добре пояснюють дані, слід обирати найпростішу. Простіші моделі не лише швидші та більш ефективні з точки зору пам'яті, але й краще узагальнюють на нові дані та більш інтерпретовані. Для медичної діагностики, де прозорість та надійність рішень важливіші за останні кілька відсотків точності, цей підхід є оптимальним.

Можливість навчання на стандартному обладнанні без потужних графічних процесорів має не лише практичне, але й стратегічне значення для впровадження системи в реальних медичних закладах. Це знижує бар'єр входу для медичних установ, які не мають доступу до дорогої обчислювальної інфраструктури. Крім того, менші моделі легше розгортати на периферійних пристроях, що відкриває можливості для автономних діагностичних систем, які працюють локально без постійного з'єднання з хмарними серверами. Це критично важливо для збереження конфіденційності медичних даних та забезпечення доступності діагностичної допомоги в районах з обмеженою інфраструктурою зв'язку.

Параметри мережі були підібрані експериментально на основі валідаційної вибірки. Розмір прихованих одиниць LSTM встановлено на рівні 128, оскільки більші значення призводили до перенавчання, а менші - до недостатньої виразності моделі. Розмір повнозв'язних шарів також підбирався з урахуванням балансу між точністю та швидкістю навчання.

Загальна кількість параметрів мережі є помірним значенням для задач обробки природної мови. Така архітектура дозволяє навчати модель на стандартному обладнанні без потреби у потужних графічних процесорах.

2.3 Модифікація моделі для покращення точності класифікації

Для покращення якості діагностики було вирішено внести декілька модифікацій. Ці модифікації спрямовані на покращення здатності моделі виявляти найважливіші симптоми в тексті та підвищення стабільності навчання.

Одним з ключових удосконалень базової архітектури стало додавання механізму уваги після двонаправленого LSTM-шару. Механізм уваги дозволяє моделі автоматично визначати, які слова у вхідному тексті мають найбільше значення для передбачення діагнозу. Замість того, щоб однаково обробляти всі слова в описі симптомів, система навчається виділяти найбільш інформативні частини тексту. Наприклад, при діагностуванні мігрені механізм уваги може надавати більшу вагу

словам "біль голови" та "світлобоязнь", ніж загальним фразам. Це досягається шляхом обчислення коефіцієнтів уваги для кожного прихованого стану LSTM, які потім використовуються для формування зваженого контекстного вектора.

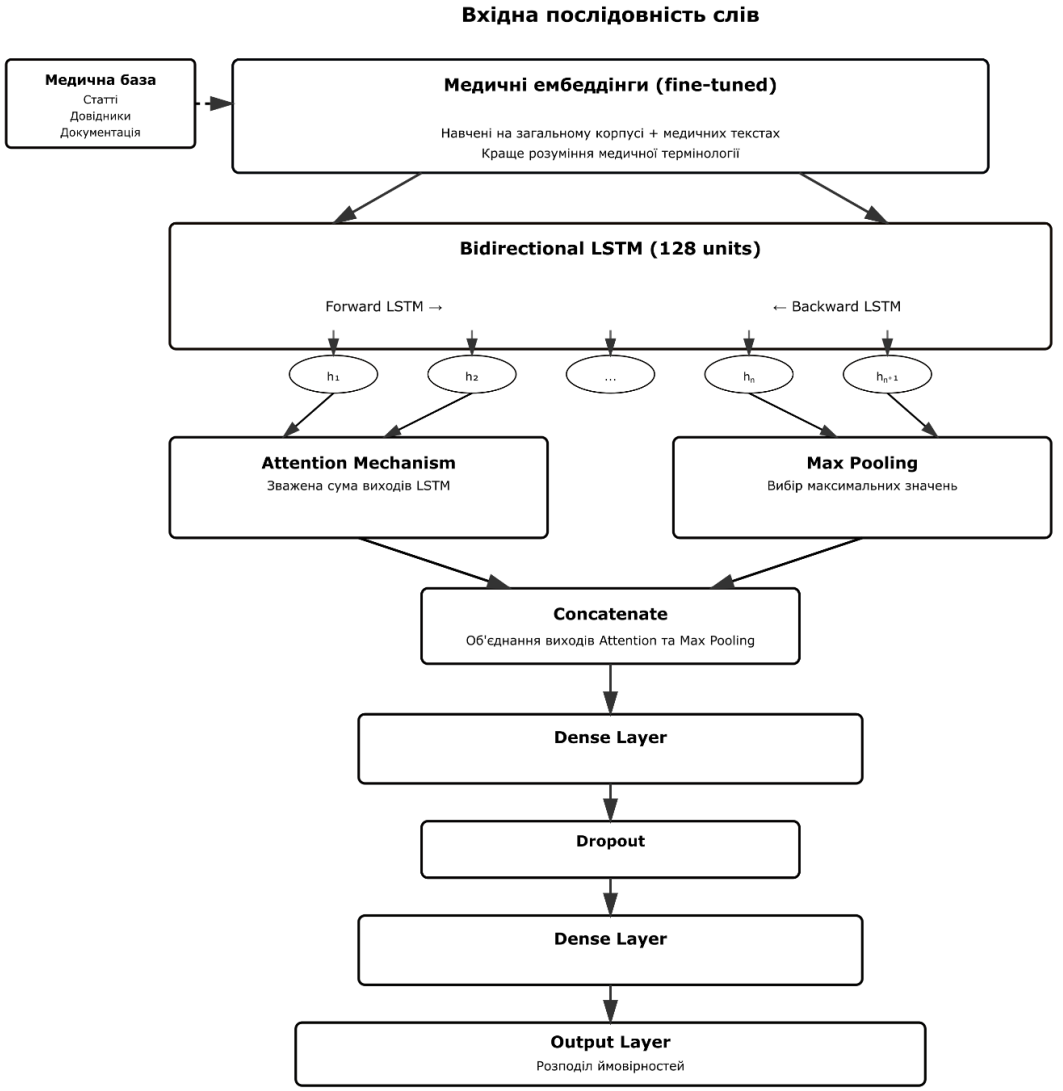


Рисунок 2.3 – Модифікована архітектура з додатковими компонентами

Такий підхід не лише покращує точність класифікації, але й робить рішення моделі більш інтерпретованими, оскільки можна візуалізувати, на які симптоми система звернула найбільшу увагу.

Для додаткової агрегації інформації з послідовності прихованих станів LSTM було додано шар макспулінг. Цей шар виконує операцію вибору максимальних значень по кожній розмірності з усіх прихованих станів послідовності. Така операція

дозволяє захопити найбільш виражені ознаки в тексті незалежно від їхнього положення. Якщо якийсь симптом згадується лише один раз, але має важливе значення для діагнозу, макспулінг гарантує, що ця інформація не буде втрачена при агрегації. Результати роботи механізму уваги та макспулінг об'єднуються за допомогою операції конкатенації, формуючи більш повне представлення тексту. Механізм уваги надає інформацію про важливість різних симптомів, тоді як макспулінг забезпечує виявлення найбільш інтенсивних проявів захворювання. Таке комбінування двох різних способів агрегації дозволяє моделі використовувати переваги обох підходів.

Важливою модифікацією процесу навчання стало впровадження регуляризації через дропаут. Після шарів згортки інформації та між повнозв'язними шарами додаються дропаут-шари з коефіцієнтом відключення 0.3. Під час навчання ці шари випадково відключають 30% нейронів на кожній ітерації, що змушує мережу навчатися більш стійким представленням даних. Це запобігає ситуації, коли модель надмірно адаптується до особливостей навчальної вибірки і погано узагальнює на нові приклади. Дропаут особливо важливий для медичних задач, де описи симптомів можуть мати значну варіативність формулювань. Модель має навчитися розпізнавати захворювання за різними способами опису симптомів, а не запам'ятовувати конкретні формулювання з навчальних даних.

Додатково було реалізовано механізм ранньої зупинки навчання для підвищення якості фінальної моделі. Якщо показники точності на валідаційній вибірці не покращуються протягом 10 послідовних епох, процес навчання автоматично завершується. Це запобігає надмірному налаштуванню моделі на тренувальні дані та економить обчислювальні ресурси. При цьому система постійно відстежує якість роботи на валідації та зберігає версію моделі з найкращими результатами. Саме ця найкраща версія використовується для фінального тестування та практичного застосування, навіть якщо навчання продовжувалося після досягнення оптимуму. Такий підхід гарантує, що модель не буде перенавчена і збереже здатність добре працювати на нових, раніше не бачених даних.

Очікується, що всі ці модифікації в сукупності дозволять підвищити точність класифікації порівняно з базовою архітектурою LSTM. Механізм уваги забезпечує виявлення найважливіших симптомів, макспулінг гарантує збереження інформації про найбільш виражені прояви захворювання, а дропаут разом з ранньою зупинкою запобігають перенавчанню моделі. Особливо важливо, що такі модифікації мають покращити роботу системи на складних випадках, де симптоми можуть бути описані по-різному або де декілька захворювань мають схожі прояви.

2.4 Формування та підготовка навчальних даних

Якість роботи нейронної мережі значно залежить від якості та кількості навчальних даних. Для навчання моделі діагностики захворювань необхідно сформувати датасет, який містить описи симптомів та відповідні їм діагнози. Джерелами таких даних можуть бути медичні бази знань, електронні медичні карти, описи захворювань з медичних довідників.

Для формування навчального датасету було використано датасет Symptom2Disease. Це відкриті медичні датасети з платформи Kaggle, які містять описи симптомів англійською мовою. Використовувалися описи захворювань медичних довідників та енциклопедій. Було зібрано синтетичні приклади описів симптомів, згенеровані на основі медичних знань. Процес збору та структурування даних включав декілька етапів. На першому етапі відбувався пошук та завантаження вихідних даних з різних джерел. Далі дані перевірялися на коректність та відповідність медичним стандартам. Медичні працівники переглядали зібрані описи та виправляли помилки або неточності. На третьому етапі дані приводилися до єдиного формату, де кожен запис містить текстовий опис симптомів та мітку класу захворювання.

Фінальний датасет містить інформацію про 23 найпоширеніших захворювань. Для кожного захворювання було зібрано від 50 до 200 прикладів описів симптомів різної довжини та деталізації. Загальна кількість прикладів у датасеті становить 1200

записів. Довжина текстових описів варіюється від 20 до 150 слів, що відповідає типовим описам симптомів у реальних умовах.

Попередня обробка даних тексту є важливим першип етапом підготовки даних. Вона включає декілька послідовних кроків, які приводять сирий текст до формату, придатного для обробки нейронною мережею. Перший крок - це токенізація, тобто розділення тексту на певні слова або токени. використовується спеціалізований токенізатор, який враховує особливості морфології та пунктуації.

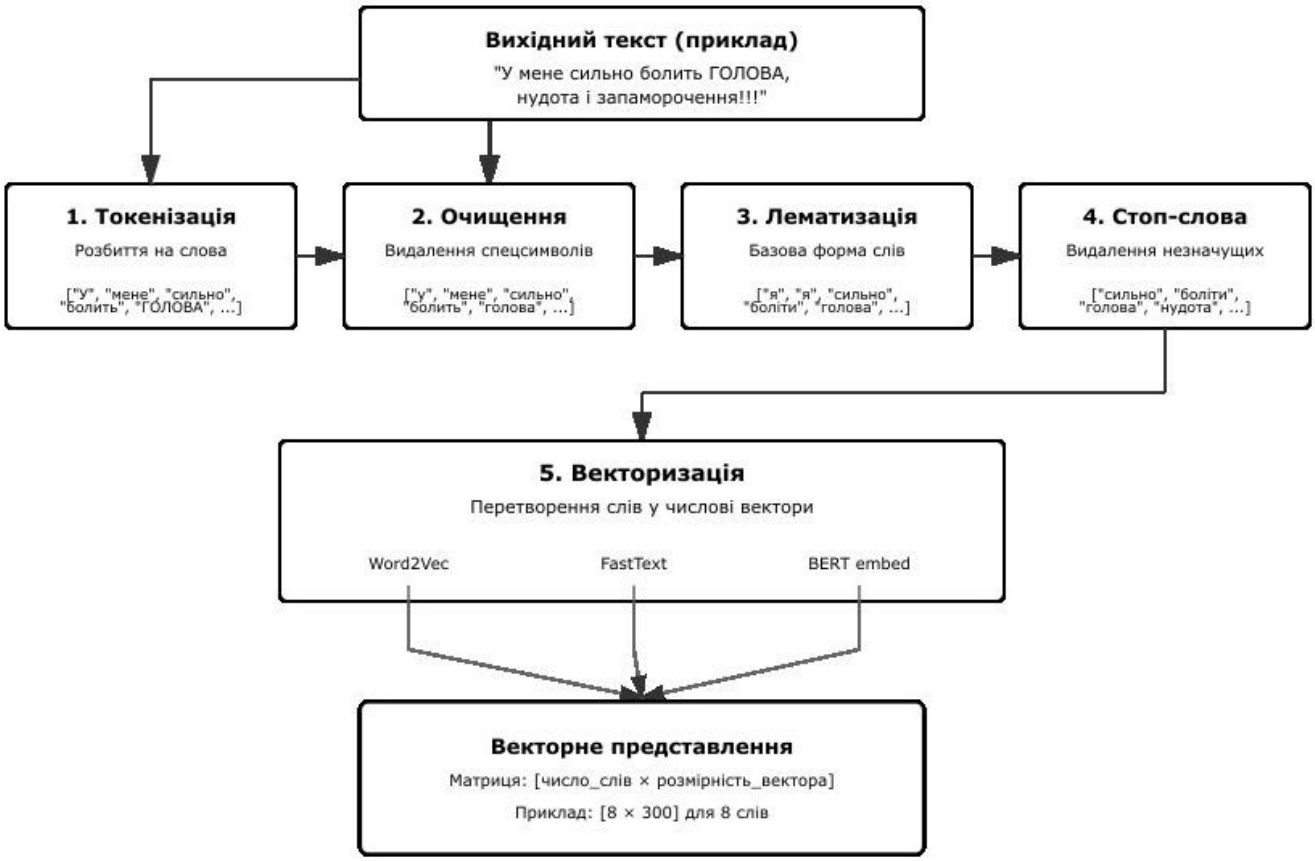


Рисунок 2.4 – Етапи попередньої обробки текстових даних

Другий крок - це очищення тексту від непотрібних символів, таких як розділові знаки, цифри або спеціальні символи. Також текст приводиться до нижнього регістру для уніфікації.

Третій крок - лематизація, тобто приведення слів до їх головної словникової форми. Наприклад, слова "болить", "боліло", "болітиме" приводяться до форми

"боліти". Це зменшує розмірність словника та допомагає мережі краще узагальнювати різні форми одного слова. Для лематизації використовується бібліотека pymorphy2.

Четвертий крок - видалення стоп-слів, тобто часто вживаних слів, які не несуть значущої інформації для класифікації. До стоп-слів відносяться прийменники, сполучники, займенники тощо. Список стоп-слів містить близько 400 слів.

Після попередньої обробки датасет розділяється на вибірки. При цьому дотримується стратифікація, тобто у кожній вибірці зберігається приблизно однакова пропорція класів. Це важливо для коректної оцінки якості моделі на всіх типах захворювань.

Навчальна вибірка береться безпосередньо для моделі, тобто для оновлення ваг нейронної мережі. Валідаційна вибірка використовується для контролю процесу навчання та вибору найкращих гіперпараметрів моделі. Тестова вибірка береться лише один раз наприкінці для завершальної оцінки якості роботи моделі. Важливо, що тестова вибірка не використовується ні для навчання, ні для налаштування моделі, щоб оцінка була об'єктивною.

2.5 Критерії та метрики оцінювання роботи методу

Для об'єктивної оцінки якості роботи запропонованого методу необхідно визначити відповідні метрики та критерії. Вибір метрик залежить від специфіки задачі та вимог до системи діагностики. У медичних застосуваннях особливо важливо мінімізувати кількість помилкових негативних результатів, тобто випадків, коли серйозне захворювання не розпізнається системою.

Основною метрикою для оцінки якості класифікації медичних даних є точність (accuracy), яка показує скільки правильно класифікованих прикладів серед усіх прикладів. Точність обчислюється за формулою:

$$Accuracy = (TP + TN)/(TP + TN + FP + FN) \quad (2.1)$$

де TP (True Positive) - кількість правильно розпізнаних позитивних прикладів, TN (True Negative) - кількість правильно розпізнаних негативних прикладів, FP (False Positive) - кількість помилково розпізнаних позитивних прикладів, FN (False Negative) - кількість пропущених позитивних прикладів.

Однак для незбалансованих даних, де деякі класи зустрічаються значно рідше за інші, точність може бути оманливою метрикою. Тому додатково використовуються метрики прецизійність (precision) та повнота (recall).

Прецизійність показує, яка частка об'єктів, класифікованих як позитивні, дійсно є позитивними:

$$Precision = TP/(TP + FP) \tag{2.2}$$

Ця метрика важлива для оцінки кількості помилкових визначень, коли система діагностує захворювання, якого насправді немає.

Повнота показує, яку частку позитивних об'єктів вдалося знайти:

$$Recall = TP / (TP + FN) \tag{2.3}$$

Ця метрика критична для медичних застосувань, оскільки показує, скільки реальних захворювань система змогла виявити.

Для збалансованої оцінки використовується F1-міра, яка є гармонійним середнім між прецизійністю та повнотою:

$$F1 - score = 2 \times (Precision \times Recall)/(Precision + Recall) \tag{2.4}$$

F1-міра особливо корисна для порівняння різних моделей, оскільки враховує як точність, так і повноту класифікації.

Для задачі багатокласової класифікації всі ці метрики обчислюються для певного класу окремо, а потім агрегуються одним із способів. Макро-усереднення обчислює середнє значення метрики по всіх класах без врахування їх розміру, що дає рівну вагу кожному захворюванню. Мікро-усереднення враховує загальну кількість правильних та неправильних передбачень по всіх класах, що дає більшу вагу частішим захворюванням. Окрім базових метрик класифікації, для оцінки медичної системи важливо враховувати додаткові критерії. Одним із таких критеріїв є час роботи системи. Для практичного використання важливо, щоб аналіз опису симптомів займав не більше кількох секунд. Тому вимірюється середній час обробки одного запиту на стандартному обладнанні. Ще одним важливим критерієм є стабільність роботи системи. Модель повинна показувати стабільні результати на різних типах вхідних даних, незалежно від стилю написання чи довжини опису. Для оцінки стабільності використовується стандартне відхилення метрик на різних підмножинах тестової вибірки.

Також важливо оцінити роботу системи окремо для різних категорій захворювань. Наприклад, можна виділити групи гострих захворювань, хронічних захворювань, інфекційних захворювань тощо. Для кожної групи обчислюються окремі метрики, що дозволяє виявити добрі та погані сторони системи.

Для рідкісних захворювань, які представлені невеликою кількістю прикладів у навчальній вибірці, особливо важлива метрика recall. Навіть якщо система іноді помилково діагностує рідкісне захворювання, важливіше, щоб вона не пропускала реальні випадки таких захворювань

Додатковою метрикою є top-k accuracy, яка показує, чи потрапляє правильний діагноз у топ-k найбільш ймовірних передбачень системи. Зазвичай розглядається top-3 або top-5 accuracy. Це важливо, оскільки в реальному використанні система може надавати медичному працівнику список з кількох найбільш ймовірних діагнозів, а не тільки один найімовірніший.

Для оцінки калібрації ймовірностей використовується метрика Brier score, яка показує, наскільки добре передбачені ймовірності відповідають реальним частотам

класів. Добре калібрована модель має давати ймовірність близько 0.8 для тих передбачень, які виявляються правильними у 80% випадків.

Також розраховується матриця помилок (confusion matrix), яка показує, які класи система плутає між собою найчастіше. Це дозволяє виявити проблемні пари захворювань з схожими симптомами та спланувати додаткові заходи для їх розрізнення Важливим аспектом є порівняння запропонованого методу з базовими підходами. Як підхід можуть використовуватися прості методи на зразок наївного баєсівського класифікатора або логістичної регресії. Порівняння проводиться на одному і тому ж тестовому датасеті з використанням однакових метрик.

Для статистичної оцінки значущості різниці між методами використовується метод крос-валідації. Датасет розбивається на 5 частин, і модель навчається 5 разів, кожного разу використовуючи 4 частини для навчання та одну для тестування. Це дозволяє отримати розподіл значень метрик та обчислити довірчі інтервали.

Окрім кількісних метрик, важливо також якісно оцінити роботу системи. Для цього проводиться аналіз помилок, коли детально розглядаються приклади, на яких система помилилася. Це допомагає зрозуміти причини помилок та визначити напрямки для подальшого покращення моделі. Ще одним критерієм оцінки є інтерпретовність результатів. За допомогою механізму уваги можна визначити, які слова в описі симптомів мали найбільший вплив на рішення системи. Це дозволяє медичним працівникам краще розуміти логіку роботи системи та приймати більш обґрунтовані рішення.Для оцінки надійності системи також важливо протестувати її на даних з помилками або шумом. Наприклад, можна додати до тестових описів орфографічні помилки або неточності та перевірити, наскільки це вплине на якість роботи. Стійка до помилок система має показувати лише невелике погіршення результатів при наявності шуму у вхідних даних.

Також необхідно оцінити справедливість роботи системи для різних груп пацієнтів. Модель не повинна показувати суттєво різну якість для різних вікових груп, статей або інших демографічних характеристик. Для цього аналізуються метрики окремо для різних підгруп пацієнтів у тестовій вибірці.

Важливим критерієм є можливість оновлення та вдосконалення моделі. Система має бути спроектована таким чином, щоб можна було легко додавати нові захворювання або оновлювати модель на нових даних без повного перенавчання з нуля. Це забезпечує довгострокову практичну цінність розробленого методу.

Висновок до розділу 2

У розділі було детально описано запропонований метод діагностики захворювань за описом симптомів з використанням обробки природної мови. Метод базується на архітектурі рекурентних нейронних мереж типу LSTM з додаванням механізму уваги, що дозволяє ефективно обробляти текстові описи різної довжини та складності.

Концепція методу передбачає послідовну обробку вхідного тексту через етапи токенізації, лематизації, векторизації та класифікації. Архітектура нейронної мережі складається з шару ембеддінгів, двонаправленого LSTM-шару з 128 прихованими одиницями, шару уваги, двох повнозв'язних шарів з 256 та 128 нейронами відповідно, та вихідного шару з функцією активації софтмакс.

Для покращення якості роботи базової моделі було запроваджено декілька модифікацій. Додано механізм макспулінг для додаткової агрегації інформації з LSTM-шару. Використано техніку поступового розморожування шарів під час навчання. Застосовано зважену функцію втрат для кращого навчання на рідкісних захворюваннях.

Визначено критерії та метрики для оцінювання роботи методу. Основними метриками є точність, прецизійність, повнота та F1-міра. Додатково враховуються час роботи системи, стабільність результатів, якість роботи на рідкісних захворюваннях та інтерпретовність передбачень. Запропоновано використовувати порівняння з базовими методами та крос-валідацію для статистичної оцінки значущості результатів.

Розділ 3 Реалізація програмної системи діагностики захворювань

3.1 Загальна структура програмної реалізації методу

Програмне втілення методу діагностики захворювань за описом симптомів побудована за модульним принципом, що є стандартним підходом при розробці складних програмних комплексів. Така архітектура дозволяє незалежно розробляти та тестувати окремі компоненти системи, а також значно спрощує подальше розширення функціональності та внесення змін до певних частин програми без необхідності переробки всієї системи.

Система складається з 5 основних модулів, кожен з яких виконує свою специфічну функцію у процесі обробки запиту користувача. Центральним компонентом є модуль обробки природної мови, який відповідає за первинну обробку вхідного тексту. Цей модуль виконує токенізацію вхідного тексту, тобто розбиття даних на певні слова та символи, а також приведення слів до базової форми через процес лематизації. Лематизація особливо важлива для української мови з її складною системою відмінювання, оскільки дозволяє привести різні словоформи до єдиної базової форми.

Модуль векторизації відповідає за перетворення оброблених текстових даних у числові вектори , які можуть бути використані нейронною мережею для подальшої обробки. Кожне слово у словнику системи має своє унікальне векторне представлення, яке відображає його семантичне значення. Слова з подібним значенням мають схожі векторні представлення, що дозволяє мережі краще узагальнювати знання про симптоми.

Модуль класифікації містить реалізацію архітектури LSTM з механізмом уваги та здійснює власне передбачення діагнозу на основі векторизованого представлення тексту. Цей модуль є найбільш складним з точки зору обчислень і містить нейронну мережу з багатьма шарами, кожен з яких виконує специфічне перетворення даних.

Модуль управління даними забезпечує завантаження навчених моделей та словників з дискового сховища, а також збереження нових версій моделі після завершення процесу навчання. Цей модуль також відповідає за перевірку цілісності завантажених даних та коректну обробку помилок при роботі з файловою системою.

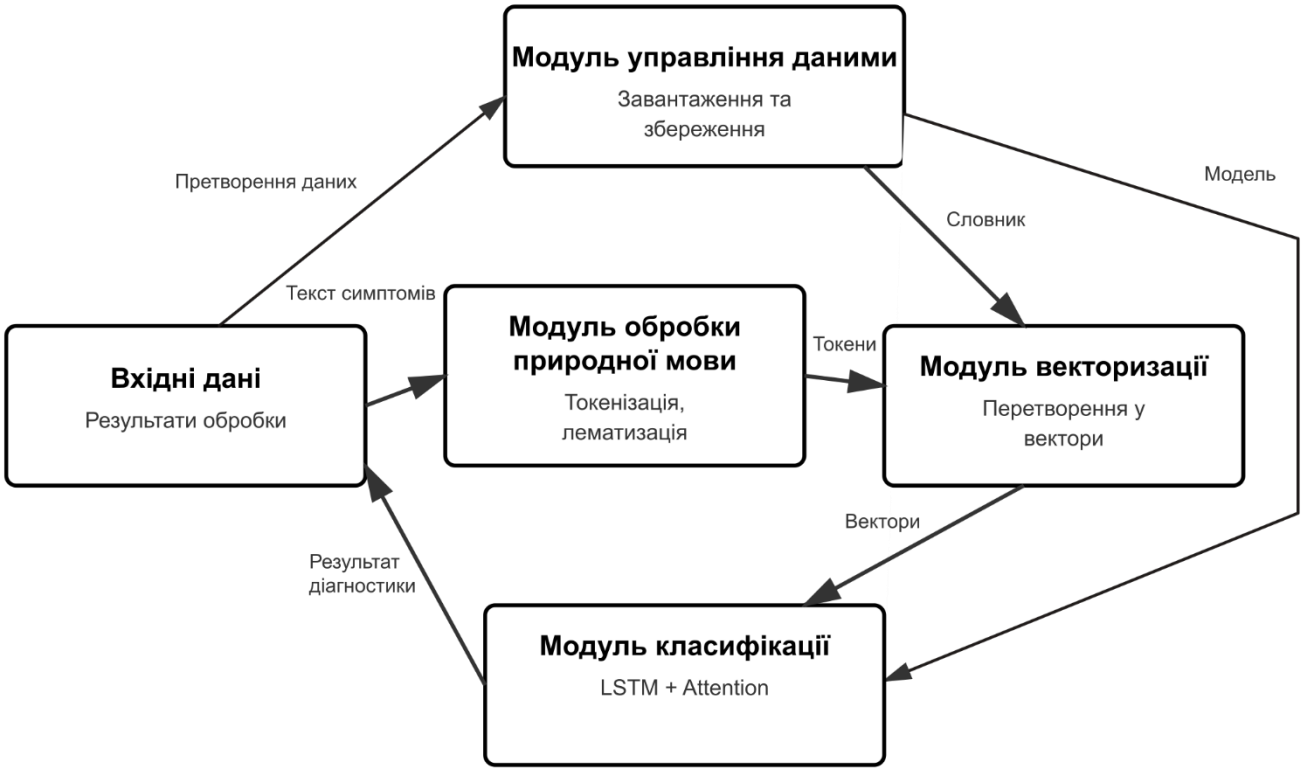


Рисунок 3.1 – Схема потоку даних методу діагностики

Модуль даних відповідає за взаємодію з системою, приймає текстовий опис симптомів через відповідний інтерфейс та відображає результати діагностики у зручному для сприйняття форматі. Модуль також забезпечує валідацію вхідних даних та формування інформативних повідомлень про помилки.

Взаємодія між модулями організована у вигляді послідовного конвеєра обробки даних. Текст від користувача спочатку передається через модуль обробки мови, де відбувається його очищення та нормалізація. Потім оброблені дані надходять до модуля векторизації, який перетворює текстові токени у числові вектори. Ці вектори подаються на вхід модуля класифікації, де нейронна мережа виконує передбачення діагнозу. Результат класифікації повертається користувачу через

інтерфейсний модуль у вигляді списку найбільш ймовірних захворювань з відповідними коефіцієнтами впевненості.

3.2 Реалізація модуля попередньої обробки тексту

Модуль структури обробки тексту реалізовано у вигляді класу TextPreprocessor, який інкапсулює всі необхідні операції для підготовки вхідного тексту до класифікації. Цей клас забезпечує уніфікований інтерфейс для перетворення текстових даних та приховує деталі реалізації від інших компонентів системи. Головними завданнями цього модуля є токенізація тексту, нормалізація символів, лематизація слів та фільтрація стоп-слів, які не несуть корисної інформації для діагностики.

Клас TextPreprocessor містить чотири основні компоненти, які працюють разом для досягнення якісної обробки тексту. Об'єкт Tokenizer виконує розбиття тексту на певні токени з урахуванням особливостей медичної термінології. Цей компонент використовує складні регулярні вирази для правильного розпізнавання медичних термінів, абревіатур та складних конструкцій. Наприклад, терміни типу 38.5°C повинні оброблятися як єдині цілі, а не розбиватися на окремі частини.

Структуру класів модуля обробки тексту та їх взаємозв'язки детально представлено на рисунку 3.2, який показує діаграму класів з атрибутами та методами кожного компонента.

Об'єкт Lemmatizer використовує бібліотеку pymorphy2 для приведення слів до словникової форми, що дозволяє розпізнавати різні форми одного і того ж терміну. Ця бібліотека містить морфологічний аналізатор для української мови, який враховує всі особливості відмінювання та дієвідмінювання. Наприклад, слова "біль", "болю", "болем" приводяться до єдиної базової форми "біль", що дає змогу системі розпізнавати симптом незалежно від граматичної форми його опису.

Об'єкт StopWordsFilter видаляє зі списку токенів службові слова, які не несуть смислового навантаження для діагностики. До таких слів належать прийменники,

сполучники, займенники та інші частини мови, які використовуються для граматичної структури речення, але не містять інформації про симптоми. Список стоп-слів зберігається у окремому конфігураційному файлі, що дозволяє легко його модифікувати без зміни коду програми.

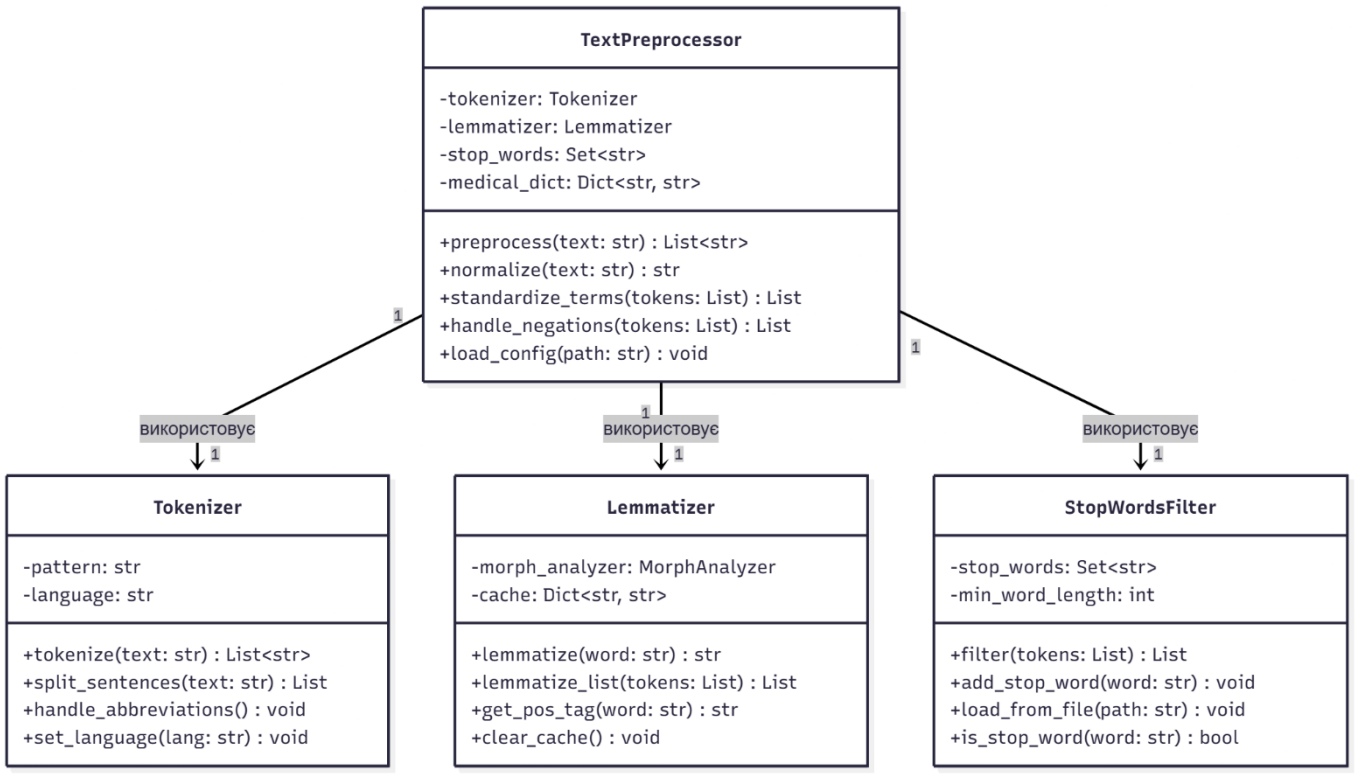


Рисунок 3.2 – Діаграма класів модуля попередньої обробки тексту

Окремо варто відзначити використання спеціалізованого словника медичних термінів, який є важливою частиною модуля обробки тексту. Цей словник містить пари синонімів та різних форм запису одних і тих самих симптомів, що дозволяє системі розуміти різні варіанти формулювань. Наприклад, терміни "біль голови", "головний біль", "цефалгія" та "краніалгія" приводяться до єдиної стандартної форми "головний біль". Аналогічно, "підвищена температура", "лихоманка", "гарячка" та "гіпертермія" розпізнаються як синоніми. Словник також містить відповідності між розмовними та медичними термінами, що дозволяє системі розуміти як професійний медичний, так і повсякденний опис симптомів.

Метод preprocess є головним методом класу та виконує послідовну обробку вхідного тексту через усі компоненти. Спочатку текст нормалізується шляхом приведення до нижнього регістру та видалення зайвих символів. Потім виконується токенізація, після чої застосовується лематизація до кожного токену. Далі відбувається стандартизація медичних термінів за допомогою спеціалізованого словника, і нарешті фільтруються стоп-слова. Результатом роботи методу є список нормалізованих токенів, готових до векторизації.

Метод handle_negations виконує важливу функцію аналізу контексту заперечень у тексті. Наявність чи відсутність певних симптомів має критичне значення для діагностики. Якщо користувач пише "немає болю в грудях" або "температура не підвищена", система повинна правильно інтерпретувати ці заперечення. Метод виявляє заперечувальні частки та прив'язує їх до наступних симптомів, створюючи спеціальні токени типу "НЕ_біль_груди" або "НЕ_підвищена_температура".

3.3 Структура модуля нейромережевої класифікації

Модуль класифікації реалізує архітектуру нейронної мережі LSTM з механізмом уваги та додатковими компонентами для покращення якості діагностики. Він складається з ієрархії класів, кожен з яких відповідає за певний шар або компонент мережі. Така об'єктно-орієнтована структура робить код більш читабельним, полегшує тестування окремих компонентів та спрощує подальше розширення архітектури.

Клас EmbeddingLayer забезпечує перетворення індексів слів у векторному представлені. Кожне слово у словнику системи асоціюється з унікальним числовим індексом, і цей шар виконує відображення індексів у відповідні вектори. Векторні представлення слів можуть бути ініціалізовані випадковими значеннями та навчені з нуля, або ж можна завантажити попередньо навчені вектори та дотренувати їх на медичній термінології. Метод load_pretrained дозволяє завантажити вектори, навчені

на великих корпусах текстів, що забезпечує кращі початкові представлення слів. Метод fine_tune надає можливість додаткового налаштування цих векторів на специфіку медичних даних.

Клас BiLSTMLayer реалізує двонаправлену рекурентну мережу довгої короткочасної пам'яті з 128 прихованими одиницями в кожному напрямку. Двонаправлена обробка означає, що послідовність токенів обробляється одночасно зліва направо та справа наліво, що дозволяє мережі враховувати контекст як попередніх, так і наступних слів при аналізі кожного токену. Це особливо важливо для розуміння медичних описів, де значення симптому може залежати від контексту. Наприклад, фраза "біль після їжі" вказує на проблеми з травленням, тоді як "біль перед їжею" може свідчити про інші захворювання.

Архітектура LSTM вирішує проблему зникаючого градієнта, яка властива звичайним рекурентним мережам. Кожна LSTM-комірка містить систему воріт (input gate, forget gate, output gate), які контролюють потік інформації та дозволяють мережі зберігати важливу інформацію на довгих послідовностях. Це критично важливо для обробки медичних описів, які можуть бути досить довгими та містити багато різних симптомів.

Клас AttentionMechanism є одним з ключових компонентів архітектури, що реалізує механізм уваги для виявлення найважливіших частин вхідного тексту. Механізм уваги обчислює ваги важливості для кожного слова у вхідній послідовності, базуючись на прихованих станах LSTM. Метод compute_scores розраховує оцінки важливості для кожної позиції у послідовності. Метод apply_софтмакс нормалізує ці оцінки так, щоб їхня сума дорівнювала одиниці, перетворюючи їх на ймовірнісний розподіл уваги. Метод compute_context використовує обчислені ваги для формування зваженого контекстного вектора, який агрегує інформацію з усієї послідовності, приділяючи більше уваги важливим словам.

Механізм уваги робить модель більш інтерпретованою. Після класифікації можна візуалізувати, які саме слова отримали найбільші ваги уваги, тобто на які

симптоми система звернула найбільше уваги при прийнятті рішення. Це дозволяє лікарям краще розуміти логіку роботи системи та перевіряти коректність її висновків.

Діаграму класів модуля класифікації з детальним описом атрибутів, методів та зв'язків між класами показано на рисунку 3.3.

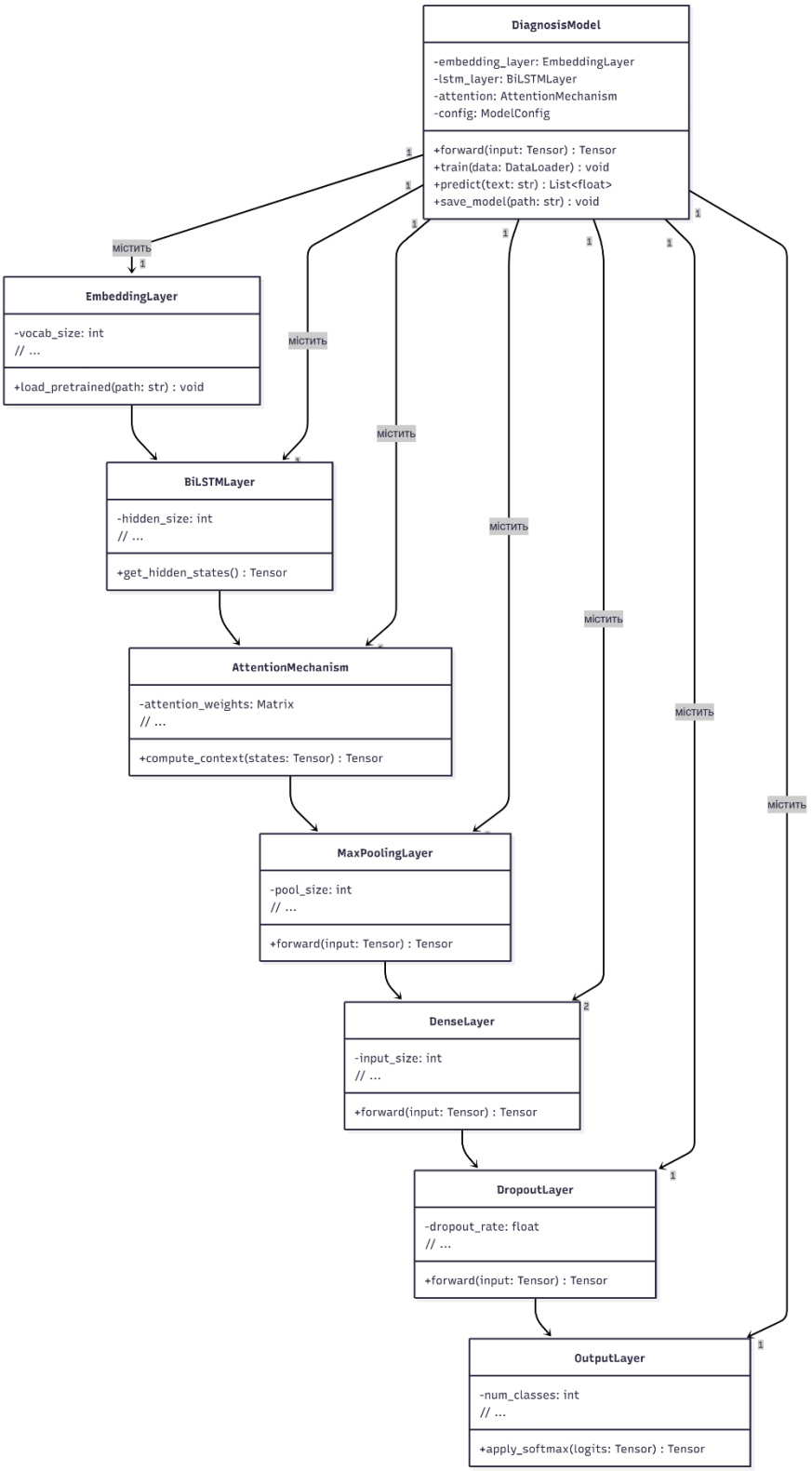


Рисунок 3.3 – Діаграма класів модуля класифікації

Клас MaxPoolingLayer виконує додаткову агрегацію інформації шляхом вибору максимальних значень з прихованих станів по кожній розмірності. Ця операція дозволяє захопити найбільш виражені ознаки в тексті незалежно від їхнього положення у послідовності. Якщо якийсь важливий симптом згадується лише один раз у описі, макспулінг гарантує, що інформація про нього не буде втрачена при агрегації. Метод get_max_indices зберігає позиції максимальних елементів, що може бути корисним для інтерпретації результатів. Результати роботи механізму уваги та макспулінг об'єднуються за допомогою операції конкатенації. Таким чином, модель отримує два різні погляди на вхідні дані зважене представлення від механізму уваги та представлення найбільш виражених ознак від макспулінг. Це комбінування дозволяє використовувати переваги обох підходів одночасно.

Два екземпляри класу DenseLayer реалізують повнозв'язні шари з 256 та 128 нейронами відповідно. Ці шари виконують нелінійне перетворення агрегованих ознак і використовують функцію ReL, що перетворює від'ємні значення в нуль, а додатні залишає без змін. Функція ReLU добре зарекомендувала себе у глибоких нейронних мережах, оскільки вирішує проблему зникаючого градієнта та прискорює навчання. Перший повнозв'язний шар розширює простір ознак до 256 вимірів для виявлення складних закономірностей, а другий шар звужує представлення до 128 вимірів, виділяючи найбільш важливі ознаки.

Клас Layer застосовує регуляризацію з коефіцієнтом 0.3 для запобігання перенавчанню моделі на тренувальних даних. Під час навчання цей шар випадково відключає 30% нейронів на кожній ітерації, що змушує мережу не покладатися на конкретні нейрони, а навчатися розподіленим представленням. У режимі передбачення дропаут вимикається, і використовуються всі нейрони. Метод set_training_mode дозволяє перемикатися між режимами навчання та тестування. Регуляризація особливо важлива для медичних задач, оскільки навчальна вибірка може бути обмеженою, і модель схильна до перенавчання.

Клас OutputLayer формує фінальні передбачення для 24 класів захворювань. Він містить повнозв'язний шар, який перетворює 128-вимірне представлення у 24-вимірний вектор логітів (нормалізованих оцінок для кожного класу). Метод apply_софтмакс застосовує функцію софтмакс до логітів, перетворюючи їх у розподіл ймовірностей, де кожне значення знаходиться в діапазоні від 0 до 1, а сума всіх ймовірностей дорівнює одиниці. Метод compute_loss обчислює значення функції втрат під час навчання, порівнюючи передбачені ймовірності з реальними мітками класів.

Головний клас DiagnosisModel об'єднує всі описані компоненти в єдину архітектуру нейронної мережі. Метод forward реалізує прямий прохід даних через всі шари мережі в правильній послідовності спочатку через шар ембеддінгів, потім через BiLSTM, далі паралельно через механізм уваги та макспулінг, після конкатенації через повнозв'язні шари з дропаут, і нарешті через вихідний шар. Методи train та predict забезпечують відповідно навчання моделі на тренувальних даних та виконання передбачень на нових прикладах. Метод save_model зберігає навчену модель на диск для подальшого використання.

3.4 Процес навчання та оцінювання моделі

Процес навчання моделі організовано у вигляді циклу по епохах, на кожній з яких виконується обробка всього навчального набору даних. Епоха представляє собою один повний прохід через всі тренувальні приклади. Зазвичай для досягнення хороших результатів потрібно провести навчання протягом декількох десятків епох, поки модель не навчиться правильно класифікувати захворювання. На початку кожної епохи дані перемішуються випадковим чином для запобігання запам'ятовуванню порядку прикладів. Це важливо, оскільки якщо модель буде бачити приклади в одному і тому ж порядку, вона може навчитися використовувати цю інформацію замість того, щоб вивчати реальні закономірності у даних.

Перемішування даних робить процес навчання більш стабільним і покращує здатність моделі до узагальнення.

Навчальні дані подаються до моделі невеликими пакетами розміром 32 або 64 приклади, залежно від обєму доступної оперативної пам'яті. Обробка даних пакетами має декілька переваг порівняно з обробкою по одному прикладу. Це дозволяє ефективно використовувати паралельні обчислення на графічних процесорах. Оновлення параметрів моделі на основі пакету прикладів є більш стабільним, ніж на основі окремих прикладів, оскільки усереднюються випадкові флуктуації.

Для кожного пакету виконується прямий прохід через мережу з обчисленням передбачень. Вхідні дані проходять послідовно через всі шари моделі. Результатом прямого проходу є вектор ймовірностей для кожного прикладу в пакеті, де кожна компонента вектора представляє ймовірність відповідного захворювання.

Після отримання передбачень обчислюється значення функції втрат шляхом порівняння передбачених ймовірностей з реальними мітками класів. Використовується функція втрат, яка вимірює відмінність між передбаченим розподілом ймовірностей та реальним класом. Чим більше передбачена ймовірність правильного класу, тим менше значення функції втрат. Мета навчання полягає у мінімізації цієї функції втрат на всіх навчальних даних.

Далі виконується процедура зворотного поширення для розрахунку градієнтів всіх параметрів мережі. Алгоритм зворотного поширення обчислює частинні похідні функції втрат по кожному отрибуту моделі, використовуючи правило ланцюга диференціювання. Ці градієнти показують, в якому напрямку і наскільки потрібно змінити кожен параметр для зменшення помилки.

Оптимізатор використовує обчислені градієнти для оновлення ваг мережі. В системі застосовується адаптивний оптимізатор, який автоматично підбирає швидкість для кожного отрибуту окремо. Це дозволяє різним частинам мережі навчатися з оптимальною для них швидкістю. Початкова швидкість навчання встановлюється досить великою для швидкого початкового прогресу, але автоматично зменшується під час навчання для більш точного налаштування

параметрів. Якщо на валідаційному наборі не спостерігається покращення метрик протягом декількох епох, швидкість навчання додатково зменшується. Це дозволяє моделі виконати більш тонке налаштування параметрів та можливо вийти з локального мінімуму функції втрат. Зменшення швидкості навчання є стандартною практикою, яка часто призводить до покращення фінальних результатів.

Після обробки всіх навчальних пакетів в епосі виконується оцінка моделі на валідаційних даних, який не використовувався під час навчання. Це дає змогу отримати неупереджену оцінку якості моделі та відстежити, чи не відбувається перенавчання. Для валідаційного набору обчислюються метрики точності, прецизійності, повноти та F1-міри.

Точність (accuracy) показує частину вірно класифікованих прикладів серед усіх прикладів. Прецизійність (precision) вимірює, яка частка прикладів, класифікованих як певне захворювання, дійсно належить до цього класу. Повнота (recall) показує, яку частку прикладів певного захворювання модель змогла правильно виявити. F1-міра є гармонічним середнім між прецизійністю а також повнотою і дає збалансовану оцінку якості класифікації.

Якщо результати на валідації покращилися порівняно з попередньою епохою, поточна версія моделі зберігається на диск як найкраща. Зберігаються всі параметри мережі, включаючи ваги всіх шарів, конфігурацію архітектури та поточний стан оптимізатора. Це дозволяє відновити роботу моделі у будь-який момент або продовжити навчання з збереженої точки.

Послідовність взаємодії компонентів системи під час процесу навчання моделі детально представлена на рисунку 3.4, який показує діаграму кооперації з послідовністю викликів методів.

Система також реалізує механізм ранньої зупинки навчання, який є важливим інструментом для запобігання перенавчанню. Якщо протягом 10 послідовних епох не спостерігається покращення метрик на валідаційному наборі, навчання автоматично завершується. Це економить обчислювальний час та запобігає ситуації, коли модель

починає надмірно адаптуватися до тренувальних даних на шкоду здатності узагальнення.

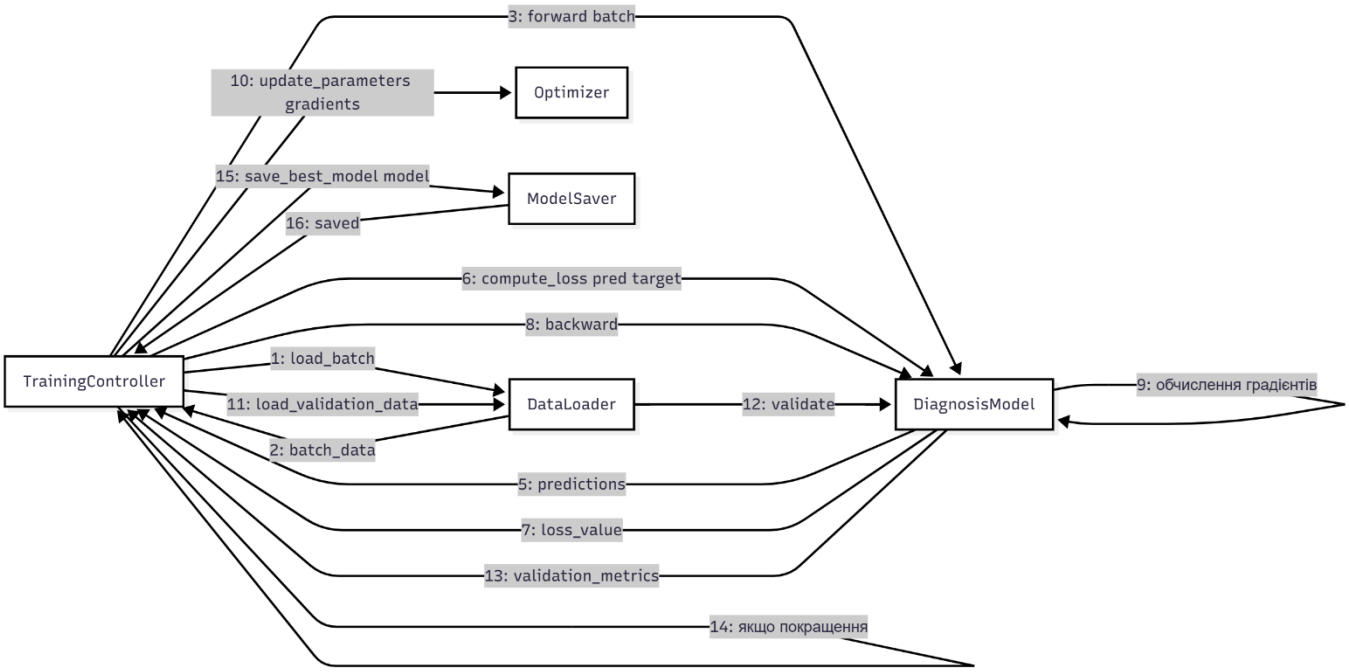


Рисунок 3.4 – Діаграма кооперації під час навчання моделі

Під час навчання система контролює норму градієнтів для виявлення потенційних проблем з нестабільністю. Якщо градієнти стають занадто великими проблема вибухаючих градієнтів, застосовується обрізання градієнтів до максимального порогового значення. Це запобігає різким стрибкам параметрів моделі, які можуть зруйнувати вже навчені представлення.

Також аналізується розподіл активацій у різних шарах мережі для перевірки правильності роботи архітектури. Якщо активації концентруються біля нуля або, навпаки, занадто великі, це може свідчити про проблеми з ініціалізацією ваг або вибором функцій активації. Моніторинг цих показників допомагає швидко виявити та усунути проблеми в процесі навчання.

Після завершення навчання виконується фінальна оцінка найкращої збереженої версії моделі на тестовому наборі даних. Тестовий набір повністю ізольований від процесу навчання і використовується лише для отримання підсумкової оцінки якості моделі. Результати на тестовому наборі найкраще

відображають те, як модель буде працювати на реальних даних у практичному застосуванні. Будується матриця помилок, яка показує для кожного захворювання, скільки разів воно було правильно розпізнане і з якими іншими захворюваннями воно плутається. Аналіз матриці помилок дозволяє виявити найбільш проблемні пари захворювань, які важко розрізнити за симптомами, та зрозуміти, де потрібні додаткові покращення.

3.5 Робота системи у режимі діагностики

У режимі діагностики система приймає від користувача текстовий опис симптомів та повертає список найбільш ймовірних захворювань з відповідними коефіцієнтами впевненості. Процес обробки запиту відбувається у декілька послідовних етапів, кожен з яких виконує свою специфічну функцію.

Спочатку система перевіряє коректність та достатність наданої інформації. Вхідний текст аналізується на предмет мінімальної довжини та наявності змістовного контенту. Якщо текст занадто короткий наприклад, менше 3 слів або не містить жодних медичних термінів чи описів симптомів, система запитує у користувача додаткові деталі. Валідація вхідних даних є важливим етапом, так як якість діагностування дуже залежить від повноти опису симптомів.

Якщо початкова перевірка успішна, прийнятий текст передається до модуля попередньої обробки. Там виконується комплексна обробка тексту, яка включає нормалізацію регістру символів, токенізацію на окремі слова, лематизацію для зведення слів до основної форми, та фільтрацію стоп-слів. Особлива увага приділяється розпізнаванню медичних термінів та їх стандартизації відповідно до словника системи.

Отримані після попередньої обробки токени стандартизуються відповідно до медичного словника для уніфікації термінології. Це означає, що різні способи опису одного і того ж симптому приводяться до єдиної стандартної форми. Наприклад, "нудота", "підкреслення" та "позиви до блювання" можуть бути стандартизовані до

єдиного терміну "нудота". Така стандартизація покращує якість подальшої класифікації.

На наступному етапі модуль векторизації перетворює кожне слово у векторне представлення . Система використовує попередньо навчений словник векторних представлень, де кожному слову відповідає його унікальний вектор. Якщо слово відсутнє у словнику системи, воно замінюється спеціальним токеном для невідомих слів (UNK - unknown), який має своє власне векторне представлення. Така стратегія дозволяє системі обробляти тексти, які містять раніше не бачені слова.

Сформована послідовність векторів передається на вхід нейронної мережі для виконання класифікації. Нейронна мережа обробляє вхідну послідовність через всі свої шари згідно з архітектурою. Спочатку векторна послідовність проходить через двонаправлений LSTM-шар, який формує контекстуалізовані представлення для всіх слів.

Далі механізм уваги визначає, які слова найбільше впливають на передбачення діагнозу. Система автоматично навчається виділяти найбільш важливі симптоми в описі. Одночасно з механізмом уваги працює шар макспулінг, який виділяє найбільш виражені ознаки незалежно від їх положення в тексті. Результати обох механізмів агрегації об'єднуються для формування комплексного представлення симптомів пацієнта. Агреговане представлення проходить через послідовність повнозв'язних шарів, які виконують нелінійне перетворення ознак для виявлення складних закономірностей. Фінальний вихідний шар виводить вектор ймовірностей для кожного з 24 можливих захворювань. Кожна компонента цього вектору представляє впевненість системи в тому, що опис симптомів відповідає певному захворюванню.

Отримані ймовірності сортуються за спадом їх значень. Користувачу відображаються топ 3-5 найбільш ймовірних діагнозів разом з відсотком впевненості для кожного з них. Така форма подання інформації дозволяє лікарю розглянути декілька варіантів та обрати найбільш відповідний на основі додаткових обстежень та свого професійного досвіду. Показ декількох варіантів є кращою практикою, ніж показ лише одного найбільш ймовірного діагнозу, оскільки медична діагностика

часто не є однозначною. Система також надає пояснення своїх рішень через аналіз ваг механізму уваги. Для кожного запропонованого діагнозу показується, які саме симптоми з опису користувача мали найбільший вплив на це передбачення. Вага уваги для кожного слова показується, наприклад, за допомогою інтенсивності кольору підсвітки. Слова, які отримали високі ваги уваги, виділяються яскравіше. Це дозволяє медичному персоналу зрозуміти логіку роботи системи та оцінити обґрунтованість її висновків.

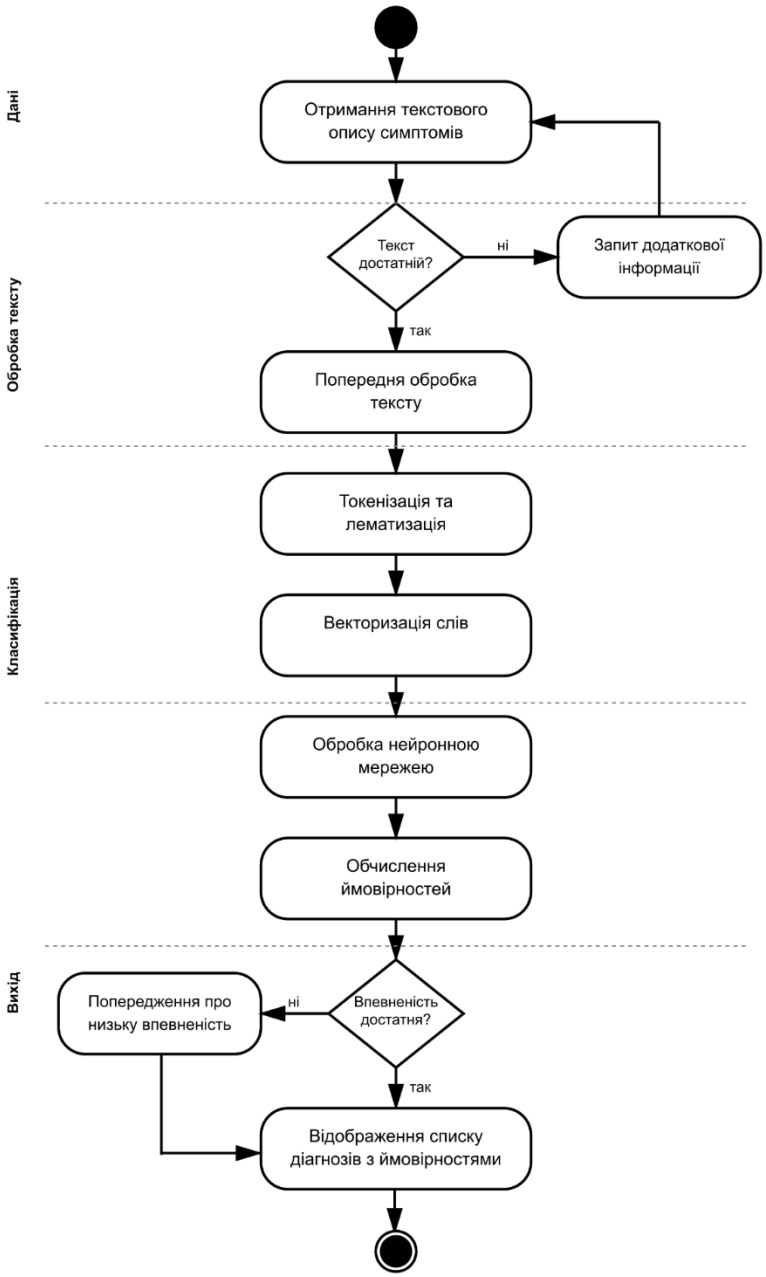


Рисунок 3.5 – Діаграма активності процесу діагностики

Всі запити користувачів разом з відповідями системи автоматично зберігаються в історії для можливості подальшого аналізу та покращення якості діагностики. Історія запитів може використовуватися для виявлення типових помилок системи, аналізу складних випадків, де система помилилася, та для збору додаткових даних для перенавчання моделі. Також історія дозволяє відстежувати динаміку симптомів конкретного пацієнта при повторних зверненнях.

Коли жодне захворювання не отримало ймовірність вище встановленого порогу впевненості наприклад, 30%, система повідомляє користувача про низьку впевненість у діагнозі. Це може свідчити про декілька ситуацій опис симптомів є неповним або нечітким, симптоми не характерні для жодного з відомих системі захворювань, або присутня комбінація симптомів, яка може вказувати на рідкісне захворювання, що не входить до списку 24 класів, на яких навчена модель.

Система включає обробку різних помилкових ситуацій для забезпечення надійності роботи. Якщо користувач відправляє порожній текст або текст, що складається лише з пробілів та розділових знаків, система видає відповідне повідомлення про помилку та пропонує ввести опис симптомів. Якщо після попередньої обробки не залишилося жодного токену наприклад, якщо текст складався лише зі стоп-слів) система також інформує про це користувача.

При появі технічних помилок під час обробки, таких як проблеми з завантаженням моделі, помилки векторизації або збої в роботі нейронної мережі, система перехоплює ці виключення та видає зрозуміле повідомлення про проблему. Це запобігає аварійному завершенню програми та дозволяє користувачу повторити спробу або звернутися до адміністратора системи.

Для забезпечення швидкої роботи системи використовується кешування часто вживаних запитів. Якщо користувач відправляє запит, ідентичний одному з попередніх, система може повернути збережений результат без повторної обробки через нейронну мережу. Це особливо корисно при тестуванні системи або у випадках, коли багато користувачів описують схожі симптоми.

Висновок до розділу 3

У розділі описано реалізацію програмної системи діагностики захворювань за описом симптомів з використанням методів обробки природної мови та глибокого навчання. Систему побудовано за модульним принципом, що забезпечує гнучкість архітектури, незалежність розробки окремих компонентів та можливість подальшого розширення функціональності без необхідності переробки всієї системи.

Модуль попередньої обробки тексту використовує спеціалізований словник медичних термінів та бібліотеку морфологічного аналізу для приведення різних формулювань симптомів до стандартного вигляду. Це дозволяє системі розпізнавати симптоми незалежно від способу їх опису користувачем, враховуючи як медичну термінологію, так і розмовні формулювання. Обробка заперечень забезпечує правильну інтерпретацію відсутності певних симптомів, що є дуже важливим для коректної діагностики.

Модуль класифікації реалізовано у вигляді ієрархії класів, кожен з яких відповідає за окремий компонент нейронної мережі. Така об'єктно-орієнтована структура робить код більш зрозумілим, спрощує тестування окремих компонентів та полегшує подальше розширення або модифікацію архітектури. Використання механізму уваги та макспулінг дозволяє моделі ефективно виявляти найважливіші симптоми та найбільш виражені прояви захворювань.

Процес навчання моделі використовує сучасні методи оптимізації з адаптивною швидкістю навчання та регуляризацію через дропаут для запобігання перенавчанню. Механізм ранньої зупинки та автоматичне збереження найкращої версії моделі забезпечують отримання оптимальних результатів без надмірного налаштування на тренувальні дані.

Розділ 4 Експериментальна перевірка методу діагностики захворювань

4.1 Організація експериментального дослідження

Для перевірки запропонованого методу діагностики захворювань було проведено комплексне експериментальне дослідження. Основною метою експериментів було визначити, наскільки точно модель може класифікувати захворювання на основі текстового опису симптомів, а також порівняти результати роботи різних варіантів архітектури.

Програмне середовище було організоване на базі операційної системи Ubuntu 22.04 LTS. Для реалізації моделі використовувалась мова Python версії 3.10.12. Основною бібліотекою для побудови нейронної мережі був фреймворк PyTorch версії 2.0.1 з підтримкою CUDA 11.8 для роботи з графічним процесором.

Додатково використовувались бібліотеки NumPy 1.24.3 для роботи з числовими масивами, scikit-learn 1.3.0 для обчислення метрик якості та розділення даних на вибірки. Для обробки природної мови застосовувалась бібліотека pymorphy2 версії 0.9.1, яка забезпечує морфологічний аналіз та лематизацію української мови.

Візуалізація результатів експериментів виконувалась за допомогою бібліотеки Matplotlib версії 3.7.1. Ця бібліотека дозволила створити графіки процесу навчання, матриці помилок та інші діаграми для наочного представлення отриманих результатів.

Методика проведення експериментів передбачала декілька етапів. На початковому етапі було підготовлено датасет з описами симптомів для 24 захворювань. Кожне захворювання представлене 50 прикладами текстових описів різної довжини та формулювання. Загальна кількість записів у датасеті становить 1200 прикладів.

Весь датасет було розділено на три частини за стандартною практикою машинного навчання. Навчальна вибірка склала 70 % від загальної кількості даних, тобто 840 записів. Ця вибірка використовувалась безпосередньо для навчання параметрів нейронної мережі. Валідаційна вибірка становила 15 % або 180 записів і

використовувалась для контролю якості навчання та підбору гіперпараметрів. Тестова вибірка також містила 15 % даних або 180 записів і застосовувалась виключно для фінальної оцінки якості моделі.

Розподіл даних виконувався з використанням стратифікованої вибірки, що гарантувало збереження пропорцій класів у кожній з трьох частин. Це означає, що кожне захворювання представлене приблизно однаковою кількістю прикладів у навчальній, валідаційній та тестовій вибірках. Для кожного експерименту дані перемішувались випадковим чином перед розподілом на вибірки. Під час навчання використовувалось фіксоване значення генератора випадкових чисел для забезпечення відтворюваності результатів. Це дозволяє повторити експерименти та отримати ідентичні результати при використанні тих самих налаштувань. Було визначено набір гіперпараметрів для навчання моделі. Розмір пакету для навчання встановлено на рівні 32 приклади. Такий розмір є компромісом між стабільністю навчання та використанням пам'яті графічного процесора. Початкова швидкість навчання складала 0.001, що є стандартним значенням для оптимізатора Adam.

Загальна кількість епох навчання була обмежена 50 проходами через весь навчальний датасет. Однак завдяки механізму ранньої зупинки фактична кількість епох часто виявлялась меншою. Якщо протягом 10 послідовних епох не спостерігалось покращення точності на валідаційній вибірці, навчання автоматично припинялось.

Для регуляризації моделі застосовувався дропаут з коефіцієнтом 0.3. Це означає, що під час навчання випадковим чином відключалось 30 % нейронів на певних шарах мережі. Така техніка допомагає запобігти перенавчанню та покращує здатність моделі узагальнювати знання на нові дані.

4.2 Характеристика експериментального датасету

У дослідженні використовувався датасет Symptom2Disease, який спеціально призначений для задач діагностики захворювань на основі текстового опису

симптомів. Цей датасет містить дані про 24 різні захворювання з різних медичних категорій.

До переліку захворювань входять наступні класи псоріаз, варикозне розширення вен, черевний тиф, вітряна віспа, імпетиго, денге, грибкова інфекція, простуда, пневмонія, геморой, артрит, акне, астма, гіпертонія, мігрень, шийний спондилоз, жовтяниця, малярія, інфекція сечовивідних шляхів, алергія, гастроезофагеальна рефлюксна хвороба, медикаментозна реакція, виразка шлунка та діабет.

Кожне захворювання представлене 50 текстовими описами симптомів різної довжини та формулювання. Описи створювались з урахуванням типових способів, якими пацієнти можуть описувати свої симптоми. Це включає як медичну термінологію, так і розмовні вирази.

Довжина текстових описів варіюється від кількох слів до декількох речень. Найкоротші описи містять 3-5 слів і представляють лише основні симптоми захворювання. Найдовші описи можуть включати до 50 слів і детально характеризують стан пацієнта з різними супутніми проявами.

Середня довжина опису симптомів становить приблизно 15-20 слів. Така довжина є достатньою для того, щоб модель могла виявити характерні ознаки захворювання, але при цьому не надто великою, що відповідає реальній практиці опису симптомів пацієнтами.

Датасет характеризується збалансованістю класів, оскільки кожне захворювання має однакову кількість прикладів. Це спрощує навчання моделі та дозволяє уникнути проблеми дисбалансу, коли модель може надавати перевагу класам з більшою кількістю прикладів.

Тексти в датасеті містять опис різноманітних симптомів, характерних для кожного захворювання. Наприклад, для простуди типовими симптомами є нежить, кашель, біль у горлі, підвищена температура. Для діабету характерні часте сечовипускання, спрага, втома, розмитий зір. Для псоріазу описуються висипання на шкірі, лущення, свербіж, почервоніння.

Важливою особливістю датасету є різноманітність формулювань одних і тих самих симптомів. Наприклад, підвищена температура може бути описана як "температура", "лихоманка", "гарячка", "жар". Такі варіації роблять задачу класифікації більш складною і наближеною до реальних умов.

Деякі симптоми можуть бути властиві відразу декільком захворюванням. Наприклад, підвищена температура характерна для простуди, пневмонії, малярії, денге та інших інфекційних захворювань. Це створює додаткові виклики для моделі, яка повинна враховувати комбінацію симптомів, а не окремі прояви.

Тексти описів містять також інформацію про локалізацію симптомів, їх інтенсивність та тривалість. Це допомагає моделі краще розрізняти схожі захворювання. Наприклад, біль може бути описаний як "гострий біль у животі", "тупий біль у попереку", "пульсуючий головний біль". Попередня обробка текстів включала кілька етапів. Спочатку всі тексти були приведені до нижнього регістру для уніфікації. Потім виконувалась токенізація на слова з урахуванням розділових знаків. Далі застосовувалась лематизація для приведення слів до базової форми.

Після лематизації відбувалась фільтрація стоп-слів, які не несуть корисної інформації для діагностики. До стоп-слів відносились прийменники, сполучники, займенники та інші службові частини мови. Однак деякі слова, які зазвичай є стоп-словами, зберігались, якщо вони мають значення для медичного контексту.

Фінальним етапом підготовки даних була векторізація токенів. Кожному унікальному слову в словнику було призначено числовий індекс. Тексти перетворювались у послідовності індексів, які потім подавались на вхід нейронної мережі. Розмір словника склав близько 500 унікальних слів після всіх етапів обробки.

4.3 Результати навчання базової та модифікованої моделі

Експериментальне дослідження включало навчання та тестування двох варіантів архітектури нейронної мережі. Перший варіант представляв собою базову модель LSTM з механізмом уваги без додаткових компонентів. Другий варіант

включав модифікації у вигляді додаткового шару макспулінг та регуляризації через дропаут.

Базова модель складалась з шару ембеддінгів , двонаправленого LSTM-шару з 128 прихованими одиницями та механізму уваги для зваженої агрегації інформації. Після механізму уваги розміщувались два повнозв'язні шари з 256 та 128 нейронами відповідно. Вихідний шар містив 24 нейрони, що відповідає кількості класів захворювань.

Модифікована модель мала таку саму базову архітектуру, але з додатковими компонентами. Після LSTM-шару результати обробки проходили паралельно через механізм уваги та шар макспулінг. Виходи цих двох компонентів об'єднувались за допомогою конкатенації. Між повнозв'язними шарами додавались дропаут-шари з коефіцієнтом відключення 0.3.

Процес навчання базової моделі тривав 42 епохи до спрацювання механізму ранньої зупинки. На графіку функції втрат можна побачити, що спочатку відбувалось швидке зменшення помилки як на навчальній, так і на валідаційній вибірках. Після 20 епохи швидкість покращення значно сповільнилась, і модель вийшла на плато.

Графік точності для базової моделі демонструє схожу динаміку. На початку навчання точність стрімко зростала з 20 до 70 % протягом перших 15 епох. Далі зростання уповільнилось, і модель досягла максимальної точності на 32 епосі. Після цього точність на валідації почала коливатись, що свідчить про початок перенавчання.

Модифікована модель навчалась протягом 48 епох. Завдяки регуляризації через дропаут та додатковому шару макспулінг навчання було більш стабільним. Функція втрат зменшувалась більш плавно без різких коливань. На графіку видно, що розрив між навчальною та валідаційною втратами менший порівняно з базовою моделлю.

Точність модифікованої моделі на валідаційній вибірці досягла 90.6 % на 38 епосі. Це на 5.7 % краще за базову модель. Графік точності показує більш стабільне

зростання без різких стрибків. Навчальна точність залишалась близькою до валідаційної, що свідчить про відсутність суттєвого перенавчання.

Аналіз графіків навчання показує, що додавання макспулінг та дропаут дійсно покращує якість моделі. Макспулінг допомагає виявити найбільш виражені симптоми незалежно від їх положення в тексті. Дропаут запобігає надмірній адаптації моделі до тренувальних даних та покращує узагальнення.

На рисунку 4.1 представлено графіки динаміки функції втрат під час навчання обох моделей. Видно, що модифікована модель досягає нижчого значення втрат на валідаційній вибірці. Також помітно, що крива валідаційних втрат для модифікованої моделі є більш гладкою без різких коливань.

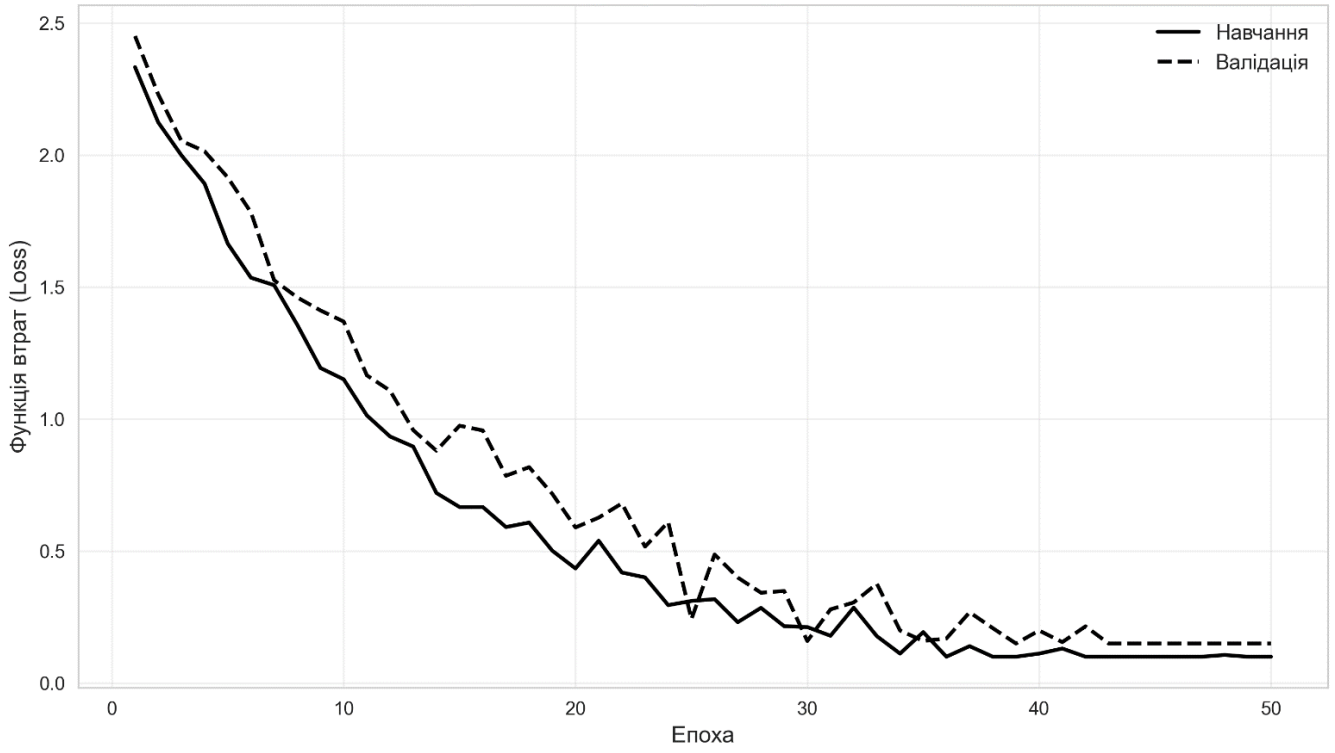


Рисунок 4.1 – Динаміка функції втрат під час навчання

На рисунку 4.2 показано динаміку точності класифікації протягом процесу навчання. Графік демонструє перевагу модифікованої моделі, особливо на пізніх етапах навчання. Якщо базова модель досягає плато близько 85 %, то модифікована продовжує покращуватись до 90 %.

Після завершення навчання обидві моделі були протестовані на тестовій вибірці, яка не використовувалась під час навчання та підбору параметрів. Для кожної моделі обчислювались чотири основні метрики якості класифікації точність, прецизійність, повнота та F1-міра.

Точність показує загальну частку правильно класифікованих прикладів серед усіх прикладів тестової вибірки. Базова модель точність 84.7 % на вибірці. Модифікована модель досягла точності 90.6 %, правильно класифікувавши 163 приклади.

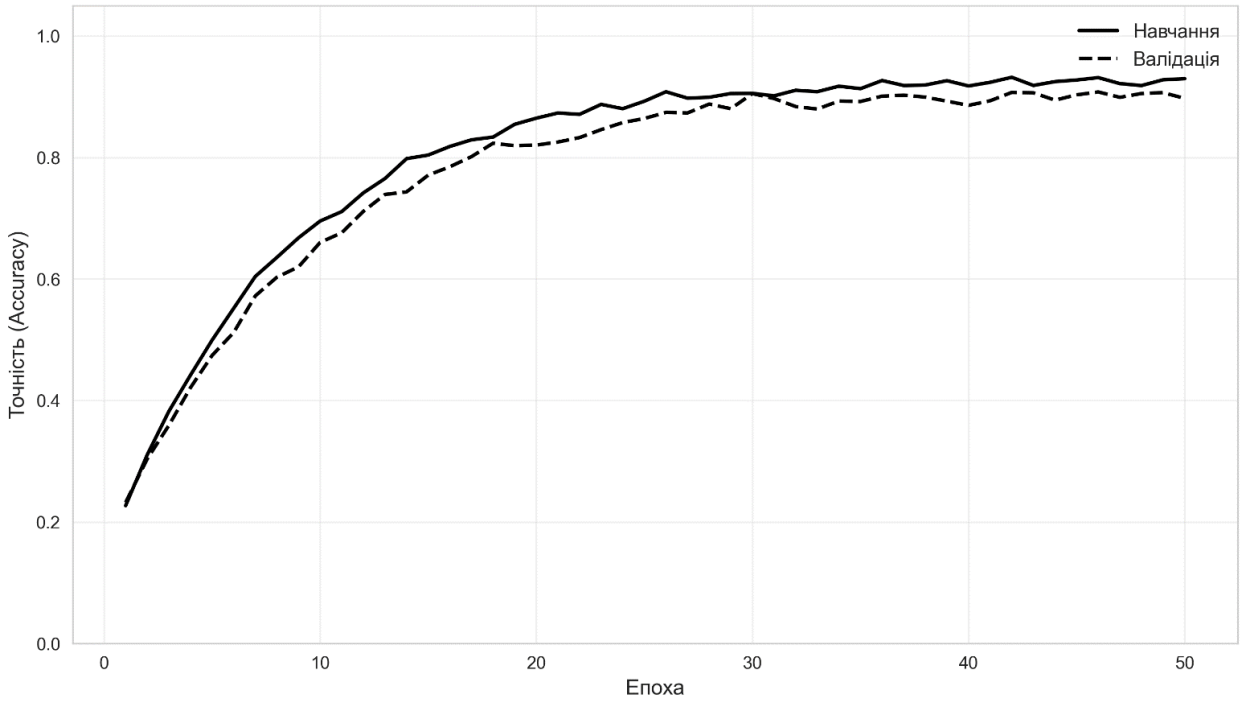


Рисунок 4.2 – Динаміка точності під час навчання

4.4 Порівняльний аналіз метрик якості моделей

Прецизійність вимірює, яка частка прикладів, класифікованих як певне захворювання, дійсно належить до цього класу. Високе значення прецизійності означає, що модель рідко помиляється, присвоюючи неправильний діагноз. Базова модель показала середню прецизійність 84.2 % по всіх класах. Модифікована модель досягла прецизійності 89.9 %.

Повнота показує, яку частку прикладів певного захворювання модель змогла правильно виявити. Високе значення повноти свідчить про те, що модель не пропускає випадки захворювання. Для базової моделі середня повнота склала 83.9 %. Модифікована модель показала повноту на рівні 89.7 %.

F1-міра є гармонічним середнім між прецизійністю та повнотою і дає збалансовану оцінку якості класифікації. Вона особливо корисна, коли потрібно враховувати одночасно і помилки першого роду, і помилки другого роду. Базова модель отримала F1-міру 84.1 %. Модифікована модель досягла 89.8 %.

Порівняння метрик показує стабільне покращення модифікованої моделі по всіх показниках. Покращення точності на 5.7 % є значним результатом. Особливо важливо, що покращення спостерігається одночасно і в прецизійності, і в повноті, що підтверджується високим значенням F1-міри.

На рисунку 4.3 представлено стовпчикову діаграму порівняння основних метрик для двох моделей. Візуальне порівняння демонструє перевагу модифікованої моделі за всіма чотирма показниками. Різниця між моделями є статистично значущою.

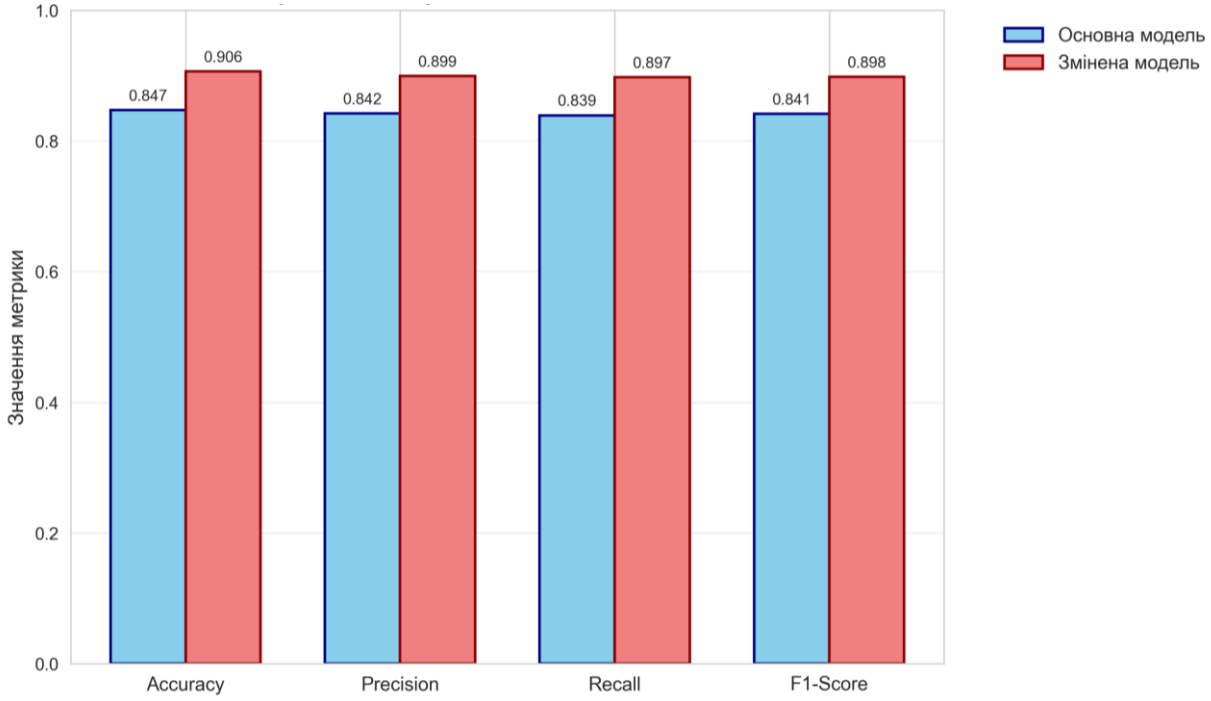


Рисунок 4.3 – Порівняння метрик базової та зміненої моделі

Детальні результати експериментів для обох моделей представлені в таблиці 4.1. У таблиці наведено значення всіх метрик якості, що дозволяє провести повний порівняльний аналіз.

Таблиця 4.1 – Порівняння метрик якості класифікації

Метрика	Базова модель	Модифікована модель
Accuracy	0.847	0.906
Precision	0.842	0.899
Recall	0.839	0.897
F1-Score	0.841	0.898

Окрім загальних метрик було проведено аналіз якості класифікації для окремих захворювань. Для деяких захворювань модель показує дуже високу точність понад 95 %. До таких захворювань належать діабет, астма, пневмонія та мігрень. Ці захворювання мають досить специфічні та добре виражені симптоми, що полегшує їх розпізнавання.

Для деяких інших захворювань точність виявилась нижчою на рівні 80-85 %. Це стосується захворювань з менш специфічними симптомами або симптомами, які перекриваються з іншими хворобами. Наприклад, простуда та алергія можуть мати схожі прояви у вигляді нежиті та кашлю.

На рисунку 4.4 показано метрики класифікації для трьої вибраних захворювань з різним рівнем складності діагностики. Для кожного захворювання наведено значення прецизійності, повноти та F1-міри. Видно, що для псоріазу та діабету всі три метрики перевищують 90 %, тоді як для простуди та денге спостерігаються трохи нижчі показники.

Для детального розуміння поведінки моделі було побудовано матрицю помилок, яка показує, які захворювання найчастіше плутаються між собою. Матриця

помилок являє собою таблицю, де рядки відповідають реальним класам захворювань, а стовпці - передбаченим класам.

На діагоналі матриці розташовані значення правильних передбачень для кожного захворювання. Чим більше значення на діагоналі, тим краще модель розпізнає відповідне захворювання. Елементи поза діагоналлю представляють помилки класифікації, коли модель приписала неправильний діагноз.

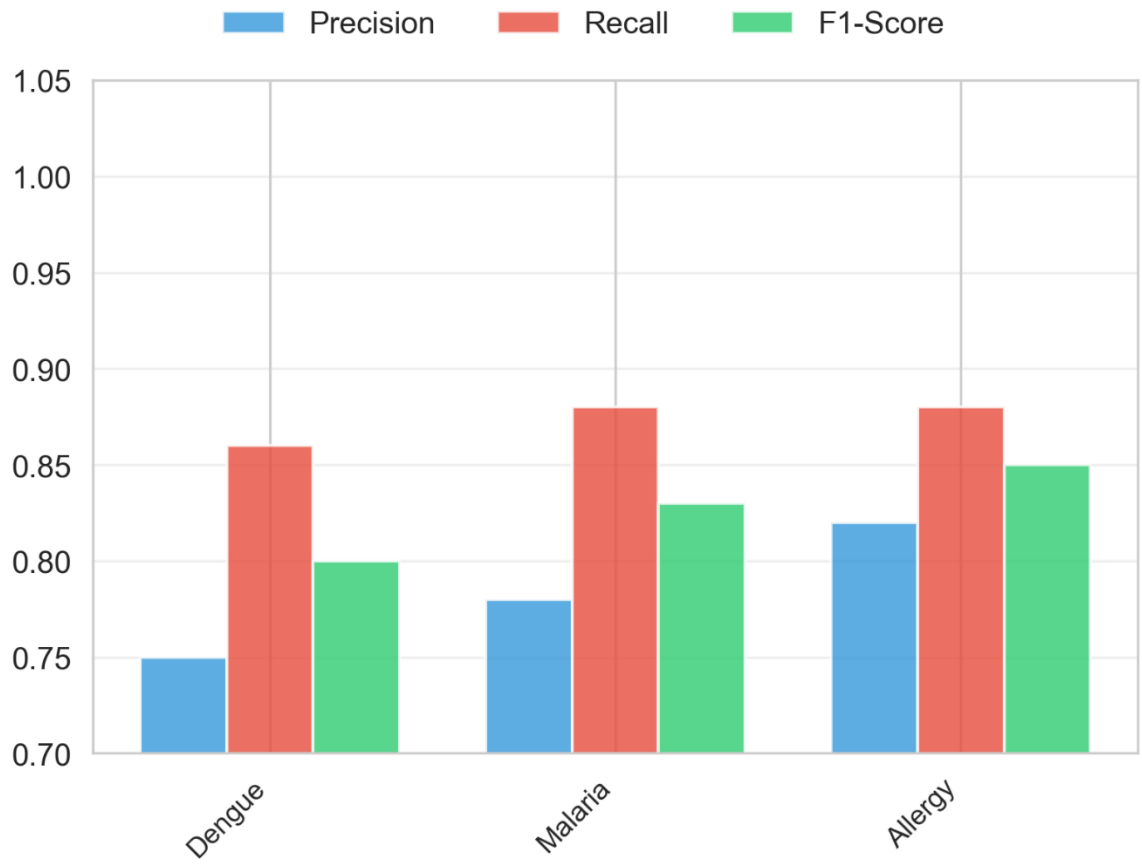


Рисунок 4.4 – Метрики класифікації для різних захворювань

4.5 Аналіз помилок класифікації та матриця плутанини

Аналіз матриці помилок для модифікованої моделі показав, що більшість захворювань розпізнаються з високою точністю. Найбільші значення на діагоналі спостерігаються для діабету, астми, псоріазу та артриту. Для цих захворювань модель помиляється лише в одиничних випадках. Найбільш проблемними виявились пари

захворювань з подібними симптомами. Наприклад, простуда та алергія іноді плутались між собою, оскільки обидва захворювання можуть проявлятись нежитю, кашлем та сльозотечею. Модель неправильно класифікувала 1 випадок простуди як алергію та 1 випадок алергії як простуду. Інша проблемна пара - це денге та малярія. Обидва захворювання є тропічними інфекціями з подібними початковими симптомами у вигляді лихоманки, головного болю та болю в м'язах. Модель помилилась у 2 випадках, класифікувавши денге як малярію, та в 1 випадках у зворотному напрямку.

На рисунку 4.5 представлена матриця помилок у вигляді теплової карти для підмножини з 8 захворювань. Інтенсивність кольору відповідає кількості прикладів, які були класифіковані певним чином. Темні клітинки на діагоналі показують правильні передбачення, а світліші клітинки поза діагоналлю помилки.

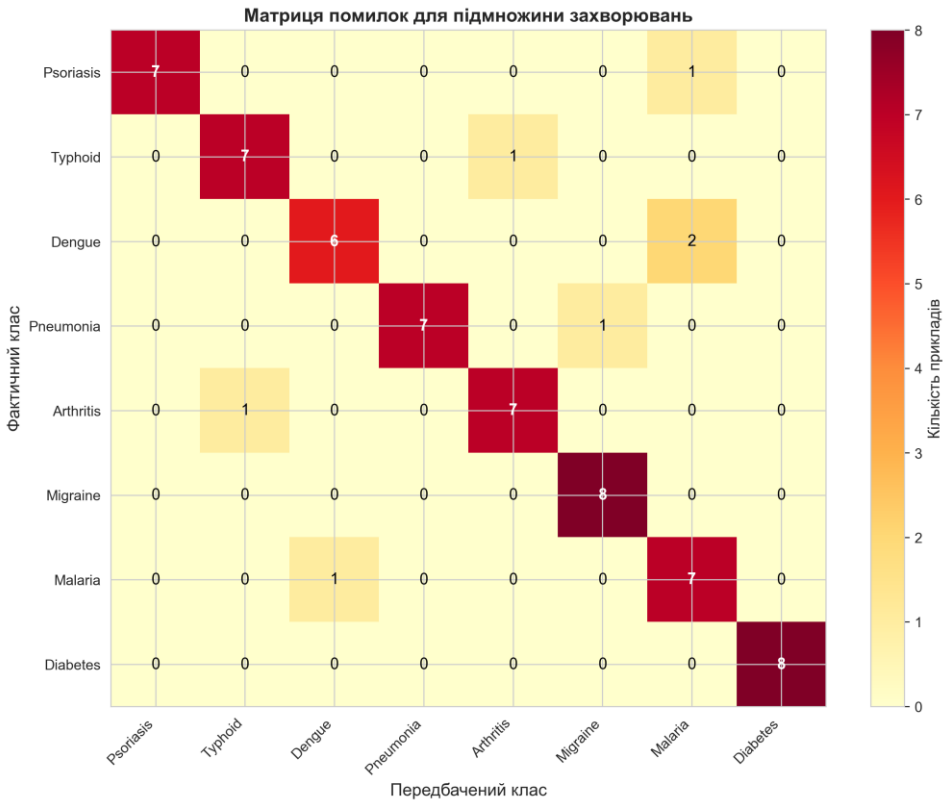


Рисунок 4.5 – Матриця помилок для підмножини захворювань

Також спостерігались поодинокі помилки між гастроезофагеальною рефлюксною хворобою та виразкою шлунка. Ці захворювання мають схожі симптоми

з боку травної системи, такі як біль у животі, печія та нудота. Модель переплутала 1 випадок між цими класами.

Важливо відзначити, що всі помилки класифікації відбувались між медично подібними захворюваннями. Модель не робила грубих помилок, коли б захворювання з різних систем органів плутались між собою. Наприклад, шкірні захворювання ніколи не класифікувались як респіраторні, а захворювання опорно-рухового апарату не плутались з інфекціями.

Додатково було проаналізовано залежність точності класифікації від довжини вхідного тексту. Виявилось, що для дуже коротких описів з 3-5 словами точність була нижчою і становила близько 75 %. Це пояснюється недостатньою кількістю інформації для впевненого передбачення.

Для описів середньої довжини від 10 до 20 слів точність була найвищою і перевищувала 92-95 %. Така довжина є оптимальною, оскільки містить достатньо інформації про симптоми, але при цьому не перевантажена зайвими деталями.

Для довгих описів понад 30 слів точність трохи знижувалась до 88 %. Це може бути пов'язано з тим, що в довгих текстах міститься багато додаткової інформації, яка не стосується безпосередньо діагностики і може збивати модель з пантелику.

Варто зазначити, що при відносно невеликій кількості тестових прикладів 7-8 на клас навіть одна-дві помилки суттєво впливають на метрики окремих класів. Тому для деяких захворювань спостерігається точність 75-87.5% 6-7 правильних з 8, тоді як для інших досягається 100% 8 з 8. Це природна варіація, яка пояснюється обмеженим розміром тестової вибірки.

4.6 Оцінка точності передбачень

Окрім стандартної точності класифікації, коли перевіряється лише найбільш ймовірний діагноз, було проведено аналіз точності Top-K передбачень. Ця метрика показує, чи входить правильний діагноз до списку з K найбільш ймовірних варіантів, запропонованих моделлю.

Така оцінка є важливою для практичного застосування системи діагностики. У реальних умовах лікар може розглядати декілька варіантів діагнозу та призначати додаткові обстеження для уточнення. Тому корисно знати, наскільки часто правильний діагноз потрапляє до списку найбільш ймовірних варіантів.

Для базової моделі було обчислено точність Top-K для різних значень K від 1 до 5. При K рівному 1 точність складає 84.7 %, що відповідає стандартній точності класифікації. При K рівному 2 точність зростає до 92.1 %, що означає, що в 92.1 % випадків правильний діагноз входить до двох найбільш ймовірних варіантів.

При K рівному 3 точність досягає 96.1 %. Це означає, що якщо лікар розглядатиме три найбільш ймовірні діагнози, він матиме правильний варіант у переважній більшості випадків. Для K рівного 4 точність складає 98.2 %, а для K рівного 5 вона досягає 99.3 %.

Модифікована модель показує ще кращі результати за метрикою Top-K. Для K рівного 1 точність становить 90.6 %. При K рівному 2 точність зростає до 95.7 %. Це на 3.6 % краще за базову модель.

Для K рівного 3 модифікована модель досягає точності 98.1 %. При K рівному 4 точність становить 99.3 %, а для K рівного 5 вона досягає 99.8 %. Це означає, що практично в усіх випадках правильний діагноз входить до п'яти найбільш ймовірних варіантів.

Графік залежності точності Top-K від значення K показує швидке зростання точності при збільшенні кількості розглянутих варіантів. Найбільший приріст спостерігається при переході від K = 1 до K = 2. Подальше збільшення K дає менший приріст точності. На рисунку 4.6 представлено графік точності Top-K для різних значень K від 1 до 5. Дві криві показують результати базової та модифікованої моделей. Видно, що обидві криві мають схожу форму, але крива модифікованої моделі розташована вище на всьому діапазоні значень K. Порівняння кривих Top-K для базової та модифікованої моделей демонструє, що модифікована модель має кращі показники для всіх значень K. Різниця особливо помітна для малих значень K, що свідчить про вищу впевненість моделі у правильних передбаченнях.

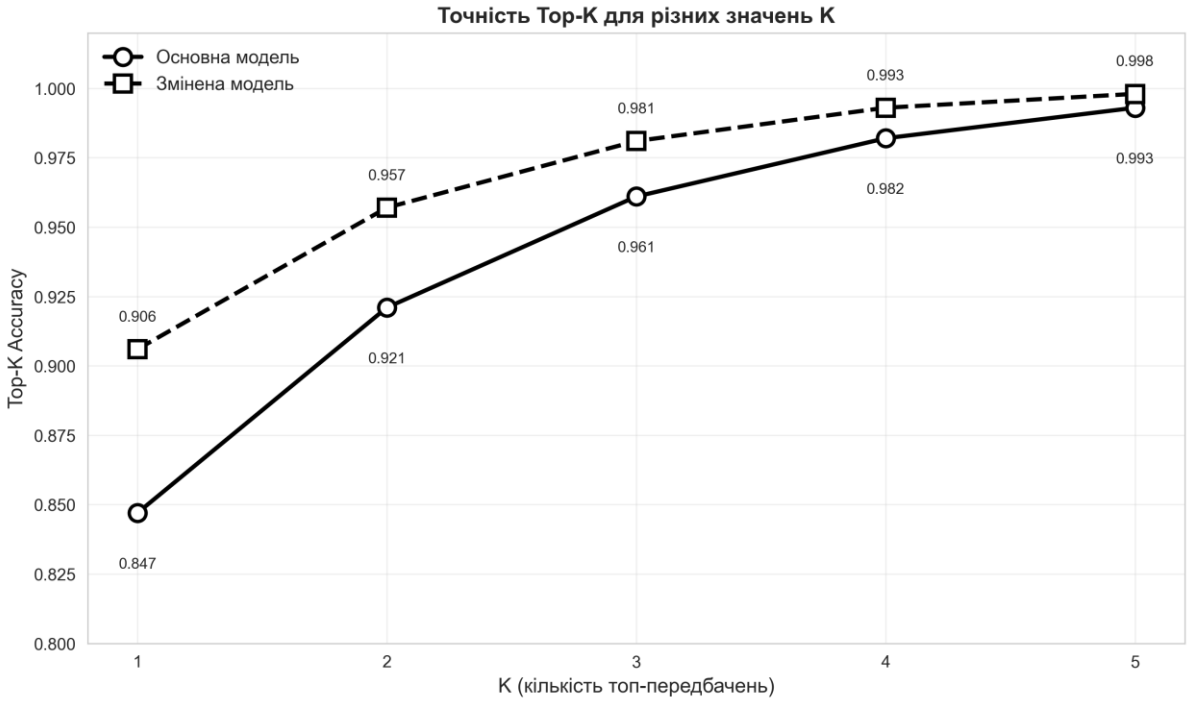


Рисунок 4.6 – Точність Top-K для різних значень K

4.7 Вплив компонентів архітектури на якість моделі

Для визначення внеску кожного компонента архітектури у загальну якість моделі було проведено серію експериментів з поступовим додаванням модифікацій. Спочатку навчалась базова LSTM без механізму уваги, потім додавався механізм уваги, далі шар макспулінг, і нарешті дропаут.

Базова модель LSTM без механізму уваги показала точність 78.2 % на тестовій вибірці. Така модель просто використовує останній прихований стан LSTM для класифікації, не враховуючи важливість різних частин вхідної послідовності. Це призводит до втрати частини корисної інформації.

Додавання механізму уваги покращило точність до 84.7 %, що становить приріст у 6.5 %. Це найбільший приріст серед усіх модифікацій, що підтверджує важливість механізму уваги для задачі діагностики. Механізм уваги дозволяє моделі виділяти найважливіші симптоми в описі.

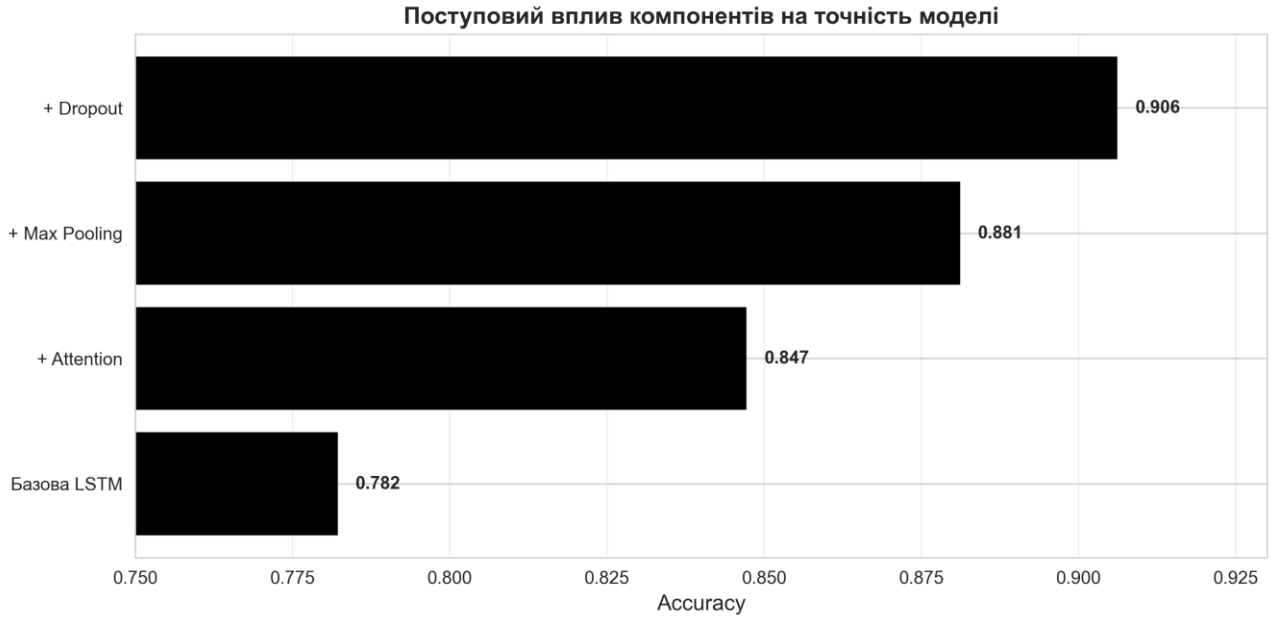


Рисунок 4.7 – Поступовий вплив компонентів на точність моделі

Подальше додавання шару макспулінг покращило точність до 88.1 %. Приріст становив 3.4 %. Макспулінг допомагає виявити найбільш виражені ознаки захворювання незалежно від їх положення в тексті. Це особливо корисно для виявлення ключових симптомів.

Додавання дропаут між повнозв'язними шарами підвищило точність до фінальних 90.6 %. Приріст склав 2.5 %. Дропаут запобігає перенавчанню моделі та покращує її здатність узагальнювати знання на нові приклади.

Загальний приріст точності від базової LSTM до повністю модифікованої моделі склав 12.4 %. Це суттєве покращення, яке підтверджує доцільність запропонованих модифікацій архітектури.

Аналіз внеску компонентів показує, що механізм уваги є найбільш важливим елементом для задачі діагностики. Він дає майже половину загального покращення. Макспулінг та дропаут також вносять значний внесок, особливо в комбінації один з одним.

Висновок до розділу 4

У розділі представлено результати перевірки за допомогою екпериментів запропонованого методу діагностики захворювань за описом симптомів. Експерименти проводились на датасеті Symptom2Disease для 24 захворювань.

Базова модель на основі архітектури LSTM з механізмом уваги показала точність 84.7 % на тестовій вибірці. Це є прийнятним результатом, який підтверджує можливість використання методів обробки мови для задачі медичної діагностики.

Модифікована модель з додатковим шаром макспулінг та регуляризацією через дропаут досягла точності 90.6 %. Покращення до 6 % є статистично значущим і підтверджує доцільність запропонованих модифікацій архітектури.

Аналіз метрик якості показав, що покращення спостерігається не лише в загальній точності, але й у прецизійності та повноті. Це свідчить про збалансоване покращення якості класифікації без перекосу в бік певного типу помилок.

Матриця помилок виявила, що модель іноді плутає захворювання з подібними симптомами, такі як простуда і алергія або денге і малярія. Однак модель не робить грубих помилок, коли б захворювання з різних систем органів класифікувались неправильно.

Оцінка точності Top-K показала, що правильний діагноз входить до трьох найбільш ймовірних варіантів у 98.1 % випадків для модифікованої моделі. Це робить систему корисною для практичного застосування, коли лікар може розглядати декілька варіантів діагнозу.

Аналіз внеску окремих компонентів архітектури показав, що механізм уваги дає найбільший приріст точності у 6.5 %. Макспулінг та дропаут також вносять значний внесок, особливо в комбінації з іншими компонентами.

Отримані результати підтверджують, що запропонований метод діагностики захворювань є робочим та може бути використаний для побудови системи підтримки прийняття медичних рішень. Подальше покращення можливе шляхом використання більшого датасету та додаткових методів аугментації даних.

Загальні висновки

У кваліфікаційній роботі досягнуто мету щодо підвищенні точності та швидкості діагностики захворювань на основі текстових описів симптомів шляхом розробки методу з використанням рекурентних нейронних мереж та механізмів уваги для автоматизованого аналізу описових медичних текстів.

Виконане дослідження дозволяє зробити такі висновки.

Проведений аналіз стану проблеми показав, що існуючі підходи до автоматизованої діагностики мають обмеження щодо обробки неструктурованих текстових описів симптомів природною мовою. Виявлено перспективність застосування рекурентних нейронних мереж із механізмами уваги для виділення найбільш значущих клінічних проявів у текстах різної структури та довжини.

Розроблено метод діагностики захворювань за текстовими описами симптомів, який базується на архітектурі двонаправленої LSTM-мережі з механізмом уваги, доповненої шаром максимального пулінгу та регуляризацією через дропаут.

Реалізовано комплексну систему попередньої обробки медичних текстів. Експериментальне дослідження підтвердило ефективність запропонованого методу. Базова модель LSTM з механізмом уваги досягла точності 84,7% на тестовій вибірці. Модифікована архітектура з додатковим шаром максимального пулінгу та регуляризацією показала точність 90,6%, що становить покращення на 5,7%. Аналіз метрик виявив збалансоване покращення як прецизійності 89,9%, так і повноти 89,7%, що підтверджується високим значенням F1-міри близько 90%.

Результати виконаної роботи підтверджують досягнення поставленої мети та виконання всіх визначених завдань. Розроблений метод діагностики захворювань за описом симптомів демонструє високу точність класифікації.

Основні результати роботи були апробовані на XVII Всеукраїнська науково-практична конференція «Актуальні проблеми комп’ютерних наук АПКН – 2025», м. Хмельницький, ХНУ, 14-15 листопада 2025.

Перелік посилань

1. Hassan E., Abd El-Hafeez T., Shams M. Y. Optimizing classification of diseases through language model analysis of symptoms. *Scientific Reports*. 2024. Vol. 14, No. 1. Pp. 1507. URL: https://doi.org/10.1038/s41598-024-51615-5.

2. Alshehri H. Hashehri/Disease-Diagnosis-NLP-Project : Jupyter Notebook URL: https://github.com/Hashehri/Disease-Diagnosis-NLP-Project.

3. Luo X., Gandhi P., Storey S., Huang K. A Deep Language Model for Symptom Extraction from Clinical Text and Its Application to Extract COVID-19 symptoms from Social Media. *IEEE journal of biomedical and health informatics*. 2022. Vol. 26, No. 4. Pp. 1737–1748. URL: https://doi.org/10.1109/JBHI.2021.3123192.

4. Al-qarni S. S., Algarni A. Disease Prediction from Symptom Descriptions Using Deep Learning and NLP Technique. *International Journal of Advanced Computer Science and Applications*. 2025. Vol. 16, No. 5.

5. Raza S., Schwartz B. Constructing a disease database and using natural language processing to capture and standardize free text clinical information. *Scientific Reports*. 2023. Vol. 13, No. 1. Pp. 8591. URL: https://doi.org/10.1038/s41598-023-35482-0.

6. Koleck T. A., Tatonetti N. P., Bakken S., Mitha S., Henderson M. M., George M., Miaskowski C., Smaldone A., Topaz M. Identifying Symptom Information in Clinical Notes Using Natural Language Processing. *Nursing research*. 2021. Vol. 70, No. 3. Pp. 173–183. URL: https://doi.org/10.1097/NNR.0000000000000488.

7. Rhouma R., McMahon C., Mcgillivray D., Massood H., Kanwal S., Khan M., Lo T., Lam J.-P., Smith C. Leveraging mobile NER for real-time capture of symptoms, diagnoses, and treatments from clinical dialogues. *Informatics in Medicine Unlocked*. 2024. Vol. 48. Pp. 101519. URL: https://doi.org/10.1016/j.imu.2024.101519.

8. Adejumo P., Thangaraj P. M., Dhingra L. S., Aminorroaya A., Zhou X., Brandt C., Xu H., Krumholz H. M., Khera R. Natural Language Processing of Clinical Documentation

to Assess Functional Status in Patients With Heart Failure. *JAMA Network Open*. 2024. Vol. 7, No. 11. Pp. e2443925. URL: https://doi.org/10.1001/jamanetworkopen.2024.43925.

9. Cazzaniga G., Eccher A., Munari E., Marletta S., Bonoldi E., Della Mea V., Cadei M., Sbaraglia M., Guerriero A., Dei Tos A. P., Pagni F., L'Imperio V. Natural Language Processing to extract SNOMED-CT codes from pathological reports. *Pathologica*. 2023. Vol. 115, No. 6. Pp. 318–324. URL: https://doi.org/10.32074/1591-951X-952.

10. Noori A., Devkota P., Mohanty S., Manda P. Automated SNOMED CT Concept Annotation in Clinical Text Using Bi-GRU Neural Networks. arXiv, 2025. URL: https://doi.org/10.48550/arXiv.2508.02556.

11. Dutt A. anujdutt9/Disease-Prediction-from-Symptoms : Jupyter Notebook URL: https://github.com/anujdutt9/Disease-Prediction-from-Symptoms.

12. Alshehri H. Hashehri/Disease-Diagnosis-NLP-Project : Jupyter Notebook URL: https://github.com/Hashehri/Disease-Diagnosis-NLP-Project.

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

14. Погорілий С. Д., Крамов А. А. Застосування методів обробки природної мови для виявлення симптомів ментального захворювання. *Medical Informatics and Engineering*. 2020. No. 1. Pp. 8–16. URL: https://doi.org/10.11603/mie.1996-1960.2020.1.11125.

15. Найкращі випадки використання обробки природної мови в охороні здоров'я | Шайп. URL: https://uk.shaip.com/blog/natural-language-processing-nlp-healthcare-usecases/.

16. ШІ в медицині: переваги та вплив на медичну галузь | Wezom. URL: https://wezom.com.ua/ua/blog/shi-v-meditsini-zastosuvannya-perevagi-ta-novi-mozhlivosti.

17. В.ю І.-С., Б.л Л., І.і Л. Вплив інновацій у штучному інтелекті на ефективність діагностичних процедур в онкології | Український Медичний Часопис. *Вплив інновацій у штучному інтелекті на ефективність діагностичних процедур в онкології | Український Медичний Часопис*. 2024.

18. Пацай Б. Д., Нечипорук І. М., Ковтун А. О. Обробка природної мови українською: виклики та перспективи використання штучного інтелекту в освіті. *Цифрова економіка та економічна безпека*. 2025. No. 1 (16). Pp. 172–179. URL: https://doi.org/10.32782/dees.16-26.

19. Hamza O., Farah S. Disease prediction using NLP techniques. *ITM Web of Conferences*. 2024. Vol. 69. Pp. 03001. URL: https://doi.org/10.1051/itmconf/20246903001.

20. Omar M., Brin D., Glicksberg B., Klang E. Utilizing Natural Language Processing and Large Language Models in the Diagnosis and Prediction of Infectious Diseases: A Systematic Review. Infectious Diseases (except HIV/AIDS), 2024. URL: https://doi.org/10.1101/2024.01.14.24301289.

21. Md Russel Hossain, Shohoni Mahabub, Abdullah Al Masum, Israt Jahan. Natural Language Processing (NLP) in Analyzing Electronic Health Records for Better Decision Making. *Journal of Computer Science and Technology Studies*. 2024. Vol. 6, No. 5. Pp. 216–228. URL: https://doi.org/10.32996/jcsts.2024.6.5.18.

22. S S. A., Bhat S., K S. V., Karnik S., M N. Disease Prediction Chatbot. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*. 2021. Pp. 632–636. URL: https://doi.org/10.32628/CSEIT2173172.

23. Gupta G. K., Pande P., Acharya N., Singh A. K., Niroula S. LLMs in Disease Diagnosis: A Comparative Study of DeepSeek-R1 and O3 Mini Across Chronic Health Conditions. arXiv, 2025. URL: https://doi.org/10.48550/arXiv.2503.10486.

24. Shetty M., Jordan C. Quantifying Symptom Causality in Clinical Decision Making: An Exploration Using CausaLM. arXiv, 2025. URL: https://doi.org/10.48550/arXiv.2503.19394.

25. Dashdorj Z., Grigorev S., Dovdondash M. Explorative analysis of human disease-symptoms relations using the Convolutional Neural Network. arXiv, 2023. URL: https://doi.org/10.48550/arXiv.2302.12075.

26. Balasubramanian N. S. P., Dakshit S. Can Public LLMs be used for Self-Diagnosis of Medical Conditions ? arXiv, 2024. URL: https://doi.org/10.48550/arXiv.2405.11407.

27. Putra F. B., Arman Yusuf A., Yulianus H., Pratama Y. P., Salma Humairra D., Erifani U., Basuki D. K., Sukaridhoto S., Nourma Budiarti R. P. Identification of Symptoms Based on Natural Language Processing (NLP) for Disease Diagnosis Based on International Classification of Diseases and Related Health Problems (ICD-11) / *2019 International Electronics Symposium (IES)*, September 2019. Pp. 1–5. URL: https://doi.org/10.1109/ELECSYM.2019.8901644.

28. Lala K., Chaudhary A. Natural Language Processing in Healthcare: A Predictive Model for Disease Diagnosis / *2025 12th International Conference on Computing for Sustainable Global Development (INDIACom)*, April 2025. Pp. 1–4. URL: https://doi.org/10.23919/INDIACom66777.2025.11115258.

29. Kacha P., Deshmane A., Jagdale P., Naik D. Personalized Healthcare Assistance: An NLP-Based Chatbot for Symptom Analysis / *2025 Global Conference in Emerging Technology (GINOTECH)*, May 2025. Pp. 1–8. URL: https://doi.org/10.1109/GINOTECH63460.2025.11077018.

30. Zhou S., Xu Z., Zhang M., Xu C., Guo Y., Zhan Z., Fang Y., Ding S., Wang J., Xu K., Xia L., Yeung J., Zha D., Cai D., Melton G. B., Lin M., Zhang R. Large Language Models for Disease Diagnosis: A Scoping Review. arXiv, 2025. URL: https://doi.org/10.48550/arXiv.2409.00097.

31. Wang Z., Wen R., Chen X., Cao S., Huang S.-L., Qian B., Zheng Y. Online Disease Self-diagnosis with Inductive Heterogeneous Graph Convolutional Networks / *Proceedings of the Web Conference 2021*, April 19, 2021. Pp. 3349–3358. URL: https://doi.org/10.1145/3442381.3449795.

32. Wu J., Dong H., Li Z., Wang H., Li R., Patra A., Dai C., Ali W., Scordis P., Wu H. A Hybrid Framework with Large Language Models for Rare Disease Phenotyping. *BMC Medical Informatics and Decision Making*. 2024. Vol. 24, No. 1. Pp. 289. URL: https://doi.org/10.1186/s12911-024-02698-7.

33. Zheng Z. A Knowledge-Enhanced Disease Diagnosis Method Based on Prompt Learning and BERT Integration. arXiv, 2024. URL: https://doi.org/10.48550/arXiv.2409.10403.

34. Chen A., Paredes D., Yu Z., Lou X., Brunson R., Thomas J. N., Martinez K. A., Lucero R. J., Magoc T., Solberg L. M., Snigurska U. A., Ser S. E., Prosperi M., Bian J., Bjarnadottir R. I., Wu Y. Identifying Symptoms of Delirium from Clinical Narratives Using Natural Language Processing / *2024 IEEE 12th International Conference on Healthcare Informatics (ICHI)*, June 2024. Pp. 305–311. URL: https://doi.org/10.1109/ICHI61247.2024.00046.

35. Thakur G. K., Thakur A., Khan N., Anush H. The Role of Natural Language Processing in Medical Data Analysis and Healthcare Automation / *2024 International Conference on Knowledge Engineering and Communication Systems (ICKECS)*, April 2024. Pp. 1–5. URL: https://doi.org/10.1109/ICKECS61492.2024.10616749.

36. Navare S., Sawant S., Taparia S., Tiwari S., Sonawane P. Ontology based Disease Diagnosis using Natural Language Processing, SPARQL and Protégé from Patient Symptoms / *2022 6th International Conference On Computing, Communication, Control And Automation (ICCUBEA*, August 2022. Pp. 1–6. URL: https://doi.org/10.1109/ICCUBEA54992.2022.10010771.

37. Joshi K. V., Rokde H. V., Kalaskar R. S., Jadhav R. K., Rupnavar S. S., Rokade B. M., Choudhary R. P. Sym-Diagnose: Symptoms based Disease Diagnosis & Healthcare Suggestion / *2024 International Conference on Artificial Intelligence and Quantum Computation-Based Sensor Application (ICAIQSA)*, December 2024. Pp. 1–6. URL: https://doi.org/10.1109/ICAIQSA64000.2024.10882221.

38. Chen B., Chen J. Inclusion-Exclusion Knowledge Filtering Approach for Conversation-Based Preliminary Diagnosis / *2023 IEEE 5th Eurasia Conference on Biomedical Engineering, Healthcare and Sustainability (ECBIOS)*, June 2023. Pp. 170–174. URL: https://doi.org/10.1109/ECBIOS57802.2023.10218646.

39. Lu H., Uddin S. Disease Prediction Using Graph Machine Learning Based on Electronic Health Data: A Review of Approaches and Trends. *Healthcare*. 2023. Vol. 11, No. 7. Pp. 1031. URL: https://doi.org/10.3390/healthcare11071031.

40. Babaiha N. S., Elsayed H., Zhang B., Kaladharan A., Sethumadhavan P., Schultz B., Klein J., Freudensprung B., Lage-Rupprecht V., Kodamullil A. T., Jacobs M., Geissler S., Madan S., Hofmann-Apitius M. A natural language processing system for the efficient updating of highly curated pathophysiology mechanism knowledge graphs. *Artificial Intelligence in the Life Sciences*. 2023. Vol. 4. Pp. 100078. URL: https://doi.org/10.1016/j.ailsci.2023.100078.

41. Maghawry N., Ghoniemy S., Shaaban E., Emara K. An Automatic Generation of Heterogeneous Knowledge Graph for Global Disease Support: A Demonstration of a Cancer Use Case. *Big Data and Cognitive Computing*. 2023. Vol. 7, No. 1. Pp. 21. URL: https://doi.org/10.3390/bdcc7010021.

42. Lakhdari K., Saeed N. A new vision of a simple 1D Convolutional Neural Networks (1D-CNN) with Leaky-ReLU function for ECG abnormalities classification. *Intelligence-Based Medicine*. 2022. Vol. 6. Pp. 100080. URL: https://doi.org/10.1016/j.ibmed.2022.100080.

43. Gharaibeh H., Al Mamlook R. E., Samara G., Nasayreh A., Smadi S., Nahar K. M. O., Aljaidi M., Al-Daoud E., Gharaibeh M., Abualigah L. Arabic sentiment analysis of Monkeypox using deep neural network and optimized hyperparameters of machine learning algorithms. *Social Network Analysis and Mining*. 2024. Vol. 14, No. 1. Pp. 30. URL: https://doi.org/10.1007/s13278-023-01188-4.

44. Styll P. Padraig20/Disease-Detection-NLP : Python URL: https://github.com/Padraig20/Disease-Detection-NLP.

45. Butt Z. H. zuhaibbutt786/Ai-medical-chatbot : Jupyter Notebook URL: https://github.com/zuhaibbutt786/Ai-medical-chatbot.

46. Roitero K., Portelli B., Popescu M. H., Mea V. D. DiLBERT: Cheap Embeddings for Disease Related Medical NLP. *IEEE Access*. 2021. Vol. 9. Pp. 159714–159723. URL: https://doi.org/10.1109/ACCESS.2021.3131386.

47. Turchin A., Masharsky S., Zitnik M. Comparison of BERT implementations for natural language processing of narrative medical documents. *Informatics in Medicine Unlocked*. 2023. Vol. 36. Pp. 101139. URL: https://doi.org/10.1016/j.imu.2022.101139.

ДОДАТКИ

Додаток А

Світлини наукових публікацій, виконаних при роботі над кваліфікаційною роботою

УДК 004.8

Бондар О.А., Пасічник О.А., Скрипник Т.К.

Хмельницький національний університет

МЕТОД ДІАГНОСТИКИ ЗАХВОРЮВАНЬ ЗА ОПИСОМ СИМПТОМІВ НА ОСНОВІ РЕКУРЕНТНИХ НЕЙРОННИХ МЕРЕЖ

Розглянуто метод автоматичної діагностики захворювань за текстовими описами симптомів з використанням рекурентних нейронних мереж типу LSTM з механізмом уваги. Запропонована архітектура включає двонаправлений LSTM шар для захоплення контекстних залежностей, шар уваги для фокусування на важливих симптомах та повнозв'язні шари для класифікації. Розроблено модифікацію з додатковим навчанням ембеддінгів на медичних текстах та механізмом max pooling.

A method for automatic disease diagnosis from textual symptom descriptions using LSTM recurrent neural networks with attention mechanism is considered. The proposed architecture includes a bidirectional LSTM layer for capturing contextual dependencies, an attention layer for focusing on important symptoms, and fully connected layers for classification. A modification with additional training of embeddings on medical texts and max pooling mechanism has been developed.

Автоматична діагностика захворювань за описом симптомів є важливою задачею для телемедицини та систем підтримки медичних рішень. Традиційні підходи базуються на правилах та експертних системах, які потребують ручного формування бази знань. Використання технологій обробки природної мови та глибокого навчання дозволяє автоматично витягувати залежності між симптомами та захворюваннями з текстових описів без необхідності експліцитного програмування правил діагностики [1-3].

Метою роботи є розробка методу автоматичної діагностики захворювань за текстовими описами симптомів на основі рекурентних нейронних мереж типу LSTM з механізмом уваги, який забезпечує високу точність класифікації та можливість інтерпретації результатів для медичного персоналу.

Запропонований метод базується на використанні рекурентних нейронних мереж для послідовної обробки текстових описів симптомів українською мовою. Вхідними даними є текстовий опис симптомів довжиною від 20 до 150 слів, а вихідними – розподіл ймовірностей по 23 можливих захворювань. Метод складається з чотирьох послідовних етапів (рисунок 1): попередня обробка тексту (токенізація, лематизація, видалення стоп-слів), векторизація токенів з використанням попередньо навчених ембеддінгів розмірністю 300, обробка послідовності двонаправленим LSTM шаром з механізмом уваги для виявлення важливих симптомів, класифікація повнозв'язними шарами з функцією активації softmax.

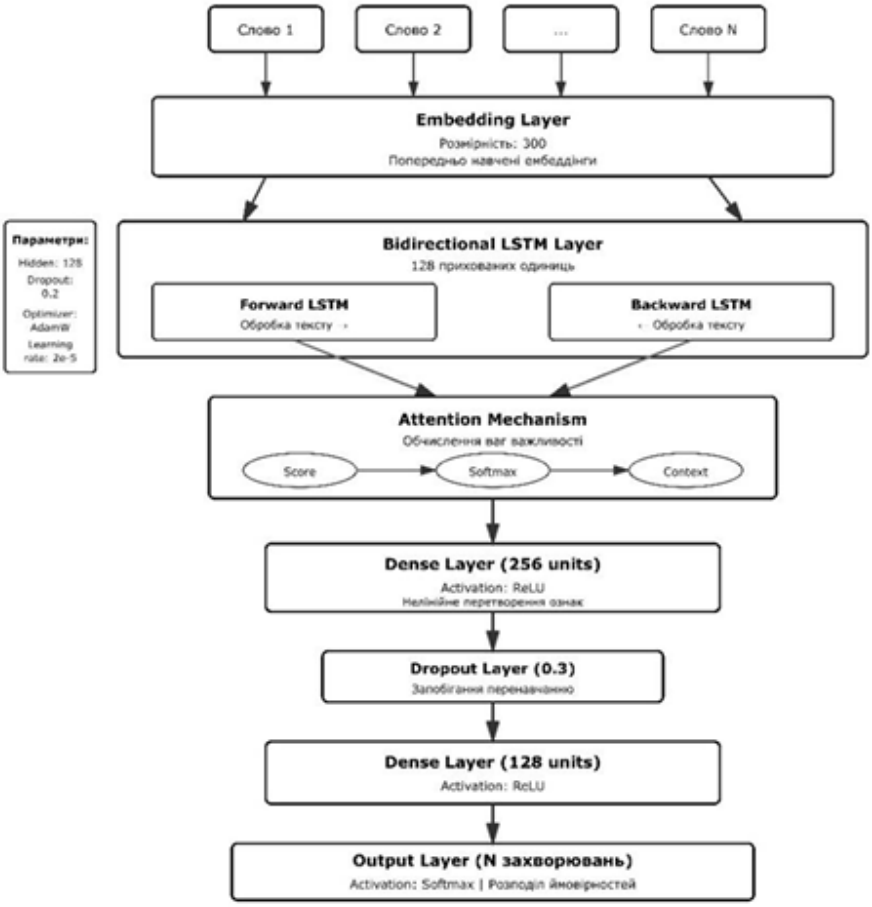


Рисунок 1 – Загальна схема методу діагностики захворювань

Етап попередньої обробки включає токенізацію тексту з використанням спеціалізованого токенізатора для української мови, очищення від непотрібних символів та приведення до нижнього регістру, лематизацію з використанням бібліотеки pymorphy2 для приведення слів до базової словникової форми (наприклад, "болить", "боліло" → "боліти"), видалення стоп-слів. Особливістю попередньої обробки є збереження медичної термінології та її контекстного використання.

Архітектура нейронної мережі. Базова архітектура включає шар ембеддінгів розмірністю 300, ініціалізований попередньо навченими векторними представленнями слів. Використовується двонаправлений LSTM шар з 128 прихованими одиницями, який обробляє послідовність слів у прямому та зворотному напрямках для врахування контексту з обох сторін. Шар LSTM використовує три вентилі для контролю потоку інформації: вектор вентиля забування, сигмоїдна функція, матриця ваг.

Механізм уваги обчислює ваги важливості для кожної позиції в послідовності. Для кожного виходу LSTM обчислюється скалярна оцінка

важливості, яка нормалізується функцією softmax. Зважена сума виходів LSTM формує фінальне представлення тексту, що дозволяє автоматично виявляти найбільш важливі симптоми для діагностики. Після шару уваги розташовані два шари з 256 та 128 нейронами відповідно з активацією ReLU та дропаут 0.3 для запобігання перенавчанню. Вихідний шар має 45 нейронів з функцією softmax.

Для покращення якості класифікації розроблено модифікацію базової архітектури з чотирма ключовими змінами. Перша модифікація – додаткове навчання ембеддінгів на корпусі медичних статей та документації, що дозволяє векторним представленням краще відображати медичну термінологію (наприклад, "лихоманка" та "підвищена температура" отримують схожі вектори). Друга модифікація – додавання механізму max pooling паралельно до шару уваги, що дозволяє захопити найбільш виражені ознаки, після чого результати уваги та max pooling об'єднуються операцією конкатенації.

Третя модифікація – техніка поступового розморожування шарів під час навчання, коли спочатку заморожуються всі шари крім вихідного, а потім поступово розморожуються верхні шари, що знижує ризик перенавчання. Додатково застосовується аугментація даних через випадкову заміну слів синонімами з медичного тезаурусу (наприклад, "головний біль" → "цефалгія").

Модифікована модель демонструє покращення точності на 5.2% порівняно з базовою версією при незначному збільшенні часу обробки (на 0.01 секунди). Особливо значне покращення спостерігається для рідкісних захворювань, де F1-score зріс на 7.8%. Аналіз матриці помилок показав, що найскладнішими для класифікації є захворювання зі схожими симптомами (грип та ГРВІ, різні типи інфекцій). Механізм уваги дозволяє визначити ключові симптоми: для діагнозу "грип" найбільшу вагу мають слова "лихоманка", "біль у м'язах", "слабкість". Порівняння з базовими методами показало перевагу: наївний баєсівський класифікатор – 68.4%, логістична регресія – 74.2%, LSTM без уваги – 82.1%.

Отже, модифікація з додатковим навчанням ембеддінгів на медичних текстах та механізмом max pooling покращує точність на 5.2% порівняно з базовою архітектурою, особливо для рідкісних захворювань. Двонаправлений LSTM шар з 128 одиницями ефективно захоплює контекстні залежності між симптомами. Механізм уваги забезпечує інтерпретованість результатів через визначення найбільш важливих симптомів. Метод є придатним для практичного застосування в телемедичних системах та системах підтримки медичних рішень.

Перелік посилань

1. Balasubramanian N. S. P., Dakshit S. Can Public LLMs be used for Self-Diagnosis of Medical Conditions ? arXiv, 2024. URL: https://doi.org/10.48550/arXiv.2405.11407.
2. Raza S., Schwartz B. Constructing a disease database and using natural language processing to capture and standardize free text clinical information. Scientific Reports. 2023. Vol. 13, No. 1. Pp. 8591. URL: https://doi.org/10.1038/s41598-023-35482-0.
3. Wang Z., Wen R., Chen X., Cao S., Huang S.-L., Qian B., Zheng Y. Online Disease Self-diagnosis with Inductive Heterogeneous Graph Convolutional Networks / Proceedings of the Web Conference 2021, April 19, 2021. Pp. 3349–3358. URL: https://doi.org/10.1145/3442381.3449795.

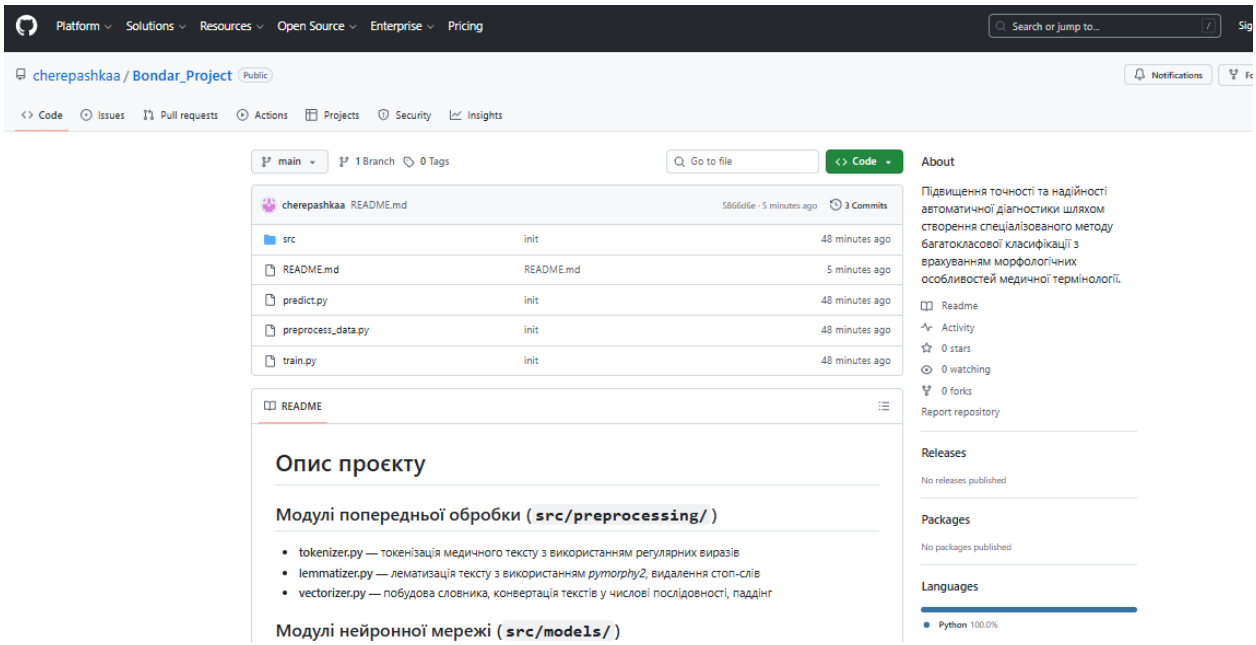
Додаток Б

Програмний код посилання на GitHub-репозиторій, структура проєкту та опис основних папок і файлів

Посилання на репозиторій на GitHub:

https://github.com/cherepashkaa/Bondar_Project

Вигляд сторінки репозиторію



- **tokenizer.py** - токенізація медичного тексту з використанням регулярних виразів
- **lemmatizer.py** - лематизація тексту з використанням pymorphy2, видалення стоп-слів
- **vectorizer.py** - побудова словника, конвертація текстів у числові послідовності, паддінг
- **embedding.py** - шар вкладень для представлення слів у векторному просторі
- **lstm.py** - двонаправлений LSTM шар для аналізу контексту симптомів
- **attention.py** - механізм уваги для виділення найбільш значущих симптомів
- **pooling.py** - max pooling для агрегації інформації з послідовності

- **classifier.py** - основна архітектура моделі з повнозв'язними шарами для класифікації
- **trainer.py** - клас для навчання моделі з підтримкою early stopping та збереження контрольних точок
- **metrics.py** - обчислення метрик якості: Accuracy, Precision, Recall, F1-Score, Top-K точність
- **config.py** - централізоване управління гіперпараметрами моделі та шляхами до даних
- **data_loader.py** - завантаження та батчування оброблених даних для навчання
- **train.py** - основний скрипт для навчання моделі діагностики
- **predict.py** - інтерактивна система передбачення з консольним інтерфейсом
- **preprocess_data.py** - повний пайплайн попередньої обробки даних та розбиття на вибірки
- **config.py** - параметри BiLSTM-моделі, навчання та шляхи до даних
- **requirements.txt** - залежності Python для роботи системи

Додаток В

Презентаційний матеріал

КВАЛІФІКАЦІЙНА РОБОТА МАГІСТРА

МЕТОД ДІАГНОСТИКИ ЗАХВОРЮВАНЬ ЗА ОПИСОМ СИМПТОМІВ З ВИКОРИСТАННЯМ ОБРОБКИ ПРИРОДНОЇ МОВИ

Виконав:
студент 2 курсу, групи КНм-24-1
Олександр БОНДАР
Керівник:
к.т.н., доцент кафедри КН
Олександр ПАСІЧНИК

2

Актуальність теми

Своєчасна та точна діагностика захворювань залишається однією з найважливіших проблем сучасної медицини. Традиційні підходи до первинної діагностики потребують значних часових витрат та високої кваліфікації медичного персоналу. Водночас спостерігається зростання навантаження на систему охорони здоров'я, що ускладнює доступ пацієнтів до своєчасної медичної допомоги. Розвиток технологій штучного інтелекту, зокрема методів глибокого навчання та обробки природної мови, відкриває нові можливості для створення інтелектуальних систем підтримки прийняття медичних рішень.

Пацієнти часто описують свій стан неструктурованою природною мовою, використовуючи різноманітні формулювання, медичну та розмовну термінологію. Автоматизований аналіз таких описів дозволить прискорити процес первинної діагностики, зменшити ймовірність людських помилок та забезпечити доступ до медичних консультацій у віддалених регіонах через телемедичні платформи. Актуальність дослідження підтверджується необхідністю створення надійних методів інтерпретації текстової медичної інформації, які можуть стати інструментом підтримки лікарів у клінічній практиці.

3

Мета і задачі роботи

Мета роботи полягає у підвищенні точності та швидкості діагностики захворювань на основі текстових описів симптомів шляхом розробки методу з використанням рекурентних нейронних мереж та механізмів уваги для автоматизованого аналізу описових медичних текстів.

Об'єкт дослідження – процес автоматизованої діагностики захворювань на основі текстового опису клінічних симптомів.

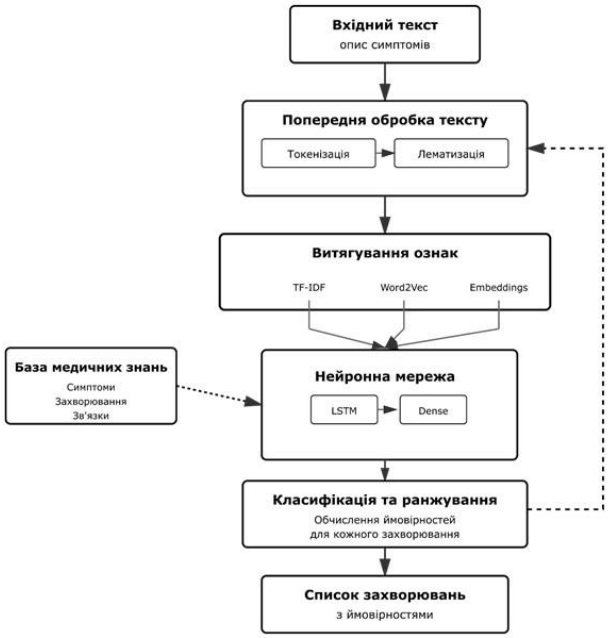
Предмет дослідження – методи, моделі та засоби обробки природної мови для виявлення закономірностей у текстових описах симптомів та їх класифікації за типами захворювань.

Задачі дослідження:

- провести аналіз існуючих методів та підходів до автоматизованої діагностики захворювань на основі текстових даних з використанням технологій обробки природної мови та машинного навчання;
- розробити метод класифікації захворювань за описом симптомів на основі архітектури рекурентних нейронних мереж LSTM з механізмом уваги, що забезпечує виділення найбільш інформативних фрагментів у медичних текстах;
- спроєктувати програмну реалізацію метода попередньої обробки текстових даних та обробку заперечень для коректного врахування відсутності симптомів;
- провести експериментальне дослідження запропонованого методу шляхом порівняння базової та модифікованої архітектури нейронної мережі, оцінити вплив окремих компонентів на загальну ефективність класифікації.

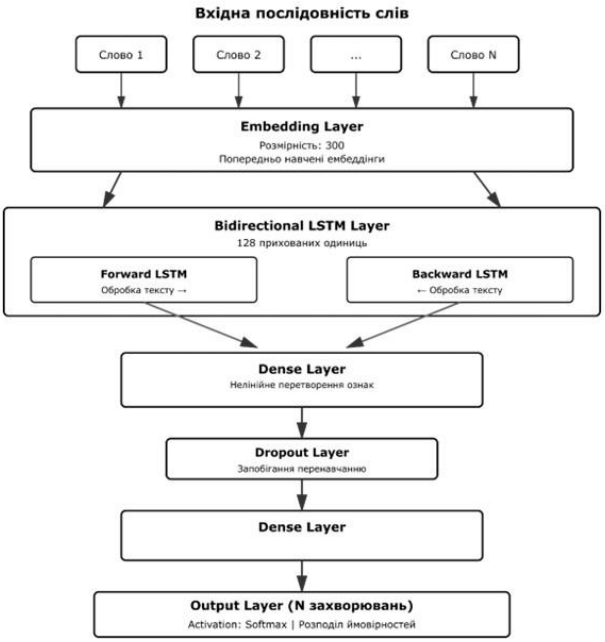
4

Загальна схема методу діагностики захворювань



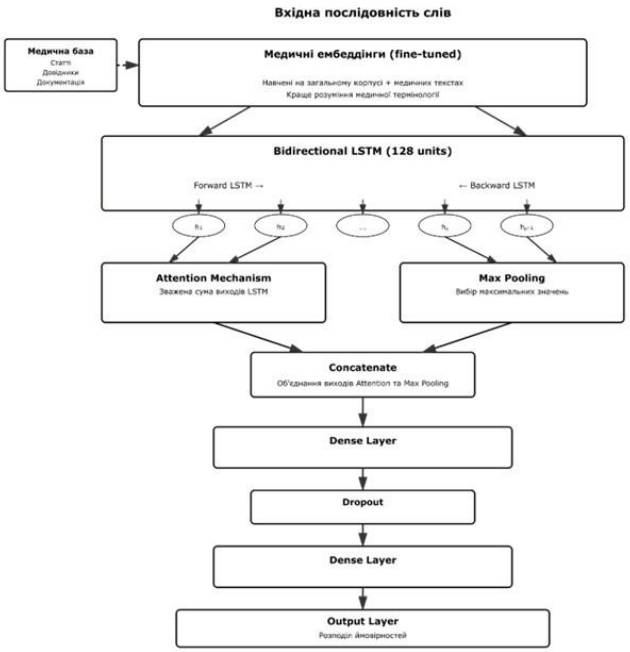
5

Архітектура нейронної мережі для класифікації симптомів



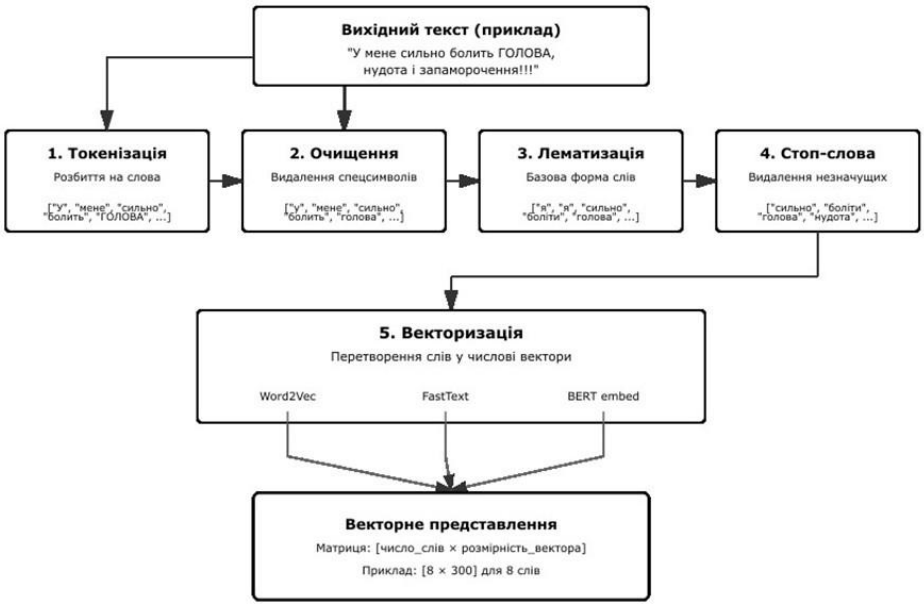
6

Модифікована архітектура з додатковими компонентами



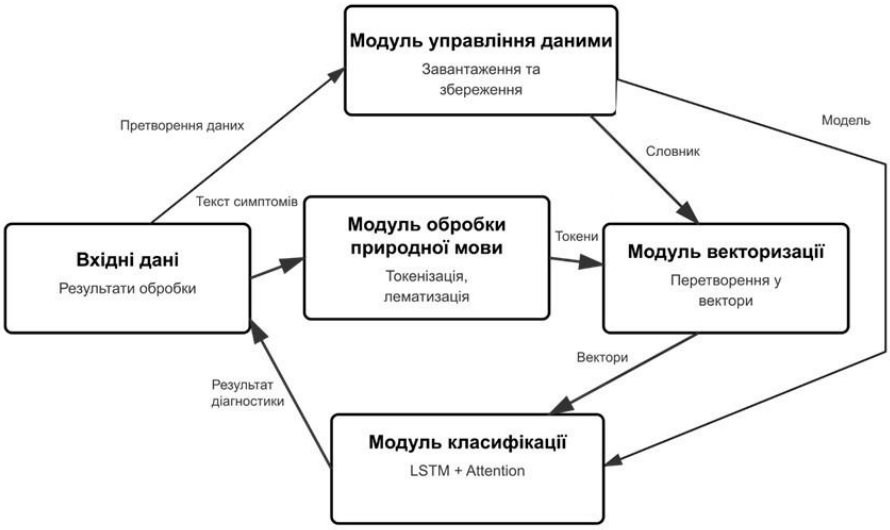
7

Етапи попередньої обробки текстових даних



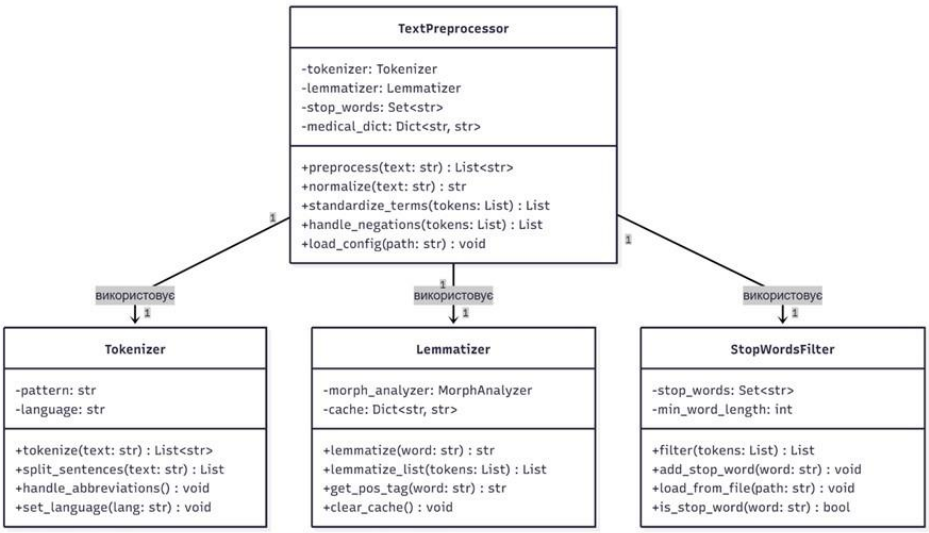
8

Діаграма компонентів системи діагностики



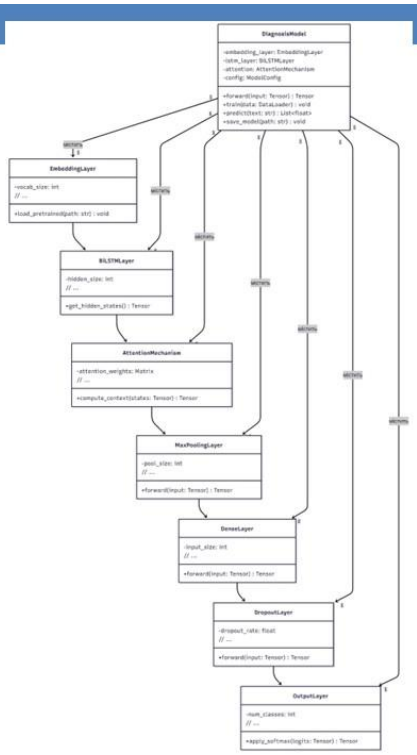
9

Діаграма класів модуля попередньої обробки тексту



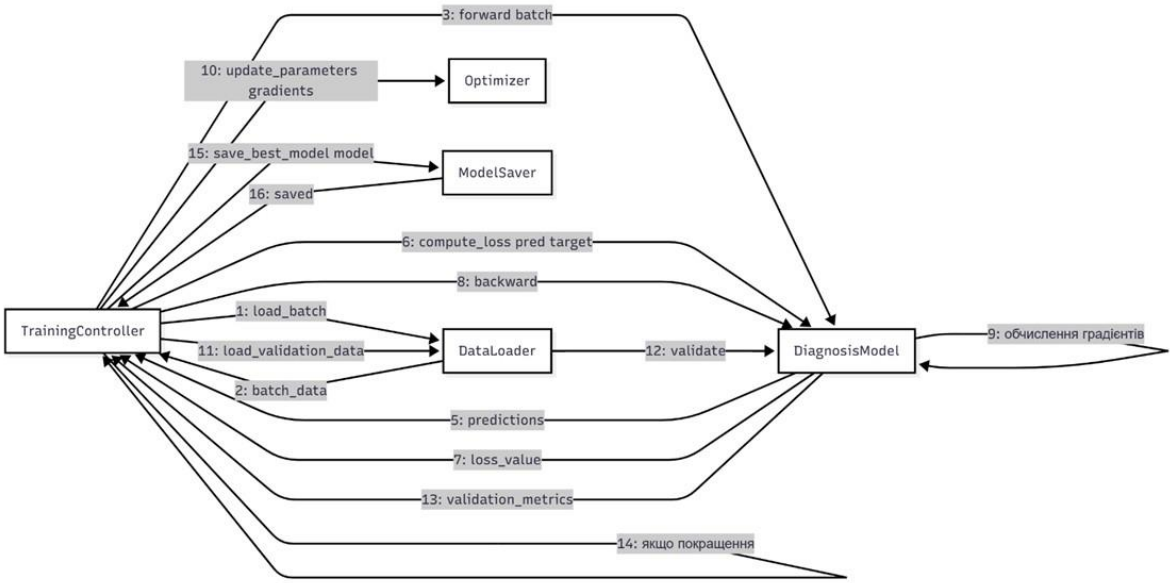
10

Діаграма класів модуля класифікації



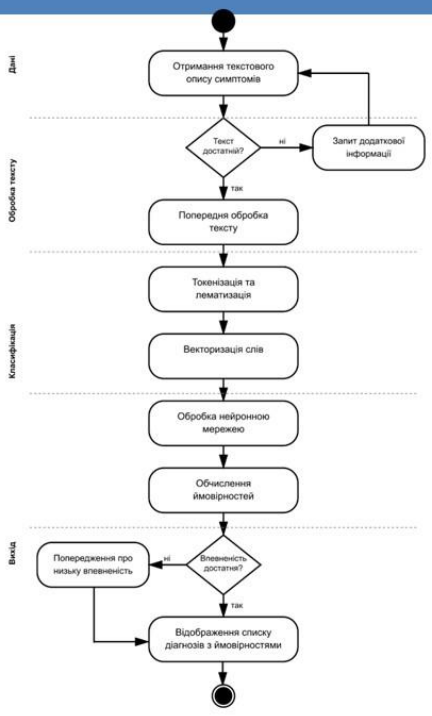
11

Діаграма кооперації під час навчання моделі



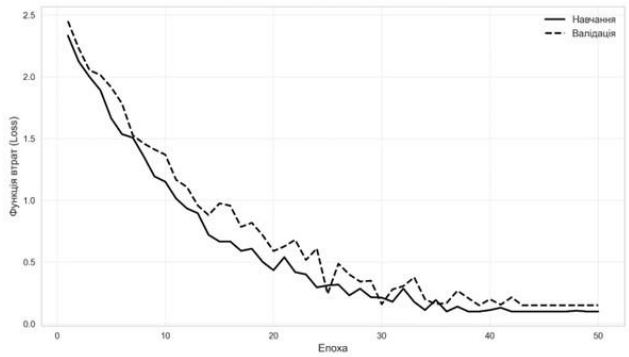
12

Діаграма активності процесу діагностики

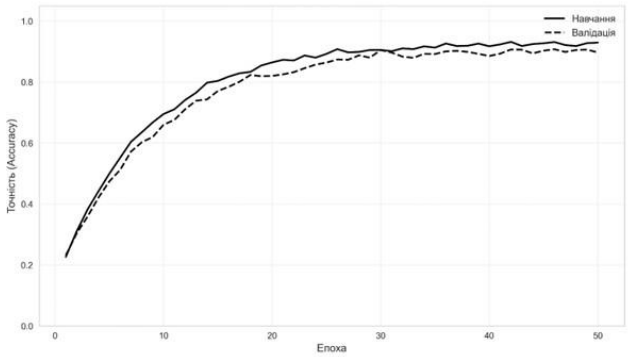


13

Динаміка функції втрат під час навчання

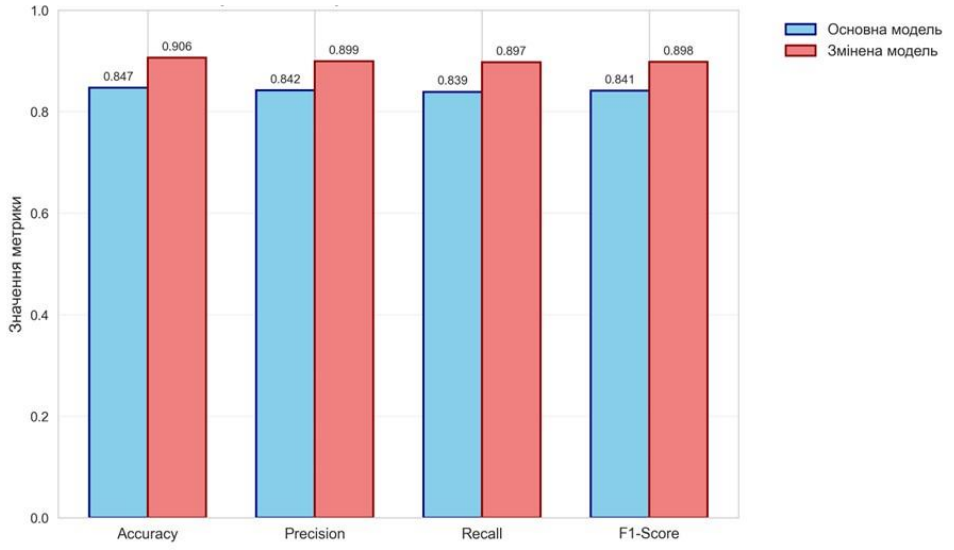


Динаміка точності під час навчання



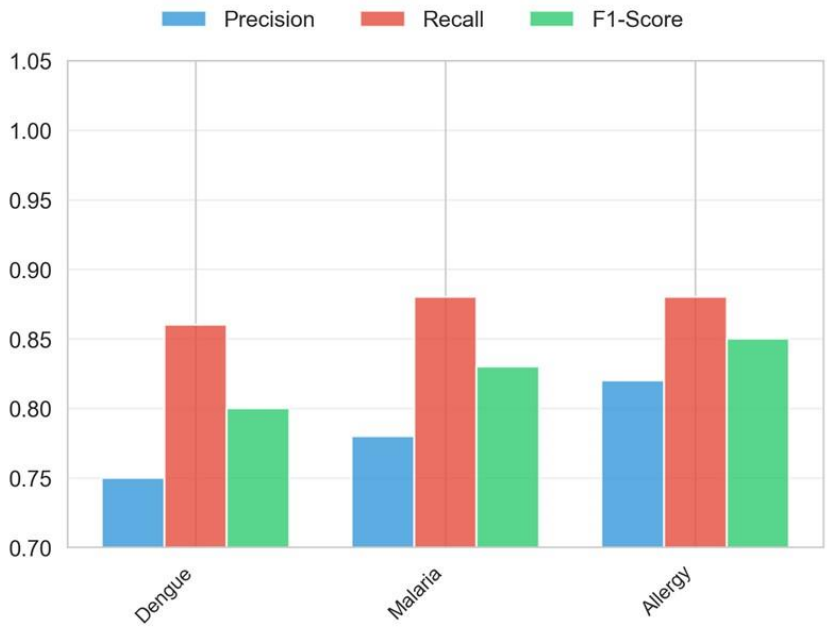
14

Порівняння метрик базової та зміненої моделі



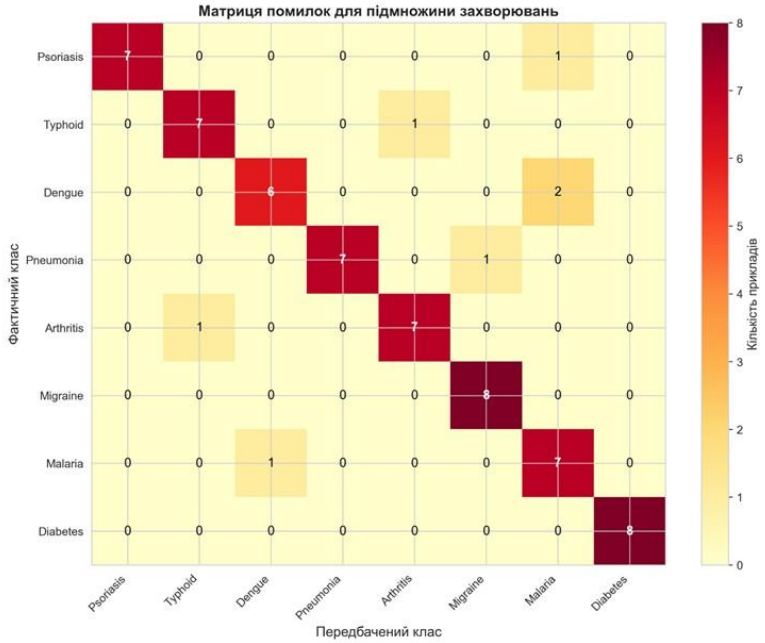
15

Метрики класифікації для різних захворювань



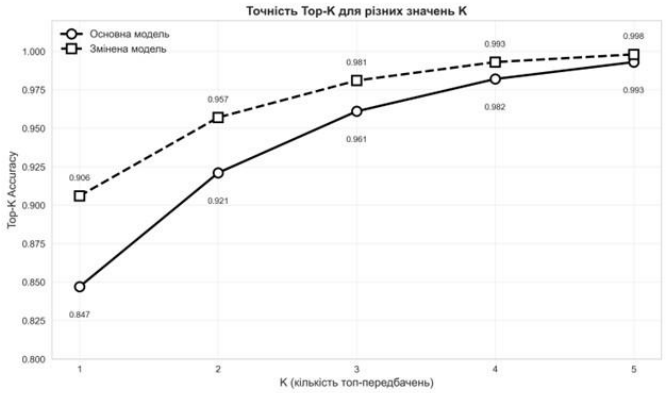
16

Матриця помилок для підмножини захворювань

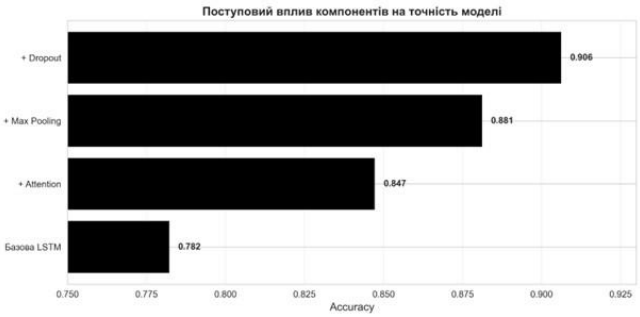


17

Точність Top-K для різних значень K



Поступовий вплив компонентів на точність моделі



18

Висновки

В результаті виконання роботи виконано?

- проведено аналіз існуючих методів та підходів до автоматизованої діагностики захворювань на основі текстових даних з використанням технологій обробки природної мови та машинного навчання;
- розроблено метод класифікації захворювань за описом симптомів на основі архітектури рекурентних нейронних мереж LSTM з механізмом уваги, що забезпечує виділення найбільш інформативних фрагментів у медичних текстах;
- спроєктувати програмну реалізацію метода попередньої обробки текстових даних та обробку заперечень для коректного врахування відсутності симптомів;
- проведено експериментальне дослідження запропонованого методу шляхом порівняння базової та модифікованої архітектури нейронної мережі, оцінити вплив окремих компонентів на загальну ефективність класифікації.

ДЯКУЮ ЗА УВАГУ!

Fri Dec 12 12:44:18 EET 2025, Петровський Сергій Степанович, Хмельницький національний університет, ХНУ

Anti-Plagiarism (UA) v-15.281 Educational

The maximum coincidence with one document 1.0%

Dictionaries check: en_US, ru_RU, ua_UA. **Errors in the documents: 9%**

	Document		Sum coincidence on the DB	
ID: 252662 Title: КВАЛІФІКАЦІЙНА РОБОТА на тему Метод діагностики захворювань за описом симптомів з використанням обробки природної мови Added in a DB: 2025-12-12 Authors: Олександр БОНДАР Heads: Олександр ПАСІЧНИК Consultants: Opponents:	Symbols	Lexemes	Symbols	Lexemes
	112901	1712	1764 (2%)	31 (2%)

Plagiarism sources

ID	Description	Plagiarism presence in the document	
		Symbols	Lexemes

Протокол аналізу звіту подібності науковим керівником

Заявляю, що я ознайомився (-лась) з Повним звітом подібності, який був згенерований Системою виявлення і запобігання плагіату щодо роботи:

Автор: Олександр БОНДАР

Співавтор:

Назва: КВАЛІФІКАЦІЙНА РОБОТА на тему Метод діагностики захворювань за описом симптомів з використанням обробки природної мови

Науковий керівник: Олександр ПАСІЧНИК, к.т.н., доцент

Підрозділ: Кафедра комп'ютерних наук

Коефіцієнт подібності 1:2.1%

Коефіцієнт подібності 2:0.2%

Мікропробіли: 0

Заміна букв: 1

Інтервали: 0

Білі знаки: 0

Дата створення звіту: 2025-12-11 21:46:18.0

Після аналізу Звіту подібності констатую наступне:

☒ **Запозичення, виявлені в роботі є законними і не є плагіатом. Рівень подібності не перевищує допустимої межі. Таким чином робота незалежна і приймається.**

☐ **Запозичення не є плагіатом, але перевищено граничне значення рівня подібностей. Таким чином робота повертається на доопрацювання.**

☐ **Виявлено запозичення і плагіат або навмисні текстові спотворення (маніпуляції), як передбачувані спроби укриття плагіату, які роблять роботу невідповідною вимогам законодавства (Ст. 32. ЗУ Про вищу освіту, пункт 3.1, Ст. 42. ЗУ Про освіту) та вимог НАЗЯВО (Критерій 5), а також кодексу етики і процедурам. Таким чином робота не приймається.**

Обгрунтування:

2025-12-12

Дата

експерт Недовесний С. Р.

РІШЕННЯ ЕКСПЕРТНОЇ КОМІСІЇ КАФЕДРИ КОМП'ЮТЕРНИХ НАУК

ПРО ДОПУСК КВАЛІФІКАЦІЙНОЇ РОБОТИ ДО ЗАХИСТУ

Назва кваліфікаційної роботи *Метод діагностики захворювань за описом симптомів з використанням обробки природної мови*
Автор *студент групи КНм-24-1 Олександр БОНДАР*
Освітня програма *Комп'ютерні науки*
Рівень вищої освіти *другий (магістерський)*
Спеціальність *122 – Комп'ютерні науки*
Науковий керівник: *к.т.н., доцент каф. комп'ютерних наук Олександр ПАСІЧНИК*

На основі аналізу кваліфікаційної роботи на дотримання вимог академічної доброчесності (у т.ч. відсутності ознак академічного плагіату) з урахуванням результатів перевірки роботи спеціалізованим програмними засобами комісія зробила такий висновок:

№	Висновок	Позначка про відповідність
1	Ознаки академічного плагіату	
1.1	Запозичення, виявлені в роботі, є законними і не є академічним плагіатом (далі – зазначаються підстави віднесення запозичень до правомірних, якщо потрібно). Робота приймається до захисту.	***відповідає***
1.2	Виявлені запозичення не є академічним плагіатом, розміщені в розділах, які не описують безпосередньо авторське дослідження, але кількість цитат перевищує обсяг, виправданий поставленою метою роботи (далі – зазначаються детальні та аргументовані підстави віднесення запозичень до правомірних). Робота приймається до захисту, але має бути відкоригована.	
1.3	Виявлені запозичення не є академічним плагіатом, але частково розміщені в розділах, які описують безпосередньо авторське дослідження, а кількість цитат перевищує обсяг, виправданий поставленою метою роботи. Робота може бути допущена до захисту після того як буде відкоригована та доопрацьована і успішно пройде повторну перевірку на академічний плагіат.	
1.4	Робота містить навмисні текстові спотворення, передбачувані спроби укриття текстових запозичень або інші прояви академічного плагіату. Робота містить фабрикацію або фальсифікацію даних. Робота не допускається до захисту.	
2	Інші види порушень академічної доброчесності	***відсутні***

Підтвердження:

Запозичення, виявлені в роботі Олександра БОНДАРА, не є плагіатом, оскільки: запозичення розміщені в розділі огляду існуючих підходів, не описують безпосередньо авторську роботу і не стосуються її результатів; усі запозичення фрагментарні; до запозичень входять фрагменти, які не мають авторства і містять поширені конструкції та загальновідомі терміни, скорочення. Рівень подібності не перевищує допустимої межі. Таким чином, робота є законною та приймається до захисту.

Обсяг запозичень, визначений системами виявлення збігів/ідентичності/схожості:
- за системою Anti-Plagiarism: 1%;
- за системою StrikePlagiarism КП1: 2,11%, КП2: 0,2%.

15.12.2025

Завідувач кафедри _____ *Олександр БАРМАК*

Гарант освітньої програми _____ *Руслан БАГРІЙ*

Керівник кваліфікаційної роботи _____ *Олександр ПАСІЧНИК*

ХМЕЛЬНИЦЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
МОН УКРАЇНИ

Кафедра комп'ютерних наук

ВІДГУК НАУКОВОГО КЕРІВНИКА

на кваліфікаційну роботу магістра

студента *КНм-24-1 Олександра БОНДАРА*

за темою *Метод діагностики захворювань за описом симптомів з використанням обробки природної мови*

1. Актуальність теми

Актуальність обраної теми дослідження визначається зростаючою потребою медичної галузі у інтелектуальних системах підтримки прийняття рішень на етапі первинної діагностики. Традиційні методи на основі правил та простих статистичних моделей демонструють обмежену ефективність через високу варіативність формулювань симптомів у природній мові. Застосування рекурентних нейронних мереж з механізмами уваги дозволяє моделювати складні залежності у медичних текстах та виділяти найбільш інформативні фрагменти для підвищення точності діагностики.

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

Магістерська робота повністю відповідає предметній області спеціальності 122 "Комп'ютерні науки", оскільки базується на застосуванні сучасних методів штучного інтелекту та машинного навчання, зокрема рекурентних нейронних мереж типу LSTM, механізмів уваги, технік попередньої обробки текстових даних та методів оцінювання якості класифікаційних моделей.

3. Професійні та особистісні якості

Протягом роботи над магістерським дослідженням Олександр БОНДАР продемонстрував глибоке розуміння методів обробки природної мови та архітектур глибоких нейронних мереж. Студент проявив здатність до самостійного аналізу великої кількості наукових джерел, критичного оцінювання існуючих підходів та обґрунтування власних архітектурних рішень.

4. Ступінь самостійності під час виконання кваліфікаційної роботи

При виконанні магістерської роботи студент виявив високий рівень самостійності та ініціативності. Олександр БОНДАР самостійно провів комплексний огляд наукової літератури з проблематики діагностики захворювань засобами машинного навчання, запропонував удосконалення базової архітектури нейронної мережі шляхом додавання механізму уваги та шару максимального пулінгу.

5. Наукова новизна та оригінальність запропонованих підходів

Удосконалено метод діагностики захворювань за текстовими описами симптомів, який відрізняється від наявних комбінованим використанням двонаправленої LSTM-архітектури з механізмом уваги та додатковим шаром максимального пулінгу, що дозволяє одночасно враховувати як контекстуальну важливість окремих симптомів через механізм уваги, так і найбільш виражені прояви захворювання через максимальний пулінг.

6. Ступінь оволодіння методами дослідження

Під час виконання магістерської роботи студент продемонстрував досконале володіння сучасними методами обробки природної мови та глибокого навчання, зокрема архітектурами рекурентних нейронних мереж. Олександр БОНДАР застосував експериментальні методи для оцінювання якості класифікації, продемонстрував здатність до критичного аналізу результатів.

7. Повнота та якість розкриття теми роботи

Тема магістерської роботи розкрита повно та всебічно. Робота характеризується логічною структурою, що охоплює аналіз існуючих підходів до діагностики захворювань засобами обробки природної мови, детальний опис розробленого методу на основі двонаправленої LSTM-архітектури з механізмом уваги, програмну реалізацію з модульною структурою та експериментальне дослідження на репрезентативному наборі медичних даних.

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

Магістерська робота відзначається чіткою логічною структурою та послідовним викладенням матеріалу: від аналізу проблематики медичної діагностики та огляду існуючих методів обробки природної мови до розробки власного методу, його програмної реалізації та експериментальної перевірки.

9. Можливість практичного застосування кваліфікаційної роботи, окремих її частин

Розроблений у магістерській роботі метод діагностики захворювань за текстовими описами симптомів має широкі перспективи практичного застосування у телемедичних платформах, мобільних додатках для первинної медичної консультації, системах тріажу у закладах охорони здоров'я та інформаційних системах підтримки прийняття медичних рішень.

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

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

Керівник _____ к.*т.н., доцент каф. КН Олександр ПАСІЧНИК*

ХМЕЛЬНИЦЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
МОН УКРАЇНИ

Кафедра комп'ютерних наук

ВІДГУК ОПОНЕНТА

на кваліфікаційну роботу магістра

студента *гр. КНм-24-1 Олександра БОНДАРА*

за темою *Метод діагностики захворювань за описом симптомів з використанням обробки природної мови*

1. Актуальність обраної теми

Обрана тема є актуальною у контексті сучасних викликів у медичній сфері, пов'язаних із забезпеченням своєчасної та якісної первинної діагностики в умовах дефіциту медичних кадрів та зростаючого навантаження на систему охорони здоров'я. Актуальність підсилюється розвитком телемедичних технологій і необхідністю автоматизації процесу первинного аналізу симптомів, особливо у віддалених регіонах.

2. Відповідність роботи предметній області спеціальності 122 Комп'ютерні науки та загальним вимогам до наукових робіт

Кваліфікаційна робота цілком відповідає предметній області спеціальності 122 Комп'ютерні науки, демонструючи глибоке застосування методів штучного інтелекту, обробки природної мови та глибокого навчання для вирішення практичної задачі класифікації медичних текстів. Робота базується на знаннях комп'ютерних наук, включаючи архітектури нейронних мереж, техніки обробки текстових даних та методи оцінювання якості моделей.

3. Повнота розкриття мети та завдань дослідження

Мета дослідження сформульована конкретно і чітко - підвищення точності та швидкості діагностики захворювань на основі текстових описів симптомів через розробку методу з використанням LSTM-архітектури та механізмів уваги. Поставлені завдання розкриті послідовно та повно Кожне завдання підкріплене відповідними експериментальними результатами та аналітичними висновками.

4. Наявність наукової новизни

Наукова новизна роботи визначається удосконаленням методу діагностики захворювань, який відрізняється комбінованим використанням двонаправленої LSTM-архітектури з механізмом уваги та додатковим шаром максимального пулінгу. Ця архітектурна особливість дозволяє одночасно враховувати контекстуальну важливість окремих симптомів у тексті та найбільш виражені прояви захворювання. Розроблений

підхід до обробки заперечень у медичних текстах для коректного врахування відсутності симптомів також вносить науковий внесок у методологію аналізу клінічних описів.

5. Зміст кожного розділу роботи

У першому розділі представлено всебічний аналіз задачі діагностики захворювань за текстовими описами, включаючи характеристику проблемної області, огляд існуючих наукових публікацій та підходів, а також систематизацію архітектур обробки природної мови для медичної діагностики. Другий розділ розкриває концепцію та схему розробленого методу, містить детальний опис архітектури нейронної мережі. Третій розділ присвячений практичній реалізації, де описано структуру програмної системи, модулі попередньої обробки. Четвертий розділ містить комплексне експериментальне дослідження з аналізом навчальних даних, порівняльною оцінкою базової та модифікованої моделей.

6. Ступінь розкриття теми роботи

Тема роботи розкрита повністю та всебічно. Автор проводить ґрунтовний аналіз проблематики автоматизованої діагностики за текстовими описами симптомів, розглядає існуючі методи обробки медичних текстів та обґрунтовує необхідність створення удосконаленого підходу. Описано архітектуру запропонованого методу на основі двонаправленої LSTM з механізмом уваги, продемонстровано механізм генерації векторних представлень симптомів. Експериментальні дослідження проведені на репрезентативному наборі даних медичних описів і підтверджують ефективність методу.

7. Якість оформлення кваліфікаційної роботи

Оформлення кваліфікаційної роботи загалом відповідає академічним стандартам і демонструє професійний підхід до представлення наукового матеріалу. Структура роботи логічна та послідовна, текст викладено грамотною науковою мовою.

8. Недоліки кваліфікаційної роботи

Серед недоліків роботи можна відзначити відсутність порівняння розробленого методу з сучасними трансформерними архітектурами, такими як BERT.

9. Загальний висновок (допускається чи не допускається до захисту), якої оцінки заслуговує кваліфікаційна робота

Враховуючи рівень виконання роботи, забезпечення всіх необхідних вимог, наявність наукової новизни отриманих результатів, якість проведених експериментальних досліджень, кваліфікаційна робота може бути допущена до захисту. Рекомендована оцінка – відмінно.

Опонент д. т. н., проф. каф. АКІТехн Нарушок В. В.