МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ «ЛЬВІВСЬКА ПОЛІТЕХНІКА»

Патерега Юрій Ігорович

УДК 004.942

ІНФОРМАЦІЙНА ТЕХНОЛОГІЯ ОПРАЦЮВАННЯ ПЕРСОНАЛІЗОВАНИХ ДАНИХ ДЛЯ АНАЛІЗУ СТАНУ ОСОБИ

Спеціальність 05.13.06 – інформаційні технології

АВТОРЕФЕРАТ

дисертації на здобуття наукового ступеня

кандидата технічних наук

Львів – 2024

Дисертацією є рукопис.

Робота виконана на кафедрі систем автоматизованого проектування Національного університету "Львівська політехніка" Міністерства освіти і науки України.

Науковий керівник:
кандидат технічних наук, доцент
Мельник Михайло Романович,
Національний університет «Львівська політехніка»,
доцент кафедри систем автоматизованого проектування.

Офіційні опоненти:

доктор технічних наук, професор
Корнага Ярослав Ігорович,
Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського», декан факультету інформатики та обчислювальної техніки, професор кафедри інформаційних систем та технологій;

доктор технічних наук, професор
Мулеса Оксана Юріївна,
ДВНЗ «Ужгородський національний університет» Міністерства освіти і науки України, професор кафедри програмного забезпечення систем.

Захист відбудеться "04" листопада 2024 р. о 15:00 годині на засіданні спеціалізованої вченої ради Д 35.052.14 Національного університету "Львівська політехніка" за адресою: 79013, м. Львів, вул. Митрополита Андрея, 3, IV навчальний корпус, кафедра САП, ауд. 322.

З дисертацією можна ознайомитися у Науково-технічній бібліотеці Національного університету "Львівська політехніка" за адресою: 79013, м. Львів, вул. Професорська, 1.

Автореферат розісланий "02" вересня 2024 р.

Вчений секретар
спеціалізованої вченої ради
кандидат технічних наук, доцент Анатолій БАТЮК

ЗАГАЛЬНА ХАРАКТЕРИСТИКА ДОСЛІДЖЕННЯ

Актуальність теми.

Зростання обсягів даних є одним із критично важливих трендів у сучасному світі. Щороку організації генерують і зберігають все більше інформації, що обумовлює необхідність створення та використання ефективних систем опрацювання та використання даних у процесах підготовки та прийняття персоналізованих рішень. Розроблення систем опрацювання персоналізованих даних є актуальним та важливим завданням для багатьох галузей, включаючи бізнес, освіту та інформаційні технології. Особливо значущим є застосування таких систем у медичній галузі.

Персоналізований підхід до діагностики та лікування, заснований на аналізі великих обсягів медичних даних, дає змогу виявляти приховані закономірності, більш точно прогнозувати розвиток захворювань та відповідно оптимізувати лікувальні заходи. Для ефективного розв'язання цих завдань застосовують ряд методів штучного інтелекту та машинного навчання, включаючи нейронні мережі, дерева рішень, методи опорних векторів, Random forest та інші. Це сприяє розвитку персоналізованої медицини, де лікування адаптується до індивідуальних генетичних, фізіологічних та інших особливостей кожного конкретного пацієнта.

Вагомий внесок у розвиток методів класифікації та прогнозування стану пацієнтів зробили вчені Бідюк П.І., Нільсон Н., Ang L.M., Seng K.P., Tang Yan, Wang Tsai. Зокрема, ними розроблені підходи до класифікації характеристик стану пацієнта. Важливі результати для виявлення асоціативних правил оцінки значущості параметрів стану пацієнта отримані авторами Арсеньєвим Ю.Н., Дюком В.А., Мельниковою Н.І., Субботіним С.О., Hunyadi D. та Runger G.C. Вчені Кореневський А.Н., Мужичик А.В., Бодянський Є. В., Зайченко Ю.П., Ткаченко Р.О., Мисник А.В., Attallah O., Lei Zhang зосередили свої зусилля на розробленні рішень на основі аналізу поточного стану пацієнта та прогнозуванні його змін.

Використання методів машинного навчання для персоналізованого опрацювання медичних даних пацієнтів є актуальним науковим завданням, спрямованим на підвищення точності діагностики та оптимізацію лікування, що в результаті має забезпечити покращення якості медичних послуг, особливо при захворюваннях головного мозку.

Зв'язок роботи з науковими програмами, планами, темами.

Дисертаційні дослідження виконувалися відповідно до пріоритетних напрямків науково-дослідних робіт Національного університету “Львівська політехніка” та координаційних планів Міністерства освіти і науки України. Дослідження проведені в межах науково-дослідної роботи кафедри систем штучного інтелекту Національного університету «Львівська політехніка», а саме: «Методи та засоби обробки, консолідації та аналізу персоналізованої медичної інформації» (номер державної реєстрації № 0120U00025). Її результати є складовою частиною проєктів, які виконувалися в межах держбюджетних науково-дослідних робіт та грантових проєктів: «Effectiveness of Medicine E-Learning Distance Courses» / «Ефективність дистанційних курсів медицини» реєстраційний номер на порталі Європейської Комісії 2022-1-IT02-KA220-HED-000087665.

Мета і завдання дослідження.

Метою дисертаційної роботи є розроблення та вдосконалення інформаційної технології опрацювання персоналізованих даних для покращення процесів аналізу стану особи і підвищення точності класифікації таких даних для ефективнішого лікування пацієнтів.

Для досягнення поставленої мети сформульовано та вирішено наступні завдання:

1. Провести аналіз існуючих інформаційних технологій підтримки прийняття лікарських рішень, визначити переваги та недоліки у зв'язку зі сферою їхнього застосування, відповідно до існуючих протоколів і стандартів, таких як GDRP та HL7.
2. Розробити інформаційну технологію аналізу стану особи.
3. Розробити метод класифікації персоналізованих даних шляхом введення етапу аугментації для опрацювання персоналізованих медичних даних пацієнтів.
4. Удосконалити метод персоналізації медичних даних особи внаслідок введення ансамблю обраних моделей класифікації, які забезпечують кращі результати класифікації, та ансамблевого голосування, що дасть змогу підвищити точність прогнозування стану особи при захворюваннях головного мозку.
5. Розробити архітектуру інформаційної системи опрацювання персоналізованих даних аналізу стану особи.
6. Апробувати на основі отриманих результатів інформаційну технологію опрацювання персоналізованих даних аналізу стану особи.

Об'єктом дослідження є процеси збору, обробки та аналізу персоналізованих даних про стан особи.

Предметом дослідження є методи машинного навчання - decision tree, random forest, k-nearest neighbors, ada boost, stacking, SMOTE, grid search, ResNet та CNN – для класифікації та пошуку рішень ідентифікації стану особи; структурні та об'єктноорієнтовані методи програмування – для розроблення інформаційної технології опрацювання персоналізованих даних для аналізу стану особи.

Наукова новизна отриманих результатів.

Полягає у розв'язанні актуального наукового завдання удосконалення процесу опрацювання персоналізованих даних внаслідок підвищення точності класифікації та зменшення кількості ітерацій в процесі машинного навчання шляхом застосування аугментації до навчальної вибірки.

Отримано такі нові наукові результати:

Вперше

– побудовано узагальнену модель інформаційної технології опрацювання персоналізованих даних для аналізу стану особи шляхом консолідації мультимодальних даних, яка дає змогу покращити процес ідентифікації стадії захворювання та пошук рішень для ефективного лікування;

– розроблено метод класифікації персоналізованих медичних даних шляхом введення етапу аугментації при опрацюванні медичної інформації про стан особи, що дало можливість збільшити обсяг та різноманітність навчальної вибірки, зменшити ризик перенавчання і забезпечити узагальнення моделей класифікації.

Удосконалено метод персоналізації даних особи, який, на відміну від наявних, використовує ансамбль моделей класифікації та ансамблеве голосування, що дало змогу підвищити точність прогнозування стану особи.

Практична цінність роботи полягає у досягненні таких результатів:

– створено комплекс моделей, методів, алгоритмів і програм, які покладені в основу функціонування інформаційної технології опрацювання персоналізованих даних для аналізу стану особи. Розроблено алгоритм обробки персоналізованих медичних даних особи для аналізу її стану, що дає змогу формалізувати процес підготовки даних пацієнтів з різними патологіями. Розроблено архітектуру інформаційної системи опрацювання персоналізованих даних, на основі якої реалізована прикладна інформаційна система опрацювання персоналізованих даних для аналізу стану особи;

– наукові результати дисертаційної роботи впроваджено при виконанні науково-дослідної роботи кафедри систем штучного інтелекту Національного університету «Львівська політехніка» за темою «Методи та засоби обробки, консолідації та аналізу персоналізованої медичної інформації» (номер державної реєстрації № 0120U00025) (акт впровадження від 14.06.2024 р.) та у лікувальний процес під експертизою Львівської асоціації алергологів, імунологів, імунореабілітологів (акт впровадження від 19.06.2024 р.);

– отримані результати досліджень використовуються в освітньому процесі Національного університету «Львівська політехніка» при підготовці фахівців першого (бакалаврського) рівня спеціальності 122 *Комп'ютерні науки* (акт впровадження від 09.04.2024 р.)

Особистий внесок здобувача.

Усі наукові результати дисертаційної роботи автор отримав самостійно. У працях, опублікованих у співавторстві, здобувачеві належать: [1] – розроблено алгоритм обробки персоналізованих даних особи для аналізу стану особи; [2] – розроблена архітектура інформаційної системи опрацювання персоналізованих даних для аналізу стану особи; [3] – удосконалено метод персоналізації медичних даних особи; [4] – розроблено метод класифікації персоналізованих даних; [5] – узагальнено модель інформаційної технології опрацювання персоналізованих даних; [6] – проведено аналіз існуючих сенсорів та давачів щодо збору клінічних даних особи; [7] – проведений аналіз існуючих рішень щодо опрацювання даних; [8] – розроблено програмні рішення опрацювання персоналізованих даних для аналізу стану особи; [9] – розроблено структури системи збору та опрацювання інформації; [10] – розроблено структури системи збору текстових мультимодальних даних; [11] – розроблено програмні засоби опрацювання та покращення якості вхідних даних; [12] – проведено аналіз відомих моделей клітин нейронних осциляторів та описано їх застосування; [13] – класифіковано застосування алгоритмів штучної еволюції.

Апробація результатів дисертації.

Основні результати наукових досліджень неодноразово доповідалися та обговорювалися на міжнародних науково-технічних конференціях, зокрема: «САПР у проектуванні машин. Питання впровадження та навчання»: XVIII Міжнар. укр.-пол.

наук.-техн. конф. CADMD'2010, 14–16 жовт. 2010, Львів, Україна; «Перспективні технології і методи проектування МЕМС»: 6-та міжнар. конф. MEMSTECH 2010, 20–23 квіт. 2010, Поляна, Україна, «Computer science and information technologies»: V Intern. sci. and techn. conf. CSIT 2010, 14–16 Oct. 2010, Lviv, Ukraine; під час виконання госпдоговору з благодійною організацією «Львівська асоціація алергологів, імунологів, імунореабілітологів» № 201-2023 від 30.10.2023 р. «Розроблення реєстру пацієнтів зі спадковим ангіоневротичним набряком (САН)».

Публікації.

Основні результати дисертації опубліковано у 13 наукових працях, зокрема: 5 статей – у наукових фахових виданнях України; 1 стаття – у науковому періодичному виданні іншої держави; 1 колективна монографія, 6 тез міжнародних науково-технічних конференцій.

Структура та обсяг дисертації.

Дисертаційна робота викладена на 152 сторінках та складається із змісту, вступу, чотирьох розділів, в яких містяться 57 рисунків, 17 таблиць, списку використаних джерел із 140 найменувань та 3 додатків.

ОСНОВНИЙ ЗМІСТ ДИСЕРТАЦІЙНОЇ РОБОТИ

У вступі обґрунтовано актуальність теми дисертаційної роботи, сформульовано мету та основні завдання досліджень, визначено наукову новизну роботи і практичне значення отриманих результатів, показано зв'язок роботи з науковими програмами, планами та темами. Подано відомості про апробацію результатів роботи, особистий внесок автора та його публікації.

У першому розділі проведено огляд наявних систем опрацювання персоналізованих даних, здійснено порівняльний аналіз рішень, які застосовуються для розв’язання медичних проблем за допомогою комп'ютерних технологій. Також подано опис та узагальнений аналіз використання поширених методів штучного інтелекту у медичних даних, висвітлено перспективи використання машинного навчання у діагностиці. Виокремлено особливості підходу, набору даних та задачі, що вирішуються кожним з розглянутих джерел. Проведено критичний аналіз вибраних наукових джерел, сформульовано та обґрунтовано основні проблеми, що виникають у наявних дослідженнях. Наприкінці розділу сформульовано мету та завдання подальшого наукового дослідження.

Аналіз особливостей опрацювання персоналізованих даних про особу дає змогу оцінити їх функціональність, ефективність та ступінь відповідності потребам користувачів. Цей аналіз є вирішальним для розуміння того, які можливості вже присутні на ринку, а також для ідентифікації прогалин і можливостей для подальшого вдосконалення.

Сьогодні існує широкий спектр систем опрацювання персоналізованих даних про пацієнтів, які використовуються у медичній сфері. Деякі з них включають:

- *електронні медичні записи (EMR),* які дають змогу збирати, зберігати та опрацьовувати медичні дані про пацієнтів, включаючи історії захворювань, результати обстежень, рецепти та інше;

- *системи управління клінічними даними (CDMS),* котрі дають змогу лікарням і медичним установам керувати клінічними даними, включаючи планування прийому пацієнтів, розподіл ресурсів та ведення медичної документації;
- *системи аналізу даних про здоров'я*, які використовують аналітичні технології для опрацювання великих обсягів медичних даних з метою виявлення тенденцій, виявлення ризиків та розробки індивідуальних стратегій лікування;
- *медичні портали та додатки для смартфонів,* що дають змогу пацієнтам отримувати доступ до особистих медичних даних, записуватися на прийоми, спілкуватися зі своїм лікарем та вести контроль за своїм здоров'ям;
- *системи геномної медицини,* використовують генетичні дані пацієнтів для розробки індивідуальних планів лікування та профілактики захворювань.

Визначено проблеми, що пов'язані з опрацюванням персоналізованих даних, які унеможливлюють досягнення високої точності і продуктивності моделей машинного навчання та методів штучного інтелекту. Відсутність адекватного розуміння та застосування ефективних методів аугментації даних для різних модальностей може вести до наступних проблем:

1. *Перенавчання (overfitting)* – моделі можуть стати занадто специфічними до навчального набору даних, що призводить до некоректного узагальнення їх передбачень на нові дані.

2. *Недостатня кількість даних* – відсутність достатньої кількості та якості даних може призвести до низької ефективності моделей та їх невідповідності для вирішення практичних завдань.

3. *Витрати часу та обчислювальних ресурсів* – без застосування оптимальних методів аугментації даних навчання моделей може стати витратним як за часом, так і за обчислювальними ресурсами.

Проведено аналіз особливостей опрацювання персоналізованих даних, який дав змогу встановити основні типи та методи, які застосовуються для їх аналізу. Результати наведено у таблиці 1.

Таблиця 1. Аналіз рішень, які використовуються в процесах класифікації

Інструменти дослідження	Типи даних	Висновок дослідження
Систематичний огляд	Різні види даних	Систематичний аналіз виявляє потенціал машинного навчання у сфері медичного обслуговування пацієнтів і проводить огляд останніх наукових досліджень.
Метод dropout	Зображення з ImageNet	Застосування глибоких згорткових нейронних мереж дає змогу досягти високої точності у класифікації зображень.
Логістична регресія, дерева рішень і нейронні мережі	Клінічні дані	Створення та валідація множинних моделей прогнозування відновлення, одужання та якості життя у пацієнтів.
ResNet	Радіологічні характеристики та клінічні змінні	Створення глибокої моделі навчання для передбачення мікроваскулярної інвазії у пацієнтів з гепатоцелюлярною карциномою, що була підтверджена в різних медичних установах.

Методи та засоби штучного інтелекту втілилися у моделях діагностики та лікування найрізноманітніших захворювань, включаючи онкологічні та кардіологічні захворювання. Однак, існують певні виклики та обмеження у використанні штучного інтелекту в медицині, такі як необхідність створення розширених наборів даних, ризик помилкових діагнозів та проблеми з етикою і конфіденційністю даних пацієнтів. У цілому, використання новітніх підходів до опрацювання даних у медицині є значним кроком до покращення якості діагностики та лікування пацієнтів.

Таблиця 2. Аналіз рішень, які використовуються в процесах аугментації даних

Інструменти дослідження	Розв'язувані задачі
Згорткові нейронні мережі (CNN), логістична регресія, аугментація даних.	Використання методології налаштування гіперпараметрів аугментації даних для підвищення ефективності глибоких згорткових нейронних мереж у класифікації зображень.
Згорткові нейронні мережі (CNN), рекурентні нейронні мережі з довготривалою короткочасною пам'яттю (LSTM), логістична регресія, метод опорних векторів (SVM).	Покращення раннього виявлення хвороби Паркінсона за допомогою методів машинного навчання та розширення набору даних за допомогою аугментації.
Генеративні змагальні мережі (GAN), згорткові нейронні мережі (CNN).	Розширення обсягу навчального набору та досягнення вищої точності за допомогою глибокого навчання.
Пошуковий аналіз даних (EDA)	Створення та аналіз методу генерації тексту, який спрямований на підвищення ефективності класифікаторів для текстів різної довжини шляхом розширення тренувального набору даних.
Аугментація даних, процес Маркова	Модель загального розширення даних як Марківський процес з урахуванням згенерованих ядер.
Згорткова нейронна мережа (CNN), мережа ResNet-20, стохастичний градієнтний спуск (SGD)	Оцінка впливу складності моделі на ефективність у завданнях з обмеженими тренувальними даними.

Дослідження наявних рішень процесу аугментації даних свідчать про можливість неефективного використання обчислювальних ресурсів у процесі розроблення моделей, що може негативно вплинути на їх продуктивність та точність. Це підкреслює важливість проведення досліджень і розробки методів аугментації даних для різних модальностей, що допоможе розв'язати вищезазначені проблеми та досягти кращих результатів у побудові і застосуванні моделей машинного навчання та штучного інтелекту (табл. 2).

На підставі проведеного аналізу визначено завдання досліджень, спрямовані на вдосконалення процесів опрацювання персоналізованих даних для аналізу стану особи, зокрема при нейродегенеративних захворюваннях, таких як хвороба Альцгеймера. Розроблені методи також були апробовані на інших типах медичних даних, що підтверджує їх універсальність та потенціал застосування в різних галузях медицини.

Ключовими напрямками досліджень визначено: розробку узагальненої моделі інформаційної технології аналізу стану особи; створення методу класифікації

персоналізованих даних з використанням етапу аугментації для підвищення ефективності при роботі з обмеженими наборами даних; удосконалення методу персоналізації медичних даних шляхом впровадження ансамблю моделей класифікації та ансамблевого голосування для підвищення точності діагностики.

У другому розділі представлено концептуальну модель, яка формалізує процес обробки персоналізованих даних. Ця модель консолідує етапи збору даних від периферійних пристроїв та інших джерел, передачі даних, а також етапи їх опрацювання та аналізу. Вона забезпечує комплексний підхід до роботи з персоналізованими даними, враховуючи вимоги бібліотеки шаблонів C-CDA, яка встановлює додатковий рівень вимог до базового стандарту HL7. Проаналізовано наявні давачі та протоколи, які забезпечують збір та передачу клінічних даних до місця їх опрацювання. Узагальнено модель інформаційної технології обробки персоналізованих даних для представлення стану пацієнта, беручи до уваги основні параметри його загального стану та визначені характеристики. Розроблено алгоритм опрацювання персоналізованих даних для аналізу стану пацієнта під час діагностування хвороб головного мозку та їх ускладнень, що дає змогу формалізувати процес підготовки даних пацієнтів, структуруючи етапи виконання попередньої підготовки даних, включаючи опрацювання зображень клінічних досліджень, пошук дублікатів, балансування та нормалізацію.

Після проведеного аналізу переваг і недоліків наявних систем синхронізовано етапи опрацювання інформації про особу в одному програмному модулі, який розв’язує задачу збору інформації, опрацювання даних, класифікації за характеристиками стану особи, валідації даних, контролю за відповідністю результатів та прогнозування наступних станів.

На рис. 1 подано концептуальну модель функціонування комплексної системи опрацювання та аналізу персоналізованої інформації. Дана концептуальна модель процесу є узагальненим поданням процесу, що описує його основні елементи, взаємозв'язки та функціональні характеристики. Головною метою концептуальної моделі є надання загального розуміння процесу при використанні для подальшої деталізації, аналізу та проєктування.

На відміну від відомих досліджень які пропонували формальну модель стану пацієнта, що спрямована на пошук оцінки його стану та рішень щодо оптимізації процесу одужання, для об'єктивної оцінки поточного стану пацієнта нами розроблено узагальнену модель, яка дає змогу представити пацієнта як систему, що відображає зв'язки між ключовими етапами процесу опрацювання даних та надає інформацію про його стан. Враховуючи потреби користувачів, система сприяє покращенню ефективності лікування завдяки більш точній та своєчасній інформації про стан пацієнта. Отже, модель стану пацієнта може бути представлена як система, що об'єднує різні елементи, подані у вигляді множин, які взаємозалежні та залежні від умов оцінювання.

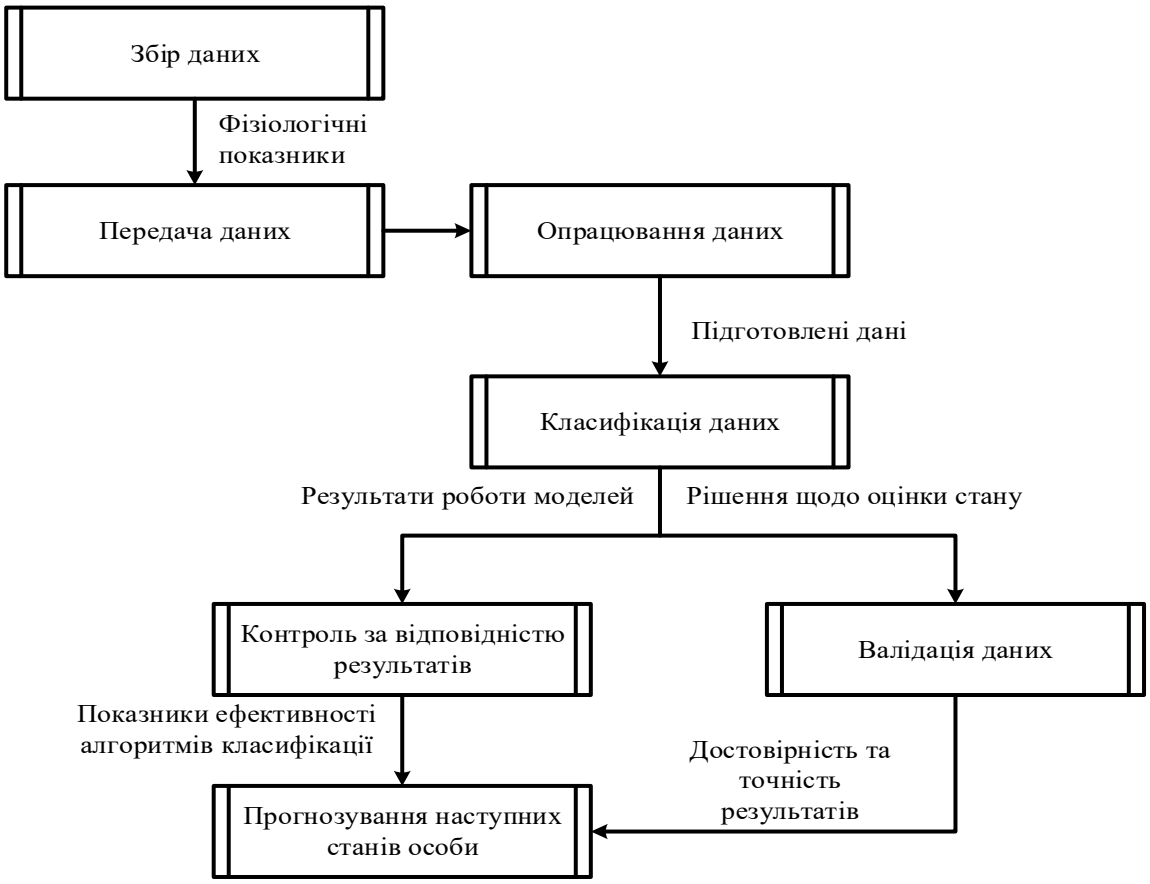


Рис. 1. Концептуальна модель процесу опрацювання персоналізованих даних

Формалізовано складові елементи інформаційної моделі системи аналізу стану особи. Необхідні давачі, які використовуються в дослідженні, подані множиною *S*, що формалізує пристрої, які забезпечують збір та передачу клінічних даних відповідними сенсорами. Отже, множина давачів для збору даних подана виразом:

$$S = \{s_1, s_2, \dots, s_n\}; \quad (1)$$

Показники давачів представлені множиною *C*, що формалізує дані особи після збору клінічних даних відповідними сенсорами, які належать до множини *S*, що відображені у формулі 2. Наприклад, температура, рівень електролітів, артеріальний тиск тощо. Кожен елемент із множини давачів *S* відповідає певному елементу з множини клінічних даних *C*. Тобто, давач s_1 відповідає клінічному показнику c_i. Отже, множина сенсорних даних відображена у формулі 2:

$$C = \{c_1, c_2, \dots, c_n\}; \quad (2)$$
$$S \to C; \text{де } \; s_i \to c_i.$$

Фізіологічні та антропометричні дані представлені множиною *A*, що формалізує дані особи після збору історії хвороби, наприклад: вік, супутні захворювання, шкідливі звички (паління, зловживання алкоголем). Отже, множина антропометричних та фізіологічних даних подана виразом 3, де *m* визначається кількістю необхідних показників, відповідно до протокольних рішень діагностованої патології:

$$A = \{a_1, a_2, \dots, a_m\}; \quad (3)$$

Тоді множина персоналізованих даних P становить об'єднання інструментальних та антропометричних показників:

$$C \cup A = P;$$
$$P = \{p_1, p_2, \dots, p_m\},\ k \le m + n. \quad (4)$$

База знань представлена у вигляді набору правил D. Припускається, що D - це набір персоналізованих рішень, який має скінченний розмір r =rank(D).

$$D = \{d_1, d_2, \dots, d_r\}; \quad (5)$$

Для прийняття персоналізованих рішень використано продукційні правила множини D. При цьому встановлюється залежність між множиною персоналізованих даних P та валідацією стану пацієнта V:

$$D: P \to V(K). \quad (6)$$

Зазначимо, що елемент множини персоналізованих рішень P є комплексним елементом, що складається із елементів множини клінічних даних C та антропометричних A, і подано як кортеж:

$$p_i = < C, \mathrm{A} >, \quad (7)$$

Тоді V – це множина вислідних показників стану особи, що залежить від множини гіперпараметрів класифікатора K.

Вибір правил здійснюється на основі багатокритеріального вибору, де: $V = \{v_1, v_2, \dots, v_q\}$ – векторна оцінка одержаних станів особи з урахуванням персоналізованих параметрів пацієнта та гіперпараметрів класифікаторів:

$$D(V) = \{x \in P \mid \forall y \in V(k_1, k_2, \dots, k_r) \quad (\forall i \in \{1,\dots, r\}[\, x_i \ge y_i])\}. \quad (8)$$

Прикладом правил є рішення щодо визначення оцінки стану визначеного класу захворювання на основі обраних давачів, клінічних даних, гіперпараметрів та рішень.

Отже, формалізоване представлення структури ключових елементів інформаційної технології опрацювання персоналізованих даних для аналізу стану особи, яка забезпечує пошук оцінки її стану та пошук рішень щодо покращення процесу ідентифікації стадії захворювання, подана у вигляді кортежу, як:

$$I = \langle S, \mathrm{C}, A, P, K, V, D \rangle, \quad (9)$$

де S – множина давачів для збору клінічних даних, C – множина клінічних вхідних даних системи, що характеризують стан пацієнта, які отримують з пристроїв клінічних досліджень, A – множина антропометричних даних про особу, P – множина персоналізованих даних, що залежить від клінічних та антропометричних даних особи, K – це множина гіперпараметрів класифікаторів хвороби особи, V – це множина вислідних показників стану особи, що залежить від множини гіперпараметрів класифікатора, D – множина правил, які визначаються з урахуванням персональних даних особи та вислідних показників стану особи.

Сформульована модель забезпечує комплексний підхід до аналізу стану пацієнта. Інтеграція різноманітних типів даних, включаючи дані з сенсорів, у рамках єдиної системи дає змогу досягти гнучкості та універсальності у порівнянні з підходами, що зосереджуються на окремих аспектах аналізу медичних даних.

Модель була застосована при аналізі стану пацієнтів з хворобою Альцгеймера, враховуючи клінічні й антропометричні показники, результати магнітно-резонансної

томографії головного мозку та іншу персоналізовану інформацію. Це дало змогу створити повнішу картину стану пацієнта та потенційно підвищити точність класифікації стадії захворювання.

Запропонована структура моделі демонструє адаптивність до різних сценаріїв у медичних системах, що особливо важливо в контексті персоналізованої медицини. Додаткові експерименти, проведені з використанням різноманітних медичних даних, включаючи рентгенівські знімки легень та клінічні показники пацієнтів з серцево-судинними захворюваннями, підтвердили потенціал моделі для аналізу стану особи при різних патологіях. Ці дослідження продемонстрували здатність моделі адаптуватися до різних типів вхідних даних та клінічних сценаріїв, що підкреслює її універсальність та широкі можливості застосування в різних галузях медицини. Подальші дослідження спрямовані на валідацію моделі на великих клінічних наборах даних та оптимізацію алгоритмів прийняття рішень з використанням передових методів машинного навчання.

Для попереднього опрацювання персоналізованих даних розроблено *алгоритм обробки персоналізованих даних особи* для аналізу стану особи, який формалізує процес підготовки даних пацієнтів з різними патологіями через послідовність визначених кроків:

1. Завантаження зображення клінічних досліджень у систему опрацювання даних. Для кожного файлу зображення завантажується за допомогою бібліотеки opencv-python, після чого обчислюється хеш-сума зображення методом SHA-256 із бібліотеки hashlib. Якщо хеш-сума зображення вже існує у словнику хеш-сум завантажених зображень, це свідчить про те, що зображення є дублікатом. У протилежному випадку, зображення додається до масиву зображень для подальшої обробки, а його хеш-сума та ім'я зберігаються у словнику. Крім того, ця функція додатково зберігає лічильники для підрахунку статистики дублікатів та загальної кількості зображень.

2. Перевірка на наявність дублікатів у наборі даних. Для виявлення дублікатів використано алгоритми хешування для швидкої перевірки ідентичності зображень та метрики схожості.

Набір клінічних даних описується множиною $\mathrm{C} = \{\mathrm{c}_i\}_{i=1}^{n}$. У випадку зображень $\mathrm{c}_i = (x_i, y_i)$.

3. Балансування набору даних зображень. Проблема незбалансованості набору даних зображень виникає, коли кількість зображень у різних класах неоднакова. Це може спричинити перегин уваги моделі навчання на користь категорій з більшою кількістю зображень, що своєю чергою може знизити точність класифікації для менш представлених категорій.

Набір даних зображень:

$$\mathrm{C} = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$$

де x_i – це зображення, а y_i – відповідна мітка класу. Припустимо, що існує L_C класів, і кожен клас c має n_c зображень.

Балансування набору даних:

- Визначення кількості зображень у кожному класі:

Для кожного класу визначено кількість зображень n_c: $n_c = \sum_{i=1}^{n} I(y_i = c)$

де I – індикаторна функція, яка дорівнює 1, якщо $y_i = c$ і 0 в іншому випадку.

- Визначення цільової кількості зображень у кожному класі:

Нехай N – бажана кількість зображень у кожному класі після балансування, де N – середнє значення кількості зображень у всіх класах: $N = \frac{\sum_{j=1}^{L_C} n_c}{L_C}$

- Балансування класів

Є кілька підходів до балансування набору даних:

Oversampling (перевибірка) – метод, що включає додавання копій наявних зображень з недостатньо представлених класів.

Undersampling (недовантаження) – метод, який полягає у видаленні деяких зображень з надмірно представлених класів.

Для кожного класу c з n_c<N, додамо $N-n_c$ копій наявних зображень:

$$C' = \{C \cup (x_i, y_i) | y_i = c, \text{ копій} = N - n_c\}$$

Балансування набору даних зображень є важливим кроком для забезпечення рівномірного представлення всіх класів. Це допомагає створити більш надійну і справедливу модель, яка неупереджена до певних класів через дисбаланс даних.

Таким чином, на цьому етапі опрацювання необхідно провести аналіз через додаткові підзадачі балансування даних: аналіз повноти даних, вилучення викидів, кодування категоріальних ознак, масштабування атрибутів, аугментація даних.

Для збалансування набору даних додано симульовані зображення до менш представлених класів. Це було реалізовано за допомогою аугментації даних шляхом, який детально описано у 3-му розділі.

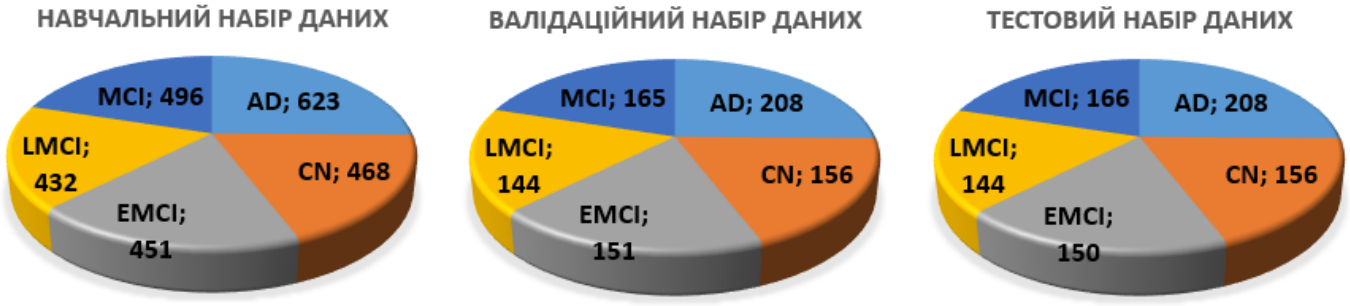
4. Нормалізація зображень.

Для стандартизації шкали яскравості зображень МРТ головного мозку на п'ять стадій хвороби Альцгеймера застосована нормалізація. Цей процес виконаний з метою кращого контролю над впливом різних джерел світла та інших відмінностей між зображеннями. Нормалізація полягала у приведенні значень пікселів до діапазону від 0 до 1, що здійснювалося за допомогою стандартної формули (10):

$$normalized\_X = \frac{X - \min(X)}{\max(X) - \min(X)} \quad (10)$$

де X – оригінальне зображення, $\min(X)$ – мінімальне значення пікселя зображення, $\max(X)$ – максимальне значення пікселя зображення.

Отримані дані розділені на тренувальний, валідаційний та тестовий набори. Тренувальний набір використаний для навчання окремих моделей класифікації, валідаційний – для налаштування гіперпараметрів моделей та оцінки їхньої якості, а тестовий – для оцінки загальної ефективності ансамблю моделей на раніше небачених даних. Розподіл даних збережено з балансом класів у кожному наборі у відношенні 60:20:20. На рис. 2 показано розподіл набору даних та кількість зображень для кожного класу у нових наборах.



а) б) в)

Рис. 2. Діаграми розподілу зображень для кожного класу в різних наборах даних: (а) – для навчального, (б) – для валідаційного, (в) – для тестового.
(AD - хвороба Альцгеймера, CN – когнітивна норма, EMCI – раннє легке когнітивне порушення, LMCI – пізнє легке когнітивне порушення, MCI – легке когнітивне порушення).

У третьому розділі розроблено метод класифікації персоналізованих даних шляхом введення етапу аугментації. Проведено аналіз застосування процесу аугментації даних різних модальностей, що дало змогу виявити, що надмірно спотворені аугментовані дані можуть призвести до некоректного передбачення класу, удосконалено метод персоналізації медичних даних шляхом введення ансамблю моделей класифікації та ансамблевого голосування. Досліджено два типи ансамблевого голосування – жорстке голосування (hard voting) і м'яке голосування (soft voting) та оцінено їх вплив на точність класифікації.

Розроблено метод класифікації персоналізованих даних шляхом впровадження етапу аугментації, що забезпечило розширення обсягу та різноманіття навчальних даних, сприяло кращому узагальненню моделей і зменшенню ризику перенавчання. Даний метод включає наступні кроки:

1. Збір даних. Необхідно отримати набір даних, який містить вектори ознак та їхні відповідні мітки класів. Набір даних у розрізі даного дослідження - це персоналізовані дані особи $P = (p_1, p_2, \ldots, p_n)$, де n – кількість спостережень.
2. Попередня обробка. Вона включає очищення, нормалізацію або створення нових ознак для забезпечення якості та придатності даних для моделювання. Задачу попередньої обробки, що враховує процес аугментації, можна формально представити як застосування набору перетворень. $T=\{t_1,t_2,\ldots,t_n\}$ до кожного зразка даних p_i. Кожна трансформація t_j визначається функцією $t_j:P\rightarrow P$.

Аугментація даних дає змогу розширити обсяг та різноманітність навчального набору шляхом застосування різноманітних перетворень до наявних даних. Це сприяє покращенню здатності моделі до узагальнення та зниженню ризику перенавчання.

3. Визначення найбільш релевантних особливостей є кроком у виборі ознак, які мають значний вплив на кінцевий результат моделі. Цей процес допомагає покращити продуктивність моделі та зменшити обчислювальну складність шляхом виключення нерелевантних або зайвих ознак.

Нехай матриця даних

$$P \in R_n \times p, \tag{11}$$

де n – кількість спостережень (зразків), p – кількість ознак (функцій).

Вектор міток класів $y \in R_n$ – мітки класів для кожного зразка зазвичай представляються у вигляді вектора, де кожен елемент відповідає конкретному зразку і містить ідентифікатор або назву класу, до якого він належить.

4. Підбір моделі. На цьому етапі важливо вибрати належний класифікатор, враховуючи особливості даних та вимоги проблеми. Алгоритми класифікації включають логістичну регресію, метод опорних векторів (SVM), дерева рішень, Random Forest, метод k-найближчих сусідів (KNN) і штучні нейронні мережі.

Вибір алгоритму класифікації є ключовим етапом у процесі машинного навчання, що суттєво впливає на точність, ефективність та узагальнення моделі. Крім того, важливо визначити гіперпараметри моделі, такі як кількість дерев у Random Forest або параметр регуляризації в SVM.

5. Навчання моделі полягає у розділенні позначеного набору даних на набір для навчання та набір для перевірки (або тестування). Використовуючи навчальний набір, класифікатор «навчається» встановлювати взаємозв'язки між вхідними ознаками та мітками класу, оптимізуючи його внутрішні параметри. Залежно від обраного алгоритму, цей етап може включати оптимізацію функції втрат, побудову границь прийняття рішень або вивчення ваг ознак.

Навчання моделі проводиться з використанням навчального набору згідно виразу:

$$y' = f(P;\theta) \tag{12}$$

де P – вхідні дані, θ – параметри моделі, y' – передбачення моделі.

Кількість етапів навчання моделі Q залежить від підготовлених даних, включаючи використання методів аугментації даних. Ці методи дають змогу розширити обсяг та різноманітність навчального набору, застосовуючи різні перетворення до існуючих даних. Це сприяє покращенню здатності моделі до узагальнення та зменшує ризик перенавчання. Якість синтезованих зображень визначається за допомогою набору перетворень, які застосовуються до вихідних даних $T=\{t_1,t_2,\ldots,t_n\}$.

Отже, кількість етапів навчання моделі на навчальному наборі персональних даних розширеними синтезованими даними визначається із залежності:

$$\mathrm{Q} = \text{Number of Epochs} \times \frac{\text{Size of Training Set} \times K_{transf}}{\text{Batch Size}}, \tag{13}$$

$$K_{transf} = \sum_{i=1}^{n} \frac{p_i}{t_i}/n, \text{ де } K_{transf} \in (0,1),$$

де Number of Epochs – кількість епох, Size of Training Set – розмір навчальної вибірки, Batch Size – розмір даних (зразків), що використовується для одного кроку (ітерації) навчання, K_{transf} – коефіцієнт перетворень.

Оцінка моделі полягає в перевірці її продуктивності та здатності до узагальнення за допомогою тестового набору даних, що містить невидимі екземпляри. Показники,

такі як точність (precision), повнота (recall) та F1-score, використовують для кількісної оцінки успішності класифікатора у передбаченні правильних класів для кожного зразка. Точність є основним критерієм для багатьох задач класифікації і визначає частку правильно класифікованих зразків у тестовому наборі даних.

Удосконалено метод персоналізації медичних даних за допомогою введення ансамблю моделей класифікації та ансамблевого голосування, що підвищило точності прогнозування стану особи. Він включає наступні кроки:

1. *Збір, підготовка, вибір ознак персональних даних*, як детально описано в методі удосконалення класифікації.

2. *На етапі вибору моделі* проведено процес підбору гіперпараметрів, опрацьовані моделі Decision Tree, Random Forest, SVM та MLP та проаналізовано у четвертому розділі ефективність моделей на такого типу даних і визначено, що Random Forest, SVM та MLP показали найкращі результати метрик.

Математичне подання процесу класифікації з використанням моделей Random Forest, SVM та MLP включає опис кожної з моделей та їх інтеграцію в ансамблеву модель класифікації.

2.1. Random Forest складається з ансамблю дерев рішень $h_1(p), h_2(p), \dots, h_t(p)$, де t – кількість дерев у лісі.

2.1.1. Побудова дерева

Для кожного дерева b:

Вибрано випадковий набір ознак Pt з загального простору ознак y.

Навчено дерево $h_t(p)$ на випадковій підмножині навчальних даних Pt.

Класифікація:

Результат класифікації отримано шляхом голосування дерев:

$$y' = mode(h_1(p), h_2(p), \dots, h_t(p)) \quad (14)$$

2.2. Support Vector Machine (SVM) здійснює намагання знайти гіперплощину, яка максимально відділяє класи одного від одного.

Нехай $\{(p_i, y_i)\}_{1=n}^{n}$ – набір навчальних даних, де $p_i \in R_n$ – вектор ознак, а $y_i \in \{-1,1\}$ – мітка класу.

$$min_{w,t} \frac{1}{2} ||w||^2 \quad (15)$$

з умовою: $y_i(w \cdot x_i + t) \geq 1, \ \forall i = 1,\dots,n$

2.3. Multi-Layer Perceptron (MLP) складається з одного або декількох шарів нейронів. Основні компоненти MLP включають вхідний шар, приховані шари та вихідний шар.

2.3.1. Передача сигналу (Feedforward):

Для шару l: $a_l = \sigma(W_l a_l - 1 + b_l)$,

де: al - активація нейронів шару l, Wl- матриця вагів між шарами $l-1$ та l, b_l - вектор зміщення для шару l, σ – функція активації (наприклад, сигмоїдна).

2.3.2. Зворотне поширення помилки (Backpropagation):

Оновлення ваг здійснюється за допомогою алгоритму зворотного поширення помилки:

$$\mathrm{W}^l := \mathrm{W}^l - \eta \frac{\partial \mathrm{L}}{\partial \mathrm{W}^l},\ \mathrm{b}^l := \mathrm{b}^l - \eta \frac{\partial \mathrm{L}}{\partial \mathrm{b}^l},$$

де η - швидкість навчання, L - функція втрат.

3. *Ансамблева модель.* Ансамблеве голосування об'єднує результати класифікації від різних моделей (Random Forest, SVM, MLP).

3.1. Hard Voting:

$$y' = mode(y'RF, y'SVM, y'MLP), \quad (16)$$

де: $y'RF$– прогноз моделі Random Forest, y'SVM - прогноз моделі SVM, y'MLP - прогноз моделі MLP.

Soft Voting:

$$y' = \arg max_y \left(\sum_m w_m P_m\left(y|p\right)\right), \quad (17)$$

де: $P_m(y|p)$ - ймовірність класу y для зразка p за моделлю m, w_m - ваговий коефіцієнт для моделі m.

4. *Узагальнена модель.* Кінцеве рішення ансамблевої моделі формується на основі об'єднаних результатів індивідуальних моделей, що сприяє підвищенню точності класифікації та зменшенню ризику перенавчання. У такому математичному описі процесу класифікації використано моделі Random Forest, SVM та MLP, включаючи їхні алгоритми навчання та інтеграцію в ансамблеву модель класифікації за допомогою методу голосування. Результати порівняльного аналізу чотирьох моделей наведено у таблиці 3.

Таблиця 3. Зведені результати тестування досліджуваних моделей

Модель	Метрики	Стадії хвороби				
		AD	CN	EMCI	LMCI	MCI
Decision Tree	Precision	0.87	0.85	0.90	0.79	0.95
	Recall	0.91	0.84	0.87	0.84	0.89
	F1-score	0.89	0.84	0.88	0.81	0.92
	Accuracy	0.87				
Random Forest	Precision	0.96	0.93	0.94	0.88	0.98
	Recall	0.94	0.90	0.97	0.93	0.95
	F1-score	0.95	0.92	0.96	0.91	0.96
	Accuracy	0.94				
SVM	Precision	0.95	0.92	0.93	0.92	0.96
	Recall	0.93	0.94	0.97	0.90	0.95
	F1-score	0.94	0.93	0.95	0.91	0.95
	Accuracy	0.94				
Multi-Layer Perceptron	Precision	0.95	0.94	0.95	0.91	0.97
	Recall	0.96	0.93	0.97	0.92	0.95
	F1-score	0.95	0.94	0.96	0.92	0.96
	Accuracy	0.95				

Отже, серед чотирьох досліджених моделей найвищі показники ефективності класифікації продемонстрували Random Forest, Multi-Layer Perceptron та SVM. Ці методи об’єднані в один ансамбль для досягнення стабільніших результатів. Метод Decision Tree показав найнижчі результати класифікації.

У четвертому розділі розроблено архітектуру інформаційної системи підтримки прийняття медичних рішень, яка базується на аналізі стану особи на основі опрацювання персоналізованих медичних даних. Представлено функціональну схему інформаційної системи, проведено аналіз ефективності роботи моделей класифікації та здійснено пошук найкращих гіперпараметрів. Проведено порівняльний аналіз застосування наявних моделей класифікації, проаналізовано наявні методи побудови ансамблевого навчання та впроваджено використання VotingClassifier. У рамках цього дослідження проведений порівняльний аналіз розглянутих у третьому розділі методик та розробленого ансамблю моделей. Подано результати імплементації інформаційної системи для супроводу процесу збору та аналізу стану особи.

У межах даного дослідження розроблено інформаційну систему для обробки персоналізованих медичних даних, ключовим компонентом якої є вебзастосунок. Ця система забезпечує взаємодію з користувачами через API-сервіси, зберігає необхідні дані у базі даних та включає необхідні апаратні та програмні компоненти для повноцінного функціонування в медичному середовищі.

Розроблена архітектура інформаційної системи підтримки прийняття медичних рішень щодо аналізу стану особи на підставі опрацювання персоналізованих медичних даних, яка представлена на рис. 3.

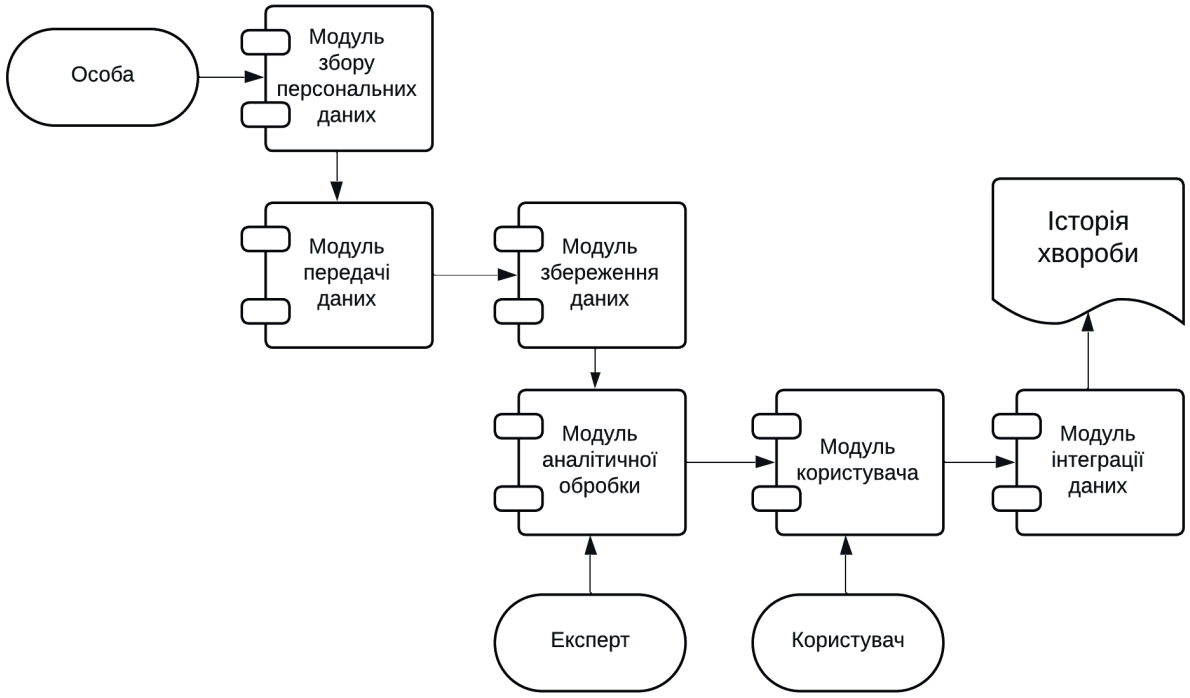


Рис. 3. Загальна архітектура інформаційної системи опрацювання персоналізованої інформації аналізу стану особи

Основними компонентами та модулями архітектури є:

1. *Модуль збору даних*. Забезпечує збір даних з різних джерел, а саме: давачі (сенсори): електрокардіографічні (ЕКГ) давачі, пульсоксиметри, глюкометри, тонометри, температурні давачі, біохімічні давачі, давачі руху та активності, монітори сну, спірометри, капнографи, імплантовані давачі, нейродавачі.

2. *Модуль передачі даних*. Забезпечує передачу клінічних даних особи засобами використання комунікаційних технологій: бездротові протоколи: Bluetooth, Wi-Fi, Zigbee, мобільні мережі: 4G, 5G.

3. *Модуль збереження даних*. Забезпечує збереження персональних даних особи на серверах обробки даних або хмарних платформах: AWS, Microsoft Azure, Google Cloud.

4. *Модуль аналітичної обробки даних.* Забезпечує опрацювання одержаних даних щодо аналізу та класифікації стану особи з використанням алгоритмів машинного навчання та штучного інтелекту.

5. *Модуль користувача*. Забезпечує використання інтерфейсних рішень візуалізації одержаних персональних даних та відповідних рішень користувачу системи: мобільні/веб додатки, медичні хаби.

6. *Модуль забезпечення інтеграції.* Забезпечення синхронізації даних з іншими існуючими медичними системами.

Функціонал розробленої інформаційної технології представлено на рис. 4.

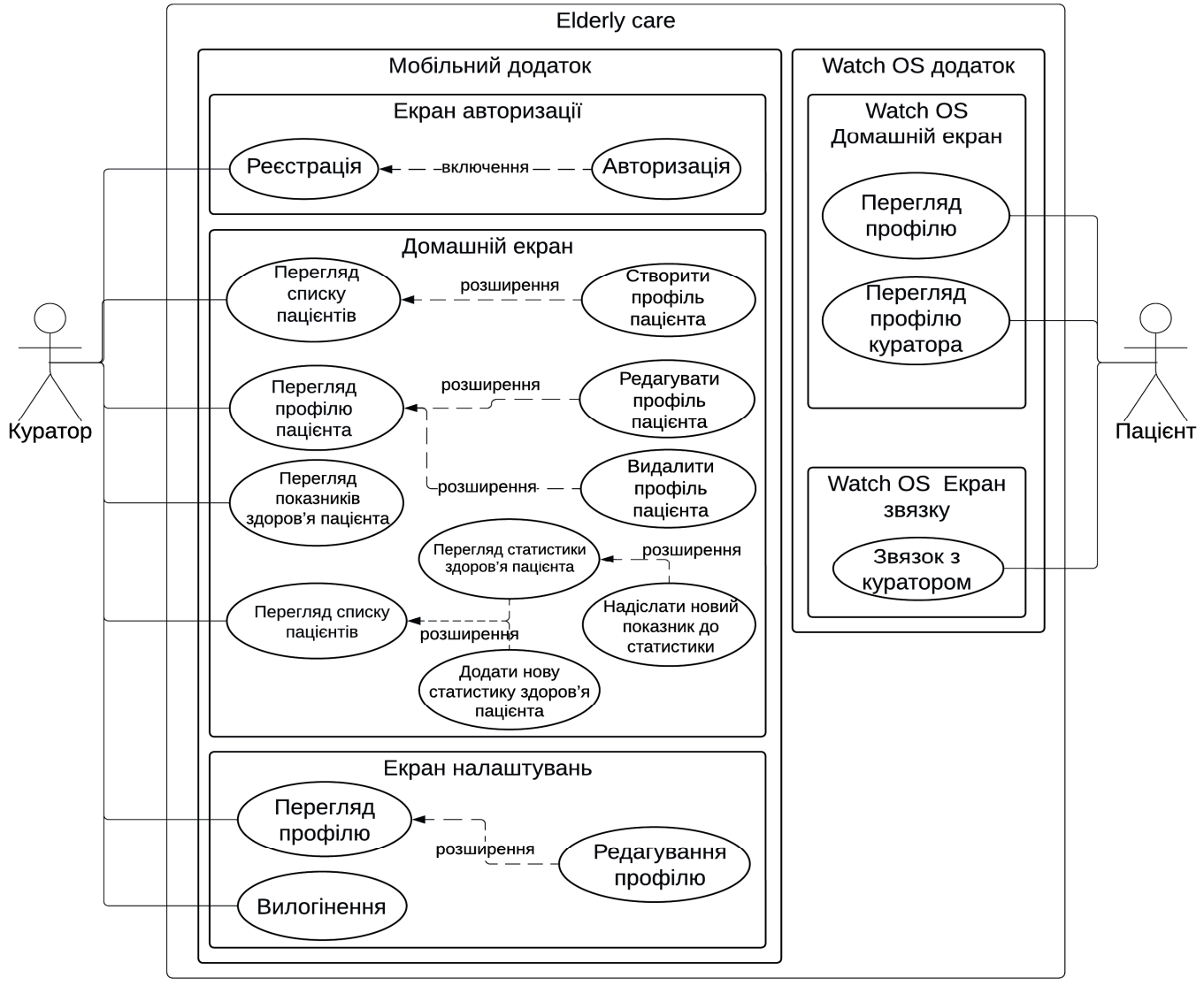


Рис. 4. Розширена UML-діаграма використання інформаційної технології

Функціональна схема інформаційної технології включає наступний функціонал, представлений у діаграмі з двома акторами - «Куратор» (Supervisor) і «Пацієнт» (Patient). Кожен актор має доступ до відповідної підсистеми: «Куратор» - до мобільного застосунку (Mobile App), а «Пацієнт» – до застосунку для смарт-годинника (WatchOS App). Методи використання розділені на логічні групи, що відображаються на екранах застосунків.

Запропоновані рішення імплементовані в межах розробки інформаційної системи аналізу та супроводу пацієнта. Система складається з двох основних частин: серверна складова (back-end) та клієнтська складова (front-end).

Для оцінки результатів тренування та тестування моделей проаналізовано роботу базових моделей з гіперпараметрами за замовчуванням на тестових даних.

На основі знайдених гіперпараметрів побудовано нові покращені моделі і досягнуто кращих результатів порівняно з базовими моделями (табл. 4).

Таблиця 4. Результати кращих моделей

Model	Precision	Recall	F1-score	Accuracy
Random Forest Classifier(покращена)	0.93 для 0 класу та 0.89 для 1 класу	0.94 для 0 класу та 0.86 для 1 класу	0.9	0.91
Random Forest Classifier(базова)	0.93 для 0 класу та 0.87 для 1 класу	0.94 для 0 класу та 0.86 для 1 класу	0.86	0.9
Stacking Classifier(базова)	0.94 для 0 класу та 0.87 для 1 класу	0.93 для 0 класу та 0.88 для 1 класу	0.87	0.91

Отже, модель Random Forest Classifier, яка покращена методом Grid Search, дає кращі результати для поставленого завдання. Вона продемонструвала дуже стабільні результати для обох класів і досягла значення F1-score 90%, що є дуже хорошим результатом.

Використовуючи описані вище моделі та методи машинного навчання, проведено передбачення цільового класу текстів та зображень на оригінальних наборах даних та аугментованих. Також проведено порівняльний аналіз отриманих результатів на оригінальних і штучно збільшених даних.

Аналіз результатів текстового набору даних з використанням вищезазначених методів аугментації показав, що завдяки кожному методу аугментації досягнуто балансу класів, а також збільшено розмірність кожного класу удвічі. Апробовано різні типи аугментації для текстових даних в аугментованому наборі, зокрема із заміною синонімів, випадковими перестановками та випадковими вставками, що дозволило підвищити точність та F1-score на 9%.

При застосуванні аугментації на зображеннях рентгенівських знімків точність (precision) у визначенні пневмонії значно підвищилась, досягаючи 88%. Чутливість (recall) також зросла до 93%, забезпечуючи майже повне виявлення всіх випадків пневмонії. Порівняно з попередніми результатами F1-score зросла до 90%.

Загалом, точність моделі значно зросла, досягаючи 87%. Це свідчить про те, що модель значно ефективніше розпізнавала стан хворих як при пневмонії, так і здорових пацієнтів.

Для проведення узагальненого порівняльного аналізу основних метрик класифікації, що стосується створених ансамблів з типами голосування hard voting та soft voting з метою вибору найкращого ансамблю, їх результати об'єднано в одну таблицю 5.

Таблиця 5. Об’єднані результати тестування ансамблів

Ансамбль	Метрики	Стадії хвороби				
		AD	CN	EMCI	LMCI	MCI
Hard Voting	Precision	0.98	0.95	0.93	0.91	0.97
	Recall	0.96	0.95	0.99	0.92	0.93
	F1-score	0.97	0.95	0.96	0.92	0.95
	Accuracy	0.95				
Soft Voting	Precision	0.98	0.95	0.94	0.94	0.97
	Recall	0.97	0.97	0.99	0.91	0.94
	F1-score	0.97	0.96	0.96	0.92	0.95
	Accuracy	0.96				

Узагальнюючи результати класифікації, встановлено, що обидва методи - hard voting та soft voting є досить ефективними. Розглянувши результати для обох типів ансамблів, зроблено такі висновки:

Hard voting: precision для всіх класів знаходиться у діапазоні від 0.90 до 0.98, що свідчить про високу точність моделі у класифікації, показник recall змінюється від 0.92 до 0.99, підтверджуючи добру здатність моделі виявляти дійсні екземпляри для кожного класу, загальна точність досягає 0.95, що свідчить про загальну ефективність моделі.

Soft voting: precision для всіх класів також знаходиться у діапазоні від 0.94 до 0.98, демонструючи високу точність класифікації, показник recall варіюється від 0.91 до 0.99, що підтверджує здатність моделі виявляти дійсні екземпляри для кожного класу, загальна точність підвищується до 0.96, що вказує на високу ефективність моделі.

Отже, обидва типи ансамблів продемонстрували добрі результати з високими значеннями precision та recall для більшості класів. Soft voting виявився дещо кращим за hard voting у забезпеченні точності та здатності виявляти дійсні екземпляри. Загалом, обидва типи ансамблів ефективно працюють у класифікації досліджуваного набору даних.

Зазначені методи, застосовані первісно для класифікації стадій хвороби Альцгеймера, також були використані для аналізу інших захворювань, демонструючи універсальність та гнучкість запропонованих підходів. Це підтверджує потенціал розширеного застосування розробленої класифікаційної системи у медичній діагностиці різних захворювань, відкриваючи нові можливості для її використання.

ВИСНОВКИ

У дисертаційній роботі розв'язано актуальне наукове завдання удосконалення процесу опрацювання персоналізованих даних внаслідок підвищення точності класифікації та зменшення кількості ітерацій в процесі машинного навчання шляхом застосування аугментації до навчальної вибірки.

1. Узагальнено модель інформаційної технології опрацювання персоналізованих даних для аналізу стану особи, яка забезпечує його ефективне визначення та дає змогу покращити процес ідентифікації стадії захворювання.

2. Розроблено новий метод класифікації персоналізованих медичних даних шляхом введення етапу аугментації при опрацюванні медичної інформації про особу, що дало можливість збільшити обсяг і різноманітність навчальної вибірки, покращити узагальнення моделей та зменшити ризик перенавчання.

3. Удосконалено метод персоналізації медичних даних особи, який, на відміну від наявних, передбачає введення ансамблю моделей класифікації та ансамблевого голосування, що дало змогу підвищити точність прогнозування результатів оцінювання стану особи, особливо при захворюваннях головного мозку.

4. Розроблено архітектуру інформаційної системи опрацювання персоналізованих даних, яка дозволяє аналізувати різні терапевтичні захворювання. На базі цієї розробки створено систему, оснащену механізмами обробки індивідуальних даних і прогнозування стану пацієнта з урахуванням специфіки різних класів захворювань.

5. Реалізовано прикладну інформаційну систему опрацювання персоналізованих даних для аналізу стану особи. Здійснено числові експерименти класифікації різних стадій захворювань, зокрема хвороби Альцгеймера. Дослідження показали підвищення загальної точності прогнозування з 0,90 (базова модель) до 0,96 (ансамбль з soft voting). Показники precision покращились з діапазону 0,87-0,93 до 0,94-0,98, а recall – з 0,86-0,94 до 0,91-0,99. Отримані результати свідчать про підвищення ефективності розробленої моделі класифікації та її здатність точно розпізнавати різні стадії захворювань. Модель також продемонструвала ефективність при аналізі інших медичних даних.

6. Результати дисертаційного дослідження впроваджені при виконанні науково-дослідної роботи кафедри систем штучного інтелекту Національного університету «Львівська політехніка» за темою «Методи та засоби обробки, консолідації та аналізу персоналізованої медичної інформації» та у лікувальний процес під експертизою Львівської асоціації алергологів, імунологів, імунореабілітологів.

СПИСОК ПУБЛІКАЦІЙ ЗА ТЕМОЮ ДИСЕРТАЦІЙНОЇ РОБОТИ

1. Melnykova N., Paterega Iu. Imbalanced data: a comparative analysis of classification enhancements using augmented data. Intellektuelles Kapital–die Grundlage

für innovative Entwicklung: Innovative Technologie, Informatik, Sicherheitssysteme, Verkehrsentwicklung, Physik und Mathematik. Monografische Reihe «Europäische Wissenschaft». Buch 28. Teil 3 = Intellectual capital is the foundation of innovative development: Innovative technology, Computer science, Security systems, Transport development, Physics and mathematics, Agriculture. Monographic series «European Science». Book 28. Part 3 : monograph. Karlsruhe: ScientificWorld-NetAkhatAV, 2024. P. 54–72. DOI: 10.30890/2709-2313.2024-28-00-01.

2. Bokhonko A., Melnykova N., Patereha Yu. Comparative analysis of data augmentation methods for image modality. Вісник Тернопільського національного технічного університету. 2024. № 1 (113). С. 16–26. / https://visnyk.tntu.edu.ua/index.php?art=762.

3. Patereha Yu., Melnyk M. Prediction of the occurrence of stroke based on machine learning models. Комп'ютерні системи проектування. Теорія і практика. 2024. Вип. 6, № 1. С. 17–27. https://doi.org/10.23939/cds2024.01.017

4. Paterega I. Main strategies for autonomous robotic controller design. Радіоелектроніка та інформатика. 2011. Вип. 4. С. 36–41. https://openarchive.nure.ua/entities/publication/505fca9b-c6b2-445c-9c23-e4338981c56d.

5. Патерега Ю. І. Особливості використання штучних нейронних осциляторів у робототехніці. Науковий вісник НЛТУ України. 2010. Вип. 20.13. С. 322–331.

6. Тимощук П. В., Патерега Ю. І. Штучні нейронні осцилятори. Вісник Національного університету "Львівська політехніка". Серія: Комп’ютерні системи проектування. Теорія і практика. 2009. № 651. С. 40–45.

7. Nykoniuk M., Melnykova N., Patereha Yu., Sala D., Cichoń D. Classification of patients with the development of Alzheimer's disease using an ensemble of machine learning models. CEUR Workshop Proceedings. 2023. Vol. 3609 : 6th Intern. conf. on informatics and data-driven medicine IDDM 2023, Bratislava, Slovakia, 17-19 Nov. 2023. P. 198–216. (Scopus) DOI: 10.30890/2709-2313.2024-28-00-017. / https://ceur-ws.org/Vol-3609/short4.pdf

8. Paterega Yu. I. Analysis of neural network controller for mobile robot navigation / // САПР у проектуванні машин. Питання впровадження та навчання : матеріали XVIII Міжнар. укр.-пол. наук.-техн. конф. CADMD'2010, 14–16 жовт. 2010, Львів, Україна / Нац. ун-т «Львів. політехніка». – Л.: Вежа і Ко, 2010. – С. 91–92.

9. Paterega Yu. Artificial neural oscillators in robotics. Perspective Technologies and Methods in MEMS Design MEMSTECH'2010 : proc. of the 6th Intern. conf., 20-23 Apr. 2010. P. 123– 130. (Scopus).

10.Paterega Yu. Izhikelich's model of spiking neurons // Computer science and information technologies : proc. of the V Intern. sci. and techn. conf. CSIT 2010, 14–16 Oct. 2010, Lviv, Ukraine / Lviv Polytechnic Nat. Univ. – Lviv : Publ. House Vezha and Co, 2010. – P. 32–33.

11. Tymoshchuk P. V., Paterega Yu. I. Mathematical models of spiking neurons // Computer science and information technologies : proc. of the V Intern. sci. and techn. conf. CSIT 2010, 14–16 Oct. 2010, Lviv, Ukraine / Lviv Polytechnic Nat. Univ. – Lviv : Publ. House Vezha and Co, 2010. – P. 47–48.

12. Tymoshchuk P. V., Paterega Y. I. Implementation of artificial neural oscillators. 5th Intern. Conf. on Perspective Technologies and Methods in MEMS Design MEMSTECH 2009, 22-24 Apr. 2009. P. 149–154 (Scopus).

13. Paterega I. Artificial evolution mechanisms in robot navigation. 2011 11th International Conference “The Experience of Designing and Application of CAD Systems in Microelectronics” CADSM 2011, 23-25 Febr. 2011. P. 281–286 (Scopus).

АНОТАЦІЯ

Патерега Ю. І. Інформаційна технологія опрацювання персоналізованих даних для аналізу стану особи. – Кваліфікаційна наукова праця на правах рукопису. Дисертація на здобуття наукового ступеня кандидата технічних наук за спеціальністю 05.13.06 – Інформаційні технології. – Національний університет «Львівська політехніка» Міністерства освіти і науки України, Львів, 2024.

Дисертаційна робота присвячена розробленню та вдосконаленню інформаційної технології опрацювання персоналізованих даних для покращення процесів аналізу стану особи і підвищення точності класифікації таких даних для ефективнішого лікування пацієнтів.

У *першому* розділі розглянуто існуючі системи опрацювання персоналізованих даних, виділено їхні переваги та недоліки, що дало змогу визначити напрямки вдосконалення таких систем шляхом консолідації процесів збору, обробки та аналізу індивідуальних даних осіб та класифікації стану особи. Проведено порівняльний аналіз існуючих комп’ютерних технологій підтримки медичних рішень, які використовують комп’ютерні технології. Надано опис і стислий аналіз поширених методів застосування методології штучного інтелекту для роботи з медичними даними та перспектив застосування алгоритмів машинного навчання у діагностиці різних захворювань. Сформульовано та аргументовано основні проблеми наявних досліджень у сфері використання методів і засобів штучного інтелекту в медицині. Головною проблемою існуючих рішень є недостатня точність класифікації етапів захворювання, на що впливають ключові характеристики стадій хвороби, зокрема індивідуальні особливості пацієнта. До проблем також належать особливості класифікації стадій хвороби під час навчання та тестування моделей на малих наборах даних, що може призвести до неточностей класифікації. Іншою проблемою є недостатні результати тестування моделі. На основі виокремлених проблем сформульовано мету та завдання дослідження.

У *другому* розділі представлено концептуальну модель, яка формалізує процес опрацювання персоналізованих даних. Ця модель консолідує етапи збору даних від периферійних пристроїв, передачі даних, а також етапи їх обробки та аналізу. Вона забезпечує комплексний підхід до роботи з персоналізованими даними, враховуючи вимоги бібліотеки шаблонів C-CDA, яка встановлює додатковий рівень вимог до базового стандарту HL7. Також проаналізовано наявні давачі та протоколи, що забезпечують збір і передачу даних до місця їх опрацювання. Узагальнено модель інформаційної технології обробки персоналізованих даних для представлення стану пацієнта, беручи до уваги основні параметри його загального стану та визначені характеристики. Запропоновано алгоритм обробки персоналізованих даних для аналізу стану пацієнта під час діагностування хвороб головного мозку та їх

ускладнень, що дає змогу формалізувати процес підготовки даних пацієнтів, структуруючи етапи виконання попередньої підготовки даних, включаючи обробку зображень клінічних досліджень, пошук дублікатів, балансування та нормалізацію.

У *третьому* розділі запропоновано метод класифікації персоналізованих даних шляхом введення етапу їх аугментації, що дало змогу збільшити обсяг та різноманітність навчальної вибірки та збалансувати її, покращуючи узагальнення моделей і знижуючи ризик перенавчання. Проаналізовано застосування процесу аугментації для даних різних модальностей, що дало змогу виявити утворюване спотворення даних, яке призводить до неправильного передбачення класу. Удосконалено метод персоналізації медичних даних шляхом впровадження ансамблю моделей класифікації та ансамблевого голосування, що підвищило точність прогнозування результатів. Досліджено два ансамблі моделей з різними типами голосування: жорстким (hard voting) та м'яким (soft voting).

У *четвертому* розділі дисертаційного дослідження розроблено архітектуру інформаційної системи підтримки прийняття медичних рішень на основі опрацювання персоналізованих медичних даних. Представлено функціональну схему цієї інформаційної системи. Проведено порівняльний аналіз застосування наявних моделей класифікації, де найкращі результати показала модель багатошарового перцептрона (MLP). Для створення ансамблю обрано моделі Random Forest, SVM та MLP.

Проаналізовано існуючі методи побудови ансамблевого навчання, обґрунтовано переваги запропонованого методу та експериментально підтверджено його ефективність порівняно з існуючими підходами. Наведено результати впровадження розробленої інформаційної системи, яка реалізує сформульований метод для супроводу процесу збору та аналізу стану особи.

Ключові слова: персоналізовані дані особи, методи машинного навчання, Random Forest, Support Vector Machine, Multi-Layer Perceptron, Soft Voting, аугментація даних, класифікація ансамблюванням.

SUMMARY

Paterega Iu. I. Information Technology for Processing Personalized Data for Analyzing the State of a Person. On the Rights of Manuscript: Dissertation for the Degree of Candidate of Technical Sciences in the specialty 05.13.06 – Information Technologies. - Lviv Polytechnic National University, Ministry of Education and Science of Ukraine, Lviv, 2024.

The dissertation work is devoted to the development and improvement of information technology for processing personalized data to enhance the processes of analyzing a person's condition and increasing the accuracy of such data classification for more effective patient treatment.

In the first chapter, existing systems for processing personalized data are examined, and their advantages and disadvantages are highlighted, which allows us to determine the directions for improving such systems by consolidating the processes of collecting, processing, and analyzing individual data of persons and classifying the state of a person. A comparative analysis of existing computer technologies for supporting medical decisions is conducted. A description and brief analysis of standard methods of applying artificial intelligence methodology to work with medical data and prospects for using machine

learning algorithms in diagnostics are provided. The main problems of existing research in using artificial intelligence methods and tools in medicine are formulated and argued. The main problem of existing solutions is insufficient accuracy in classifying disease stages, which is influenced by key characteristics of disease stages, particularly individual patient features. The problems also include peculiarities of disease stage classification during training and testing models on small datasets, which can lead to classification inaccuracies. Another problem is insufficient model testing results. Based on the identified problems, the purpose and objectives of the study are formulated.

The second chapter presents a conceptual model that formalizes the process of processing personalized data. This model consolidates the stages of data collection from peripheral devices and other sources, data transmission, as well as the stages of their processing and analysis. It provides a comprehensive approach to working with personalized data, taking into account the requirements of the C-CDA template library, which establishes an additional level of requirements for the basic HL7 standard. Available sensors and protocols that ensure the collection and transmission of clinical data to the place of their processing are also analyzed. The model of information technology for processing personalized data to represent the patient's condition is generalized, taking into account the main parameters of their general condition and defined characteristics. An algorithm for processing personalized data for analyzing a patient's condition during the diagnosis of brain diseases and their complications is proposed, which allows formalizing the process of preparing patient data, structuring the stages of preliminary data preparation, including processing images of clinical studies, searching for duplicates, balancing and normalization.

In the third chapter, a method for classifying personalized data is proposed by introducing the stage of their augmentation, which increases the volume and diversity of the training sample and balances it, improving the generalization of models and reducing the risk of overfitting. The application of the augmentation process for data of different modalities is analyzed, which revealed that excessively distorted augmented data may lead to incorrect class predictions. The method of personalizing medical data has been improved by introducing an ensemble of classification models and ensemble voting, which increased the accuracy of result prediction. Two model ensembles with different voting types are investigated: hard voting and soft voting.

In the fourth chapter of the dissertation research, the architecture of an information system for supporting medical decision-making based on processing personalized medical data is developed. The functional scheme of this information system is presented. A comparative analysis of the application of existing classification models is conducted, where the multilayer perceptron (MLP) model showed the best results. Random Forest, SVM, and MLP models were selected to create the ensemble.

Existing methods for building ensemble learning are analyzed, the advantages of the proposed method are substantiated, and its effectiveness compared to existing approaches is experimentally confirmed. The results of the developed information system that implements the formulated method for supporting the collection and analysis of a person's condition are presented.

Keywords: personalized data of a person, machine learning methods, Random Forest, Support Vector Machine, Multilayer Perceptron, Soft Voting, data augmentation, ensemble classification.

Тираж здійснено у друкарні
Львівського національного медичного університету імені Данила Галицького
79010, м. Львів, вул. Пекарська, 69

Підписано до друку 29.08.2024 р.
Формат 60×84/16. Папір офсетний. Умовн. друк. арк. 0,9.
Тираж 120 прим.