МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ «ЛЬВІВСЬКА ПОЛІТЕХНІКА»
МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ «ЛЬВІВСЬКА ПОЛІТЕХНІКА»

Кваліфікаційна наукова праця
на правах рукопису

Патерега Юрій Ігорович

УДК 004.942

ДИСЕРТАЦІЯ
ІНФОРМАЦІЙНА ТЕХНОЛОГІЯ ОПРАЦЮВАННЯ ПЕРСОНАЛІЗОВАНИХ ДАНИХ ДЛЯ АНАЛІЗУ СТАНУ ОСОБИ

05.13.06 – Інформаційні технології

Подається на здобуття наукового ступеня кандидата технічних наук
Дисертація містить результати власних досліджень. Використання ідей, результатів і текстів інших авторів мають посилання на відповідне джерело
_____ Патерега Ю.І.

Науковий керівник – Мельник Михайло Романович, кандидат технічних наук, доцент

Львів – 2024

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 наявні 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, що 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, що 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.

ABSTRACT

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.

СПИСОК ПУБЛІКАЦІЙ ЗДОБУВАЧА ЗА ТЕМОЮ ДИСЕРТАЦІЇ

Стаття у науковому виданні іншої держави:

1. 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.

Статті у наукових фахових виданнях України:

1. 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.

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

3. 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.

4. 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;

5. 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.

Монографічне видання:

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: 10.30890/2709-2313.2024-28-00-017.

Тези доповіді конференцій:

1. Analysis of neural network controller for mobile robot navigation / Yu. I. Paterega // 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.

2. 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).

3. Izhikelich's model of spiking neurons / Yu. Paterega // 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.

4. Mathematical models of spiking neurons / P. V. Tymoshchuk, Yu. I. Paterega // 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.

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

6. 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).

ЗМІСТ

ПЕРЕЛІК ПРИЙНЯТИХ СКОРОЧЕНЬ

ПД – персоналізовані дані;

ЕМЗ – електронний медичний запис (EMR – Electronic Medical Record);

ЕЗЗ – електронний запис здоров'я (EHR - Electronic Health Record);

CDMS – Clinical Data Management Systems / Системи управління клінічними даними;

REDCap – Research Electronic Data Capture/ Дослідження електронного збору даних;

ШІ – штучний інтелект;

General Data Protection Regulation (GDPR) – Загальний регламент про захист даних;

Health Level Seven International (HL7) — це набір міжнародних стандартів для обміну медичних даних;

S – множина давачів для збору клінічних даних;

C – множина клінічних вхідних даних системи, що характеризують стан пацієнта, що 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 результуючих показників стану особи, що залежить від множини гіперпараметрів класифікатора;

D – множина правил, які визначаються з урахуванням персональних даних особи та результуючих показників стану особи;

N – кількість зображень у кожному класі;

Q – визначає кількість кроків навчання;

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;

Accuracy (Точність) – відношення кількості правильних прогнозів до загальної кількості прогнозів;

Precision (Точність) – відношення кількості правильних прогнозів певного класу до загальної кількості прогнозів;

Recall (Чутливість) – відношення кількості правильних прогнозів певного класу до загальної кількості прикладів цього класу у даних;

F1-Score (F1-міра) – гармонічне середнє між точністю та чутливістю, що враховує як точність, так і чутливість;

Confusion matrix (Матриця помилок) – таблиця, що відображає кількість правильних та неправильних прогнозів класифікатора для кожного класу.

ВСТУП

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 для опрацювання та використання даних у процесах підготовки і прийняття персоналізованих рішень. Споживачі зазвичай очікують індивідуалізованих підходів у вирішенні актуальних завдань у різних сферах життя, таких як електронна комерція, надання медичних послуг, освіта та ін. Розроблення систем, здатних адаптуватися до індивідуальних потреб користувачів, стає одним із ключовим факторів підвищення конкурентоспроможності [1].

Значущість високотехнологічних процесів аналітики даних важко переоцінити. Системи опрацювання персоналізованих даних дають можливість організаціям аналізувати та використовувати їх для вдосконалення продуктів і послуг, задоволення потреб клієнтів та підвищення якості надання послуг. Розвиток методів та засобів штучного інтелекту, процесів з використанням алгоритмів машинного навчання сприяє створенню більш ефективних і точних систем опрацювання персоналізованих даних, здатних навчатися на основі індивідуальних користувацьких шаблонів [2,3].

Таким чином, розроблення систем опрацювання персоналізованих даних є актуальним та важливим завданням для багатьох галузей, включаючи бізнес, медицину, освіту, інформаційні технології та ін.

Важливою особливістю роботи з персоналізованими даними є необхідність їх анонімізації - процесу, спрямованого на збереження приватності або конфіденційності інформації шляхом вилучення або шифрування ідентифікаторів, що пов’язують конкретну особу зі збереженими її персональними даними [4]. Опрацювання анонімізованих даних відповідно до вимог Загального регламенту про захист даних (GDPR) вимагає дотримання

принципу незворотності анонімізації, забезпечення належного захисту даних, а також регулярної перевірки ефективності застосованих методів анонімізації.

Умови опрацювання анонімізованих даних відповідно до вимог GDPR вказують на те, ключова характеристика таких даних полягає у тому, що вони не можуть бути використані для ідентифікації конкретної особи. Це означає, що дані повинні бути оброблені таким чином, щоб їхня ідентифікація була неможливою або практично нездійсненною, навіть якщо використовувати додаткові дані або технічні засоби. Процес анонімізації повинен бути надійним і незворотним, що забезпечує неможливість відновлення початкової інформації, яка дозволяє ідентифікувати суб'єкта даних. Методи анонімізації повинні відповідати найкращим практикам і враховувати ризики, пов'язані з можливістю реідентифікації.

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. [86, 106,108]. 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. [45,115]. 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 [87, 92]. .

Вплив підходів, базованих на використанні алгоритмів машинного навчання на догляд за пацієнтами було докладно описано у працях Девіда Бен-Ізраєлята, Семюела П. Лейтона. Автори висловлюють припущення, що застосування штучного інтелекту в медицині призведе до фундаментальних змін у медичній практиці. Зокрема, вказується на можливість використання спеціалізованих інструментів, які допомагатимуть медичним командам надавати пацієнтам більш персоналізовану допомогу. Використання методології штучного інтелекту може загалом покращити точність та швидкість діагностики,

знизити ризик появи помилок та підвищити результативність процесів лікування [5].

У медичній галузі для ефективного вирішення задач класифікації даних, опрацювання зображень, прогнозування лікарських рішень застосовуються різні методи штучного інтелекту, алгоритми машинного навчання, включаючи нейронні мережі, дерева рішень, метод опорних векторів та інші.

Використання методів машинного навчання для опрацювання медичних даних пацієнтів є актуальним науковим завданням, спрямованим на підвищення точності діагностики та оптимізацію лікування, що в результаті має забезпечити покращення якості медичних постуг, особливо при захворюваннях головного мозку [3,5].

Зв'язок роботи з науковими програмами, планами і 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, 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 робіт: «Effectiveness Of Medicine E-Learning Distance Courses» / «Ефективність дистанційних курсів медицини» реєстраційний номер на порталі Європейської Комісії 2022-1-IT02-KA220-HED-000087665.

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

Для досягнення поставленої мети було сформульовано та вирішено наступні завдання:

1. Провести аналіз існуючих інформаційних технологій підтримки прийняття лікарських рішень, визначити переваги і недоліки у зв'язку зі сферою

їхнього застосування у відповідності 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 осіб з вадами головного мозку та їхніми ускладненнями.

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 результатів стану особи при захворюваннях головного мозку.

5. Розробити архітектуру інформаційної системи опрацювання персоналізованих даних аналізу стану особи.

6. Розробити і 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 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 за рахунок 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 існуючих, 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 14.04.2024 р.) та у лікувальний процес під експертизою Львівської асоціації алергологів, імунологів, імунореабілітологів (акт впровадження від 19.01.2024 р.).

– отримані результати проведених 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.)

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 основних розділів, в яких містяться 57 рисунків, 17 таблиць, списку використаних джерел із 140 найменувань та 3 додатків.

РОЗДІЛ 1.
АНАЛІТИЧНИЙ ОГЛЯД ТЕХНОЛОГІЙ ОПРАЦЮВАННЯ ПЕРСОНАЛІЗОВАНИХ ДАНИХ

У даному розділі проведено огляд існуючих систем опрацювання персоналізованих даних, здійснено порівняльний аналіз рішень, які застосовуються для вирішення медичних проблем за допомогою комп'ютерних технологій. Також подано опис та узагальнений аналіз використання поширених методів штучного інтелекту при опрацюванні медичних даних, висвітлено перспективи використання машинного навчання при медичній діагностиці. 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.

Матеріали розділу опубліковані у роботах автора [131-133, 136, 137, 139].

1.1. Аналіз систем опрацювання персоналізованих даних

Кожен рік спостерігається зростання обсягів генерованих та збережених медичних даних. У зв'язку з цим стає критично важливим створення систем, які можуть ефективно обробляти ці дані та використовувати їх для прийняття рішень при діагностуванні захворювань.

Очікування споживачів у сфері програмних продуктів, викликані необхідністю персоналізованих досвідів у всіх сферах життя, включаючи електронну комерцію, медичні послуги, освіту та інші, змушують розробників створювати системи, які можуть адаптуватися до індивідуальних потреб користувачів. Це стає вирішальним фактором у конкурентному середовищі [1,4,6].

Персоналізація є підходом або процесом, що спрямований на створення продуктів або послуг, які враховують індивідуальні потреби, вподобання, характеристики або контекст конкретної особи [4].

У медицині персоналізація включає створення індивідуальних планів лікування для кожного пацієнта на основі історії його захворювання, умов життя, праці та інших факторів. У маркетингу персоналізація представляється у вигляді персоналізованих рекламних пропозицій, пропозицій товарів або послуг, які відповідають індивідуальним інтересам та потребам кожного клієнта. Ідея персоналізації полягає у забезпеченні індивідуального підходу та найвищої відповідності між продуктом або послугою і конкретними потребами або очікуваннями користувача [2, 3].

Особливості персоналізації даних відповідно до Загального регламенту про захист даних (GDPR) стосуються мінімізації, що дозволяє обробляти лише ті дані, які необхідні для досягнення конкретних цілей. Це передбачає, що збір і використання інформації мають бути обмежені лише тими параметрами, які є безпосередньо важливими для виконання відповідних завдань.

Відповідно до GDPR, особа, що надає дані, повинна бути поінформована про те, як її персональні дані будуть використовуватися, і надавати на це свою згоду. Це підкреслює необхідність прозорості у процесі персоналізації, щоб користувачі точно розуміли, з якою метою та як саме обробляються їхні дані.

Окрім того, GDPR надає власникам даних права на доступ до своїх даних, їхнє виправлення, видалення та обмеження обробки. Це означає, що процес персоналізації повинен бути гнучким і дозволяти користувачам здійснювати контроль над своїми персональними даними відповідно до їхніх побажань.

Також GDPR встановлює суворі вимоги до захисту даних від несанкціонованого доступу, що є критично важливим у контексті персоналізації. Організації, які обробляють персоналізовані дані, зобов'язані забезпечити їхню безпеку, щоб уникнути витоку або неправомірної обробки.

Таким чином, персоналізація даних у відповідності з GDPR передбачає дотримання принципів мінімізації, прозорості, захисту даних, а також забезпечення прав користувачів на доступ і контроль над своїми персональними даними.

Персоналізовані системи мають ряд переваг:

- Покращена користувацька задоволеність: персоналізовані системи забезпечують індивідуально налаштований досвід для кожного користувача, що робить їх привабливішими та зручнішими при використанні.
- Підвищення ефективності: завдяки врахуванню індивідуальних потреб і вподобань респондентів персоналізовані системи здатні забезпечувати точніші рекомендації, що сприятиме збільшенню конверсії та індивідуального підходу.
- Покращення рішень: застосування персоналізованих методів аналізу даних дає змогу виявляти нові зв'язки та закономірності, що може привести до виявлення нових можливостей та покращення прийняття рішень.
- Автоматизація: персоналізовані системи можуть автоматизувати та оптимізувати багато аспектів взаємодії з респондентом, що зменшить витрати часу і зусиль.
- Покращення лояльності: індивідуально налаштований досвід сприяє покращенню взаємодії із замовником або послугою, що збільшить лояльності користувачів та їхнього повторного звернення.
- Зниження ризиків: аналіз індивідуальних потреб та вподобань респондентів може допомогти у виявленні потенційних проблем або ризиків та запобігти їх виникненню.
- Глибше розуміння респондента: за допомогою персоналізованих систем можна отримати глибше розуміння аудиторії, її потреб та уподобань, що оправдовує її очікування.

Значимість аналізу даних у сучасному світі перетворює їх у надзвичайно цінний ресурс для проведення досліджень персональних даних респондентів та прогнозування їхніх станів. Системи обробки персоналізованих даних допомагають аналізувати та використовувати ці дані для удосконалення продуктів або послуг, підвищення рівня задоволеності клієнтів і підтвердження ефективності персоналізованих рішень.

Прогрес у розвитку штучного інтелекту та машинного навчання сприяє створенню ефективніших і точніших систем опрацювання персоналізованих даних, які можуть навчатися на основі індивідуальних користувацьких шаблонів.

Отже, відсутність ефективних рішень щодо покращення точності класифікації стану хвороби особи спонукає до розробки систем опрацювання персоналізованих даних. Це, в свою чергу, визначає актуальність їхнього застосування, зокрема, у наданні медичних послуг та в інших галузях, де важливим завданням є прогнозування персоналізованих рішень [5].

Огляд існуючих систем опрацювання персоналізованої інформації 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.

На сьогоднішній день існує широкий спектр систем обробки персоналізованих даних про пацієнтів, які використовуються в медичній сфері [7, 8]. 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.

Ці системи постійно розвиваються і вдосконалюються з метою забезпечення ефективнішого індивідуалізованого медичного супроводу.

1.1.1. Системи електронних медичних записів

Існує багато компаній та продуктів, які надають системи електронних медичних записів (ЕМЗ), що використовуються у медичних установах по всьому світу. Деякі з найпопулярніших і відомих ЕМЗ включають [9, 10]:

1 Allscripts Healthcare Solutions є провідним виробником систем ЕМЗ, які використовуються для автоматизації медичних процесів та забезпечення доступу до даних пацієнтів.

2 MEDITECH є популярним постачальником ЕМЗ для лікарень та інших медичних установ, який пропонує широкий спектр функцій та можливостей.

3 Athenahealth надає хмарні рішення для управління медичними записами, практиками та фінансами.

Функціонал ЕМЗ відповідає потребам різних медичних установ. Проведений аналіз дав змогу встановити переваги та недоліки систем електронних медичних записів (ЕМЗ) [10].

Переваги включають усунення необхідності у фізичному зберіганні великої кількості паперових документів, що економить місце та дає змогу швидше знаходити й оновлювати інформацію. Також зручний доступ до медичних даних сприяє легшому обміну інформацією та супроводу роботи з пацієнтами, запобігаючи медичним помилкам завдяки кращому доступу до інформації про пацієнта, медикаментів та алергій. Автоматизація багатьох процесів, таких як ведення записів пацієнтів, призначення лікарями лікування, підвищує ефективність, зменшуючи час, необхідний для виконання рутинних завдань.

Недоліки включають високу вартість впровадження, яка охоплює витрати на програмне забезпечення, обладнання та навчання персоналу. Існують також проблеми з безпекою, зокрема ризик порушення конфіденційності медичних даних через кібератаки або несанкціонований доступ, що вимагає додаткових заходів безпеки. Технічні труднощі вимагають часу для навчання та адаптації персоналу, що може тимчасово збільшити навантаження на робочий процес.

Крім того, можливість втрати доступу через технічні або мережеві проблеми обмежує обробку медичних даних та ускладнює надання медичної допомоги.

Враховуючи особливості систем електронних медичних записів ЕМЗ, вони залишаються важливим інструментом для покращення якості та доступності медичної допомоги в сучасних закладах охорони здоров'я [10].

1.1.2. Системи управління клінічними даними

Системи управління клінічними даними (Clinical Data Management Systems CDMS) використовуються у медичних установах та дослідницьких лабораторіях по всьому світу для збору, зберігання, управління та аналізу клінічних даних, забезпечуючи точність та доступність інформації для покращення результатів лікування та наукових досліджень. Зокрема, нижче наведено приклади існуючих систем у контексті інтегрованих платформ управління клінічними даними [8, 9, 11]:

1 Oracle Clinical - інтегрована система збору, обробки та аналізу даних у клінічних дослідженнях.

2 Medidata Rave - хмарна платформа електронного збору даних та моніторингу клінічних випробувань.

3 OpenClinica - надає можливості збору, управління та аналізу даних у клінічних дослідженнях.

4 REDCap (Research Electronic Data Capture) - безкоштовна платформа електронного збору та управління клінічними даними, яка часто використовується у дослідницьких проєктах.

Системи CDMS, які доступні на ринку, мають свої особливості та функціональні можливості, що відповідають потребам різних медичних установ і дослідницьких груп.

CDMS забезпечують ефективний збір, зберігання та керування клінічними даними, що гарантує їхню цілісність, конфіденційність і доступність. Зберігання даних в електронному форматі значно спрощує доступ до інформації. Багато процесів, таких як введення даних, моніторинг, перегляд та аналіз, можуть бути

автоматизовані за допомогою CDMS, що зменшує час та зусилля при керуванні даними. Крім того, ці системи сприяють підтриманню високої якості даних завдяки вбудованим правилам перевірки та автоматичним процесам контролю якості [9,11].

Однак, реалізація та підтримка CDMS може бути вартісною, особливо для невеликих організацій або дослідницьких груп. Крім того, їх використання може бути складним, особливо для користувачів без технічного досвіду, що призведе до неефективного використання або помилок при зборі даних. Також потрібно витратити час та зусилля на навчання персоналу, що вплине на продуктивність на початкових етапах. Залежність від технологічної інфраструктури впливає на надійність та доступність даних. Системи управління клінічними даними залишаються важливим інструментом для управління персоналізованими даними в медичних установах та дослідницьких лабораторіях.

1.1.3. Системи аналізу даних

На ринку програмних рішень існують наступні відомі розробники систем аналізу даних про здоров'я, які використовуються у медичних установах, дослідницьких лабораторіях і компаніях, що займаються охороною здоров'я [12, 13]:

1 Epic Systems Corporation - це один з найпопулярніших у світі розробник систем електронних медичних записів і систем аналізу даних про здоров'я. Він пропонує широкий спектр рішень, включаючи системи управління пацієнтами, аналітику даних та інструменти управління клінічними процесами.

2 Cerner Corporation - ще один провідний постачальник систем електронних медичних записів та аналітичних рішень для охорони здоров'я. Пропонує рішення для управління клінічними даними, аналізу фінансової продуктивності та підвищення якості надання медичної допомоги.

3 IBM Watson Health пропонує інноваційні рішення для аналізу медичних даних, використовуючи штучний інтелект та машинне навчання, надає

інструменти для діагностики, прогнозування захворювань та оптимізації клінічних процесів.

З точки зору переваг аналізу даних, такі системи значно підвищують ефективність медичних процесів, даючи змогу лікарям швидше і точніше аналізувати великі обсяги медичної інформації та приймати обґрунтовані рішення щодо діагностики й лікування. Завдяки аналізу даних можна покращити якість медичної допомоги, оскільки виявлення тенденцій та патернів у здоров'ї пацієнтів запобігає виникненню захворювання і підвищує рівень надання медичної допомоги. Аналітичні системи також сприяють ефективнішому використанню ресурсів організацій охорони здоров'я та зменшенню витрат на медичні послуги.

Однак існують певні недоліки. Так, збір та аналіз медичних даних створюють ризики для конфіденційності та безпеки інформації про пацієнтів. Технічні проблеми, пов’язані з несправністю обладнання та проблемами з мережею, можуть спричиняти перебої у роботі систем аналізу даних. Крім того, використання даних пацієнтів для аналізу та дослідження може викликати етичні питання.

1.1.4. Медичні портали та додатки для смартфонів

До наступної групи систем збору та опрацювання персоналізованих даних належать медичні портали і додатки для смартфонів. Вони надають користувачам доступ до медичної інформації, послуг та інструментів для збереження і відстеження їхнього здоров'я [13, 14]. До таких належать:

1 Fitbit: додаток призначений для відстеження фізичної активності і сну, має функції моніторингу пульсу та інших параметрів здоров'я.

2 MyChart: додаток дає змогу пацієнтам переглядати свої медичні записи, результати аналізів, рецепти та здійснювати візити до лікаря в онлайн-режимі.

3 Ada: додаток для смартфонів, який дає змогу користувачам на основі симптомів отримати рекомендації щодо подальших кроків, включаючи консультацію з лікарем.

Переваги таких додатків та порталів полягають у зручному доступі до медичної інформації та доступності, можливості моніторингу здоров'я і ефективному керуванні медичними записами. Проведений аналіз додатків та платформ дав змогу встановити основні їх недоліки, якими є недостатні достовірність інформації і питання захисту конфіденційності даних та обмежені можливості деяких додатків.

1.1.5. Системи геномної медицини

Для розуміння того, які дані необхідно враховувати при розробленні інформаційної технології опрацювання персоналізованих даних для аналізу стану особи проведено дослідження систем геномної медицини. Це комплексні програмні та апаратні засоби, які використовуються для аналізу та інтерпретації геномної інформації з метою діагностики, лікування і профілактики захворювань. До відомих систем геномної медицини належать [15, 16]:

1 Next Generation Sequencing (NGS) Platforms - системи, які дають змогу проводити швидке та високоякісне секвенування геномів та екзотів, а також отримувати великий обсяг генетичних даних для подальшого аналізу.

2 Bioinformatics Software - програмні засоби для обробки та аналізу генетичних даних, які включають у себе алгоритми ідентифікації варіантів геному, асоційованих з різними захворюваннями, інструменти для визначення функціонального значення цих варіантів.

3 Clinical Decision Support Systems (CDSS) - використовують геномні дані для надання індивідуальних рекомендацій щодо діагностики, лікування та профілактики захворювань на основі клінічних протоколів та наукових даних, але основною метою CDSS є покращення якості медичної допомоги, зменшення помилок, підвищення ефективності лікування та підтримки прийняття обґрунтованих медичних рішень.

4 Personalized Medicine Platforms: платформи, які використовують геномні дані разом із клінічною інформацією про пацієнтів для розробки індивідуальних стратегій лікування та профілактики захворювань.

Переваги систем геномної медицини включають покращену можливість діагностики та лікування захворювань, індивідуалізовані підходи до пацієнтів, збільшення ефективності медичних процедур і зменшення побічних ефектів лікування. Однак, до їх недоліків можна віднести високі витрати на розробку і впровадження таких систем, складність інтерпретації генетичних даних та етичні питання, пов'язані зі зберіганням і використанням особистої геномної інформації.

1.2. Аналіз досліджень опрацювання персоналізованих даних

1.2.1. Наявні рішення щодо класифікації даних

Використання методології штучного інтелекту в галузі медицини набуло активного розвитку з настанням ери великих обсягів медичних даних. В цілому, застосування моделей глибокого навчання у медицині може є важливим кроком у покращенні якості діагностики та лікування пацієнтів. Проте, при зборі персоналізованих даних необхідно дотримуватися обережного підходу та враховувати етичні та правові аспекти [7, 8].

У зв’язку із правовими нормами щодо захисту персональних даних необхідно забезпечити конфіденційність, яка полягає у вилученні або шифруванні ідентифікаторів, що пов’язують особу зі збереженими даними. Особиста ідентифікаційна інформація, така як імена, номери полісів соціального страхування та адреси, має бути зашифрована за допомогою процесу анонімізації даних, який забезпечує збереження інформації і гарантує анонімність джерела.

Анонімізація або деідентифікація є актуальною, коли дані щодо стану здоров’я використовуються для вторинних цілей. Під вторинними розуміють цілі, непов'язані з наданням допомоги пацієнтам, такі як дослідження, охорона здоров’я, сертифікація чи акредитація, а також маркетинг. Більшість законів про конфіденційність в усьому світі ґрунтуються на згоді. Якщо пацієнти надають дозвіл, то їхні дані можуть використовуватися для цілей, які вони санкціонують. Проте, у випадках анонімізованих даних згода не потрібна. Загалом, анонімізовані дані більше не вважаються особистою інформацією про здоров’я,

і вони не підпадають під законодавство про конфіденційність [3-5].

У медичній галузі для ефективного вирішення завдань класифікації даних, опрацювання зображень, прогнозування лікарських рішень застосовуються різні методи штучного інтелекту, алгоритми машинного навчання, включаючи нейронні мережі, дерева рішень, метод опорних векторів та інші [5,18].

Вплив підходів, базованих на використанні алгоритмів машинного навчання на догляд за пацієнтами докладно описано у працях Девіда Бен-Ізраєля, Семюела П. Лейтона. [5]. Автори висловлюють припущення, що застосування штучного інтелекту в медицині приведе до фундаментальних змін у медичній практиці. Зокрема, вони вказують на можливість використання спеціалізованих інструментів, які допомагатимуть медичним командам надавати більш персоналізовану допомогу пацієнтам. Використання методології штучного інтелекту може загалом покращити точність та швидкість діагностики, знизити ризик появи помилок та підвищити результативність лікування пацієнтів [5,18].

У сучасному світі понад 16000 рецензованих наукових праць, які щорічно публікуються в галузі штучного інтелекту, свідчать про активний розвиток цієї галузі, а також про розширення використання інтелектуальних технологій у різних сферах життя. Застосування штучного інтелекту вже трансформує наше повсякденне існування через такі інноваційні рішення, як розпізнавання зображень, розпізнавання мовлення, переклад природньої мови, робототехніка та автономні автомобілі.

У 2012 році модель глибокого навчання AlexNet виявилася проривною в галузі великомасштабного візуального розпізнавання на базі ImageNet. Це стало першим випадком, коли глибоке навчання проявило суттєву перевагу над традиційними методами машинного навчання у задачах розпізнавання зображень. Основною особливістю моделі є використання функцій активації ReLU (Rectified Linear Unit), що допомагають уникнути проблеми зниклих градієнтів, та методу dropout для регуляризації, що дає змогу уникнути перенавчання. Це відкрило перспективи для подальших досліджень у галузі глибокого навчання при обробці зображень і розвитку нових моделей, таких як

VGG, ResNet та Inception [13].

Дослідження Семюела П. Лейтона зі співавт. "Розробка та валідація багатопараметричних моделей прогнозування", відіграє значну роль у розвитку медичної діагностики, оскільки вони презентували алгоритм прогнозування стану здоров'я пацієнта, готовий для клінічної перевірки. Для обробки значної кількості клінічних даних автори використовували методи машинного навчання та статистичного аналізу, розглядали різні підходи до статистичного моделювання, включаючи логістичну регресію, дерева рішень і нейронні мережі. Вони акцентували увагу на важливості перевірки розроблених моделей на зовнішніх наборах даних для забезпечення їхньої достовірності та універсальності [20]. Дана робота продовжує підхід, запропонований Кутсулерісом та його колегами, за допомогою аналізу базової інформації про пацієнта з використанням методів машинного навчання. П. Лейтон та його команда застосували цей підхід для прогнозування не лише одного (функціонального) результату, але й портфоліо клінічних, соціальних та професійних кінцевих точок. Основний висновок цього дослідження полягає в тому, що точність 70% можна досягти, використовуючи підхід перехресної перевірки, запропонований у 2015 році Стейєрбергом і Харреллом [20, 21]. В результаті їхньої роботи розроблено серію моделей прогнозування, які були успішно протестовані та валідовані на різних клінічних даних. Ці моделі можуть стати корисним інструментом для лікарів при прогнозуванні ризику захворювання та розроблення індивідуального плану лікування кожного пацієнта.

Підтвердженням популярності моделей на основі ResNet є публікація в Pubmed (2021) «Прогнозування мікросудинної інвазії при гепатоцелюлярній карциномі: модель глибокого навчання, перевірена в лікарнях». У роботі показано розробку моделей глибокого навчання на основі ResNet, що є одним із підтипів архітектури згорткової нейронної мережі (CNN), для прогнозування наявності мікросудинної інвазії (MCI) у пацієнтів з гепатоцелюлярною карциномою (ГЦК [22].

Наведені результати продемонстрували, що модель глибокого навчання на основі ResNet була високоефективною у прогнозуванні наявності мікросудинної інвазії з контрольною точністю розпізнавання 87%. Модель також перевірена на зовнішньому наборі даних з лікарні, де вона продемонструвала таку ж високу точність у прогнозуванні MCI. Це дослідження підкреслює потенціал моделей глибокого навчання для розробки точних і надійних прогностичних моделей медичної діагностики, а також важливість їх перевірки щодо забезпечення надійності [22].

1.2.2. Наявні рішення щодо аугментації даних

У сучасному світі обсяги доступних даних постійно збільшуються, проте у багатьох випадках якість і кількість цих даних можуть бути недостатніми для ефективного навчання різноманітних моделей машинного навчання та штучного інтелекту. Отже, виникає потреба у розробці ефективних методів аугментації даних з метою підвищення продуктивності та точності цих моделей [30, 31].

Аугментація даних визначається як процес генерації нових даних на основі наявних, що допомагає розширити спектр представлення різних аспектів та характеристик даних. Важливість аугментації даних у різних галузях, таких як комп'ютерне діагностування зору, розпізнавання мови та аналітика даних, ймовірно, матиме критичне значення для розробки точних і ефективних моделей.

Проблеми, пов'язані з даними, 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 зумовлювати наступні проблеми [31, 32]:

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.

У даному контексті стаття [5] є корисною для поточної роботи, оскільки її основний зміст стосується аналізу методів аугментації даних, зокрема, на прикладі дуже обмежених їх наборів, що є важливим аспектом для різних модальностей. У цьому дослідженні розглядається використання аугментації даних з метою збільшення обсягу навчальних даних для глибоких штучних нейронних мереж, зокрема для згорткових нейронних мереж (CNN). Автори проводять оцінку різних поширених стратегій аугментації даних, щоб надати дослідникам інструменти для обґрунтованого вибору найоптимальніших методів для їх власних наборів даних.

У результатах, що представлені у статті [33], відображено дослідження методів аугментації даних для охорони здоров'я, де збір великих наборів даних може бути ускладненим з етичних міркувань чи проблем з дотриманням конфіденційності. Автори розглядають використання генеративних суперницьких мереж (GAN) для аугментації даних, що може сприяти підвищенню точності класифікації зображень у галузі охорони здоров'я. Основна мета статті полягає в розробленні нового методу аугментації даних, спрямованого на збільшення розміру навчального набору та підвищення точності за допомогою глибокого навчання. Автори порівнюють ефективність класифікаторів зображень, застосовуючи стандартні методи аугментації та GAN.

Отже, ці результати пропонується використати при розробленні методу удосконалення процесу класифікації персоналізованих даних, оскільки вони досліджують використання GAN для аугментації даних у галузі охорони здоров'я та демонструють покращення точності класифікації зображень. Результати дослідження підтверджують перевагу GAN над стандартними методами аугментації, що є актуальним у різних модальностях та специфічних галузях застосування, де збір великих наборів даних є проблематичним.

Огляд методів аугментації даних також представлено у дослідженнях [31,

32]. Як вказують автори, аугментація застосовується для підвищення ефективності функціонування глибоких згорткових нейронних мереж у завданнях комп'ютерного зору, зокрема, в умовах обмеженого обсягу даних. Головна мета дослідження полягає у систематизації існуючих методів аугментації даних, що охоплюють геометричні трансформації, зміни в колірному просторі, використання фільтрів, змішування зображень, випадкове видалення, аугментацію у просторі ознак, адверсарне тренування, генеративно-суперечливі мережі, перенесення стилів та мета-навчання. Особлива увага приділяється методам аугментації, що базуються на генеративно-суперечливих мережах (GAN). Результати даного дослідження запропоновано використати при розробленні формалізованої моделі інформаційної технології. Окрім того, ознайомлення з різними підходами до аугментації та їх можливим впливом на покращення результатів моделей глибокого навчання, а також можливості розширення обмежених наборів даних для використання потенціалу великих обсягів даних дають змогу покращити процес попередньої підготовки анонімізованих даних [30, 33]. Обговорення інших аспектів аугментації даних, таких як аугментація під час тестування, вплив роздільної здатності, розмір фінального набору даних та послідовне навчання, також будуть використані при розробленні формалізованої моделі інформаційної технології.

1.3. Аналіз особливостей опрацювання персоналізованих даних

Проведено аналіз особливостей опрацювання персоналізованих даних, який дав змогу встановити основні типи та методи, які застосовуються для їх аналізу [23-28]. Результати наведено у таблиці 1.1.

Таблиця 1.1

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.

Аналіз існуючих рішень та ефективність класифікації різних типів персоналізованих даних, є важливим у подальшому розвитку та вдосконаленні методів діагностики у медицині [41, 42, 44].

Таблиця 1.2

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 за рахунок 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.

Методи і 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 та конфіденційністю даних пацієнтів. У цілому, використання новітніх підходів до опрацювання даних у медицині є значним кроком до покращення якості діагностики та лікування пацієнтів [35-37, 46-48, 50].

Дослідження існуючих рішень процесу аугментації даних (табл.1.2) 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 вирішити вищезазначені проблеми та досягти кращих результатів у побудові і застосуванні моделей машинного навчання та штучного інтелекту [31, 32, 37, 38, 40, 46, 51].

Отже, методи аугментації даних сприяють збільшенню обсягу та різноманітності навчальних даних, забезпечуючи краще узагальнення моделей і зменшуючи ризик перенавчання. Проте, аугментація даних зазвичай є залежною від їх модальності, тому важливо дослідити та зрозуміти, які методи аугментації є ефективними для різних типів даних, таких як табличні дані, зображення та інше.

1.4. Постановка проблеми та формулювання задач дослідження

Існуючі системи опрацювання персоналізованих даних реалізують синхронізацію декількох підходів до обробки та аналізу. Проте, існують особливості, які передбачають додатковий збір даних з інформаційних джерел, виключаючи вплив людського фактору, що забезпечує об'єктивність та незалежність передачі даних до джерела.

Застосування моделей машинного навчання має значний потенціал у сфері медичного догляду за пацієнтами. Вони забезпечують високу точність у прогнозуванні рішень, що може бути корисним для діагностики та лікування у різних медичних галузях.

Розроблення і перевірка багатофакторних моделей прогнозування виходу на ремісію, одужання та кращу якість життя у пацієнтів є важливим напрямком досліджень. Такі моделі можуть прогнозувати результати лікування, враховуючи різноманітні клінічні, психологічні та соціальні фактори. Використання таких моделей у різних медичних установах зможе підвищити їхню надійність у клінічній практиці.

1.5. Висновки до розділу 1

Проаналізовано існуючі системи обробки персоналізованих даних, які дали змогу встановити основні їх переваги та недоліки, які пропонується врахувати при удосконаленні таких систем шляхом синхронізації рішень щодо збору, обробки та аналізу індивідуальних даних особи.

Проведено порівняльний аналіз існуючих рішень вирішення медичних проблем за допомогою комп'ютерних технологій. Результати аналізу літературних джерел свідчать, що штучний інтелект успішно використовується у таких галузях медицини, як онкологія, кардіологія, неврологія та інші.

Подано опис та стислий аналіз поширених методів роботи штучного інтелекту з медичними даними, описано перспективи застосування машинного навчання у діагностиці. Також аналізуються проблеми та виклики, з якими може зіткнутися використання мережевих технологій у медичній практиці.

Виокремлено особливості підходу, набору даних та задач кожного з розглянутих джерел. У більшості відібраних робіт для досягнення цілей використовуються методи глибокого навчання, зокрема, згорткові нейронні мережі. Проведено критичний аналіз відібраних літературних джерел – визначено особливості реалізації поставлених завдань, переваги та недоліки існуючих досліджень.

Сформульовані основні проблеми наявних досліджень. Головною проблемою є класифікація лише на дві, три або чотири стадії хвороби, в той час як шляхи лікування безпосередньо залежать від її стадії. Інші проблеми включають навчання та тестування на малих обсягах даних, а також відсутність

тестування розробленої моделі на інших наборах даних для перевірки її ефективності.

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.

Сформульована мета та завдання дослідження на основі виокремлених проблем. Робота спрямована на розробку підходу для класифікації стадій хвороби з використанням ансамблю моделей машинного навчання на основі аналізу персоналізованих даних пацієнтів.

РОЗДІЛ 2.
РОЗРОБЛЕННЯ МОДЕЛІ ПРОЦЕСУ ОПРАЦЮВАННЯ ПЕРСОНАЛІЗОВАНИХ ДАНИХ

У розділі розроблено концептуальну модель, що формалізує процес опрацювання персоналізованих даних. Ця модель консолідує етапи збору даних від периферійних пристроїв та інших джерел, передачі даних, а також етапи їх опрацювання та аналізу. Проаналізовано існуючі давачі та протоколи, які забезпечують збір і передачу клінічних даних до місця їх опрацювання. Узагальнено модель інформаційної технології опрацювання 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.

Матеріали розділу опубліковані у роботах автора [132, 133, 139, 140].

2.1. Розроблення концептуальної моделі процесу опрацювання персоналізованих даних

Аналіз систем обробки персоналізованої медичної інформації показав, що для покращення медичних послуг важливо враховувати потреби користувачів цих систем. Система збору персоналізованих даних від давачів є комплексним рішенням, яке об'єднує апаратні та програмні компоненти для збору, зберігання, аналізу та візуалізації медичних даних. Вона включає кілька ключових елементів.

Давачі, такі як біометричні пристрої вимірювання пульсу, артеріального тиску, рівня глюкози, температури тіла та рівня кисню, забезпечують постійний моніторинг фізичного стану пацієнта. Периферійні пристрої, такі як смарт-годинники, фітнес-трекери та медичні браслети, відстежують рівень фізичної

активності і стан здоров'я в реальному часі. Імплантовані давачі, такі як кардіостимулятори, глюкометри, використовують для тривалого моніторингу.

Комунікаційні технології, зокрема, бездротові протоколи (Bluetooth, Wi-Fi, Zigbee, MQTT та інші) та мобільні мережі (4G, 5G), здійснюють передачу даних від давачів до базової станції або хмарної платформи. Інтерфейс збору даних зазвичай представлений мобільними додатками, які збирають інформацію від пристроїв та базових станцій, об'єднують ці дані і передають у центральну систему.

Серверна інфраструктура та хмарні сервіси, такі як AWS, Microsoft Azure і Google Cloud, відповідають за обробку та зберігання великих обсягів даних, забезпечуючи доступність цих даних з будь-якої точки. Аналітичні інструменти, які використовують алгоритми машинного навчання та штучного інтелекту, аналізують дані, виявляють патерни, прогнозують стан здоров'я пацієнта та надають рекомендації. Панелі моніторингу та візуалізації відображають ці дані у формі, зручній для медичного персоналу [30].

Системи безпеки та конфіденційності (шифрування даних та контроль доступу), забезпечують захист медичної інформації. Інтеграційні механізми, як API, дають змогу інтегрувати систему, що розробляється з іншими медичними системами, такими як електронні медичні записи (EMR) та клінічні інформаційні системи (CIS) [31, 32]. Серед переваг цієї системи можна виділити підвищену точність діагностики, що досягається завдяки постійному моніторингу стану здоров'я пацієнта. Також важливим є можливість отримувати дані в режимі реального часу і оперативно реагувати на зміни стану пацієнта. Індивідуальне налаштування рекомендацій і планів лікування забезпечує персоналізований підхід до кожного пацієнта. Крім того, система сприяє покращенню ефективності лікування за рахунок точнішої та своєчаснішої інформації про стан пацієнта. Однак, є й недоліки. Високі витрати на впровадження за рахунок витрат на обладнання, програмне забезпечення та інтеграцію, становлять серйозну проблему. Проблема з конфіденційністю даних виникає через необхідність забезпечення високого рівня їх захисту. Додатково можуть виникати технічні

труднощі з сумісністю між різними пристроями та програмним забезпеченням, а також необхідність постійного обслуговування, включаючи оновлення програмного забезпечення та надання технічної підтримки.

Загалом, системи збору персоналізованих даних від давачів є потужним інструментом для сучасної медицини, що сприяє покращенню якості медичних послуг та підвищенню рівня здоров'я пацієнтів [11]. У наукових працях Мельникової Н.І.запропоновано формальну модель стану пацієнта, що забезпечує пошук оцінки його стану та рішень щодо оптимізації процесу одужання, а також узагальнену модель інформаційної технології опрацювання персоналізованих даних для аналізу стану особи, пошук рішень щодо покращення процесу ідентифікації стадії захворювання [45].

Доцільним є синхронізувати етапи опрацювання інформації про особу в одному програмному модулі, який вирішуватиме задачу збору інформації, опрацювання даних, класифікації за характеристиками стану особи, валідації даних, контролю за відповідністю результатів та прогнозування наступних станів.

Отже, враховуючи, що концептуальна модель процесу є узагальненим уявленням про процес, який описує його основні елементи, взаємозв'язки та функціональні характеристики, то головною метою концептуальної моделі є надання загального розуміння процесу при використанні для подальшої деталізації, аналізу, проєктування та удосконалення. На рисунку 2.1. зображена концептуальна модель, яка описує принципи функціонування комплексної системи опрацювання та аналізу персоналізованих даних.

Для кожного поданого процесу визначені характеристики:

1. Збір даних:

- давачі: периферійні пристрої, імплантовані давачі, смартфони, інші медичні пристрої.

- типи даних: вимірювання фізіологічних показників (пульс, артеріальний тиск, рівень глюкози в крові), активність, параметри сну, екологічні дані.

2. Передача даних:

- мережеві технології: Bluetooth, Wi-Fi, мобільні мережі.

- безпека передачі: шифрування даних, аутентифікація користувачів.

Рис.2.1 Концептуальна модель процесу опрацювання персоналізованих даних

3. Опрацювання даних:

- інтеграція даних: збір даних з різних джерел у єдину базу.

- попередня обробка: фільтрація шумів, нормалізація даних, усунення пропусків.

4. Класифікація за характеристиками стану особи:

- алгоритми класифікації: машинне навчання, нейронні мережі, алгоритми прийняття рішень.

- класифікаційні категорії: стан здоров’я (норма, передзахворювання, захворювання), ризики (низький, середній, високий).

5. Валідація даних:

- перевірка точності: контроль достовірності та точності даних.
- Аналіз аномалій: виявлення та корекція аномальних значень.

6. Контроль за відповідністю результатів:

- верифікація моделей: перевірка ефективності алгоритмів класифікації та прогнозування.
- зворотний зв'язок: оцінка результатів фахівцями, коригування моделей на основі їхньої оцінки.

7. Прогнозування наступних станів пацієнта:

- моделі прогнозування: часові ряди, регресійні моделі, глибоке навчання.
- індивідуальні плани дій: рекомендації щодо профілактики та лікування на основі прогнозів.

2.2. Застосування різнотипових давачів для збору медичних даних

Процес збору давачами медичних даних є невід'ємною частиною сучасної медицини, оскільки він дає змогу отримувати важливу інформацію про стан пацієнта в режимі реального часу [38, 39]. Нижче подано типи існуючих давачів для збору персоналізованих даних (рис.2.1).

Давачі для збору медичних даних	Електрокардіографи (ЕКГ) давачі
	Пульсоксиметри
	Глюкометри
	Тонометри
	Температурні давачі
	Біохімічні давачі
	Давачі руху та активності
	Монітори сну
	Спірометри
	Капнографи
	Імплантовані давачі «Нейродавачі»

Рис.2.2 Типи існуючих давачів для збору персоналізованих даних

Електрокардіографічні (ЕКГ) давачі: використовуються для моніторингу електричної активності серця. Ці давачі можуть бути вбудовані в переносні пристрої, такі як смарт-годинники або спеціалізовані медичні прилади.

Пульсоксиметри: вимірюють рівень насичення киснем у крові та пульс. Їх застосовують і у лікарнях, і в персональних моніторах здоров'я.

Глюкометри: використовують для вимірювання рівня глюкози в крові, які можуть бути бездротовими і синхронізуватися зі смартфонами.

Тонометри: прилади для вимірювання артеріального тиску, які зберігають історію вимірювань і передають дані лікарям через інтернет.

Температурні давачі: використовуються для вимірювання температури тіла, бувають у вигляді термометрів або вбудовані в смарт-пристрої.

Біохімічні давачі: вимірюють різні хімічні показники крові або інших біологічних рідин.

Давачі руху та активності: використовуються для моніторингу фізичної активності пацієнтів, їх рухливості, кількості пройдених кроків та спалених калорій.

Монітори сну: аналізують якість та тривалість сну, визначаючи фази сну та виявляють можливі розлади, такі як апное.

Спірометри: використовують для вимірювання об'єму та швидкості повітря, що видихається, допомагаючи оцінити функцію легень.

Капнографи: вимірюють рівень вуглекислого газу у видихуваному повітрі, що важливо для пацієнтів з респіраторними проблемами.

Імплантовані давачі: використовуються для постійного моніторингу внутрішніх органів та систем (серцеві стимулятори, давачі глюкози).

Нейродавачі: Використовуються для вимірювання електричної активності мозку (наприклад, ЕЕГ) і можуть бути важливими для діагностики та моніторингу неврологічних станів.

2.3. Формалізація моделі інформаційної технології опрацювання персоналізованих даних для аналізу стану особи

Визначення індивідуальних характеристик, необхідних для вирішення завдань персоналізації, залежить від ключових факторів ідентифікації особи. У медицині для формалізованого представлення особи розглядаються основні параметри його загального стану разом із конкретними характеристиками.

Під час діагностування стану хворого експерти встановили, що ключовим показником ідентифікації стану є урахування різноманітних аспектів загального стану, таких як фізіологічні показники, результати лабораторних досліджень, аналіз запальних процесів тощо. Це підтверджує важливість аналізу конкретних показників у певний момент часу [30]. Наприклад, на перебіг хвороб, пов'язаних з серцево-судинними захворюваннями, впливають показники артеріального тиску, індекс маси тіла, місце проживання та інші зовнішні фактори.

Тому розроблення моделей аналізу процесу обробки персоналізованих даних та прогнозування поточного стану стає надзвичайно важливим. Ця задача характеризується такими аспектами: множинними критеріями, залежністю від часу та гетерогенністю вхідних даних.

Багатокритеріальність визначається впливом різних факторів, зокрема, спадковістю, харчуванням, палінням і т. д.

Для об'єктивної оцінки поточного стану пацієнта необхідно побудувати його формальну інформаційну модель, яка дає змогу представити його як систему, що відображає зв'язки між пріоритетними ознаками процесу опрацювання даних та надає інформацію про його стан. Враховуючи потреби користувачів, система сприяє покращенню ефективності лікування за рахунок більш точної та своєчасної 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.

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, які подані на рис. 2.2, граничні значення яких визначаються відповідною номенклатурою пристрою. Отже, множина давачів для збору даних подана виразом:

$$S = \{s_1, s_2, \dots, s_n\}; ; \tag{2.1}$$

Показники давачів представлені множиною *C*, що формалізує дані особи після збору клінічних даних відповідними сенсорами, які належать до множини *S*, що відображені у формулі 2.1. 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_1. Отже, множина сенсорних даних отримана з давачів відображена у формулі 2.2:

$$C = \{c_1, c_2, \dots, , c_n\}; \tag{2.2}$$

$$S \to C; \text{де} \;\; s_i \to c_i.$$

Фізіологічні та антропометричні дані представлені множиною *A*, що формалізує дані особи після збору історії хвороби. Наприклад: вік, супутні захворювання, шкідливі звички (паління, зловживання алкоголем, тощо). Отже, множина антропометричних та фізіологічних даних подана виразом 2.3, де *m* визначається кількістю необхідних показників, відповідно до протокольних рішень давгностованої патології:

$$A = \{a_1, a_2, \dots, a_m\}; \tag{2.3}$$

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

$$C \cup A = P;$$

$$P = \{p_1, p_2, \dots, p_m\}, k \leq m + n. \tag{2.4}$$

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

$$D = \{d_1, d_2, \dots, d_r\}; \tag{2.5}$$

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

$$D: P \to V(K). \tag{2.6}$$

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

$$p_i =< C, A >, \tag{2.7}$$

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

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

$$D(V) = \{x \in P \mid \forall y \in V(k_1, k_2, \ldots, k_r) \quad (\forall i \in \{1,\ldots, r\}[\, x_i \geq y_i])\}. \qquad (2.8)$$

Прикладом правил є рішення щодо 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:

$$I = \langle S, C, A, P, K, V, D \rangle, \qquad (2.9)$$

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 –множина гіперпараметрів класифікаторів хвороби особи, V –множина результуючих показників стану особи, що залежить від множини гіперпараметрів класифікатора, D – множина правил, які визначаються з урахуванням персональних даних особи та результуючих 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. Подальші дослідження спрямовані на валідацію моделі на великих клінічних наборах даних та оптимізацію алгоритмів прийняття рішень з використанням передових методів машинного навчання.

Для збору, передачі, обробки та прогнозування даних особи з метою аналізу її стану необхідно застосовувати індивідуальний підхід. Це дозволить врахувати унікальні особливості пацієнта та його індивідуальні характеристики.

2.4. Особливості препідготовки та опрацювання вхідних персоналізованих даних

Попереднє опрацювання даних є невід'ємною частиною машинного навчання, яке відіграє ключову роль у підготовці даних для створення моделей. Це перевірений метод вирішення таких реальних проблем, як неповнота, непослідовність, відсутність трендів та наявність численних помилок у сучасних даних.

Виокремлення клінічних та антропометричних даних формує множину необхідних даних для персоналізації інформації про пацієнта. Це включає вибір важливих ознак і елементів з набору даних, що сприяє індивідуальному підходу лікаря при оцінці стану пацієнта.

Приклад 1.

Для проведення досліджень використано набір даних, який був отриманий з Kaggle [19] і містить зображення МРТ головного мозку пацієнтів, розділені на п'ять стадій захворювання Альцгеймера (AD - хвороба Альцгеймера, CN – когнітивна норма, EMCI – раннє легке когнітивне порушення, LMCI – пізнє легке когнітивне порушення, MCI – легке когнітивне порушення).

Зазначений набір даних складається з 2953 файлів у форматі JPEG розміром 208x176. Вихідні зображення вже розділені на тренувальні та тестові частини, але для зручності аналізу і обробки вирішено їх об'єднати в один набір. Приклади зображень для кожної з п'яти стадій відображені на рис. 2.3.

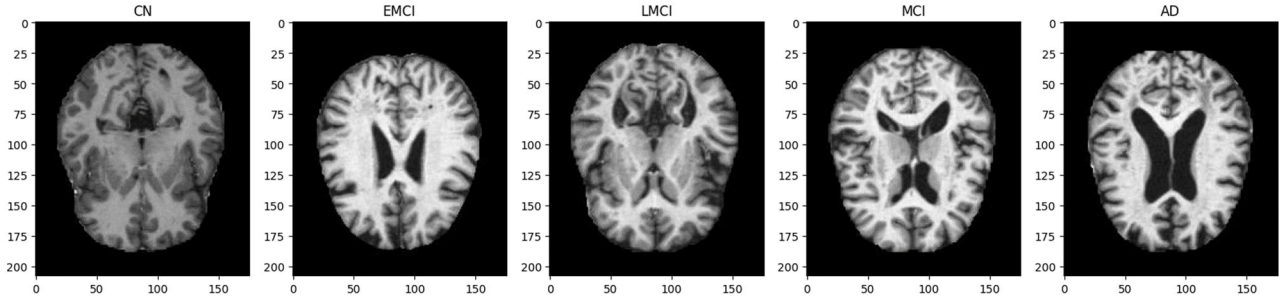


Рис. 2.3 Зображення головного мозку для кожної стадії хвороби в наборі даних

На зображеннях видалені немозкові тканини, щоб зосередитися лише на структурах мозку. Для обробки набору даних здійснено кроки, зображені на рис. 2.4.

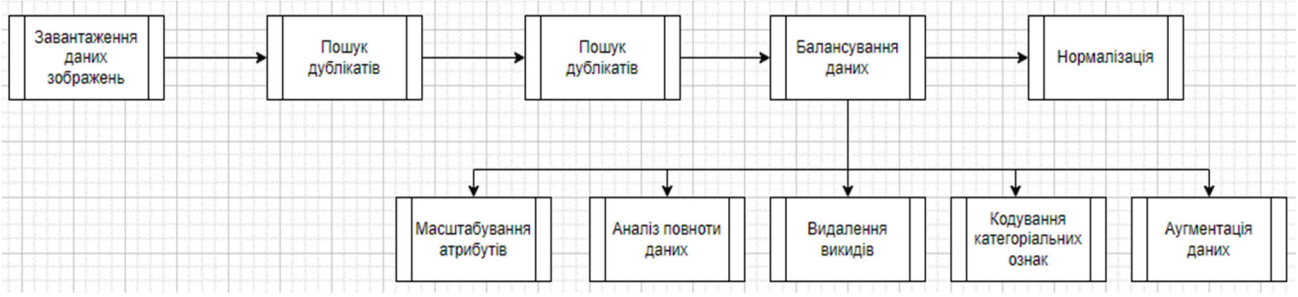


Рис. 2.4 Етапи алгоритму попередньої обробки набору даних зображень МРТ

2.4.1. Алгоритм обробки персоналізованих даних особи для аналізу стану особи

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 в 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 [20].

Вважаючи, що множина вхідних даних від давачів *S*, що формують множину клінічних даних *C* з розділу 2, п.2.3 залежність буде наступною:

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

де для прикладу x_i - це зображення, а y_i - відповідна мітка класу.

2. Перевірка на наявність дублікатів у наборі даних

Перевірка наявності дублікатів у наборі даних є важливим етапом підготовки даних, оскільки допомагає забезпечити точність та надійність моделі. Це особливо важливо для зображень, оскільки наявність дублікатів може негативно позначитися на процесі навчання моделі та її здатності до узагальнення.

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

Дублікати можуть бути визначені, як ідентичні зображення або зображення, які є однаковими з точністю до певної міри схожості.

Для виявлення дублікатів використано хешування для швидкої перевірки ідентичності зображень та метрики схожості. Для обчислення хеш-коду для кожного зображення x_i використовується наступні функція:

$$h_i = \mathrm{hash}(x_i)$$

Для збереження хеш-коду в множині H та перевірки на наявність дублікатів використано залежність:

$$H = \{h_i | i = 1,2,\dots,N\}$$

Якщо кардинальність множини H (тобто кількість унікальних елементів) менша за N, це означає, що у наборі даних є дублікати.

На наступному кроці використано метрики схожості, такі як MSE (Mean Squared Error), SSIM (Structural Similarity Index), або інші для обчислення подібності між кожною парою зображень (x_i,x_j):

$$\text{sim}(x_i,x_j)=\text{similarity\_metric}(x_i,x_j).$$

Визначено поріг схожості τ. Якщо $\text{sim}(xi,xj)\geq\tau$, то x_i і x_j вважаються дублікатами.

Перевірка наявності дублікатів у наборі даних виконується за допомогою різних методів, в залежності від вимог до точності та швидкості. Просте хешування є ефективним способом точного визначення ідентичних зображень, тоді як метрики схожості можуть бути використані для виявлення зображень, які схожі, але не ідентичні.

Алгоритм SHA-256 є одним з найпоширеніших методів хешування, який використовують для пошуку дублікатів фотографій. Однією з його головних переваг є те, що він є більш безпечним і надійним, оскільки генерує довший хеш-код (256 бітів) порівняно з MD5 (128 бітів). Це означає, що ймовірність зіткнення (колізії) при генерації хеш-коду для різних даних дуже низька. Таким чином, шанси на те, що дві різні фотографії мають однаковий SHA-256 хеш-код, дуже малі [22,23].

Під час перевірки зображень обчислюється хеш-значення кожного з них. Якщо такий хеш ще не зустрічався, зображення додається до списку для подальшої роботи. У випадку знаходження ідентичного хешу, таке зображення вважається дублікатом. Для візуалізації кількості зображень по кожній стадії хвороби та кількості дублікатів можна використовувати стовпчасту діаграму, яка подана на рис. 2.5.

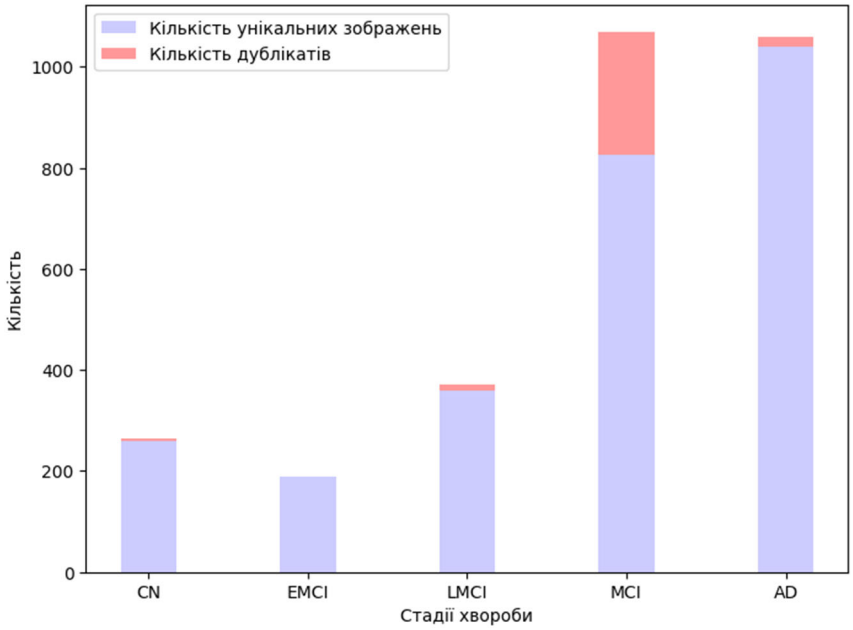


Рис.2.5 Стовпчаста діаграма про кількість зображень кожної стадії хвороби

Після виконаних операцій отримано наступний розподіл зображень по класах:

AD – 1039; MCI – 827; LMCI – 720; EMCI – 752; CN – 780.

На основі аналізу результатів діаграми видно, що клас MCI має найбільшу кількість дублікатів. Крім того, цей набір даних є незбалансованим.

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 може знизити точність класифікації для менш представлених категорій [26, 27].

Тому балансування набору даних зображень є важливим етапом підготовки даних для навчання моделей машинного навчання. Його суть полягає у вирівнюванні кількості зображень у різних класах. Це допомагає уникнути упередженості моделі до певного класу, яке може виникнути внаслідок дисбалансу в даних.

Формалізація задачі:

Нехай є набір даних зображень

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

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:

1. Визначення кількості зображень у кожному класі

Для кожного класу визначено кількість зображень:

$$n_c: n_{\text{c}} = \sum_{i=1}^{n} I(y_i = c)$$

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

2. Визначення цільової кількості зображень у кожному класі

Нехай N - бажана кількість зображень у кожному класі після балансування. Визначимо N як середнє значення кількості зображень у всіх класах:

$$N = \frac{\sum_{j=1}^{L_C} n_c}{L_C}$$

3. 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 видалення деяких зображень з надмірно представлених класів.

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

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

Комбінований підхід

Можна комбінувати обидва підходи для досягнення балансування, якщо деякі класи мають значно менше зображень, а інші - більше.

Основні етапи процесу балансування:

1 Обчислення кількості зображень для кожного класу n_c.

2 Визначення цільової кількості зображень N.

3 Для кожного класу c: Якщо $n_c<N$, додати копії наявних зображень (oversampling). Якщо $n_c>N$, випадковим чином вибрати N зображень (undersampling).

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 обробки необхідно провести аналіз через додаткові підзадачі балансування даних:

- аналіз повноти даних;

- вилучення викидів;

- кодування категоріальних ознак;

- масштабування атрибутів;

- аугментація даних.

Для збалансування набору даних потрібно додавати симульовані зображення до менш представлених класів. Це реалізовано за допомогою аугментації даних шляхом відображення зображень відносно вертикальної осі та повороту на певний кут. Детальний опис цього підходу до покращення процесу обробки даних розглянуто у розділі 3.

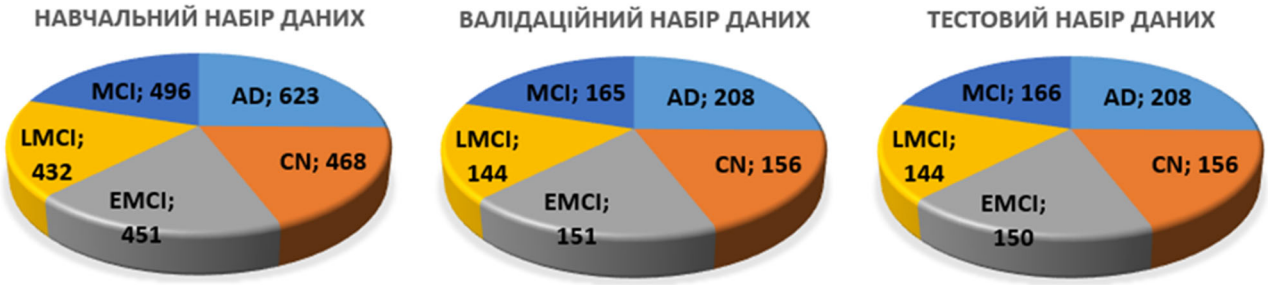
4. Нормалізація зображень

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 [34]. Нормалізація полягала у приведенні значень пікселів до діапазону від 0 до 1, що здійснювалося за допомогою стандартної формули (2.10):

$$normalized_X = \frac{X-min(X)}{max(X)-min(X)} \quad (2.10)$$

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 не бачених даних. Розподіл даних збережено з балансом класів у кожному наборі у відношенні 60:20:20. На рис.2.6 показано розподіл набору даних та кількість зображень для кожного класу у нових наборах.



а) б) в)

Рис.2.6 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).

2.4.2. Принципи етапу збалансування вхідних персоналізованих даних

Опрацювання незбалансованих даних включає в себе низку операцій, які забезпечують якісну обробку та підготовку даних для подальшого аналізу [38, 39].

Представлено приклад іншого набору даних, який використовують для вирішення певного завдання — прогнозування ризиків виникнення інсульту. Цей набір даних був отриманий з веб-сайту компанії DataHack Analytics Vidhya.

Приклад 2. У даному прикладі розглядаються етапи алгоритму забезпечення балансу даних. Цей процес включає такі ключові підзадачі збалансування даних: 1. Аналіз повноти даних. 2. Видалення викидів. 3. Кодування категоріальних ознак. 4. Масштабування атрибутів.

Набір даних містить: 11 колонок з різними показниками та параметрами пацієнтів і 4981 рядок зі значеннями показників та параметрів для кожного пацієнта.

Таблиця 2.1

Персоналізовані дані стану особи

Назва колонки	Тип(значення) колонки	Опис колонки
1	2	3
gender	Стрічка(Male, Female, Other)	Стать пацієнта
age	Число	Вік пацієнта
hypertension	Число(1, 0)	Наявність гіпертонії у пацієнта (1 - так, 0 - ні).
heart_disease	Число(1, 0)	Наявність хвороби серця у пацієнта (1 - так, 0 - ні).
ever_married	Стрічка(Yes, No)	Сімейний стан пацієнта (Yes - одружений, No - неодружений).
work_type	Стрічка(children, Govt_job, Never_worked, Private, Selfemployed)	Тип роботи пацієнта.
Residence_type	Стрічка(Urban, Rural)	Тип місцевості проживання пацієнта.
avg_glucose_level	Число з плаваючою комою	Середнє значення рівня глюкози в крові пацієнта.
bmi	Число з плаваючою комою	Індекс маси тіла пацієнта.
smoking_status	Стрічка (formerly smoked, never smoked, smokes, unknown)	Статус курця пацієнта.
stroke	Число(1, 0)	Наявність інсульту у пацієнта (1 - так, 0 - ні).

Згідно даних табл.2.1, основні характеристики особи включають параметри 'gender', 'age', 'hypertension', 'heart_disease', 'ever_married', 'work_type', 'Residence_type', 'avg_glucose_level', 'bmi', 'smoking_status' і 'stroke'.

Набір даних має значну незбалансованість, оскільки значення без інсульту зустрічається 4733 рази, а з інсультом лише 248 разів. Цей фактор обов'язково слід врахувати під час попереднього опрацювання даних.

Загалом, у наборі даних є різноманітні типи інформації:

Категоріальні ознаки:

- Номінальні: 'gender', 'ever_married', 'work_type', 'Residence_type', 'smoking_status'.

- Бінарні: 'hypertension', 'heart_disease', 'stroke'.

Числові ознаки:

- Неперервні: 'avg_glucose_level', 'bmi'.

- Дискретні: 'age'.

Візуалізація персоналізованих даних представлена на рисунках 2.7 і 2.8.

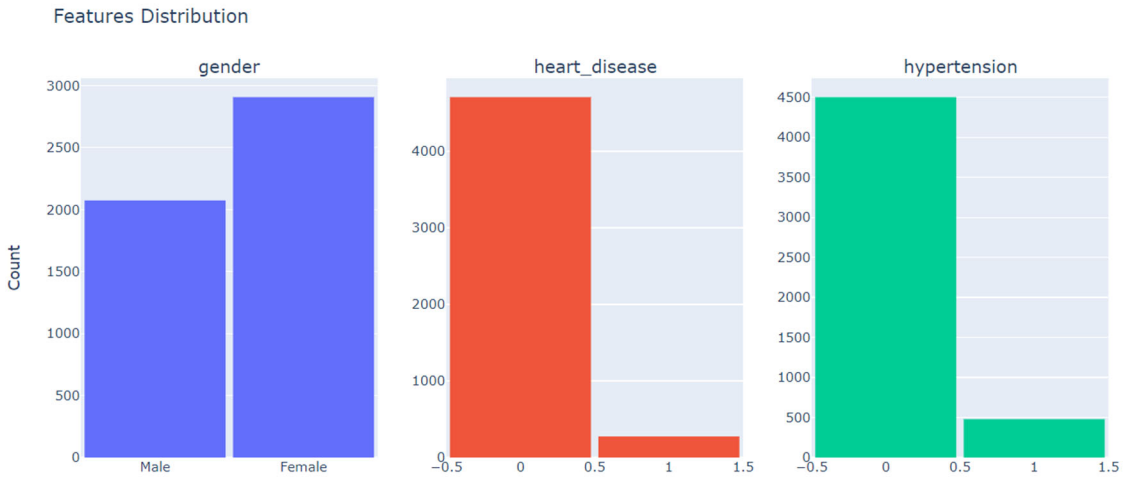


Рис.2.7 Розподіл ознак – gender, heart_disease та hypertension

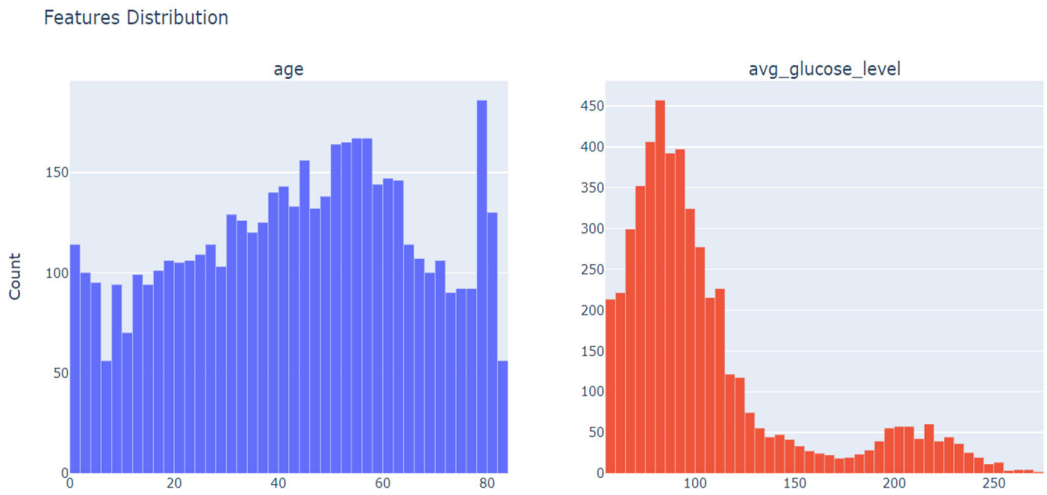


Рис. 2.8 Розподіл ознак – age та avg_glucose_level

З аналізу рисунків 2.7 та 2.8 зроблено наступні висновки:

- Розподіл віку людей у наборі даних нагадує рівномірний розподіл.

- В наборі даних є дещо більше людей віком старших 40 років та молодших 70 років.

- Кількість людей віком 10-15 років та більше 85 років є найменшою у цьому наборі даних.

- Показник середнього рівня глюкози найчастіше зустрічається на рівнях 85-90, а найрідше - на рівнях 175-180 та 250+.

- Частота середнього рівня глюкози значно зменшується після показника 100.

- Кількість людей, які не мали інсульту, значно переважає кількість тих, у кого він був. Співвідношення між людьми без інсульту та людьми з інсультом становить 9:1.

Для досягнення якісних результатів моделі дуже важливо збалансувати даний набір даних. Крім того, статистичні характеристики числових ознак представлені в таблиці 2.2:

Таблиця 2.2

Статистичні характеристики числових ознак

	age	avg_glucose_level	bmi
count	4981	4981	4981
mean	43.42	105.94	28.50
std	22.66	45.08	6.79
min	0.08	55.12	14
25%	25	77.23	23.7
50%	45	91.85	28.1
75%	61	113.86	32.6
max	82	271.74	48.9

З таблиці 2.2 можна зробити висновки:

- Максимальний вік людини у наборі даних становить 82 роки.
- Середній вік людини у наборі даних складає 43 роки.
- Максимальний рівень глюкози в людей становить 271.74.
- Середній рівень глюкози в людей складає 105.94.
- Мінімальний рівень глюкози в людей становить 55.12.
- Максимальний індекс маси тіла в людей становить 48.9.
- Середній індекс маси тіла в людей складає 28.50.
- Мінімальний індекс маси тіла в людей становить 14.

Ця статистика також вказує на високу дисперсію та наявність багатьох викидів для таких характеристик, як індекс маси тіла та середній рівень глюкози.

1. Аналіз повноти даних

Для етапу аналізу повноти даних спочатку проведено перевірку даних на наявність відсутніх значень (Null):

```
# check if we have null values in data set
df.isnull().sum()
```

```
gender               0
age                  0
hypertension         0
heart_disease        0
ever_married         0
work_type            0
Residence_type       0
avg_glucose_level    0
bmi                  0
smoking_status       0
stroke               0
dtype: int64
```

Рис. 2.9 Перевірка даних на наявність відсутніх значень (Null)

З аналізу рис. 2.9 випливає, що у наборі даних значення з пропусками відсутні. Це свідчить про те, що немає необхідності виконувати операції з видаленням або усередненням відсутніх даних.

2. Видалення викидів

Викиди - це значення, що відрізняються від загального шаблону або очікуваного розподілу в наборі даних. Поява викидів може бути спричинена помилками вимірювання, технічними аномаліями або недостовірними даними. Для аналізу викидів необхідно визначити ознаки, які мають викиди [40].

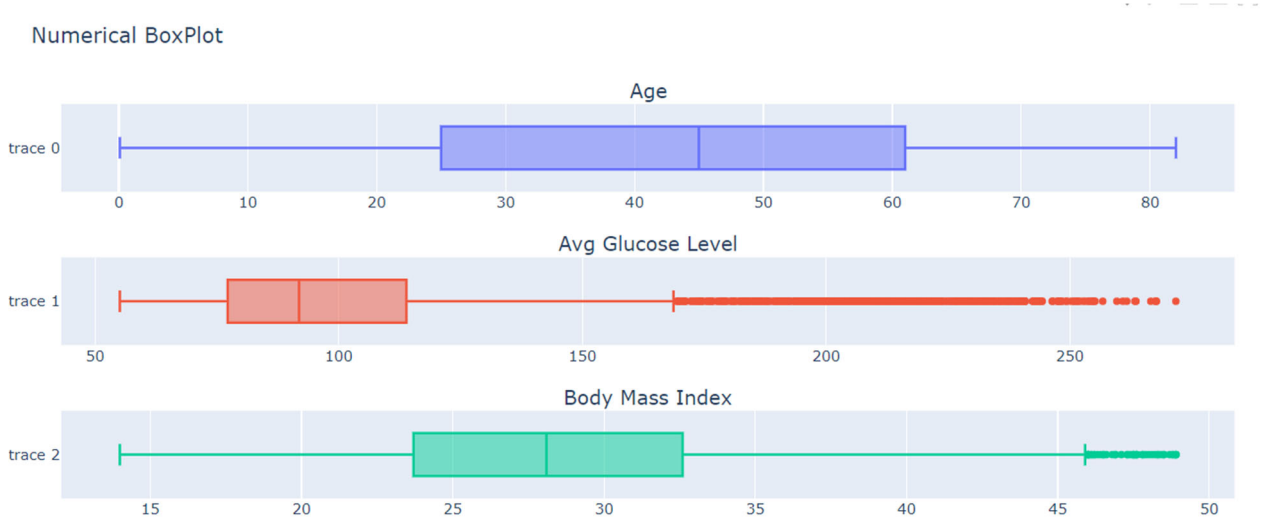


Рис. 2.10 Діаграма числових ознак

З аналізу рис. 2.10 можна зробити висновок, що в наборі даних спостерігаються викиди в характеристиках, таких як середній рівень глюкози та індекс маси тіла. При дослідженні точкового графіку для середнього рівня глюкози та індексу маси тіла (рис. 2.11), видно, що ці дані мають викиди, які необхідно опрацювати та видалити за допомогою методу міжквартильного розмаху.

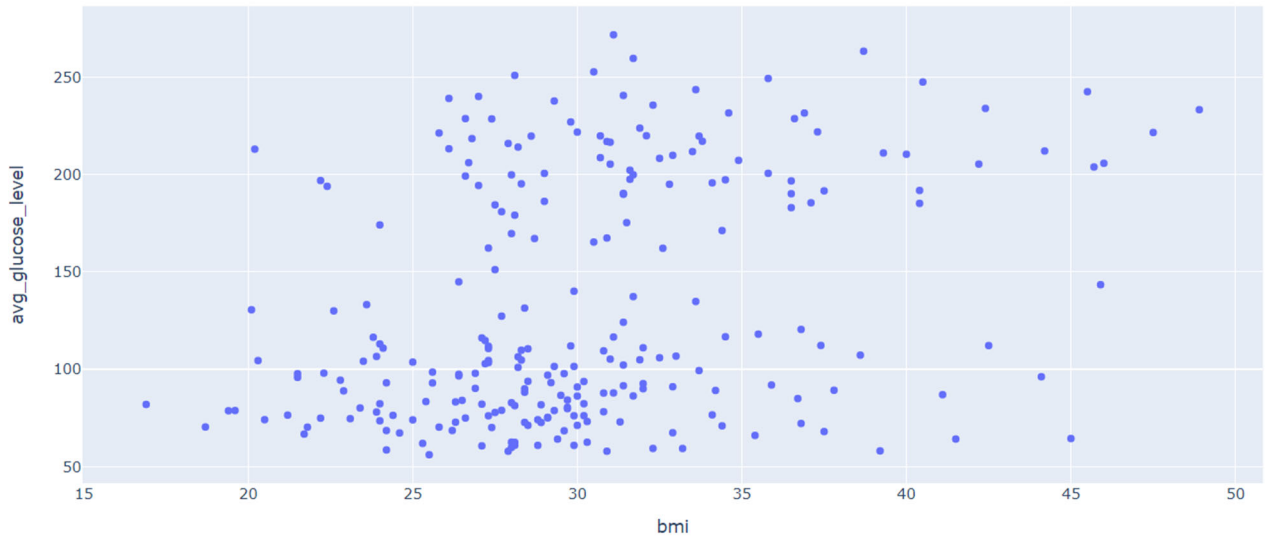


Рис. 2.11 Точковий графік avg_glucose_level та bmi

На рис. 2.12 – рис. 2.13 зображені діаграма та точковий графік після видалення викидів:

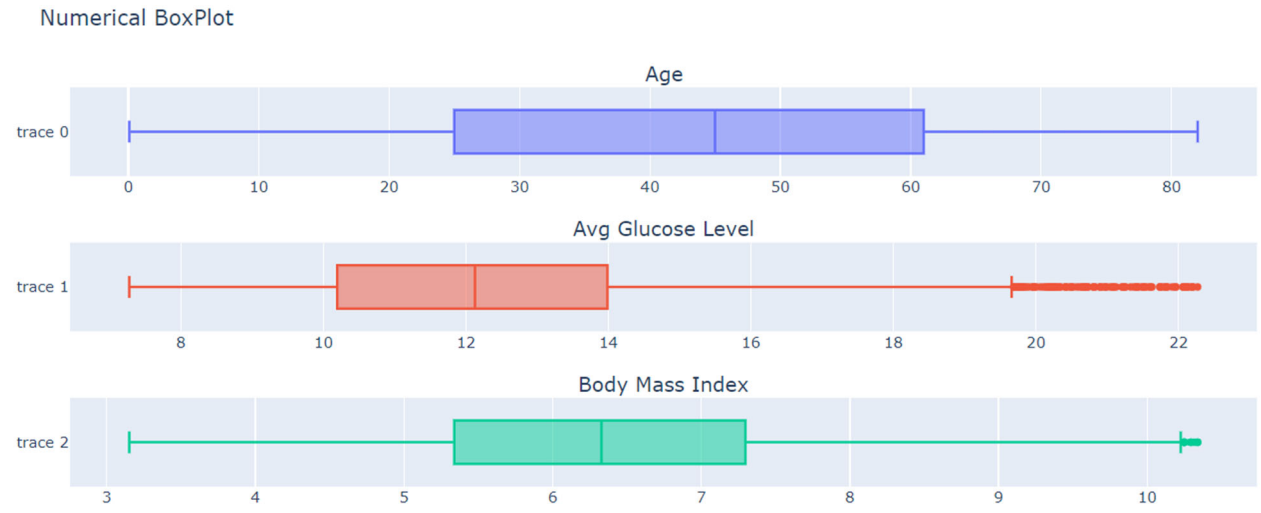


Рис. 2.12 Графік числових ознак після видалення викидів

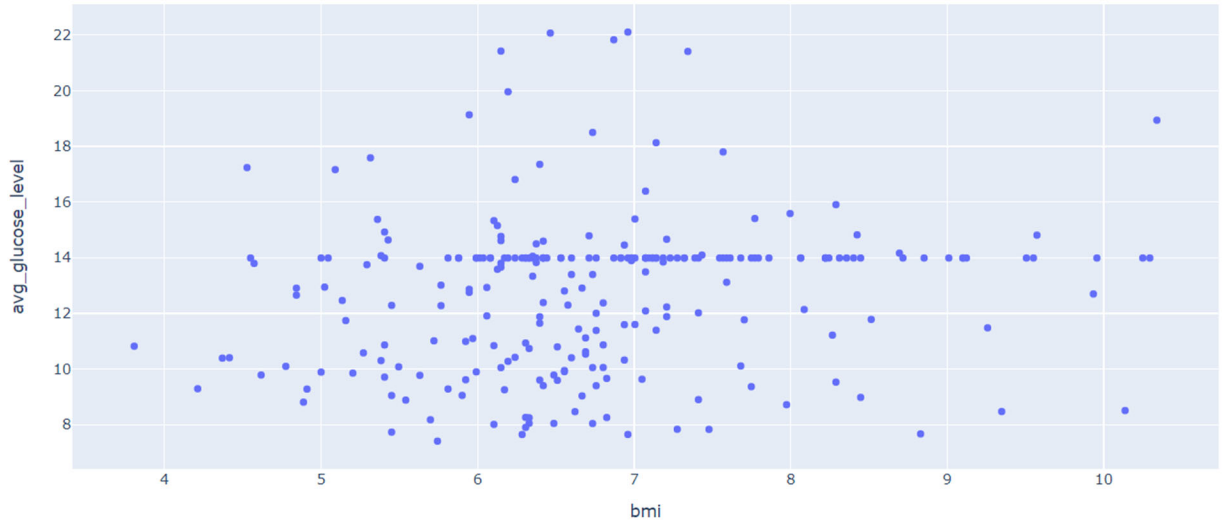


Рис. 2.13 Точковий графік avg_glucose_level та bmi після видалення викидів

Із зображених даних зроблено висновок про успішність операції з вилучення викидів: дані структуровані і не містять викидів.

3. Кодування категоріальних ознак

Кодування категоріальних ознак у наборі даних передбачає перетворення категоріальних ознак, які можна представити у бінарному вигляді (стать, сімейний стан, місце проживання). Під час цього кодування замінюються такі значення:

gender: male на 1, female на 0.

ever_married: Yes на 1, No на 0.

Residence_type: urban на 1, rural на 0.

Результати даної операції зображені на рис 2.14:

	age	hypertension	heart_disease	work_type	avg_glucose_level	bmi	smoking_status	stroke	gender	ever_married	Urban
0	67.0	0	1	Private	105.943562	36.6	formerly smoked	1	1	1	1
1	80.0	0	1	Private	105.920000	32.5	never smoked	1	1	1	0
2	49.0	0	0	Private	105.943562	34.4	smokes	1	0	1	1
3	79.0	1	0	Self-employed	105.943562	24.0	never smoked	1	0	1	0
4	81.0	0	0	Private	105.943562	29.0	formerly smoked	1	1	1	1

Рис. 2.14 Результати кодування ознак – gender, ever_married та Residence_type

Кодування звичайних категоріальних ознак (тип роботи та статус куріння) виконано за допомогою методу One Hot Encoding. Результати цього методу представлені на рис. 2.15:

married	Urban	work_type_Private	work_type_Self-employed	work_type_children	smoking_status_formerly smoked	smoking_status_never smoked	smoking_status_smokes
1	1	1	0	0	1	0	0
1	0	1	0	0	0	1	0
1	1	1	0	0	0	0	1
1	0	0	1	0	0	1	0
1	1	1	0	0	1	0	0

Рис. 2.15 Результати кодування ознак – work_type та smoking_status

4. Масштабування атрибутів

Масштабування атрибутів за допомогою методу Min-Max Scaler враховує всі числові ознаки, крім закодованих категоріальних (оскільки вони вже знаходяться в діапазоні від 0 до 1).

Результати масштабування атрибутів представлено на рис. 2.16:

	age	hypertension	heart_disease	avg_glucose_level	bmi
0	0.816895	0	1	0.447548	0.708464
1	0.975586	0	1	0.447341	0.579937
2	0.597168	0	0	0.447548	0.639498
3	0.963379	1	0	0.447548	0.313480
4	0.987793	0	0	0.447548	0.470219

Рис. 2.16 Результати масштабування атрибутів – age, avg_glucose_level та bmi

Для вирішення проблеми незбалансованості набору даних за допомогою методу SMOTE, опрацьовано дані stroke. Кількість даних у колонці stroke перед застосуванням методу SMOTE представлена на рис. 2.17, а після застосування - на рис. 2.18:

```
0    4733
1     248
Name: stroke, dtype: int64
```

Рис. 2.17 Кількість даних stroke перед застосування методу SMOTE

```
0    4733
1    2366
Name: stroke, dtype: int64
```

Рис. 2.18 Кількість даних stroke після застосування методу SMOTE

З аналізу рисунків видно, що після використання методу SMOTE дані стали більш збалансованими, оскільки кількість осіб без інсульту стала значно більшою за рахунок додавання синтетичних екземплярів[55, 56]. Після виконання всіх попередніх операцій підготовки даних (за винятком розбиття даних) розглянено матрицю кореляції для виявлення можливих зв'язків перед початком побудови моделей машинного навчання. Матриця кореляції подана на рис. 2.19:

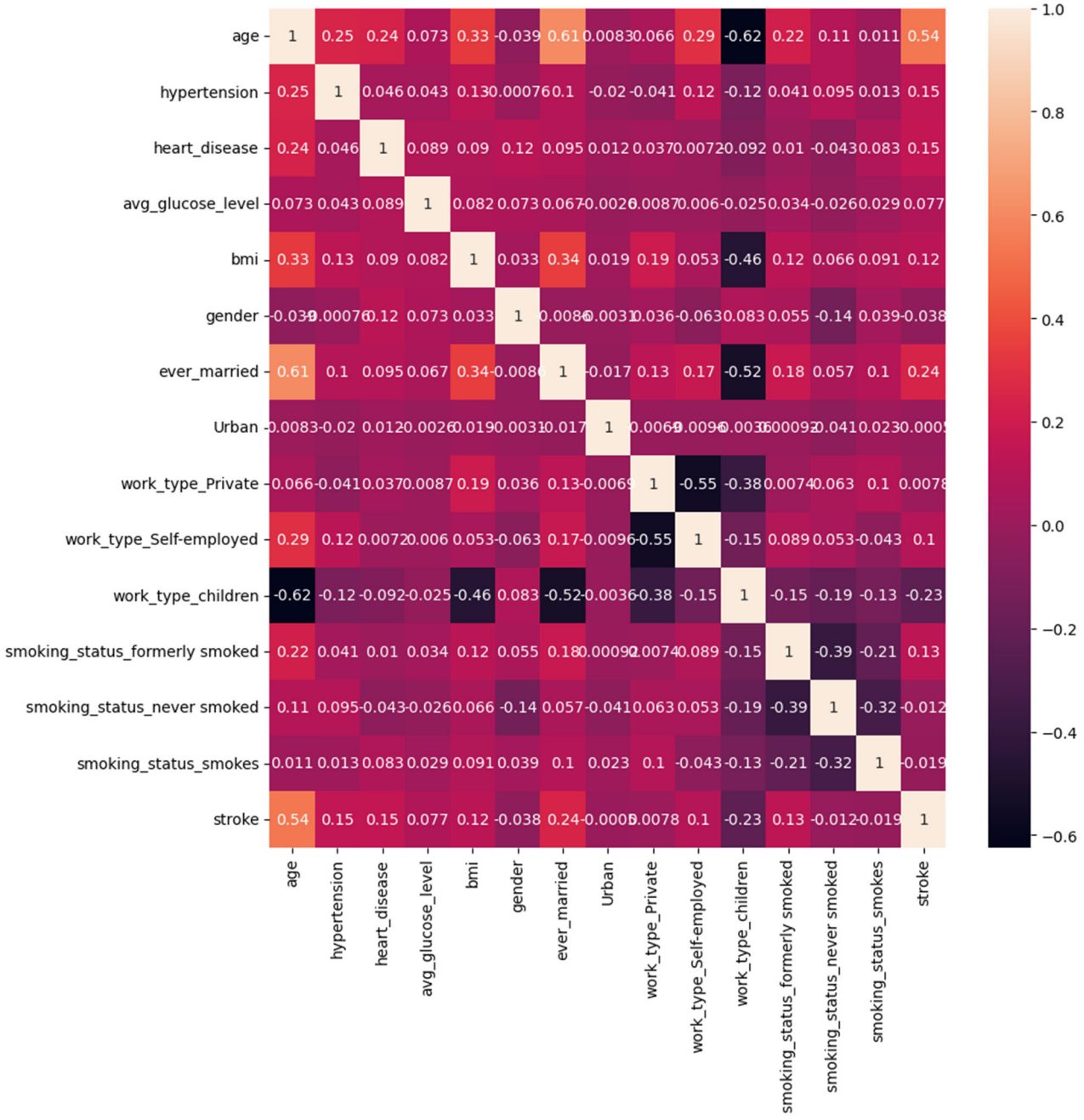


Рис. 2.19 Матриця кореляції даних у наборі даних

Таким чином, ймовірність виникнення інсульту найбільше корелює з віком. Також спостерігається певна кореляція між інсультом та сімейним статусом, хоча вона не настільки сильна, але все ж вища, ніж у випадку інших ознак. Також слабка кореляція виявляється між інсультом та гіпертонією і серцевими проблемами. Інші характеристики практично не впливають на ймовірність виникнення інсульту.

2.5. Висновки до розділу 2

У цьому розділі розроблено концептуальну модель, що формалізує процес опрацювання персоналізованих даних шляхом інтеграції етапів збору даних від периферійних пристроїв, передачі даних та їх опрацювання й аналізу. Це дозволяє забезпечити комплексний підхід до роботи з персоналізованими даними. На відміну від відомих досліджень які пропонували формальну модель стану пацієнта, що спрямована на пошук оцінки його стану та рішень щодо оптимізації процесу одужання, 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 завдяки більш точній та своєчасній інформації.

Проведено аналіз існуючих давачів, які забезпечують збір та передачу даних до центральної точки збору та опрацювання. Підготовлена формалізована модель інформаційної технології опрацювання персоналізованих даних, яка враховує основні параметри загального стану особи та її 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.

РОЗДІЛ 3.
РОЗРОБЛЕННЯ МЕТОДІВ АНАЛІЗУ ТА ОПРАЦЮВАННЯ ПЕРСОНАЛІЗОВАНИХ ДАНИХ

У даному розділі розроблено метод класифікації персоналізованих даних шляхом введення етапу аугментації. Проведено аналіз застосування процесу аугментації для 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.

Матеріали розділу опубліковані у роботах автора [128, 130, 138].

3.1. Задача класифікації

Класифікація є одним з методів навчання, спрямованим на розпізнавання (класифікацію) екземплярів або зразків у один із декількох заздалегідь визначених класів або категорій на основі їхніх особливостей або характеристик [52, 53].

Задача класифікації передбачає створення набору функцій або правил, які відображають екземпляри вхідних даних, представлених як вектори ознак, на відповідні мітки класу. Ці функції мають охоплювати основну структуру та зв’язки в даних, що дає змогу ефективно розрізняти та відокремлювати класи.

Математично це можна представити наступним чином:

Вхідні дані: вхідними даними для проблеми класифікації зазвичай є вектор $X = (x1, x2, \ldots, x_n)$, що містить числові, категоріальні значення, що представляють ознаки або атрибути об’єкта чи спостереження. Цей вектор належить простору ознак, позначеному як X.

Вихідні дані (класи): вихідними результатами проблеми класифікації є дискретна змінна *Y*, яка має набір попередньо визначених категорій або класів

$C = \{C_1, C_2, \ldots, C_k\}$. Мета полягає в тому, щоб знайти функцію (алгоритм або модель), яка призначає деякий клас C одному із вхідних векторів X.

Набір даних: надається набір даних D, який зазвичай містить низку спостережень або зразків (випадків). Кожен екземпляр представлено вектором ознак X_i та його відповідною міткою класу Y_i, позначеною як (X_i, Y_i) для $i = 1, 2, \ldots, N$, де N — загальна кількість екземплярів у наборі даних.

Завдання класифікації полягає у тому, щоб оцінити функцію $f: X \to Y$ таким чином, щоб для будь-якого вхідного екземпляра X, $f(X)$ передбачала правильну мітку класу Y для цього екземпляра. У ідеалі функція $f(X)$ має мінімізувати помилку класифікації екземплярів даних, зберігаючи при цьому прийнятний рівень складності.

Крім того, при класифікації екземплярів може виникнути проблема недостатньої кількості даних. Відсутність даних у процесі класифікації може серйозно підірвати продуктивність та ефективність моделей класифікації [56, 57]. Обмежений обсяг даних не лише ускладнює розробку моделі, але також може спричинити певні труднощі під час процесу класифікації.

Проблеми, пов'язані з обмеженою вибіркою даних, можуть включати:

Перенавчання: з обмеженими навчальними даними моделі можуть страждати від перенавчання, що означає, що вони запам'ятовують навчальні зразки замість того, щоб узагальнювати шаблони даних. Це призводить до того, що моделі працюють добре на навчальних даних, але надають неправильні прогнози на нових даних.

Недостатнє представлення: обмежений набір даних може не охоплювати всі можливі варіанти функцій і не забезпечувати належного розподілу класів, що може призвести до неправильної моделі. Такі моделі можуть бути неефективними або дають упереджені прогнози.

Дисбаланс класів: у випадках, коли набір даних має невеликий розмір і розподіл класів незбалансований, процес навчання моделі може бути спрямований на більшість класів. Це може призвести до поганої ефективності прогнозування для меншості класів, оскільки модель недостатньо орієнтується

на їхні характеристики та шаблони.

Варіативність і невизначеність: невеликий набір даних може не включати достатньо різноманітних зразків, що призводить до високої варіабельності та невизначеності в прогнозах моделі, і як наслідок - отримання ненадійних результатів, коли модель тестується на конкретному тестовому наборі даних.

Для вирішення проблеми недостатньої кількості даних запропоновано використати аугментацію.

3.2. Розробка методу класифікації персоналізованих даних

3.2.1. Аугментація даних різних модальностей

Аугментація даних - це методика, що використовується в машинному навчанні, особливо в контексті глибокого навчання для розширення та збагачення навчального набору даних шляхом створення нових їх екземплярів за допомогою різноманітних перетворень існуючих даних. Ці перетворення зберігають вихідні мітки класів, водночас вносячи варіації та різноманітність, які можуть відображати сценарії реального світу. Метою аугментації є збільшення різноманітності навчального набору для покращення здатності моделі до узагальнення [60, 61, 62].

Аугментація виконує кілька завдань, які вирішують проблему нестачі даних,застосовуючи наступні підходи:

Збільшення розміру набору даних. Застосовуючи різні перетворення до оригінальних екземплярів у навчальному наборі даних, аугментація створює більший набір різноманітних зразків. Це знижує ризик перенавчання та дає змогу моделям вивчати більше відмінних ознак, покращуючи узагальнення на невидимих даних.

Підвищення різноманітності даних. Різноманітні варіації, введені завдяки аугментації, розширюють простір ознак і дають повніше представлення розподілу даних. Навчання на більш різноманітному наборі даних робить моделі надійнішими та адаптованими до реальних ситуацій, забезпечуючи вищу точність прогнозів.

Вирішення проблеми дисбалансу класів. Аугментація даних може допомогти усунути дисбаланс класів, створюючи більше екземплярів для недостатньо представлених класів. Це вирівнює розподіл класів і сприяє кращому вивченню особливостей, пов'язаних з класом меншості, що покращує продуктивність передбачення в усіх класах.

Незмінність до перетворень. Аугментація даних за допомогою таких операцій, як обертання, масштабування та перевертання, дає змогу моделі навчитися бути стійкою до цих перетворень і зберігати точність прогнозування, незважаючи на наявність таких змін у екземплярах даних реального світу.

Неявна регуляризація. Аугментація також слугує формою неявної регуляризації, змушуючи моделі вивчати надійні та незмінні ознаки, а не запам'ятовувати навчальні дані. Це допомагає зменшити перенавчання та покращити узагальнення моделі на нових екземплярах даних.

Залежно від модальності даних використовуються різні техніки аугментації. Нехай $P = (p_1, p_2, \ldots, p_n)$, - набір оригінальних даних, де p_i - окремий зразок даних.

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

$$t_j: P \rightarrow P. \qquad (3.1)$$

Перетворення даних.

Основні типи перетворень, які використовуються для аугментації даних:

Геометричні перетворення:

Обертання (Rotation): $t_{rot}(x)=R_\theta x$,

де R_θ - матриця обертання на кут θ.

Зсув (Translation): $t_{trans}(x)=x+\Delta x$,

де Δx - вектор зсуву.

Масштабування (Scaling): $t_{scale}(x)=S_\alpha x$,

де S_α - матриця масштабування з коефіцієнтом α.

Інтенсивні перетворення:

Зміна яскравості (Brightness Adjustment): $t_{\text{bright}}(x)=x+\beta$,

де β - параметр зміни яскравості.

Зміна контрасту (Contrast Adjustment): $t_{\text{contrast}}(x)=\gamma x$,

де γ - параметр зміни контрасту.

Просторові перетворення:

Віддзеркалення (Flipping): $t_{\text{flip}}(x)=Fx$

де F - оператор віддзеркалення.

Перспективні зміни (Perspective Transform): $t_{\text{persp}}(x)=Px$

де P - матриця перспективного перетворення.

Для кожного зразка даних x_i в навчальному наборі даних X, обирається одна або декілька трансформацій з набору T, і до зразка застосовуються ці трансформації, створюючи нові зразки даних.

3.2.2. Особливості аугментації персоналізованих різнотипних даних

Більшість сучасних досліджень у галузі машинного навчання все ще фокусуються на розв'язанні фіксованих завдань. Проте в реальному світі моделі машинного навчання можуть стикатися з непередбаченими змінами у розподілі даних при їх впровадженні. Це викликає питання про те, як адаптивно перейти від створення моделі до підтримки її роботи [65, 67].

Розглянемо дві модальності персоналізованих даних: текст та зображення. У випадку зображень для їх класифікації та сегментації важливим є вміння адаптувати моделі до внесення навіть невеликих змін у візуальні дані. Щодо текстових даних, удосконалення процесу їх заповнення в історії хвороби стає значущою складовою.

1. Аугментація текстових даних

Для проведення досліджень щодо впливу аугментації на текстові дані обрано набір даних, який містить відгуки покупців товарів на Amazon та відображає їхні настрої. Загальна кількість рядків у вибірці становила 17340.

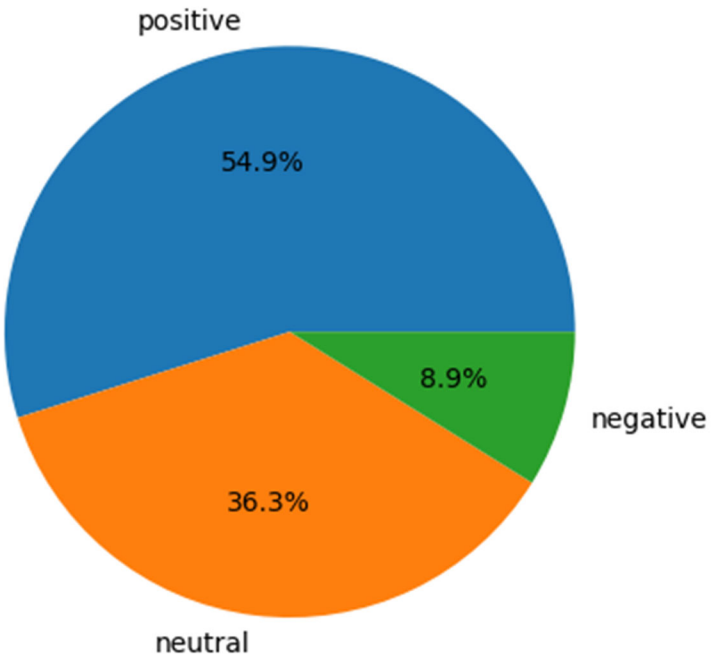


Рис. 3.1 Баланс класів у наборі даних відгуків покупців

Діаграма демонструє, що більшість відгуків у наборі даних є позитивними або нейтральними, а відсоток недійсних даних становить лише невелику частину. Такий розподіл може негативно позначитися на процесі навчання та остаточних передбаченнях моделі. Тому для цієї вибірки доцільно використовувати аугментацію [69, 70].

Кожен метод аугментації застосовувався двічі до кожного речення, використовуючи бібліотеку textaugment та її клас EasyDataAugmentation. Ця бібліотека надає різноманітні функції для маніпулювання реченнями: заміна синонімів (нові речення містять синоніми, схожі за значенням на вихідні слова), випадкове вставлення (нові слова додаються на випадкових позиціях у реченні), випадкові перестановки (слова у реченні змінюють свої позиції за випадковим порядком).

2. Аугментація зображень

Для аугментації зображень використано набір даних рентгенівських знімків легень з пневмонією та без неї. Він складався з 5863 рентгенівських зображень у форматі JPEG та розділений на дві категорії: з пневмонією та без неї. Зображення грудної клітки відібрані з медичного центру та представляли

собою рентгенограми, проведені в рамках звичайної клінічної допомоги пацієнтам [71, 72].

Перед початком аналізу усі рентгенівські знімки грудної клітки пройшли процедуру контролю якості. Цей процес включав видалення всіх знімків низької якості або нечитабельних, щоб забезпечити високу якість даних для подальшого аналізу.

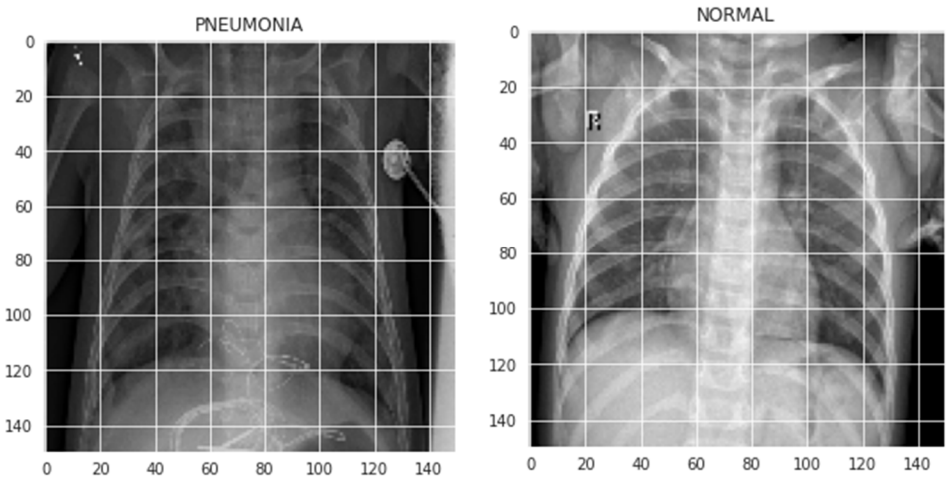


Рис. 3.2 Приклад зображень набору даних

Розподіл класів у наборі даних рентгенівських знімків грудної клітки зображено на рис. 3.3.

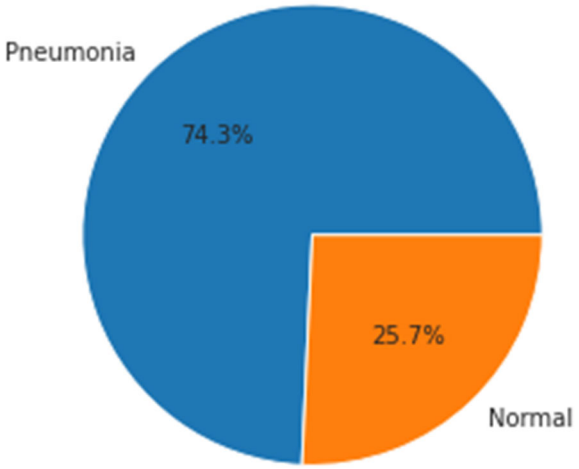


Рис. 3.3 Відповідність даних зображень набору даних до класів

Як видно з рис. 3.3, клас зображень із пневмонією переважає здорові легені. Для забезпечення балансу класів використовуються різні техніки

аугментації зображень. Однією з таких технік є використання ImageDataGenerator з бібліотеки keras. Цей інструмент працює шляхом застосування різноманітних перетворень до зображень у наборі даних, таких як масштабування, обертання, зміна приближення та перевертання.

ImageDataGenerator автоматично доповнює дані під час кожної епохи навчання, надаючи нейронній мережі різноманітні пакети даних. Це дає змогу моделі бачити різні варіації того ж зображення впродовж кожної епохи, що допомагає уникнути перенавчання та забезпечує кращу здатність моделі до узагальнення.

Після аугментації за допомогою різних комбінацій параметрів виявлено наступне: Після застосування комбінації поворотів зображень, вони можуть бути випадково перевернуті горизонтально або вертикально, а також - на певний кут (рис.3.5). Ця техніка допомагає розширити різноманітність навчального набору даних, що дає змогу моделі навчитися різноманітним шаблонам та покращує її здатність до узагальнення. Завдяки цьому підходу модель стає стійкішою до змін в орієнтації та положенні об'єктів на зображеннях, покращуючи її ефективність під час роботи з реальними даними.

```
model_rotation, history_rotation = run_test(x_train, y_train, x_val, y_val,
                                            rotation_range = 30,
                                            horizontal_flip = True,
                                            vertical_flip=True)
```

Рис. 3.4 Комбінація параметрів аугментації поворотами

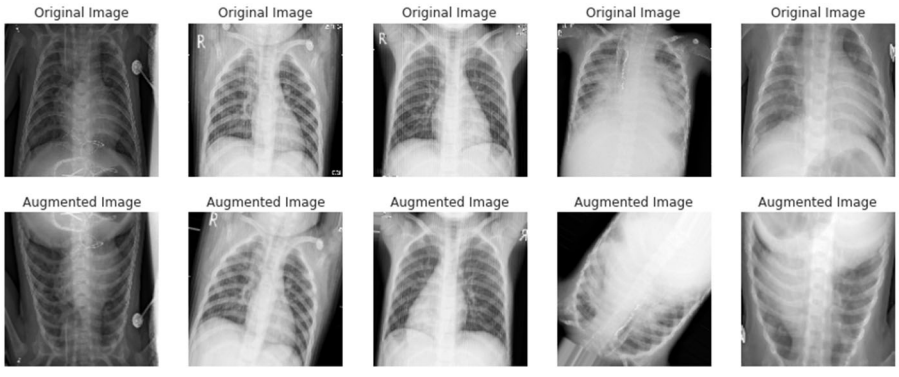


Рис. 3.5 Зображення аугментовані поворотами

Після аугментації за допомогою комбінації приближення та зсуву висоти або ширини зображень отримано наступне: у цьому випадку зображення випадково можуть бути приближені або віддалені. Окрім того, вони можуть бути довільно зміщені по ширині чи довжині від загальної частки. Це дає змогу моделі навчитися розпізнавати об'єкти на зображеннях незалежно від їх масштабу та положення, поліпшуючи загальну роботу моделі і забезпечуючи кращі результати передбачення.

```
model_zoom, history_zoom = run_test(x_train, y_train, x_val, y_val,
                                    zoom_range = 0.2,
                                    width_shift_range=0.1,
                                    height_shift_range=0.1
                                   )
```

Рис. 3.6 Комбінація параметрів аугментації приближенням та зсувами

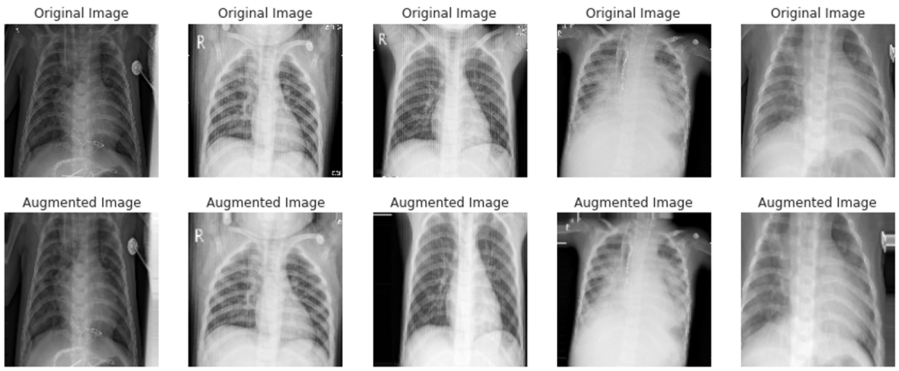


Рис. 3.7 Зображення аугментовані приближенням та зсувами

Після застосування комбінації зміни яскравості зображень, пікселі на зображеннях будуть випадковим чином змінювати свою яскравість. Це створює можливість для моделі навчитися робити передбачення на зображеннях з різними рівнями яскравості та поліпшує її здатність до узагальнення. Такий підхід робить модель стійкішою до змін в освітленні та різниці в яскравості між зображеннями, що дає змогу покращити її продуктивність при роботі з реальними даними.

```
model_brightness, history_brightness = run_test(x_train, y_train, x_val, y_val,
                                                brightness_range=(0.5, 1.2),
                                                zoom_range = 0.2
                                                )
```

Рис. 3.8 Комбінація параметрів аугментації зміною яскравості

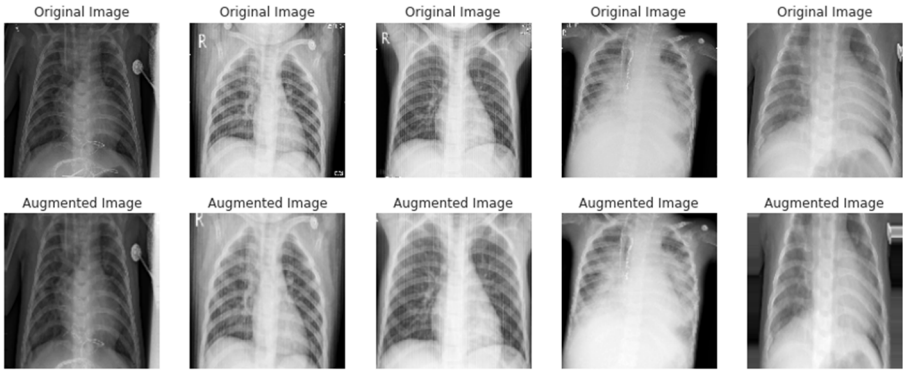


Рис. 3.9 Зображення аугментовані зміною яскравості

3.2.3. Опис методу

Процес класифікації включає наступні кроки:

1. Збір даних. Необхідно отримати набір даних, який містить вектори ознак та їхні відповідні мітки класів. Набором даних у розрізі даного дослідження вважаються 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→P.

Аугментація даних дає змогу розширити обсяг та різноманітність навчального набору шляхом застосування різноманітних перетворень до існуючих 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

$$P \in R_n \times p \tag{3.2}$$

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, опорні векторні машини (SVM), дерева рішень, випадкові ліси, метод k-найближчих сусідів (KNN) і 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. Процес визначення гіперпараметрів детально розглядається у розділі 3, підрозділі 3.3.1.

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 удосконалення функції втрат, побудову границь прийняття рішень або вивчення ваг ознак.

Навчання моделі проводиться з використанням навчального набору згідно виразу:

$$y' = f(P, \theta) \tag{3.3}$$

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

Кількість етапів навчання моделі Q залежить від підготовлених даних, включаючи використання методів їх аугментації. 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 застосовують 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:

$$Q = NumberofEpochs \times \frac{TrainingSet \times K_{transf}}{Batch}, \quad (3.4)$$

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

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.

6. Оцінка моделі полягає у 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 та визначається із залежності [73-75]:

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

де: TP - позитивні вірні результати, TN - позитивні негативні результати, FP - негативні вірні результати, FN - негативні хибні результати

Чутливість або повнота оцінює здатність моделі до правильної ідентифікації позитивних прикладів та визначається із залежності:

$$Recall = \frac{TP}{TP+FN}.$$

Точність вимірює частку правильних позитивних передбачень серед усіх позитивних передбачень та визначається із залежності:

$$Precision = \frac{TP}{TP+FP}.$$

Для визначення гармонічного середнього між точністю та повнотою, яке враховує як хибнопозитивні (False Positives), так і хибнонегативні результати (False Negatives) використано показник F1-score, який визначається із залежності:

$$F1Score = \frac{2*Precision*Recall}{Precision+Recall}.$$

3.3. Розробка методу персоналізації даних особи

3.3.1. Процедури підбору оптимальних параметрів роботи моделей

У сфері машинного навчання велике значення мають гіперпараметри, які визначаються перед початком процесу навчання моделей. Це важливі параметри, які не автоматично оптимізуються під час навчання на вхідних даних, але визначають умови та налаштування процесу навчання і структуру моделі. Вони впливають на те, як саме модель буде навчатися, і як вона буде адаптуватися до навчальних даних [63, 79].

Вибір правильних гіперпараметрів вимагає проведення ретельних експериментів та аналізів, оскільки це завдання досить складне. Оптимальне налаштування гіперпараметрів сприяє створенню моделі, яка ефективно вирішує реальні завдання та надає точні прогнози.

Існує декілька методів пошуку оптимальних значень гіперпараметрів у моделей машинного навчання, що сприяють підвищенню продуктивності та точності моделей. Серед таких методів можна виокремити Grid Search, Random Search та Bayesian Optimization.

Вибір конкретного методу пошуку гіперпараметрів для задачі класифікації залежить від кількох факторів, включаючи обсяг даних, доступні обчислювальні ресурси та час, необхідний для налаштування моделі. При невеликому обсягу

даних, у нашому випадку це 4981 рядків і 11 стовпців, ефективним рішенням було використання методу Grid Search.

Під час тренування кожної моделі використано тренувальний набір даних, а оцінка їх ефективності провена на валідаційному наборі. Для пошуку оптимальних значень гіперпараметрів використано метод Grid Search. Він отримує на вхід список параметрів, які потрібно оцінити, та їх можливі значення. Для кожної комбінації параметрів Grid Search виконує крос-валідацію і розраховує середню точність на тренувальному наборі. Оптимальними параметрами вважаються ті, які дають найвищу середню точність на тренувальному наборі. На завершальному етапі ці параметри використано для побудови остаточного класифікатора на всьому наборі даних [78-80].

Множина параметрів для класифікатора, які оцінюються за допомогою пошуку по сітці, представлена у таблиці 3.1. Серед гіперпараметрів, що розглядалися, включають criterion (критерій розбиття вузла) та max_depth (максимальна глибина дерева).

Таблиця 3.1

Сукупність гіперпараметрів для класифікатора Decision Tree

Параметр моделі	Множина значень
criterion	['gini', 'entropy']
max_depth	[2, 4, 6, 8, 10, 12]

Найбільш оптимальною комбінацією параметрів виявились *criterion='entropy'* та *max_depth=6*, яка забезпечує найкращі результати на вибірці. Ці параметри використано для подальшої класифікації за допомогою дерева рішень.

Після класифікації даних методом Decision Tree на валідаційній вибірці отримана точність 0,8713.

Дослідження методу Random Forest включало аналіз класифікатора, для якого множина параметрів, що підлягала перевірці за допомогою пошуку по сітці, відображена у таблиці 3.2. Розглянуті гіперпараметри включають

n_estimators (кількість дерев рішень) та *max_depth* (максимальна глибина дерева).

Таблиця 3.2

Сукупність гіперпараметрів для класифікатора Random Forest

Параметр моделі	Множина значень
n_estimators	[100, 200, 300, 400]
max_depth	[5, 10, 15, None]

Застосування методики Grid Search дало можливість отримати найкращий результат з параметрами *max_depth=None, n_estimators=300*. Отримані параметри використано для подальшої класифікації за допомогою випадкового лісу.

Після класифікації даних методом Random Forest на валідаційній вибірці отримано високу точність 0,9393.

Для класифікатора Multi-Layer Perceptron проведено аналіз множини параметрів за допомогою пошуку по сітці. У таблиці 3.3 відображена ця множина параметрів, що включає *hidden_layer_sizes* (кількість нейронів у кожному прихованому шарі), *activation* (функція активації) та *solver* (алгоритм оптимізації).

Таблиця 3.3

Сукупність гіперпараметрів для класифікатора MLP

Параметр моделі	Множина значень
hidden_layer_sizes	[50,50,50), (50,100,50), (100,)]
activation	['tanh', 'relu']
solver	['sgd', 'adam']

Використання методу Grid Search дозволило знайти оптимальні параметри для багатошарового персептрону, які включають *activation='tanh', hidden_layer_sizes=(50, 50, 50) і solver='sgd'*. Отримані параметри використано для подальшої класифікації даних.

Після класифікації даних методом Multi-Layer Perceptron на валідаційній вибірці отримано високу точність 0,9466.

Для класифікатора SVM проведено аналіз множини параметрів за допомогою пошуку по сітці. У таблиці 3.4 наведена ця множина параметрів, що включає C (параметр регуляризації), *gamma* (параметр для нелінійних гіперлощин) і *kernel* (тип ядра).

Таблиця 3.4

Сукупність гіперпараметрів для класифікатора SVM

Параметр моделі	Множина значень
C	[1, 10, 100]
gamma	[0.1, 0.01, 0.001]
kernel	['rbf']

Використання методу Grid Search дало найкращі результати з параметрами *C=1, gamma=0.001 і kernel='rbf'.*

Після класифікації даних методом опорних векторів зі знайденими параметрами на валідаційній вибірці отримано точність 0,9368, що підтверджує ефективність запропонованого методу.

3.3.2. Аналіз методів машинного навчання для класифікації зображень

Метою класифікації медичних зображень є автоматичне визначення їх характеристик та особливостей для подальшої класифікації на основі цих ознак. Перед створенням ансамблю моделей необхідно ретельно дослідити ефективність кожного методу класифікації окремо. Оцінка ефективності класифікаторів є особливо важливою, особливо коли класифікація включає п’ять класів. Для цього використано наступні метрики [90, 91]:

1. Accuracy (Точність) – відношення кількості правильних прогнозів до загальної кількості прогнозів, що дає змогу оцінити загальну ефективність класифікатора.

2. Precision (Точність) – відношення кількості правильних прогнозів

певного класу до загальної кількості прогнозів, що були визначені як цей клас. Ця метрика дає змогу оцінити точність класифікатора щодо визначення певного класу.

3. Recall (Чутливість) – відношення кількості правильних прогнозів певного класу до загальної кількості прикладів цього класу у даних, що дає змогу оцінити, наскільки класифікатор може визначати певний клас.

4. F1-score – гармонічне середнє між точністю та чутливістю, що враховує як точність, так і чутливість. Ця метрика дає змогу оцінити ефективність класифікатора з урахуванням нерівномірного розподілу класів.

5. Confusion matrix (Матриця помилок) - таблиця, що відображає кількість правильних та неправильних прогнозів класифікатора для кожного класу. Вона дає змогу оцінити ефективність класифікатора та зрозуміти, на яких класах він робить помилки.

Для узагальнення та порівняльного аналізу результати всіх чотирьох моделей наведено у таблиці 3.5.

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.

Таблиця 3.5

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				

3.3.3. Ансамблеве навчання: поняття та методи створення ансамблю

Ансамблеве навчання — це метод машинного навчання, який використовує кілька моделей для отримання кращих результатів передбачення.

У цьому методі кілька моделей навчаються на одній і тій же навчальній вибірці, а результати їх передбачень комбінуються для отримання кінцевого результату. Цей підхід може забезпечити більшу точність передбачення, порівняно з однією моделлю, якщо всі моделі налаштовані належним чином. Існують різні типи методів ансамблевого навчання, які відрізняються за типом моделей, вибіркою даних і функцією прийняття рішень. Деякі з найпоширеніших технік ансамблевого навчання включають Stacking, Blending та Voting [28, 29].

Stacking - це метод, спрямований на зменшення помилки узагальнення різних моделей машинного навчання. Загальна ідея полягає в створенні мета-моделі, яка складається з передбачень набору базових моделей машинного навчання, так званих "слабких учнів". Процес узагальнення стека зазвичай

включає два етапи. Під час першого етапу генерується навчальний набір даних для мета-моделі шляхом використання кратної перехресної перевірки для кожного "слабкого учня". Прогнози кожної з моделей об'єднуються, щоб створити новий навчальний набір. На другому етапі мета-модель навчається за допомогою додаткової моделі, яка відома як "останній оцінювач" або "останній учень". Цей підхід дає змогу комбінувати різні моделі для досягнення більшої точності, але при цьому вимагає значно більше обчислювальних ресурсів та часу на тренування.

Blending – це техніка, що виникла на основі Stacking Generalization, з однією відмінністю: у Blending використовується відмінності к-кратної перехресної перевірки для створення навчальних даних мета-моделі. Отже, у випадку Blending мета-модель навчається на прогнозах, отриманих на непрочитаному наборі даних, тоді як у Stacking мета-модель навчається на передбаченнях, зроблених під час к-кратної перехресної перевірки. Тренувальний набір даних розділяється на дві частини – для навчання та валідації базових моделей. Результати прогнозування моделей на валідаційних даних, які ще не використовувалися моделями, використовуються для навчання мета-моделі [88, 90, 91].

VotingClassifier – це алгоритм голосування, який об'єднує результати різних класифікаторів в один ансамбль з метою досягнення кращих результатів.

У VotingClassifier існують два види голосування: hard voting і soft voting. У hard voting класифікатори голосують за більшість класів, де кожен класифікатор вибирає свій прогнозований клас, а потім обирається клас з найбільшою кількістю голосів. Цей підхід ефективний у випадках, коли всі класифікатори мають високу точність. У soft voting голосування базується на ймовірностях, які надають класифікатори для кожного класу, а потім вибирається клас з найвищою середньою ймовірністю. Цей підхід дає змогу враховувати важливість деяких класифікаторів, які можуть мати кращі результати для певних класів. Зазвичай soft voting працює краще, коли класифікатори мають різну точність [99, 101].

Однією з переваг VotingClassifier є можливість поєднувати різні класифікатори з різними параметрами, щоб отримати більш точний результат. Він також може допомогти запобігти перенавчанню, оскільки може відкидати "шумові" прогнози від окремих класифікаторів. Проте недоліком VotingClassifier є те, що він працює повільніше, порівняно з окремими класифікаторами, оскільки використовує всі класифікатори для кожного прогнозування. Крім того, якщо всі класифікатори показують погані результати, VotingClassifier також може демонструвати незадовільні результати.

У цьому дослідженні використаний метод Voting, оскільки він не потребує складних обчислень. Крім того, Voting може бути швидким методом, оскільки він не вимагає додаткового навчання моделі ансамблю, як Stacking або Blending. Ще одна перевага Voting полягає в його гнучкості, оскільки можна експериментувати з різними комбінаціями базових моделей, а також з їх алгоритмами, гіперпараметрами та налаштуваннями. Це дає змогу знаходити оптимальне поєднання для конкретної задачі [100, 102].

У складі ансамблю використано методи SVM, Random Forest та MLP. Кожен з цих методів має свої переваги та недоліки, і їх поєднання допомагає зменшити вплив недоліків та збільшити точність класифікації. На рис. 3.9 показано діаграму утворення ансамблю з обраних моделей за допомогою VotingClassifier.

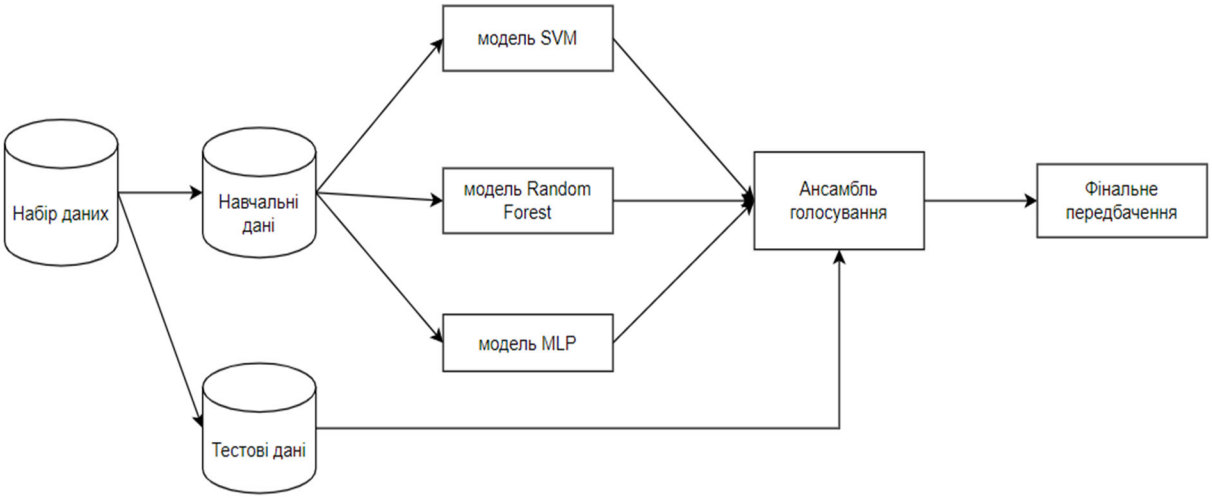


Рис. 3.9 Графічне зображення процесу формування ансамблю класифікаційних моделей

Ансамбль навчається на навчальному наборі даних, а остаточне тестування здійснюється на тестовому наборі даних. Для визначення оптимального типу голосування проводять дослідження з двома типами: hard voting та soft voting. Вибір ансамблю здійснюється відповідно до результатів цих досліджень з урахуванням обраних типів голосування.

3.3.4. Опис методу

Запропонований метод включає наступні кроки:

1. Збір, підготовка, вибір ознак персональних даних, як описано в підрозділі 3.2.3.

2. На етапі вибору моделі у підрозділі 3.3.1. продемонстровано процес підбору гіперпараметрів, опрацювання моделей Decision Tree, Random Forest, SVM та MLP та у підрозділі 3.3.2. проаналізовано 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

$h_1(p), h_2(p), \ldots, h_t(p)$, де t - кількість дерев у лісі.

2.1.1. Побудова дерева

Для кожного дерева b:

- Вибрано випадковий набір ознак P_t з загального простору ознак y.
- Навчаємо дерево $h_t(p)$ на випадковій підмножині навчальних даних P_t.

Класифікація:

Результат класифікації отримується шляхом голосування дерев:

$$y' = mode\big(h_1(p), h_2(p), \ldots, h_t(p)\big) \quad (3.5)$$

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

2.2.1. Оптимізаційна задача SVM:

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.

$$min_{w,t} = \frac{1}{2}\|w\|^2$$

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

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: a^l - активація нейронів шару l, W^l- матриця ваг 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:

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

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

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:

$$y' = mode(y'RF, y'SVM, y'MLP) \quad (3.6)$$

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

3.2 Soft Voting:

$$y' = argargmax_y \quad (3.7)$$

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.

3.4. Висновки до 3 розділу

У цьому розділі розроблено метод класифікації персоналізованих даних шляхом впровадження етапу аугментації, що розширює обсяг та різноманіття навчальних даних і сприяє кращому узагальненню моделей та зменшенню ризику перенавчання.

Проаналізовано застосування процесу аугментації для даних різних модальностей, що дало змогу виявити закономірність, що якщо аугментовані дані не були значно спотвореними, це призводило до некоректного прогнозування класу.

Тому важливим є раціональний вибір параметрів, які контролюють ступінь випадкових змін *y* даних в залежнсті від модальності даних. Невдале використання методів може призвести до суттєвого погіршення якості даних і, відповідно, результатів класифікації.

Удосконалено метод персоналізації даних шляхом впровадження ансамблю моделей класифікації та ансамблевого голосування, що забезпечує підвищення точності прогнозування стану.

Досліджені два ансамблі з різними типами голосування – hard voting і soft voting. Обрано класифікатор soft voting з найбільшою ефективністю та точністю на рівні 0,96. Отримані результати свідчать про високу ефективність моделі класифікації з високими показниками precision і recall для більшості класів, що підтверджує її здатність розпізнавати різні стадії хвороби.

РОЗДІЛ 4.
РОЗРОБЛЕННЯ АРХІТЕКТУРИ ІНФОРМАЦІЙНОЇ СИСТЕМИ ТА АПРОБАЦІЯ РЕЗУЛЬТАТІВ

У цьому 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 здійснений пошук найкращих гіперпараметрів. Проведено порівняльний аналіз застосування існуючих моделей класифікації, проаналізовано наявні методи побудови ансамблевого навчання та запропоновано використання VotingClassifier для об'єднання декількох класифікаторів в одну модель. У рамках цього дослідження проведено порівняльний аналіз розглянутих у третьому розділі методик та розробленого ансамблю моделей. Подано результати імплементації інформаційної системи для супроводу процесу збору та аналізу стану особи.

Матеріали розділу опубліковані у роботах автора [128-130, 134, 135, 138].

4.1. Побудова архітектури інформаційної системи

У рамках даного дослідження розроблено інформаційну систему, спрямовану на обробку персоналізованих даних особи, яка проходить медичне обстеження та діагностику свого стану. Обрано тип інформаційної системи - вебзастосунок, що забезпечує взаємодію з користувачами через API-сервіси [88] та зберігає необхідні дані у базі даних.

Запропонована інформаційна система опрацювання персоналізованих даних призначена для збору, зберігання, аналізу та візуалізації даних про стан здоров'я пацієнта з метою підтримки прийняття медичних рішень. Основні завдання та призначення системи включають:

Моніторинг стану здоров'я в режимі реального часу:

Постійний збір даних з давачів дає змогу відстежувати життєво важливі показники, такі як частота серцевих скорочень, рівень кисню в крові,

артеріальний тиск, рівень глюкози тощо. При виявленні критичних значень відбувається швидке реагування на зміни в стані здоров'я пацієнта завдяки оповіщенням та сигналам тривоги.

Підвищення точності діагностики:

Використовуються зібрані дані для точнішого встановлення діагнозу на основі індивідуальних показників пацієнта. Здійснюється аналіз трендів та патернів у даних для виявлення прихованих проблем зі здоров'ям.

Персоналізований підхід до лікування:

Інтеграція з електронними медичними записами (ЕМЗ) дає змогу створити повний профіль пацієнта, враховуючи історію хвороб, прийняті ліки, алергії, тощо. Розроблення індивідуальних планів лікування та рекомендацій на основі зібраних даних.

Підтримка прийняття медичних рішень:

Використання алгоритмів машинного навчання та штучного інтелекту для прогнозування можливих ризиків та ускладнень. Надання медичному персоналу аналітичних звітів та візуалізацій для покращення процесу прийняття рішень.

Покращення якості медичних послуг:

Зменшення кількості помилок у діагностиці та лікуванні завдяки точним і актуальним даним. Підвищення ефективності лікування завдяки своєчасному виявленню проблем та коригування терапії.

Забезпечення безпеки та конфіденційності даних:

Використання шифрування та інших засобів захисту для забезпечення конфіденційності медичних даних пацієнтів. Контроль доступу до даних лише для авторизованих користувачів.

Інтеграція з іншими медичними системами:

Використання API для обміну даними з іншими інформаційними системами, такими як лабораторні інформаційні системи, системи управління клінічними даними тощо. Забезпечення безперервності та узгодженості медичних даних у різних системах.

Ці завдання дають змогу забезпечити комплексний підхід до моніторингу, діагностики та лікування пацієнтів, підвищуючи ефективність медичних послуг та якість життя пацієнтів.

Запропоновано архітектуру інформаційної системи підтримки прийняття медичних рішень щодо аналізу стану особи на підставі опрацювання персоналізованих медичних даних, яка представлена на рис. 4.1.

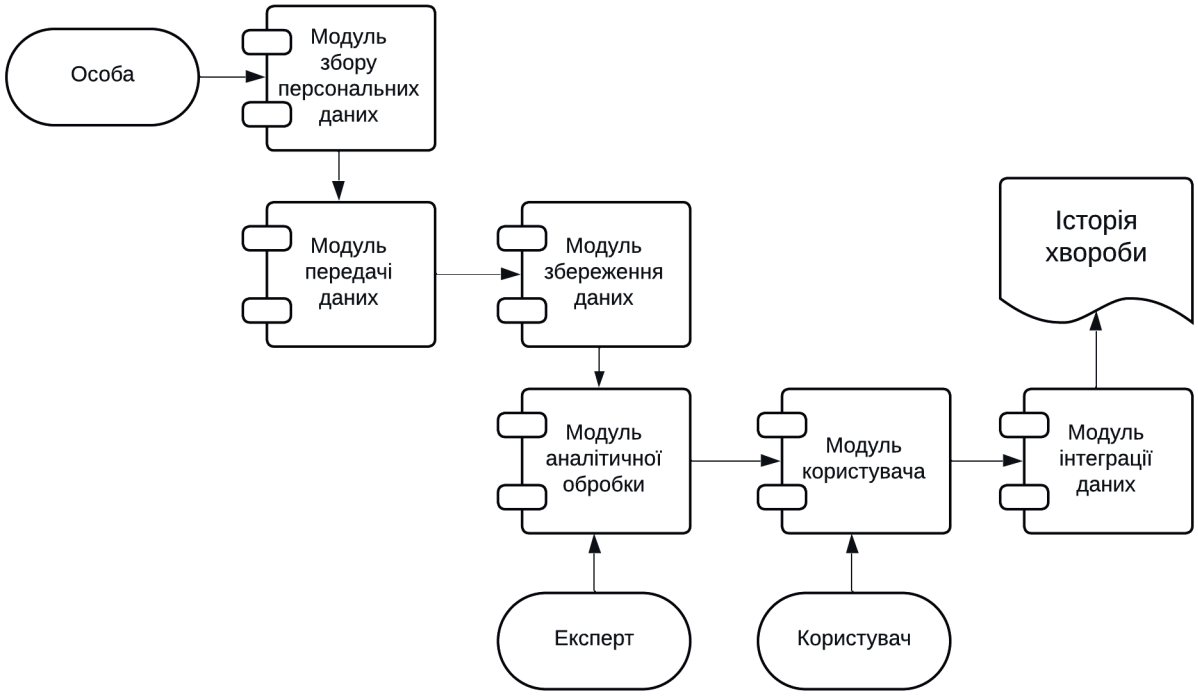


Рис. 4.1 Загальна архітектура інформаційної системи опрацювання персоналізованої інформації аналізу стану особи

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 для 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.

4.2. Проєктування прототипу інформаційної системи опрацювання персоналізованих даних особи

Система складається з двох модулів. Перший модуль – це мобільний додаток, де користувач може зареєструватися як "Куратор". " Куратор " може створювати профілі "Пацієнтів". Другий модуль – це додаток для розумного годинника. Користувач годинника за замовчуванням вважається "Пацієнтом". Він може підключитися до " Куратора" і передавати йому дані про свій стан здоров'я та статистику.

Метою системи є збір інформації про особу/пацієнта, який перебуває під супроводом особи, що наглядає, та контролює процес збору і збереження даних. Система реалізована у вигляді мобільного додатку і вирішує зазначені при аналізі проблеми:

- Збір та аналіз персоналізованої інформації для куратора;
- Збір та аналіз персоналізованої інформації про пацієнта, його стан здоров’я.

Для кожної з них існує одна чи кілька сутностей в базі даних, де буде зберігатись, вноситись та отримуватись вся необхідна інформація. Ця інформація

відображається під час використання користувачем будь-якого з розроблених застосунків, які є складовою частиною комплексної інтегрованої системи.

Отже, система містить різні типи вихідних даних, що повинні задовольняти потреби відповідних користувачів. Ці потреби та відповідні вихідні дані наведено у таблиці 4.1.

Таблиця 4.1

Функціонал інформаційної системи

Потреба	Вихідні дані
Куратор може додавати нових пацієнтів.	Форма створення профілю пацієнта
Куратор може переглядати, редагувати та видаляти профіль пацієнта.	Профіль пацієнта
Куратор може переглядати дані про стан здоров’я та статистику пацієнта.	Дані про стан здоров’я та статистика пацієнта
Куратор може додавати нові дані для ведення статистики.	Форма для створення нового типу статистичних даних
Куратор може додавати нові записи про стан пацієнта.	Форма для додавання нових даних у статистику
Куратор може обирати пацієнта, дані про якого він хоче переглянути.	Вибір пацієнта
Пацієнт може підключатись до свого куратора.	Підключення до куратора
Пацієнт може переглядати дані про свого куратора.	Форма профілю куратора

Для проєктування системи визначено наступний набір ключових даних для супроводу оцінки стану особи:

Куратор (User) - користувач мобільного застосунку в системі:

- uid: унікальний ідентифікатор

- name: повне ім’я користувача

- email: електронна пошта, вказана при реєстрації

- patients: список пацієнтів, за якими слідкує куратор

Пацієнт (Patient) - користувач застосунку для смарт-годинника:

- id: унікальний ідентифікатор

- name: ім’я

- surname: прізвище

- age: вік

- phone: контактний номер телефону

- description: додаткова інформація про пацієнта або нотатки

- heartRate: пульс пацієнта

- stepCount: кількість пройдених кроків

- data: статистичні дані пацієнта, представлені у форматі словника, де зберігаються різноманітні дані у форматі [String: Any]

Функціонал запропонованої інформаційної системи представлено на рис. 4.2.

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:

Для «Куратора»:

- Екран входу у мобільний додаток (Mobile Sign in Screen) - включає функції входу в систему та реєстрації.

- Головний екран мобільного додатка (Mobile Home Screen) - містить основні функції для куратора, такі як список пацієнтів, індикатори показників та статистика.

- Список пацієнтів (List of patients) - дає змогу переглядати та додавати нових пацієнтів.

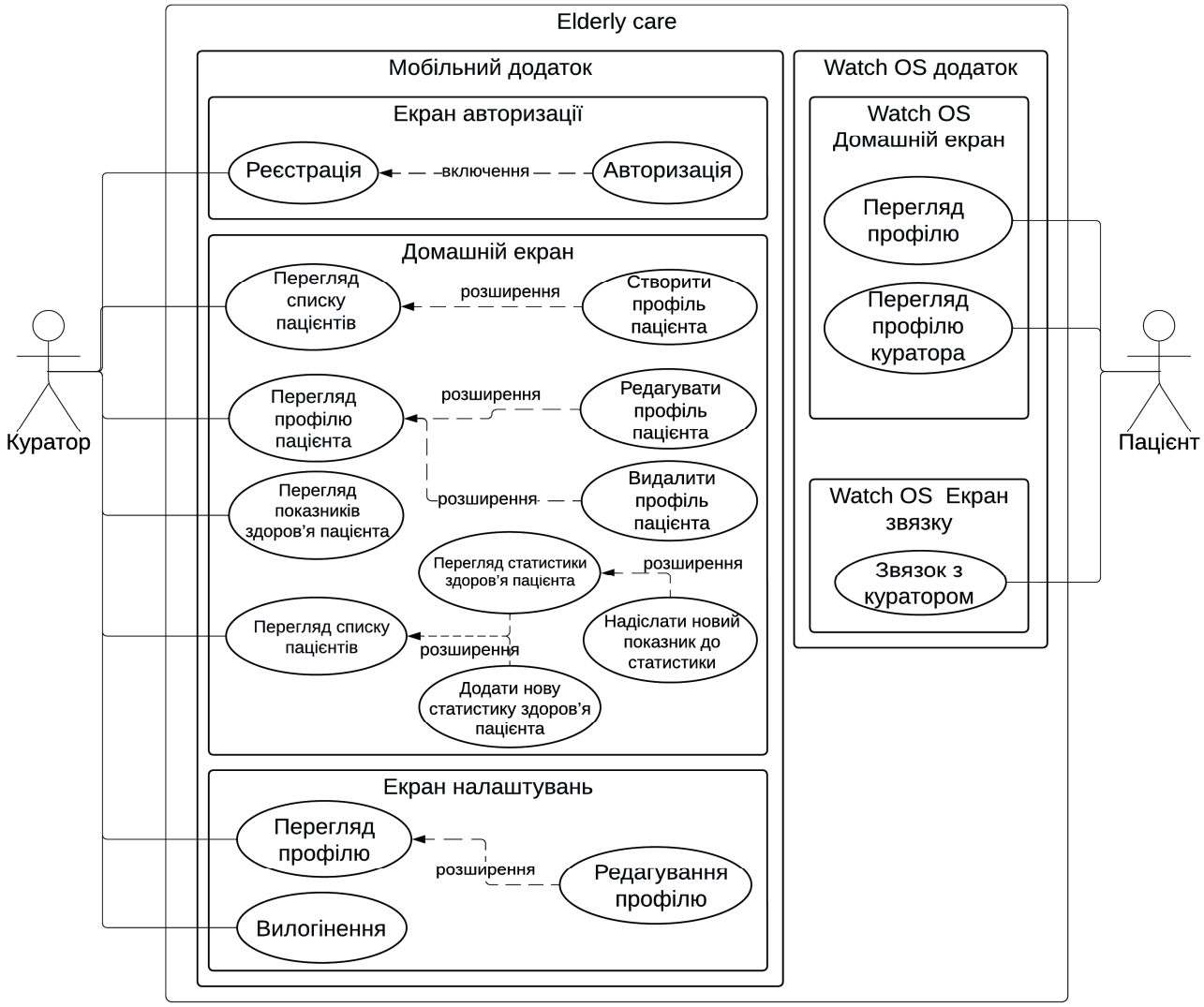


Рис. 4.2 Розширена UML-діаграма використання інформаційної технології

- Профіль пацієнта (Patient profile) - включає функції перегляду, редагування та видалення профілю пацієнта.

- Показники стану здоров'я пацієнта (Patient health indicators) - дає змогу переглядати поточні показники пацієнта.

- Статистика стану здоров'я пацієнта (Patient health statistics) - містить функції перегляду, додавання нових показників та створення нових графіків з новими даними.

- Екран налаштувань мобільного додатка (Mobile Settings Screen) - містить функції перегляду профілю та виходу з облікового запису.

Для «Пацієнта»:

- Головний екран застосунку для смарт-годинника (WatchOS Home Screen) - містить функції перегляду профілю пацієнта та куратора.

- Екран підключення до куратора (WatchOS Connecting Screen) - функція "підв'язки" пацієнта до куратора..

4.3. Оцінювання ефективності роботи запропонованих рішень щодо аналізу та опрацювання персоналізованих даних особи

Для створення програмного модулю з метою класифікації стадій хвороби Альцгеймера з використанням ансамблю моделей машинного навчання обрана мова програмування Python. Python вважається однією з найпопулярніших мов програмування для роботи з моделями машинного навчання, і це зумовлено кількома перевагами:

− Простота використання, що дає змогу швидко розробляти та розуміти код.

− Широкий вибір бібліотек і фреймворків для розробки моделей машинного навчання, що забезпечує гнучкість та можливість вибору оптимального інструменту для конкретного завдання.

− Різноманітні бібліотеки візуалізації даних, які дають змогу графічно відображати дані, результати моделей та аналізувати їх поведінку, що сприяє кращому розумінню та інтерпретації результатів.

− Незалежність від платформи, що забезпечує можливість розробляти програмне забезпечення на одній машині та використовувати його на іншій без будь-яких змін або з мінімальними модифікаціями.

− Велика спільнота користувачів, що забезпечує легкий доступ до відповідей на питання, прикладів коду та підтримки в разі виникнення проблем.

Розробка програмного модулю відбувається на безкоштовній інтерактивній платформі Google Colab, наданій компанією Google. Даний програмний модуль забезпечує доступ до безкоштовних обчислювальних ресурсів та підтримує апаратне прискорення, що дає змогу використовувати

графічні процесори. Крім того, Google Colab містить вбудовані популярні бібліотеки для машинного навчання та забезпечує легке та швидке налаштування робочого середовища.

Для розробки програмного модуля використано наступні бібліотеки та їх версії:

- numpy 1.22.4 – бібліотека для роботи з масивами даних;

- opencv-python 4.7.0.72 – бібліотека для комп'ютерного зору та обробки зображень;

- matplotlib 3.7.1 – бібліотека для візуалізації даних;

- seaborn 0.12.2 – бібліотека для візуалізації даних;

- scikit-learn 1.2.2 – бібліотека для машинного навчання, що надає засоби для класифікації та інших типів аналізу даних;

- hashlib – бібліотека для створення хеш-функцій та перевірки цілісності даних;

- scikit-image 0.19.3 – бібліотека для обробки зображень;

- joblib 1.2.0 – бібліотека для ефективного зберігання та завантаження Python-об'єктів, таких як моделі машинного навчання.

Додатково використані наступні пакети розглянутих бібліотек:

- sklearn.model_selection.train_test_split – функція розділення набору даних на тренувальну та тестову вибірки;

- skimage.transform.rotate – функцію обертання зображень;

- sklearn.model_selection.GridSearchCV – клас для пошуку оптимальних гіперпараметрів моделі;

- sklearn.svm.SVC – клас реалізації методу опорних векторів для класифікації;

- sklearn.metrics.accuracy_score – функція обчислення точності класифікаційної моделі;

- sklearn.metrics.confusion_matrix і sklearn.metrics.Confusion MatrixDisplay – пакети обчислення та візуалізації матриці помилок;

- sklearn.tree.DecisionTreeClassifier – клас реалізації моделі дерева рішень для класифікації;

- sklearn.ensemble.RandomForestClassifier – клас реалізації моделі випадкового лісу для класифікації;

- sklearn.neural_network.MLPClassifier – клас реалізації моделі штучної нейронної мережі Багатошаровий персептрон;

- sklearn.ensemble.VotingClassifier – клас об'єднання декількох класифікаторів в одну модель за допомогою голосування.

4.3.1. Дослідження ефективності обраних моделей гіперпараметрів та їхнє порівняння

Для оцінки результатів тренування та тестування моделей проаналізовано роботу базових моделей з гіперпараметрами за замовчуванням на тестових даних [109]. У таблиці 4.2 наведено результати тестування базової моделі Decision Tree Classifier.

Таблиця 4.2.

Результати базової моделі Decision Tree Classifier

class	precision	recall	F1-score	accuracy
0 (не має інсульту)	0.89	0.91	0.81	0.87
1(є інсульт)	0.82	0.80		

Аналізуючи результати таблиці 4.2, зроблено висновок, що модель забезпечила стабільну та збалансовану класифікацію даних, оскільки показники precision та recall майже однакові для обох класів. Це свідчить про те, що модель

ефективно розрізняє ці два класи. Як бачимо, показник F1-score складає 81%, що є прийнятним результатом, враховуючи складність набору даних.

Матриця помилок для запропонованої базової моделі представлена на рис. 4.3.

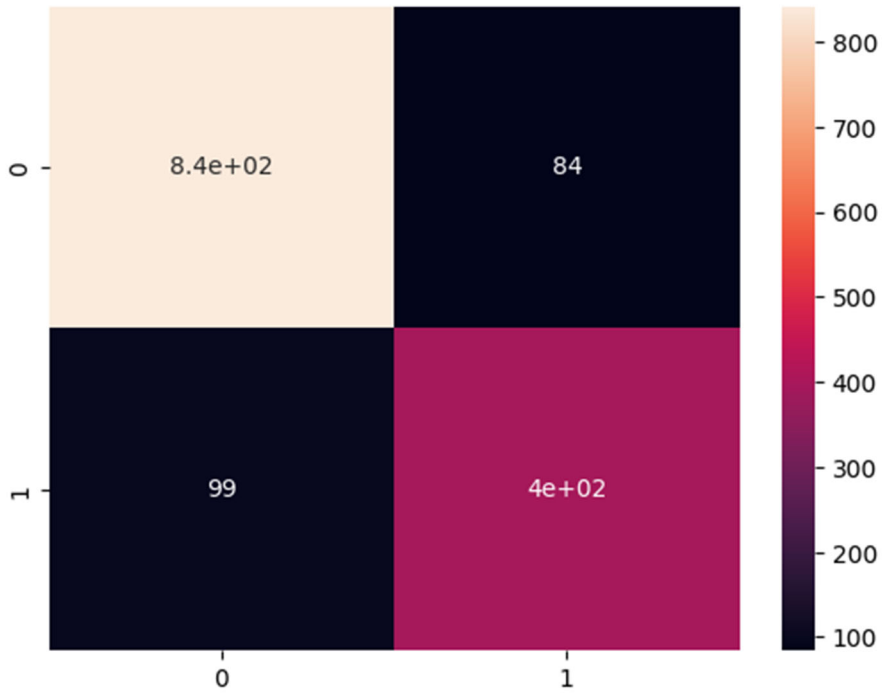


Рис. 4.3 Матриця помилок для базової моделі Decision Tree Classifier

Як видно з матриці помилок для базової моделі Decision Tree Classifier, модель практично однаково класифікувала обидва класи. Також слід додати, що не виявлено виражених проблем з певним класом. Модель не демонструє значного збурення між класами, які вона прогнозує.

Результати тестування базової моделі Random Forest Classifier подано в таблиці 4.3.

Таблиця 4.3

Результати базової моделі Random Forest Classifier

class	precision	recall	F1-score	accuracy
0 (не буде інсульту)	0.93	0.94	0.86	0.9
1 (буде інсульт)	0.87	0.86		

З аналізу цієї таблиці можна зробити такі висновки:

- модель забезпечила стабільну та збалансовану класифікацію даних, що відповідає базовій моделі Decision Tree Classifier;

- показники precision та recall практично однакові для обох класів;

- показник F1-score складає 86%, що є високим значенням і є прийнятним для подальшого вдосконалення.

Матриця помилок для цієї моделі подана на рис. 4.4.

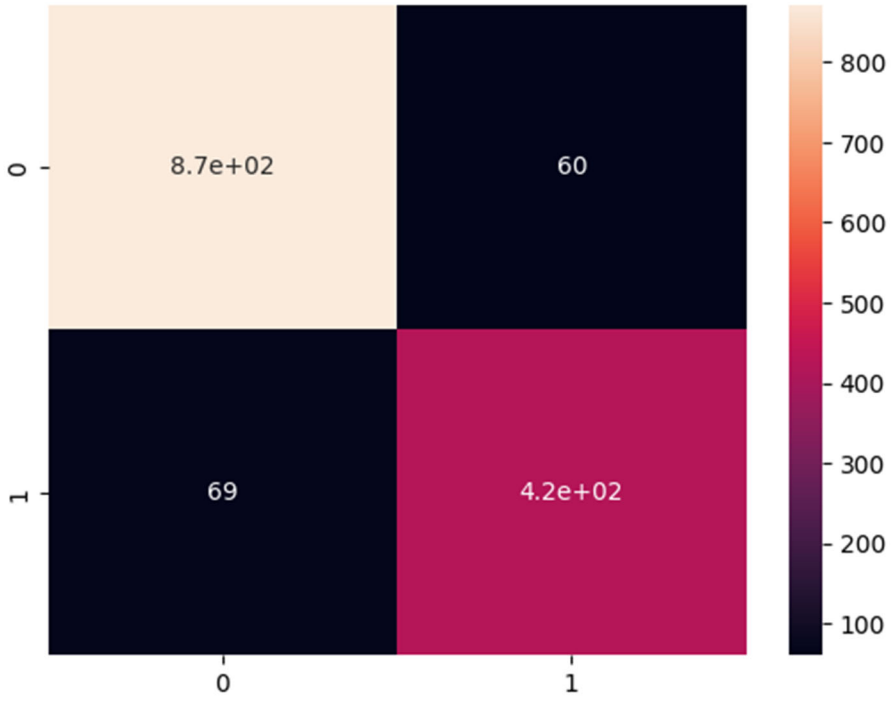


Рис. 4.4 Матриця помилок для базової моделі Random Forest Classifier

З аналізу цього графіка зроблено наступні висновки:

- Модель класифікувала обидва класи з приблизно однаковою точністю, аналогічно базовій моделі Decision Tree Classifier.

- Модель майже не робить помилок у класифікації, що підтверджується малим числом помилок: лише 69 помилок для класу 0 (не буде інсульту) та 60 помилок для класу 1 (буде інсульт).

Результати тестування базової моделі K-Neighbors Classifier представлено в таблиці 4.4.

Таблиця 4.4

Результати базової моделі K-Neighbors Classifier

class	precision	recall	F1-score	accuracy
0 (не буде інсульту)	0.84	0.95	0.82	0.86
1(буде інсульт)	0.92	0.75		

Аналізуючи результати, представлені у таблиці 4.4, зроблено такі висновки:

- модель провела класифікацію недостатньо стабільно. Показник precision майже однаковий для обох класів, але показник recall виявляє значні відмінності;

- показники F1-score та точності не є високими - 82% та 86% відповідно, що схоже на аналогічні показники базової моделі Decision Tree Classifier.

Також розглянено матрицю помилок для цієї моделі (рис. 4.5.).

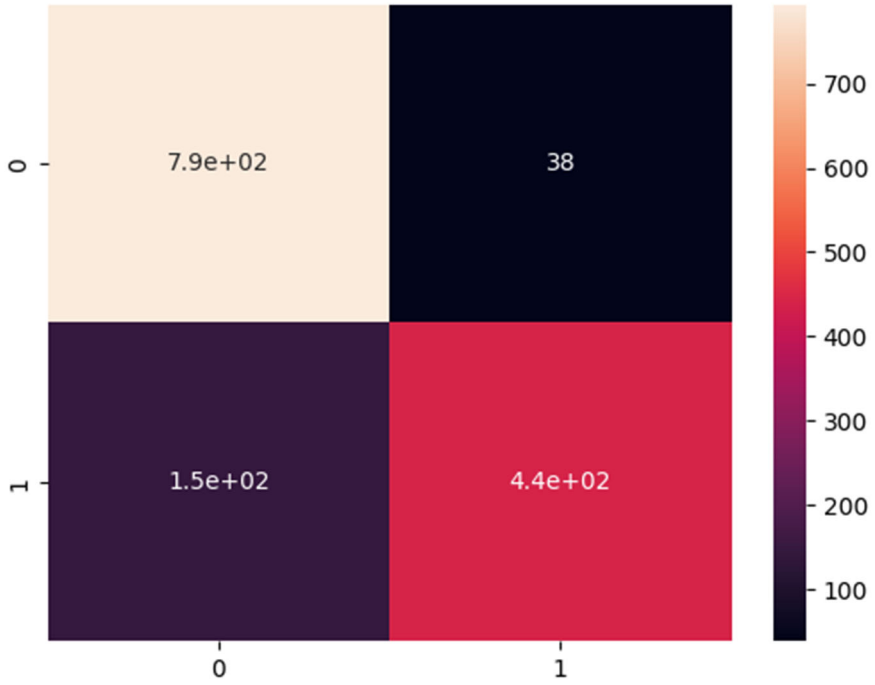


Рис. 4.5 Матриця помилок для базової моделі K-Neighbors Classifier

Як бачимо з рис. 4.5, модель приблизно однаково класифікувала клас 0 (не буде інсульту), майже так само як попередні базові моделі Decision Tree Classifier та Random Forest Classifier, та на відміну від попередніх модель трохи більше плутається в класі 1 (буде інсульт) - 38 помилок.

Результати тестування базової моделі Ada Boost Classifier наведено в таблиці 4.5.

Таблиця 4.5

Результати базової моделі Ada Boost Classifier

class	precision	recall	F1-score	accuracy
0 (не буде інсульт)	0.82	0.86	0.7	0.79
1(буде інсульт)	0.73	0.68		

З результатів таблиці 4.5 зроблено висновки:

- модель виконує не дуже стабільну класифікацію. Показник precision майже однаковий для обох класів, але показник recall значно відрізняється;

- показник F1-score невеликий - 70%.

Матриця помилок для даної моделі представлена на рис. 4.6.

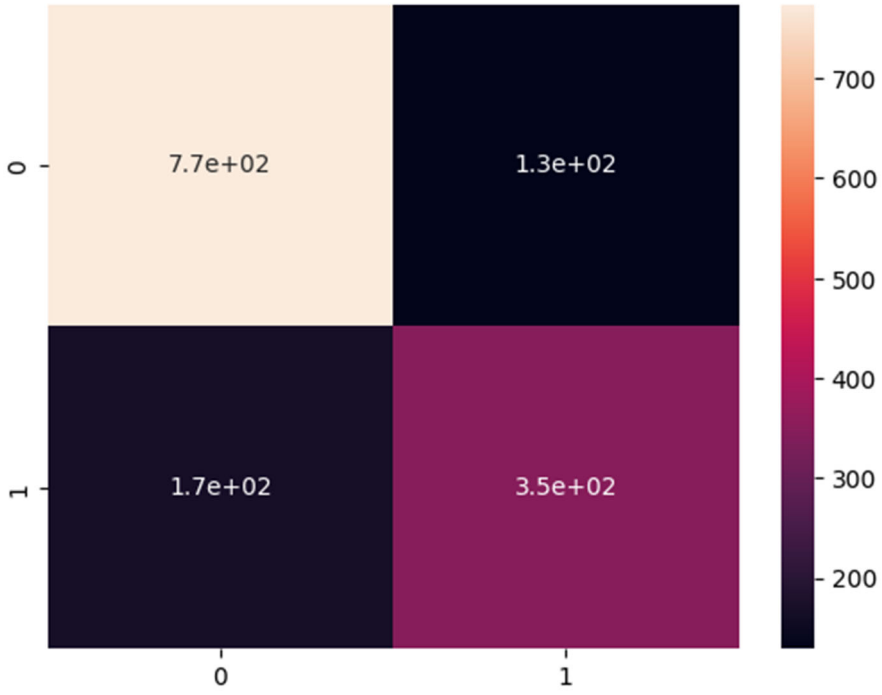


Рис. 4.6 Матриця помилок для базової моделі Ada Boost Classifier

Аналізуючи матрицю помилок для базової моделі Ada Boost Classifier, зроблено висновок, що модель робить значно більше помилок для класу 1 (індивіди, які будуть мати інсульт).

Також проведено тестування моделі Stacking Classifier, результати тестування якої наведено в таблиці 4.6.

Таблиця 4.6

Результати базової моделі Stacking Classifier

class	precision	recall	F1-score	accuracy
0 (не буде інсульту)	0.94	0.93	0.87	0.91
1 (буде інсульт)	0.87	0.88		

З результатів наведених у таблиці 4.6 зроблено висновки:

- модель здійснила стабільну класифікацію, подібну до базової моделі Random Forest Classifier. Показники precision та recall доволі однакові для обох класів;

- показник F1-score становить 87%, що є найкращим результатом серед базових моделей.

Матриця помилок для даної моделі подана на рис. 4.7.

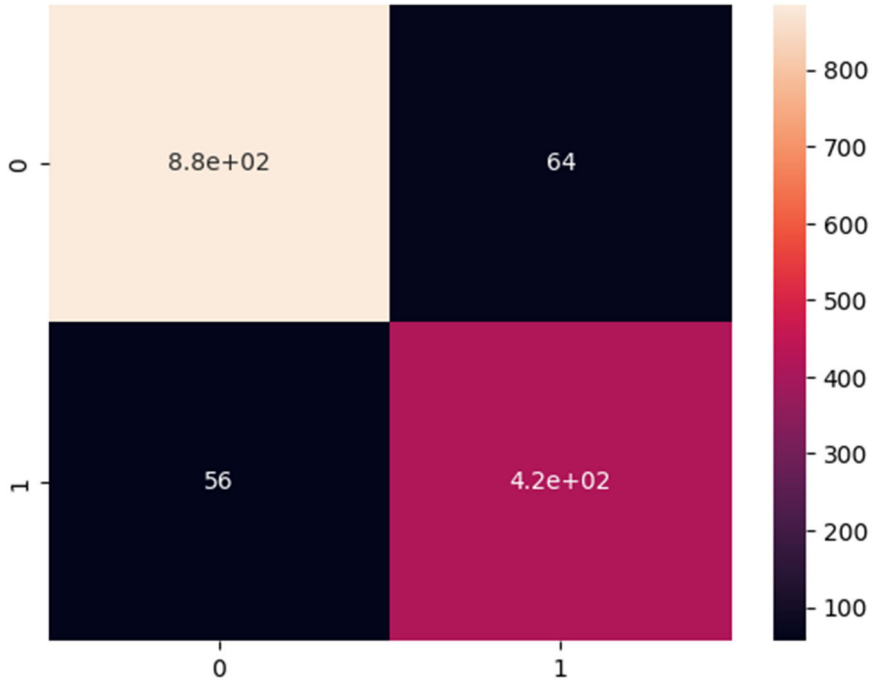


Рис. 4.7 Матриця помилок для базової моделі Stacking Classifier

Результати представлені на графіку показують, що модель добре класифікувала клас 0 (не буде інсульту) з 56 помилками та клас 1 (буде інсульт) з 64 помилками.

Отже, серед базових моделей найкраще себе показала модель Random Forest Classifier. Вона була досить стабільною і фактично не плуталась між класами. Крім того, її показник F1-score дорівнює 86%, що є досить хорошим значенням для такого набору даних.

Проведено також аналіз результатів пошуку найкращих гіперпараметрів для наших моделей за допомогою методу Grid Search.

Для моделі Decision Tree Classifier пошук здійснювався за наступними гіперпараметрами:

- splitter: best та random.

- criterion: gini, entropy та log_loss.

- max_depth: 10, 30 та 50.

- min_samples_split: 2, 4, 6, 8 та 10.

Найкращими значеннями для цих гіперпараметрів виявилися:

- splitter – best.

- criterion – log_loss.

- max_depth – 10.

- min_samples_split – 4.

Із використанням цих гіперпараметрів побудована нова покращена модель Decision Tree Classifier. Однак, результати її роботи були практично аналогічними до результатів базової моделі, що можна пояснити тим, що більшість оптимальних значень гіперпараметрів для цієї моделі є значеннями за замовчуванням [105, 107].

Для моделі Random Forest Classifier проводився пошук за такими гіперпараметрами:

- n_estimators: 50, 150, 300 та 500.

- criterion: gini, entropy та log_loss.

- max_depth: 10, 30 та 50.

Найкращими значеннями для цих гіперпараметрів виявилися:

- n_estimators – 500.

- criterion – entropy.

- max_depth – 30.

Після створення нової покращеної моделі з використанням отриманих гіперпараметрів досягнуто кращих результатів порівняно з базовою моделлю Random Forest Classifier. Результати показників покращеної моделі Random Forest Classifier наведено в таблиці 4.7, а матриця помилок для цієї моделі зображена на рис. 4.8.

Таблиця 4.7

Результати покращеної моделі Random Forest Classifier

class	precision	Recall	F1-score	accuracy
0 (не буде інсульт)	0.93	0.94	0.9	0.91
1 (буде інсульт)	0.89	0.86		

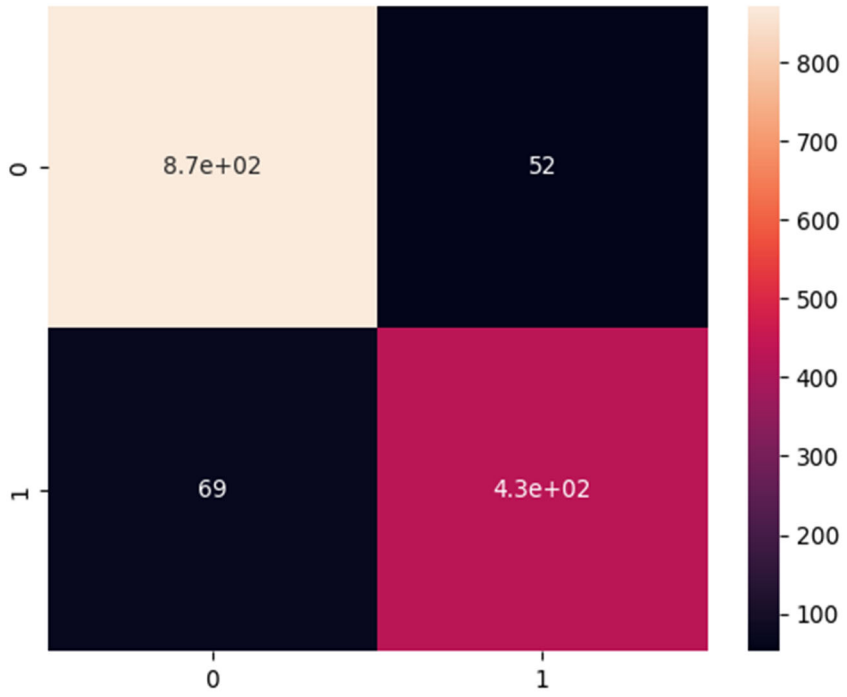


Рис. 4.8 Матриця помилок для покращеної моделі Random Forest Classifier

Як видно з отриманих результатів, показник F1-score підвищився на 4% і становив 90%, що є досить високим значенням. Також модель здійснила збалансовану класифікацію, де показники precision та recall майже однакові для обох класів, при цьому кількість помилок класифікації для класу 0 (не буде інсульту) залишилась на рівні 69, тоді як для класу 1 (буде інсульт) зменшилась до 52.

Для моделі K-Neighbors Classifier проводився пошук за такими гіперпараметрами:

- n_neighbors: 3, 5, 8 та 10.

- weights: uniform та distance.

- algorithm: auto, ball_tree, kd_tree та brute.

- metric: minkowski, manhattan та euclidean.

Найкращими значеннями для цих гіперпараметрів виявилися:

- n_neighbors – 5.

- algorithm – auto.

- metric – minkowski.

- weights – uniform.

Після створення нової покращеної моделі з використанням цих гіперпараметрів досягнуто аналогічних результатів до базової моделі K-Neighbors Classifier, оскільки більшість найкращих значень гіперпараметрів виявилися значеннями за замовчуванням.

Для моделі Ada Boost Classifier пошук здійснювався за такими гіперпараметрами:

- n_estimators: 50, 150, 300 та 500.

- base_estimator: Decision Tree Classifier (max_depth=1) та Random Forest Classifier (max_depth=1).

- learning_rate: 1, 0.1 та 0.01.

Найкращими значеннями для цих гіперпараметрів виявилися:

- n_estimators – 50.

- base_estimator – Decision Tree Classifier (max_depth=1).

- learning_rate – 0.01.

Використовуючи ці гіперпараметри, побудована нова покращена модель Ada Boost Classifier, але результати її роботи були аналогічні до результатів базової моделі Ada Boost Classifier. Точність не зросла, що можна пояснити тим, що модель досягла свого максимуму на такому складному для навчання наборі даних[109, 111].

Для моделі Stacking Classifier пошук здійснювався за такими гіперпараметрами:

- rf_n_estimators: 50, 150, 300 та 500.

- rf_criterion: gini, entropy та log_loss.

- rf_max_depth: 10, 30 та 50.

Найкращими значеннями для цих гіперпараметрів виявилися:

- rf_n_estimators – 300.

- rf_criterion – entropy.

- rf_max_depth – 30.

Створивши нову покращену модель з використанням цих гіперпараметрів, досягнуто подібних результатів до базової моделі Stacking Classifier. Це пояснюється тим, що модель також досягла свого максимуму на такому складному для навчання наборі даних.

Таким чином, найкраще справилися базові моделі Random Forest Classifier і Stacking Classifier, а також покращена методом Grid Search модель Random Forest Classifier.

На основі знайдених гіперпараметрів побудовано нові покращені моделі і досягнуто кращих результатів порівняно з базовими моделями. Результати роботи кращих моделей наведено в таблиці 4.8 і зроблено наступні висновки:

Таблиця 4.8

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

- Кращою моделлю для нашої задачі виявилася покращена методом Grid Search модель Random Forest Classifier. Вона продемонструвала дуже стабільні результати для обох класів і досягла значення F1-score 90%, що є дуже хорошим результатом.
- Також добре себе показала модель Stacking Classifier. Хоча вона виявилася менш стабільною, ніж покращена модель Random Forest Classifier, її результати також були непоганими, зі значенням F1-score 87%, що є хорошим показником.
- Серед базових моделей (Random Forest Classifier та Stacking Classifier) кращою виявилася модель Random Forest Classifier.

4.3.2. Оцінка якості аугментованих даних

Використовуючи описані вище моделі та методи машинного навчання, проведено передбачення цільового класу на оригінальних наборах даних і аугментованих для двох типів даних: текст та зображення. Також проведено порівняльний аналіз отриманих результатів на оригінальних та штучно збільшених даних.

1. Аналіз передбачень текстових даних

Для прогнозування сентименту текстових відгуків використано модель логістичної регресії. Спочатку проведено тренування моделі та здійснено оцінку результатів на оригінальному наборі даних.

	precision	recall	f1-score	support
negative	0.87705	0.22385	0.35667	478
neutral	0.72464	0.82667	0.77230	1875
positive	0.86887	0.89698	0.88270	2844
accuracy			0.80970	5197
macro avg	0.82352	0.64916	0.67055	5197
weighted avg	0.81758	0.80970	0.79449	5197

Рис. 4.9 Звіт по класифікації прогнозів на оригінальному наборі даних

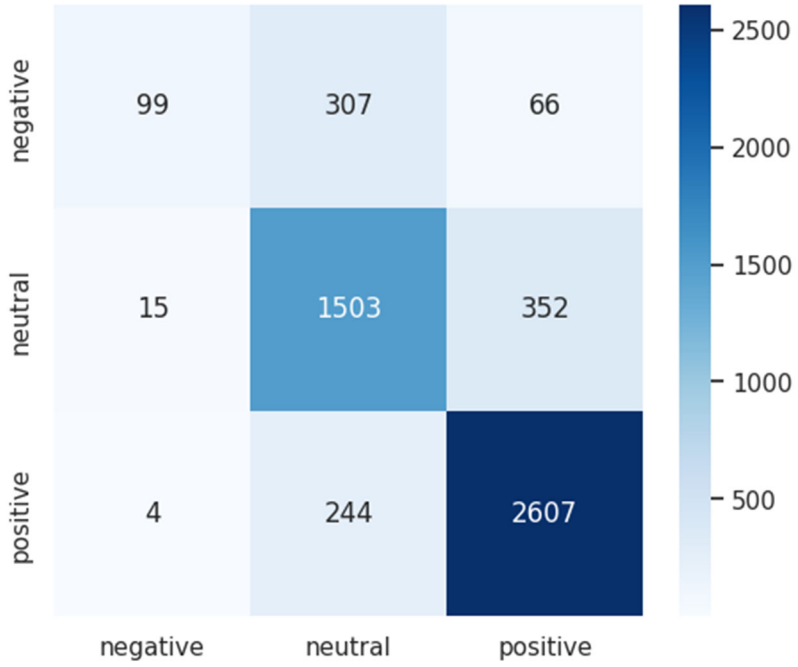


Рис. 4.10 Матриця помилок прогнозів на оригінальному наборі даних

Прогнозування сентименту тексту за результатами використання моделі отримали із загальною точністю 80.97%.

Різниця у точності обробки для різних класів емоцій досить помітна. Модель найбільш ефективно впізнавала позитивний сентимент (F1-score 88.27%). Нейтральний сентимент також розпізнавала досить добре (F1-score 77.23%).

Проте модель значно гірше впізнавала негативний сентимент, з F1-score лише 35.67%. Це свідчить про те, що модель краще розпізнає позитивні й нейтральні тексти, ніж негативні. Це також чітко видно на матриці помилок, де клас сентиментів передбачений найгірше.

Макроусереднений показник (macro avg) враховує незбалансованість класів і відображає загальну ефективність моделі при рівному обробленні всіх класів. У цьому випадку макроусереднений показник F1-score становить 67.05%, що є досить низьким, що також підтверджує вищезазначені спостереження.

Аналіз результатів текстового набору даних з використанням вищезазначених методів аугментації показав, що завдяки кожному методу аугментації тексту досягнуто балансу класів, а також збільшено розмірність

кожного класу удвічі. Це привело до наступних результатів (рис. 4.11).

	precision	recall	f1-score	support
negative	0.53249	0.73203	0.61651	459
neutral	0.80681	0.70471	0.75231	1849
positive	0.88614	0.90516	0.89555	2889
accuracy			0.81855	5197
macro avg	0.74181	0.78063	0.75479	5197
weighted avg	0.82668	0.81855	0.81994	5197

Рис. 4.11 Класифікаційний звіт прогнозів на аугментованому наборі даних з використанням заміни синонімів

Тобто, точність передбачення позитивного класу залишилася на високому рівні (89%), і повнота також є високою (91%). Це свідчить про те, що модель ефективно визначає позитивні відгуки.

Прогноз для нейтрального класу покращився: точність складає 81%, а повнота - 70%.

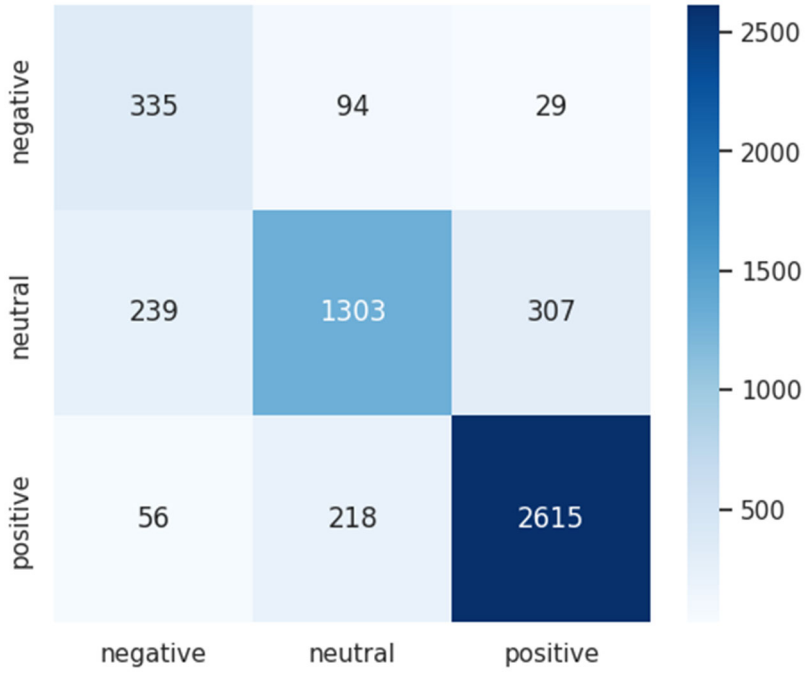


Рис. 4.12 Матриця помилок прогнозів на аугментованому наборі даних з використанням заміни синонімів

Найбільше покращення спостерігалося у прогнозуванні негативного класу.

Точність знизилася до 53%, але повнота зросла до 73%. Це означає, що модель виявляє більше негативних відгуків, навіть якщо деякі з них є помилковими.

Загальна точність моделі становить 82%. У цілому, модель почала краще прогнозувати нейтральний сентимент за рахунок балансування, в той час як інші класи отримали лише невелике покращення. Це може бути пов'язано з тим, що зміна деяких синонімів може вплинути на сентимент та змінити його тон, наприклад, з позитивного на нейтральний.

	precision	recall	f1-score	support
negative	0.83898	0.20975	0.33559	472
neutral	0.73174	0.80374	0.76606	1870
positive	0.86182	0.91313	0.88673	2855
accuracy			0.80989	5197
macro avg	0.81085	0.64221	0.66279	5197
weighted avg	0.81294	0.80989	0.79326	5197

Рис. 4.13 Класифікаційний звіт прогнозів на аугментованому наборі даних з використанням випадкових перестановок

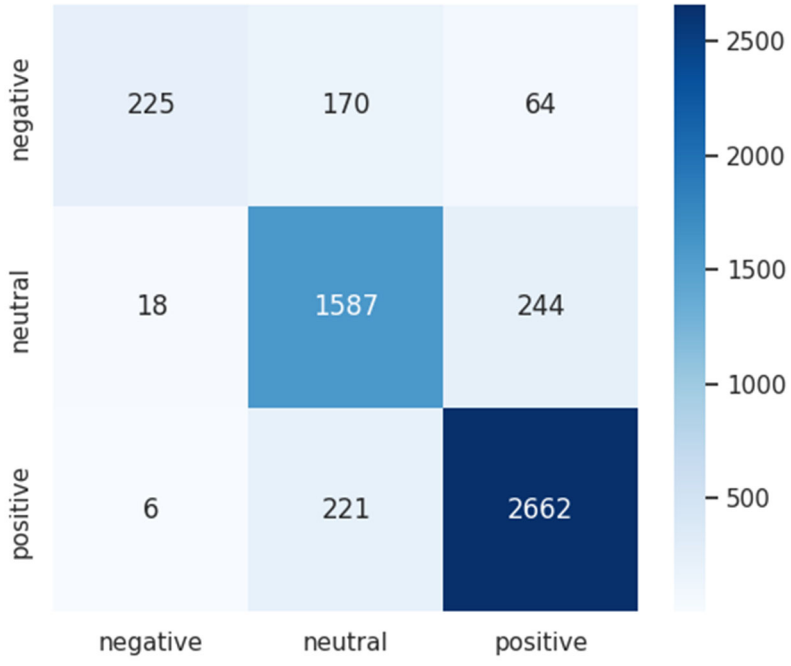


Рис. 4.14 Матриця помилок прогнозів на аугментованому наборі даних з використанням випадкових перестановок

Використання аугментації випадкових перестановок зумовило значні покращення:

Для негативного класу: точність зросла з 0.84 до 0.90, а повнота значно збільшилася з 0.21 до 0.49. F1-score також показала значне покращання, з 0.34 до 0.64.

Для нейтрального класу: точність зросла з 0.73 до 0.80, а повнота збільшилася з 0.80 до 0.86. F1-score показала покращання з 0.77 до 0.83.

Для позитивного класу: значення точності зросло з 0.86 до 0.90, а повнота залишилася майже такою ж - спочатку 0.91, потім 0.92. F1-score показала незначне покращання з 0.89 до 0.91.

Загальна точність моделі зросла з 0.81 у моделі, натренованій на оригінальному датасеті, до 0.86 у моделі, натренованій на аугментованому наборі даних.

	precision	recall	f1-score	support
negative	0.83394	0.50327	0.62772	459
neutral	0.79739	0.85776	0.82647	1849
positive	0.90242	0.91554	0.90893	2889
accuracy			0.85857	5197
macro avg	0.84458	0.75886	0.78771	5197
weighted avg	0.85900	0.85857	0.85476	5197

Рис. 4.15 Класифікаційний звіт прогнозів на аугментованому наборі даних з використанням випадкових вставлень

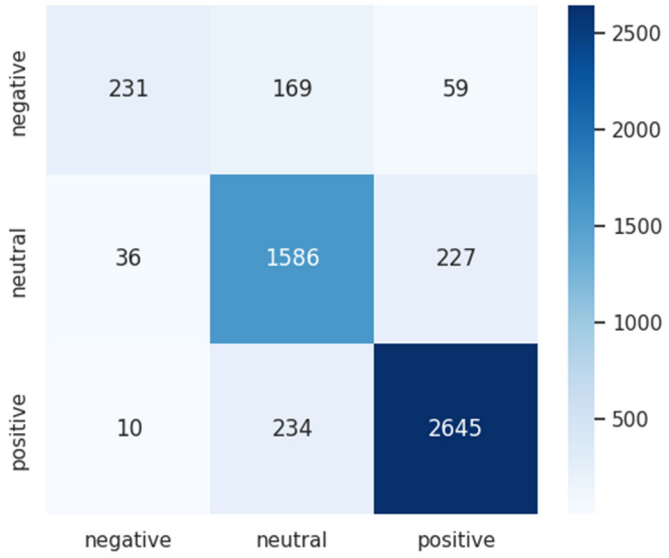


Рис. 4.16 Матриця помилок прогнозів на аугментованому наборі даних з використанням випадкових вставлень

Очевидно, що результати досить схожі з результатами, отриманими за допомогою випадкових перестановок.

Для негативного класу: точність дещо знизилась з 0.84 до 0.83, однак повнота значно зросла з 0.21 до 0.50, що привело до покращання F1-міри з 0.34 до 0.63.

Для нейтрального класу: точність зросла з 0.73 до 0.80, а повнота збільшилася з 0.80 до 0.86. Це привело до покращання F1-міри з 0.77 до 0.83.

Для позитивного класу: точність покращилась з 0.86 до 0.90, а повнота незначно зросла з 0.91 до 0.92. F1-score показала незначне покращання з 0.89 до 0.91.

Загальна точність моделі зросла з 0.81 (на оригінальному наборі даних) до 0.86 (на аугментованому наборі даних).

2. Аналіз передбачень зображень

Для виявлення хвороби пневмонії застосовано нейронну мережу CNN. Спочатку проведено тренування та оцінку результатів на оригінальному наборі зображень [114, 115]. Проведено тренування на 15 епохах із використанням функції зворотного виклику для зменшення кроку навчання.

	precision	recall	f1-score	support
Pneumonia (Class 0)	0.70	1.00	0.82	390
Normal (Class 1)	0.99	0.29	0.45	234
accuracy			0.73	624
macro avg	0.84	0.65	0.64	624
weighted avg	0.81	0.73	0.69	624

Рис. 4.17 Класифікаційний звіт прогнозів на оригінальному наборі даних зображень

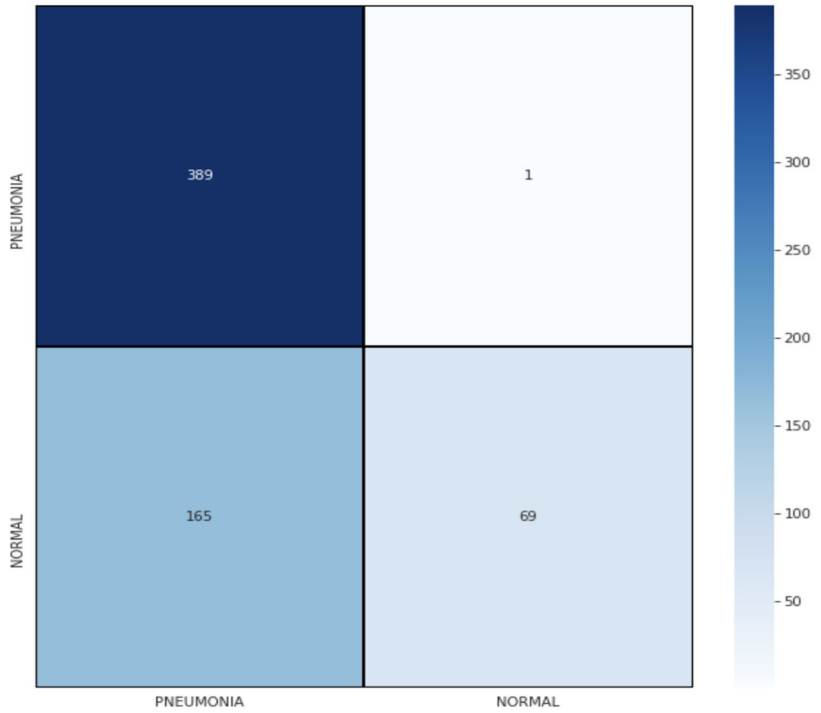


Рис. 4.18 Матриця помилок прогнозів на оригінальному наборі даних зображень

За допомогою даної моделі прогнозування пневмонії виявилося достатньо точним (Precision = 0.70) і показало високу чутливість до правильного визначення цього стану (Recall = 1.00), що привело до узагальненої метрики F1-score = 0.82.

Однак, модель виявилася менш ефективною при визначенні випадків без пневмонії (нормальний стан - клас 1), з точністю 99%, але низькою чутливістю - лише 29%, що привело до F1-score = 0.45.

Загальна точність моделі (accuracy) складає 73%. Це вище випадкового вибору, але все ще потребує покращення.

Проведено аналіз результатів прогнозування пневмонії з використання вище описаних методів аугментації. У якості аугментатора використано ImageDataGenerator з бібліотеки Keras. Він під час навчання диманічно визначає класи меншості та балансує дата сет і генерує нові штучні зразки кожного класу. З використанням аугментації отримано наступні результати:

	precision	recall	f1-score	support
Pneumonia (Class 0)	0.88	0.93	0.90	390
Normal (Class 1)	0.87	0.78	0.82	234
accuracy			0.87	624
macro avg	0.87	0.86	0.86	624
weighted avg	0.87	0.87	0.87	624

Рис. 4.19 Класифікаційний звіт прогнозів на аугментованому наборі даних зображень з використанням поворотів

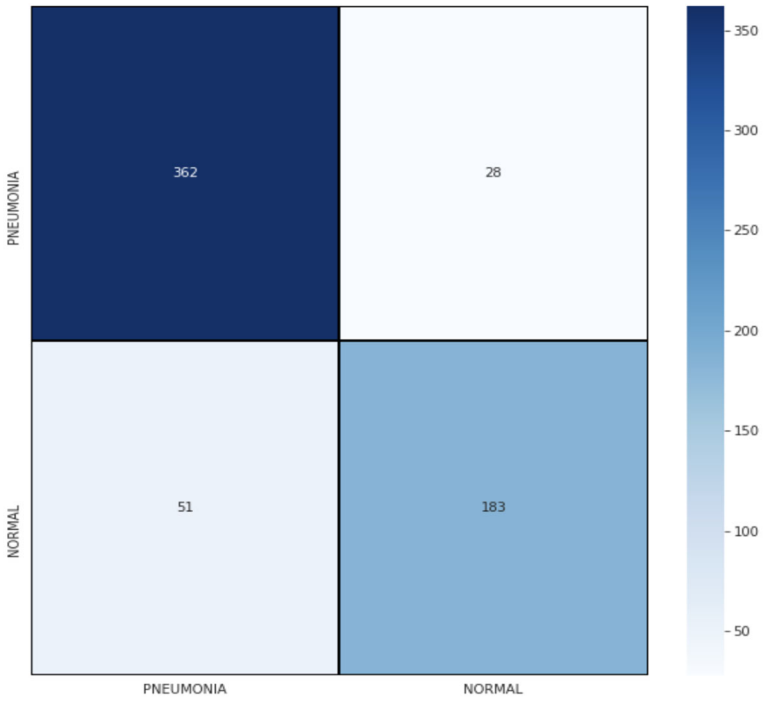


Рис. 4.20 Матриця помилок прогнозів на аугментованому наборі даних зображень з використанням поворотів

У порівнянні з оригінальним набором даних, CNN, навчена на аугментованому наборі даних з використанням поворотів, показала кращі результати.

Точність (precision) у визначенні пневмонії значно підвищилась, досягаючи 88%. Чутливість (recall) також зросла до 93%, забезпечуючи майже повне виявлення всіх випадків пневмонії. F1-міра порівняно з попередніми результатами зросла до 90%.

Щодо визначення нормального стану, модель також покращалась. Precision зросла до 87%, а recall досягла 78%. Це означає, що модель стала

набагато менш "песимістичною" і розпізнавала більшу кількість здорових пацієнтів. Середнє гармонічне цих двох показників F1-score становило 82%.

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

Випадкові приближення та зсуви висоти або ширини зображень

	precision	recall	f1-score	support
Pneumonia (Class 0)	0.93	0.94	0.94	390
Normal (Class 1)	0.90	0.89	0.89	234
accuracy			0.92	624
macro avg	0.92	0.91	0.91	624
weighted avg	0.92	0.92	0.92	624

Рис. 4.21 Класифікаційний звіт прогнозів на аугментованому наборі даних зображень з використанням зсувів та приближень

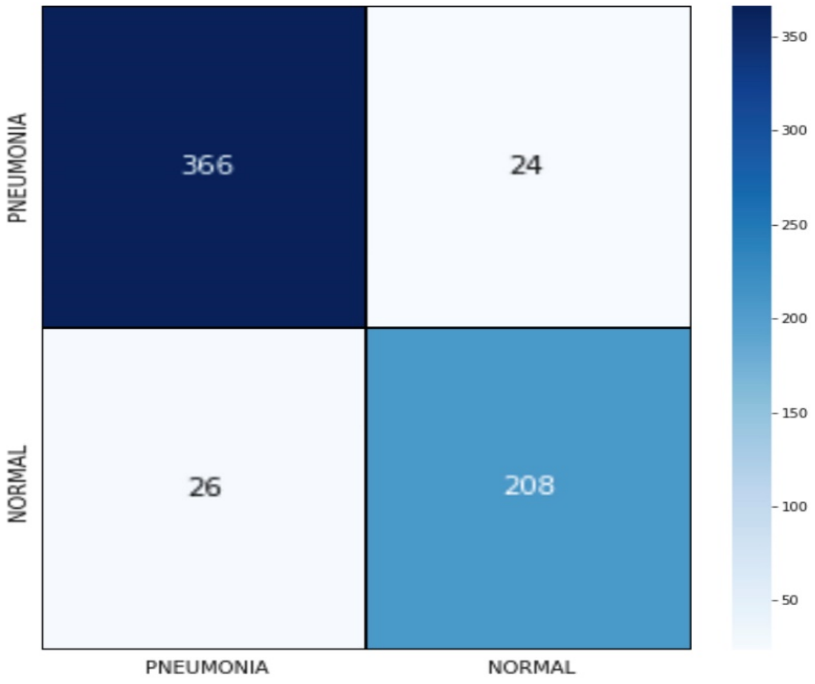


Рис. 4.22 Матриця помилок прогнозів на аугментованому наборі даних зображень з використанням зсувів та приближень

У порівнянні з оригінальним набором даних результати застосування CNN для аугментованого набору при випадкових зсувах та приближеннях показали ще кращі результати.

Для класу пневмонії precision зросла до 93%, а recall досягла 94%. Це означає, що модель визначила практично всі випадки пневмонії, і при цьому більшість передбачень були правильними. Загальний показник F1-score, який є середнім гармонічним між precision і recall, становив 94%.

При визначенні нормального стану, точність збільшилася до 90%, чутливість досягла 89%. Це означає, що модель точніше розпізнає нормальний стан пацієнтів і рідше помиляється, встановлюючи їм діагноз пневмонії. F1-score цих результатів досягла 89%.

Загальна точність моделі (accuracy) зросла до 92%, що демонструє значне покращення її ефективності у визначенні стану пацієнтів.

Випадкові зміни яскравості

	precision	recall	f1-score	support
Pneumonia (Class 0)	0.62	0.99	0.77	390
Normal (Class 1)	0.00	0.00	0.00	234
accuracy			0.62	624
macro avg	0.31	0.50	0.38	624
weighted avg	0.39	0.62	0.48	624

Рис. 4.23 Класифікаційний звіт прогнозів на аугментованому наборі даних зображень з використанням зміни яскравості

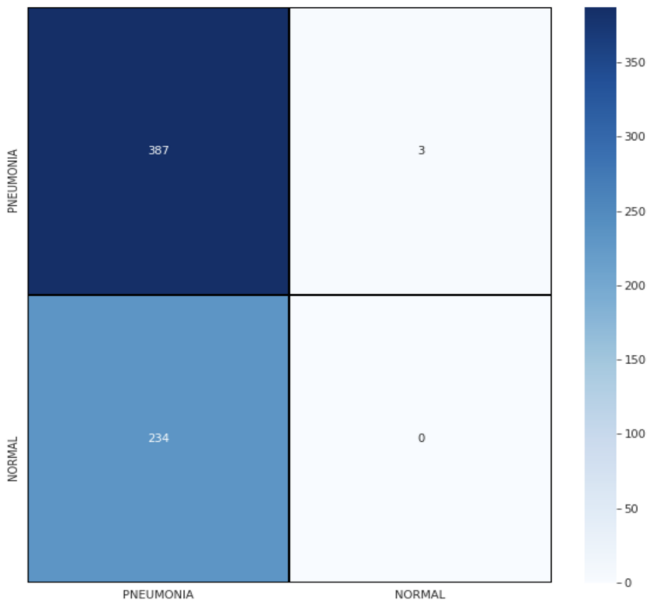


Рис. 4.24 Матриця помилок прогнозів на аугментованому наборі даних зображень з використанням зміни яскравості

У порівнянні з оригінальним набором даних, результати використання CNN на аугментованому наборі даних зі зміною яскравості значно гірші.

При виявленні пневмонії Precision зменшилася до 62%, але водночас recall зросла і склала 99%. Це означає, що модель дійшла до крайності в тому, що передбачає пневмонію у великої кількості випадків, включаючи здорових пацієнтів. Як результат, F1-score для цього класу становить 77%.

При визначенні нормального стану, модель досягла нульового значення precision, recall і F1-score. Це означає, що модель абсолютно не розпізнала здорових пацієнтів, хоча в наборі даних їх було 234.

Загальна точність моделі склала 62%, а загальна F1-score - 38%.

Такий поганий результат моделі вочевидь пов'язаний з тим, що зміна яскравості впливає на контрастність зображень, яка є ключовим фактором для розпізнавання пневмонії на рентгенівських знімках. Таким чином, використання зміни яскравості в цій конкретній задачі, найімовірніше, спотворило характеристики, на які модель повинна була звернути увагу, що призвело до дуже слабких результатів.

4.3.3. Дослідження ефективності ансамблів голосування

На основі результатів дослідження методів машинного навчання для класифікації стадій хвороби Альцгеймера у підрозділі 3.3.2, обрано три класифікатори для об’єднання в ансамбль: Random Forest, Multi-Layer Perceptron та SVM. З метою вибору кращого ансамблю, який забезпечить вищу та стабільну точність, проведено аналіз двох типів голосування – hard voting та soft voting.

Починаючи з ансамблю, що використовує hard voting, класифікація здійснювалася на валідаційних даних. Результати класифікації даних ансамблем з hard voting на валідаційній вибірці показали високу точність 0,9502. Таким чином, ансамбль здатний правильно класифікувати 95% зразків у вибірці, що є досить високим показником. Матриця помилок для отриманого ансамблю представлена на рис. 4.25.

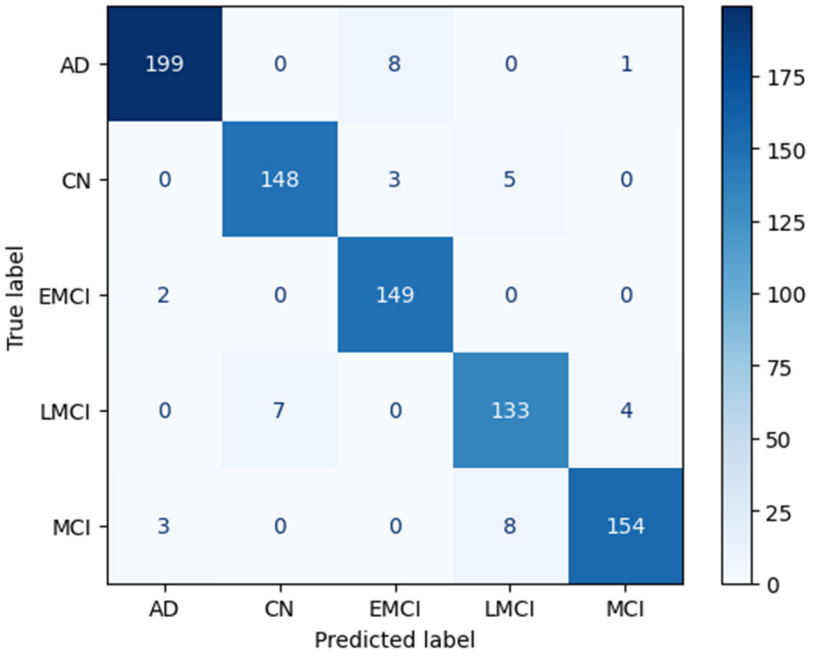


Рис.4.25 Матриця помилок ансамблю з типом голосування hard voting на валідаційній вибірці

У матриці помилок можна відзначити, що більшість помилок відбувається між класами EMCI і AD, а також між класами LMCI і MCI. Причиною цього може бути схожість даних класів між собою, що ускладнює їх розрізнення.

Порівнюючи показники Precision, бачимо, що для класу AD модель має точність 0.98, що означає, що лише 2% зразків були неправильно класифіковані як AD. Найнижчий показник мав клас LMCI – 0.91. Високі значення recall свідчать про те, що модель не пропускає багато зразків кожного класу. Наприклад, для класу EMCI модель отримала значення 0.99, що свідчить про те, що вона виявляла майже всі зразки цього класу. Показник F1-score для всіх класів є високим та близьким до 1, що є дуже добрим результатом.

Таким чином, модель має високі результати precision, recall та F1-score для класів AD, CN та MCI. Однак, для класів EMCI та LMCI модель має трохи нижчі precision та recall, але все ж є достатньо ефективною. Загалом, ці результати свідчать про те, що модель має добру здатність класифікувати дані на основі навчального набору, з яким вона була тренована.

Наступним етапом був аналіз ансамблю з типом голосування soft voting. В результаті класифікації даних ансамблем з типом голосування soft voting отримано високу точність 0,9563. Таким чином, ансамбль зміг правильно класифікувати майже 96% зразків у вибірці, що є високим результатом. Матрицю помилок ансамблю моделей представлено на рис.4.26.

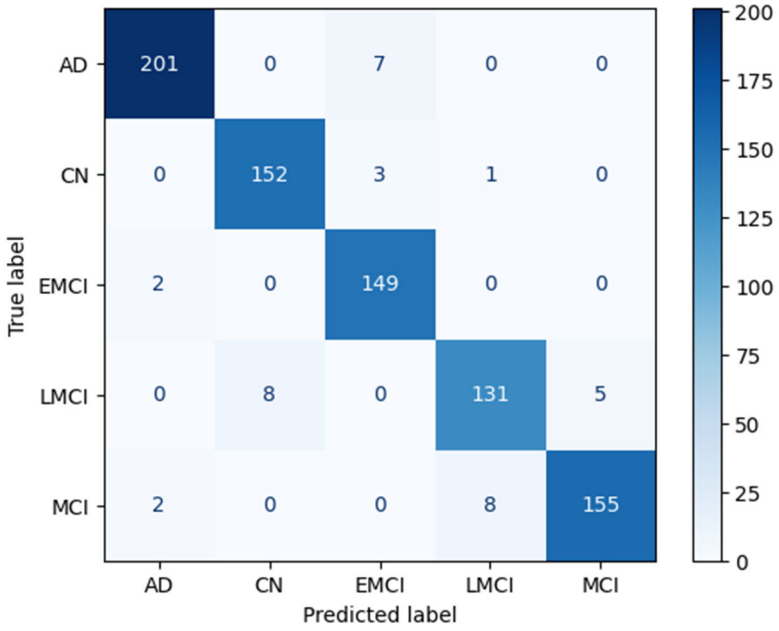


Рис. 4.26 Матриця помилок ансамблю з типом голосування soft voting на валідаційній вибірці

З аналізу матриці помилок видно, що загальні результати покращилися порівняно з ансамблем, який використовував hard voting. Кількість правильно класифікованих зображень класу AD зросла на 2 зразки, для класу CN - на 4 зразки, для класу MCI - на 1 зразок більше. Кількість правильно класифікованих зображень для класу EMCI залишилась незмінною, але для класу LMCI зменшилась на 2 зразки. Проблема класифікації класу LMCI може бути пов’язана з особливостями даних цього класу у використовуваному наборі.

Аналізуючи детальні результати класифікації за показниками precision, recall та F1-score, зроблено висновок, що класифікатор демонструє високі показники якості на досліджуваному наборі даних. Загальна точність становить 96%, що свідчить про правильну класифікацію 96% прикладів. Кожен клас має

високий показник precision, що вказує на малу кількість неправильно класифікованих прикладів. Найвищий показник precision має клас AD зі значенням 0.98, найнижчий - класи EMCI та LMCI з однаковими значеннями 0.94.

Показник recall вказує на здатність класифікатора правильно визначати деякий клас. В цілому, показники recall для всіх класів є високими, що свідчить про добру здатність моделі виявляти дійсні екземпляри кожного класу. Найвищий показник recall отриманий для класу EMCI - 0.99, а найнижчий - для класу LMCI - 0.91. З урахуванням показників precision, можна стверджувати, що модель має високу точність і здатність правильно класифікувати більшість екземплярів для всіх класів.

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, об'єднано їхні результати в одну таблицю 4.9.

Таблиця 4.9

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. Розглядаючи результати для обох типів ансамблів, зроблено висновки:

Hard voting:

- для всіх класів Precision 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:

- для всіх класів Precision 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 виявляється трохи кращим за hard voting у забезпеченні точності та здатності виявляти дійсні екземпляри. Загалом, обидва типи ансамблів ефективно працюють у класифікації даного набору даних.

4.4. Апробація програмної реалізації інформаційної системи опрацювання персоналізованих даних особи

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

Для створення серверної частини використано платформу Firebase. Firebase - це повноцінний сервіс з різноманітними інструментами для розробки мобільних додатків. Цю платформу підтримує та розвиває компанія Google, що означає, що більшість інтегрованих у нього інструментів також є продуктами

Google. Тому всередині Firebase є широкий спектр функцій, які використано при розробці клієнтської частини системи, а саме:

- Бази даних: Cloud Firestore, що є NoSQL базою даних з використанням хмарних технологій, та Firebase Realtime, яка використовує формат JSON для зберігання даних та забезпечує синхронізацію інформації між базами даних та клієнтами програми в режимі реального часу.

- Машинне навчання: функція реалізована на основі відкритих бібліотек машинного навчання компанії Google, що дає змогу розробникам, які працюють з Firebase, використовувати всі досягнення Google у цьому напрямку.

- Аутентифікація: Google Firebase пропонує різноманітні бібліотеки з готовими рішеннями для реалізації автентифікації в програмі, такі як автентифікація за номером телефону, електронною поштою або через соціальні мережі.

- Хмарні повідомлення: цей інструмент дає змогу розробникам надсилати системні повідомлення своїм користувачам.

- Хостинг: Firebase має власну систему хостингу, що дає змогу розробникам з легкістю розгортати свої додатки.

- Хмарне сховище: цей інструмент використовується для зберігання різноманітних даних у хмарі.

- Crashlytics: надає звіти про помилки в роботі системи Google Firebase, що дає змогу оперативно виявляти та виправляти проблеми.

- Моніторинг продуктивності: цей інструмент контролює продуктивність додатків на різних платформах, таких як Android та iOS.

Клієнтська частина поділена на мобільний (iOS) додаток та додаток для смарт-годинника (WatchOS). У обох випадках для розробки використовується мова програмування Swift — відкрита багатопарадигмова мова програмування загального призначення, що компілюється. Вона створена компанією Apple насамперед для розробки на її фірмових операційних системах [112].

Для розробки UI мобільного додатка використано фреймворк SwiftUI, який забезпечує необхідну інфраструктуру для програм iOS або tvOS. SwiftUI надає

архітектуру вікон та представлень для створення інтерфейсу додатка, обробку подій, підтримку анімації, роботу з документами та малюванням, а також інші необхідні функції.

Для розробки UI додатка для смарт-годинника використано фреймворк WatchKit, який, подібно до UIKit, адаптований під специфіку роботи з WatchOS. Для збору та аналізу показників та статистики пацієнта використано HealthKit — фреймворк, що забезпечує центральне сховище даних про здоров'я та фітнес на iPhone та Apple Watch. HealthKit дає змогу програмам отримувати доступ до цих даних та обмінюватися ними з дозволу користувача.

Процес роботи застосунку починається зі стартового екрана (рис. 4.27). Після цього відбувається завантаження екрану авторизації (рис. 4.28) або домашньої сторінки застосунку (рис. 4.29), залежно від авторизації користувача.

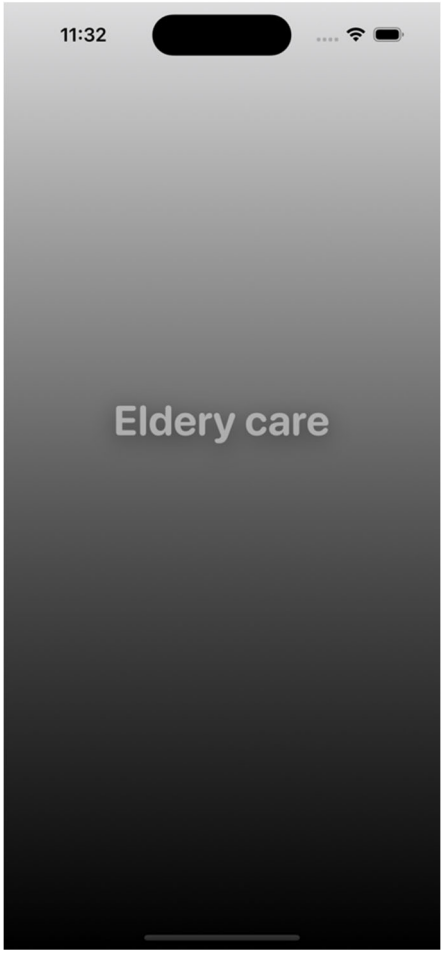


Рис. 4.27 Екран завантаження

На екрані авторизації користувач може створити новий обліковий запис, написавши своє ім'я та прізвище, електронну пошту та створивши пароль. Пароль повинен містити як цифри, так і літери і мати не менше шести символів. Якщо у користувача вже є обліковий запис, він може увійти за допомогою електронної пошти, яку він вказав під час реєстрації, а також свого паролю. Для цього йому потрібно натиснути кнопку з написом "Already have an account? Sign In".

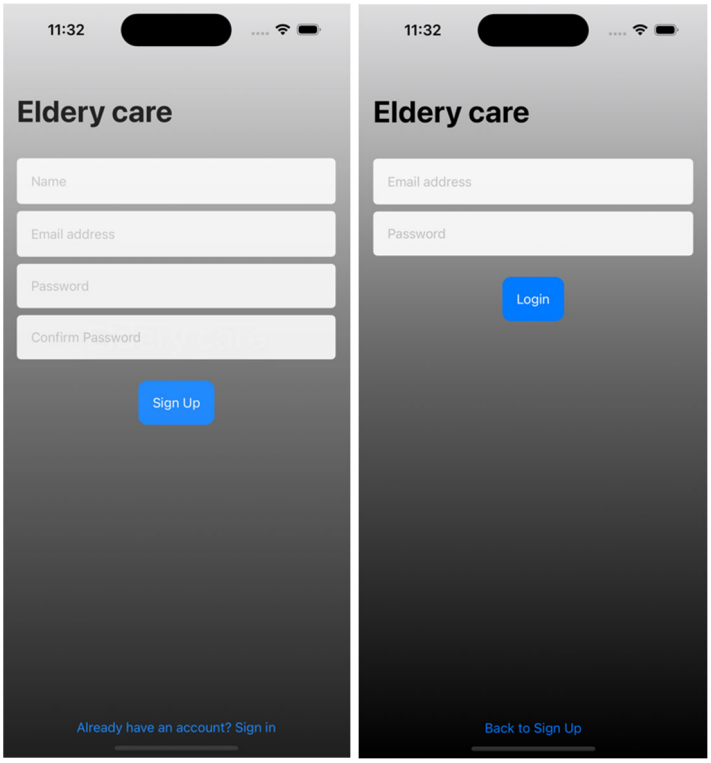


Рис. 4.28 Екрани авторизації

На головному екрані зверху розташований список пацієнтів, які підключені до поточного куратора, а також іконка "Add new patient". Після натискання на неї відкривається екран з формою, де можна додати нового пацієнта (рис. 4.29).

При натисканні на іконку з іменем потрібного пацієнта буде відображена статистика та дані про його стан здоров'я. Якщо натиснути на плашку синього кольору з ім'ям та прізвищем пацієнта, відобразиться екран з детальною інформацією про користувача. Користувач може редагувати цю інформацію та,

за потреби, видаляти користувача зі списку. Крім того, натиснувши на статистику, можна додавати нові показники до вже існуючої статистики.

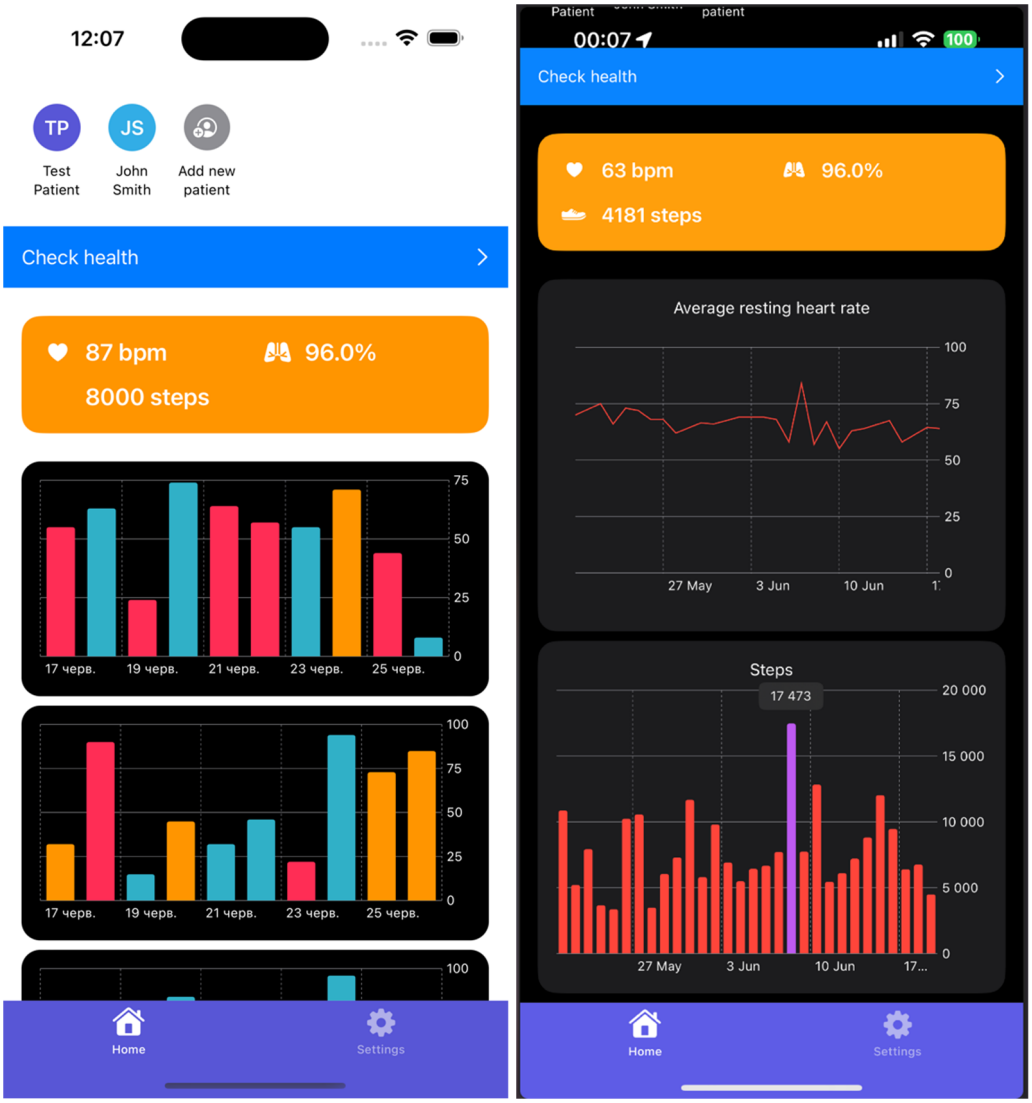


Рис. 4.29 Домашній екран застосунку

4.5. Висновки до розділу 4

У даному розділі дисертаційного дослідження розроблено архітектуру інформаційної системи для підтримки прийняття медичних рішень, що ґрунтується на аналізі стану особи на основі персоналізованих медичних даних.

Представлена функціональна схема прототипу інформаційної системи включає збір інформації про особу/пацієнта під супроводом особи, що спостерігає, а також контроль процесу збору та збереження даних. Реалізація системи відбулася у вигляді мобільного додатку, що успішно вирішує зазначені

проблеми.

Здійснено аналіз ефективності роботи моделей класифікації та проведено пошук оптимальних гіперпараметрів за допомогою методу Grid Search. Це сприяло покращенню моделі – Random Forest Classifier, яка показала хороші результати, досягаючи таких метрик точності: precision 0.9, recall 0.9, та F1-score 0.9.

На основі результатів застосування процесу аугментації даних для удосконалення процесу класифікації стану особи стане можливим проведення експериментів з оптимальнішими методами та параметрами. Можна комбінувати різні методи аугментації для визначення найефективніших способів додавання синтетичних даних для вирішення конкретної проблеми класифікації.

Проведено порівняльний аналіз існуючих моделей класифікації, під час якого модель MLP продемонструвала найкращі результати. Для подальшого вдосконалення вибрано моделі, які об’єднано для створення ансамблю: Random Forest, SVM і MLP.

Аналізуючи існуючі методи побудови ансамблевого навчання, запропоновано використання VotingClassifier, що дало змогу комбінувати різні класифікатори з різними параметрами для отримання точнішого результату.

Проведено порівняльний аналіз розглянутих у третьому розділі досліджень та розробленого ансамблю моделей. Розроблений ансамбль soft voting виявився кращим варіантом серед розглянутих моделей.

У результаті даного дослідження досягнута висока точність класифікації кожної з п'яти стадій хвороби Альцгеймера. Це досягнуто завдяки збалансуванню набору даних та поєднанню різних моделей, які взаємодоповнюються та компенсують одна одну у своїй здатності до класифікації.

Представлено результати імплементації інформаційної системи для супроводу процесу збору та аналізу стану особи.

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, що забезпечує його ефективне визначення та дає змогу покращити процес ідентифікації стадії захворювання.

2. Вперше розроблено новий метод класифікації персоналізованих медичних даних за рахунок введення етапу аугментації при опрацюванні медичної інформації про особу, що дало можливість збільшити обсяг та різноманітність навчальної вибірки, забезпечуючи краще узагальнення моделей і зменшуючи 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 та 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 ВИКОРИСТАНИХ ДЖЕРЕЛ

1 Transformacje cyfrowe i technologie innowacyjne w ekonomii [wydanie elektroniczne]: zbiór materiałów Międzynarodowej Naukowo-Praktycznej Konferencji Internetowej, Łomża – Charków, 14-15.03.2024 r. / Redakcja naukowa: Ireneusz Żuchowski, Zoia Sharlovych, Olena Dudnyk. Łomża: Międzynarodowa Akademia Nauk Stosowanych w Łomży; Charków : PISzW "Charkowski Uniwersytet Technologiczny "SHAG", Ukraina. Wydawnictwo: MANS w Łomży, Część 2. 2024. 413 s.

2 Персональні дані: використання, захист і відповідальність - що потрібно знати. ЮРЛІГА n.d. https://jurliga.ligazakon.net/news/201367_ personaln-dan-vikoristannya-zakhist--vdpovdalnst---shcho-potrbno-znati (accessed June 24, 2024).

3 Personalized Medicine n.d. https://www.genome.gov/genetics-glossary/Personalized-Medicine (accessed June 24, 2024).

4 Personalized Employee Experience Can Help with L&D Goals. Gartner, https://www.gartner.com/smarterwithgartner/3-ways-personalization-can-improve-the-employee-experience.

5 Ben-Israel D. The impact of machine learning on patient care: a systematic review / D. Ben-Israel, W. B. Jacobs, S. Casha, [et al.] // Artificial Intelligence in Medicine. — 2020. — Vol. 103. — P. 101785.

6 Best Clinical Data Management Systems include Merative Zelta, Bioclinica ICL, Dotmatics Enterprise Data Platform, Oracle Clinical One, and IQVIA Digital Site Suite / https://www.trustradius.com/clinical-data-management

7 Personalized Employee Experience Can Help with L&D Goals’. Gartner, https://www.gartner.com/smarterwithgartner/3-ways-personalization-can-improve-the-employee-experience.

8 Neumayr T, Augstein M. A Systematic Review of Personalized Collaborative Systems. Front Comput Sci 2020;2:562679. https://doi.org/10.3389/fcomp.2020.562679.

9 Mules O. Y., Snytyuk V. Y., Myronyuk I. S. “Information technology for optimizing the human resources of health care institutions,” Bulletin of Vinnytsia Polytechnic Institute.№ 6, pp. 83-90, 2019.

10 Viguera Altolaguirre C, Reddy R, Gamaldo CE, Salas RME. Developing a standardized EMR workflow for medical students and preceptors. Health Policy and Technology 2023;12:100696. https://doi.org/10.1016/j.hlpt.2022.100696.

11 Chen H, Wu Y, Zhou J, You D, Zhao Y. Identifying the association rules between adverse events and concomitant medicines in clinical trial data management using random forest. Biostatistics & Epidemiology 2023;7:e2112896. https://doi.org/10.1080/24709360.2022.2112896.

12 Chakraborty C, Ghosh U, Ravi V, Shelke Y, editors. Efficient Data Handling for Massive Internet of Medical Things: Healthcare Data Analytics. Cham: Springer International Publishing; 2021. https://doi.org/10.1007/978-3-030-66633-0.

13 Lee MJ, Bagci P, Kong J, Vos MB, Sharma P, Kalb B, et al. Liver steatosis assessment: Correlations among pathology, radiology, clinical data and automated image analysis software. Pathology - Research and Practice 2013;209:371–9. https://doi.org/10.1016/j.prp.2013.04.001.

14 Jayakumar P, Duckworth E, Bozic KJ. Value-based Healthcare: Three Ways Healthcare Systems Can Get More Usage Out of Their Patient Engagement Tools. Clin Orthop Relat Res 2021;479:2136–8. https://doi.org/10.1097/CORR.0000000000001934.

15 Auffray C, Chen Z, Hood L. Systems medicine: the future of medical genomics and healthcare. Genome Med 2009;1:2. https://doi.org/10.1186/gm2.

16 Handra J, Elbert A, Gazzaz N, Moller-Hansen A, Hyunh S, Lee HK, et al. The practice of genomic medicine: A delineation of the process and its governing principles. Front Med 2023;9:1071348. https://doi.org/10.3389/fmed.2022.1071348.

17 Suganthe R. C. Diagnosis of Alzheimer’s disease from brain magnetic resonance imaging images using deep learning algorithms / R. C. Suganthe, R. S.

Latha, M. Geetha, G. R. Sreekanth // Advances in Electrical and Computer Engineering. — 2020. — Vol. 20, No. 3. — P. 57–64.

18 Davuluri R. Improved classification model using cnn for detection of alzheimer's disease / R. Davuluri, R. Rengaswamy // Journal of Computer Science. — 2022. — Vol. 18, No. 5. — P. 415–425.

19 Angkoso C. V. Multiplane convolutional neural network (mp-cnn) for alzheimer's disease classification / C. V. Angkoso, H. P. A. Tjahyaningtijas, M. H. Purnomo, I. K. E. Purnama // International Journal of Intelligent Engineering and Systems. — 2022. — Vol. 15, No. 1. — P. 329–340.

20 Lee R, Leighton SP, Thomas L, Gkoutos GV, Wood SJ, Fenton S-JH, et al. Prediction models in first-episode psychosis: systematic review and critical appraisal. Br J Psychiatry 2022;220:179–91. https://doi.org/10.1192/bjp.2021.219.

21 Steyerberg EW, Harrell FE. Prediction models need appropriate internal, internal–external, and external validation. Journal of Clinical Epidemiology 2016;69:245–7. https://doi.org/10.1016/j.jclinepi.2015.04.005.

22 Liu S-C, Lai J, Huang J-Y, Cho C-F, Lee PH, Lu M-H, et al. Predicting microvascular invasion in hepatocellular carcinoma: a deep learning model validated across hospitals. Cancer Imaging 2021;21:56. https://doi.org/10.1186/s40644-021-00425-3.

23 Battineni G. Improved alzheimer's disease detection by mri using multimodal machine learning algorithms / G. Battineni, M. A. Hossain, N. Chintalapudi, [et al.] // Diagnostics. — 2021. — Vol. 11, No. 11. — P. 2103. https://doi.org/10.3390/diagnostics11112103.

24 Yildirim M. Classification of alzheimer's disease mri images with cnn based hybrid method / M. Yildirim, A. Cinar // Ingenierie des Systemes d'Information. — 2020. — Vol. 25, No. 4. — P. 413–418.

25 Kavitha C. Early-stage alzheimer's disease prediction using machine learning models / C. Kavitha, V. Mani, S. R. Srividhya, [et al.] // Frontiers in Public Health. — 2022. — Vol. 10. https://doi.org/10.3389/fpubh.2022.853294

26 Shahwar T. Automated detection of alzheimer’s via hybrid classical quantum neural networks / T. Shahwar, J. Zafar, A. Almogren, [et al.] // Electronics. — 2022. — Vol. 11, No. 5. — P. 721. https://doi.org/10.3390/electronics11050721

27 Abunadi I. Deep and hybrid learning of mri diagnosis for early detection of the progression stages in alzheimer’s disease / I. Abunadi // Connection Science. — 2022. — Vol. 34, No. 1. — P. 2395–2430. https://doi.org/10.1080/09540091.2022.2123450

28 Tang X. Comparing different algorithms for the course of alzheimer’s disease using machine learning / X. Tang, J. Liu // Annals of Palliative Medicine. — 2021. — Vol. 10, No. 9. — P. 9715–9724. doi: 10.21037/apm-21-2013.

29 Nebehay S. Number of people with dementia set to jump 40% to 78 mln by 2030 -WHO. Reuters 2021. [Електронний ресурс] Режим доступу: https://www.reuters.com/business/healthcare-pharmaceuticals/number-people-with-dementia-set-jump-40-78-mln-by-2030-who-2021-09-02/.

30 O. O. Abayomi-Alli, R. Damasevicius, R. Maskeliunas, i A. Abayomi-Alli, «BiLSTM with Data Augmentation using Interpolation Methods to Improve Early Detection of Parkinson Disease», в Proc. Fed. Conf. Comput. Sci. Inf. Syst., FedCSIS, Ganzha M., Maciaszek L., Maciaszek L., i Paprzycki M., Ред., Institute of Electrical and Electronics Engineers Inc., 2020, с. 371–380. doi: 10.15439/2020F188.

31 L. Taylor i G. Nitschke. Improving deep learning with generic data augmentation. 2018 IEEE symposium series on computational intelligence (SSCI), IEEE, 2018, с. 1542–1547. doi: 10.1109/SSCI.2018.8628742.

32 W. Alosaimi, M. I. Uddin. Efficient Data Augmentation Techniques for Improved Classification in Limited Data Set of Oral Squamous Cell Carcinoma, CMES Comput. Model. Eng. Sci., 2022. 131(3): 1387-1401, doi: 10.32604/cmes.2022.018433.

33 A. L. C. Ottoni, R. M. de Amorim, M. S. Novo, i D. B. Costa. Tuning of data augmentation hyperparameters in deep learning to building construction image classification with small datasets. Intl. J. Mach. Learn. Cybern., 2023. 14(1):171-186 doi: 10.1007/s13042-022-01555-1.

34 Alzheimer datasets 5 classes | kaggle [Електронний ресурс] // Режим доступу: https://www.kaggle.com/datasets/phamnguyenduytien/alzheimer-datasets-5-classes.

35 Bhavsar H. A review on support vector machine for data classification / H. Bhavsar, M. H. Panchal // International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), 2012. - Vol. 1, No. 10.- P. 185–189.

36 Cervantes J. A comprehensive survey on support vector machine classification: applications, challenges and trends / J. Cervantes, F. Garcia-Lamont, L. Rodríguez-Mazahua, A. Lopez // Neurocomputing. — 2020. — P. 189–215.

37 Kalita D. J. A survey on SVM hyper-parameters optimization techniques / D. J. Kalita, V. P. Singh, V. Kumar. — Social Networking and Computational Intelligence Lecture Notes in Networks and Systems, 2020, p. 243-256.

38 Decision trees — scikit-learn 1.2.2 documentation [Електронний ресурс] Режим доступу: https://scikit-learn.org/stable/modules/tree.html (відвідано: 20.03.2023)

39 Priyanka, D. Kumar. Decision tree classifier: a detailed survey // International Journal of Information and Decision Sciences. — 2020. Vol. 12(3) - P. 246-269.

40 Mahmood I. The role of machine learning algorithms for diagnosing diseases / I. Mahmood, A. Mohsin Abdulazeez // Journal of Applied Science and Technology Trends. — 2021. — Vol. 2. – P. 10-19.

41 Speiser J. L. A comparison of random forest variable selection methods for classification prediction modeling / J. L. Speiser, M. E. Miller, J. Tooze, E. Ip // Expert Systems with Applications. — 2019. — Vol. 134. — P. 93–101. https://doi.org/10.1016/j.eswa.2019.05.028

42 Multilayer perceptron / 2023 [Електронний ресурс] Режим доступу: https://en.wikipedia.org/w/index.php?title=Multilayer_perceptron&oldid=1147594876 (відвідано: 20.03.2023)

43 López F. Ensemble learning: stacking, blending & voting / F. López. — 2020.

44 Sklearn.ensemble.voting classifier — sci-kit learn 1.2.2 documentation [Електронний ресурс] Режим доступу: https://scikit-

learn.org/stable/modules/generated/sklearn.ensemble.VotingClassifier.html (відвідано: 20.03.2023)

45 Melnykova N., Chereshchuk L. Application of machine learning methods for predicting the risk of stroke occurrence. Proceedings of the vi international Scientific and Practical Conference. Sofia, Bulgaria. 2023. pp. 210-216. International Science Group, 2023. ISBN 9798891451926.

46 Biswas N., Uddin K. M. M., Rikta S. T., Dey S. K. A comparative analysis of machine learning classifiers for stroke prediction: A predictive analytics approach. Healthcare Analytics. 2022. Vol.2. P. 100116.

47 Mostafa S. A., Elzanfaly D. S., Yakoub A. E. A Machine Learning Ensemble Classifier for Prediction of Brain Strokes. International Journal of Advanced Computer Science and Applications (IJACSA). 2022. 13 (12):258-266.

48 Sailasya G., Kumari G. L. A. Analyzing the Performance of Stroke Prediction using ML Classification Algorithms. International Journal of Advanced Computer Science and Applications (IJACSA). 2021. 12 (6): 539-545.

49 Uchida K., Kouno J., Yoshimura S., Kinjo N., Sakakibara F., Araki H., Morimoto T. Development of Machine Learning Models to Predict Probabilities and Types of Stroke at Prehospital Stage: the Japan Urgent Stroke Triage Score Using Machine Learning (JUST-ML). Translational Stroke Research. 2022. 3(3): 370–381.

50 Mezher M. A. Genetic Folding (GF) Algorithm with Minimal Kernel Operators to Predict Stroke Patients. Applied Artificial Intelligence. 2022. 36(1): e2151179.

51 Tegistu B. S. Brain stroke prediction model using deep neural network (dnn). 2021.

52 Pitchai R., Dappuri B., Pramila P. V., Vidhyalakshmi M., Shanthi S., Alonazi W. B., Almutairi K. M. A., Sundaram R. S., Beyene I. An Artificial Intelligence-Based Bio-Medical Stroke Prediction and Analytical System Using a Machine Learning Approach. Computational Intelligence and Neuroscience. 2022. 10: e5489084.

53 Rohit A. P. V., Chowdary M. U., Ashish G. B. S., Anitha V., Sana S. ML Approach for brain stroke prediction using ist database. International Journal of Engineering Applied Sciences and Technology. 2022. 7(10):72-75.

54 Telu V., Padimi V., Ningombam D. D. Optimizing Predictions of Brain Stroke Using Machine Learning. Journal of Neutrosophic and Fuzzy Systems. 2022.Vol. 2: 31–43.

55 Abedi V., Avula V., Chaudhary D., Shahjouei S., Khan A., Griessenauer C. J., Li J., Zand R. Prediction of Long-Term Stroke Recurrence Using Machine Learning Models. Journal of Clinical Medicine. 2021. 10(6):1286.

56 Ashrafuzzaman Md., Saha S., Nur K. Prediction of Stroke Disease Using Deep CNN Based Approach. Journal of Advances in Information Technology. 2022. Vol. 13(6):604-613.

57 Sun X. Predictive model analysis of stroke disease based on machine learning. SPIE, 2023.

58 Tazin T., Alam M. N., Dola N. N., Bari M. S., Bourouis S., Monirujjaman Khan M. Stroke Disease Detection and Prediction Using Robust Learning Approaches. Journal of Healthcare Engineering. 2021. Nov.2021 e7633381.

59 Dritsas E., Trigka M. Stroke Risk Prediction with Machine Learning Techniques. Sensors. 2022. Vol.22(13):4670.

60 DataHack : Biggest Data hackathon platform for Data Scientists.

61 GOYAL C. Outlier Detection & Removal | How to Detect & Remove Outliers (Updated 2023). 2021.

62 Douzas G, Bacao F. Effective data generation for imbalanced learning using conditional generative adversarial networks. Expert Systems with Applications 2018;91:464–71. https://doi.org/10.1016/j.eswa.2017.09.030.

63 Formentin S, Mazzoleni M, Scandella M, Previdi F. Nonlinear system identification via data augmentation. Systems & Control Letters 2019;128:56–63. https://doi.org/10.1016/j.sysconle.2019.04.004.

64 Methodology for Systematic Review, DistillerSR. Дата звернення: 12, Листопад 2023. [Online]. Доступний у:

https://www.distillersr.com/resources/systematic-literature-reviews/prisma-methodology-for-systematic-review.

65 A. Rojarath, W. Songpan and C. Pong-inwong. Improved ensemble learning for classification techniques based on majority voting. 2016 7th IEEE International Conference on Software Engineering and Service Science (ICSESS), Beijing, China, 2016, pp. 107-110, doi: 10.1109/ICSESS.2016.7883026

66 M. A. Kutlugun, Y. Sirin, M. Karakaya.The effects of augmented training dataset on performance of convolutional neural networks in face recognition system. Proc. Fed. Conf. Comput. Sci. Inf. Syst., FedCSIS, 2019, Vol. 18. pp. 929–932. doi: 10.15439/2019F181.

67 K. Kim J. Jeong, «Deep learning-based data augmentation for hydraulic condition monitoring system», в Procedia Comput. Sci., Shakshuki E., Yasar A-U-H., i Malik H., Ред., Elsevier B.V., 2020, pp. 20–27. doi: 10.1016/j.procs.2020.07.007.

68 M. Bayer, M.-A. Kaufhold, B. Buchhold, M. Keller, J. Dallmeyer, i C. Reuter, «Data augmentation in natural language processing: a novel text generation approach for long and short text classifiers», Intl. J. Mach. Learn. Cybern., 2022, Vol. 14.pp. 135-150/ doi: 10.1007/s13042-022-01553-3.

69 A. Mikołajczyk, M. Grochowski.Data augmentation for improving deep learning in image classification problem. International interdisciplinary PhD workshop (IIPhDW), IEEE, 2018, S. 117–122.

70 K. Dunphy, M. N. Fekri, K. Grolinger, i A. Sadhu.Data Augmentation for Deep-Learning-Based Multiclass Structural Damage Detection Using Limited Information. Sensors, 2022. 22 (16) :6193. doi: 10.3390/s22166193.

71 R. Pappagari, J. Villalba, P. Zelasko, L. Moro-Velazquez, i N. Dehak, Copypaste: An augmentation method for speech emotion recognition. ICASSP IEEE Int Conf Acoust Speech Signal Process Proc, Institute of Electrical and Electronics Engineers Inc., 2021, pp. 6324–6328. doi: 10.1109/ICASSP39728.2021.9415077.

72 N. F. Aminuddin, Z. Tukiran, A. Joret, R. Tomari, i M. Morsin.An Improved Deep Learning Model of Chili Disease Recognition with Small Dataset. Intl. J. Adv. Comput. Sci. Appl., 2022.13(7) : 407-412. doi: 10.14569/IJACSA.2022.0130750.

73 S. T. Aroyehun, A. Gelbukh. Aggression detection in social media: Using deep neural networks, data augmentation, and pseudo labeling. Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018). 2018. p. 90–97.

74 C. Shorten, T. M. Khoshgoftaar. A survey on Image Data Augmentation for Deep Learning. J. Big Data, 2019. Vol.6 (1) doi: 10.1186/s40537-019-0197-0.

75 T. Dao, A. Gu, A. Ratner, V. Smith, C. De Sa, i C. Ré. A kernel theory of modern data augmentation. Proceedings of the 36th International Conference on Machine Learning, PMLR 97, 2019. p. 1528–1537.

76 L. Brigato, L. Iocchi. A close look at deep learning with small data. Proc. Int. Conf. Pattern Recognit., Institute of Electrical and Electronics Engineers Inc., 2020, c. 2490–2497. doi: 10.1109/ICPR48806.2021.9412492.

77 «Statistical classification», Wikipedia. 30, Жовтень 2023. Дата звернення: 12, Листопад 2023. [Online]. Доступний у: https://en.wikipedia.org/w/index.php?title=Statistical_classification&oldid=1182690049

78 T. Dietterich. Overfitting and undercomputing in machine learning. ACM Comput. Surv., 1995. 27(3):326–327. doi: 10.1145/212094.212114.

79 C. Shorten, T. M. Khoshgoftaar, B. Furht, «Text Data Augmentation for Deep Learning», J Big Data. 2021. 8:101. doi: 10.1186/s40537-021-00492-0.

80 A. Jain, P. R. Samala, D. Mittal, P. Jyoti, & M. Singh. (2023). SpliceOut: A Simple and Efficient Audio Augmentation Method. arXiv.org. Retrieved November 12, 2023, from https://arxiv.org/abs/2110.00046v2.

81 «Logistic Regression - an overview | ScienceDirect Topics». Дата звернення: 12, Листопад 2023. [Online]. Доступний у: https://www.sciencedirect.com/topics/computer-science/logistic-regression

82 «A Guide to Convolutional Neural Networks — the ELI5 way | Saturn Cloud Blog». Дата звернення: 12, Листопад 2023. [Online]. Доступний у: https://saturncloud.io/blog/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way/

83 U. Kiran, «MFCC Technique for Speech Recognition», Analytics Vidhya. Дата звернення: 12, Листопад 2023. [Online]. Доступний у: https://www.analyticsvidhya.com/blog/2021/06/mfcc-technique-for-speech-recognition/

84 Alexandre, «Perceptron: Concept, function, and applications», Data Science Courses | DataScientest. Дата звернення: 12, Листопад 2023. [Online]. Доступний у: https://datascientest.com/en/perceptron-definition-and-use-cases

85 T. Srivastava, «12 Important Model Evaluation Metrics for Machine Learning Everyone Should Know (Updated 2023)», Analytics Vidhya. Дата звернення: 12, Листопад 2023. [Online]. Доступний у: https://www.analyticsvidhya.com/blog/2019/08/11-important-model-evaluation-error-metrics/

86 Y. Tang, Y. Wang, K. M., and Cooper, L. Li. Towards big data Bayesian network learning-an ensemble learning based approach'. IEEE International Congress on Big Data. IEEE, 2014, pp.355-357.

87 Ткаченко Р. О., Ізонін І. В., Данилик В. М., Михалевич В. Ю. Стекінг нейроподібної структури МПГП з RBF шаром на підставі генерування випадкового кортежу її гіперпараметрів для завдань прогнозування. Український журнал інформаційних технологій. 2021, т. 3, № 1. С. 49–55.

88 FL. Seixas, Flávio Luiz, et al. A Bayesian Network Decision Model for Supporting the Diagnosis of Dementia, Alzheimer's Disease and Mild Cognitive Impairment. Computers in Biology and Medicine, Aug. 2014, vol. 51, pp. 140–58. https://doi.org/10.1016/j.compbiomed.2014.04.010.

89 Perova I. Bodyanskiy Ye. Fast medical diagnostics using autoassociative neuro-fuzzy memory // International Journal of Computing. -2017. -16 (1). P. 34-40. https://doi.org/10.47839/ijc.16.1.869.

90 Bhatt, Chintan, et al., editors. Internet of Things and Big Data Technologies for Next Generation Healthcare. Springer International Publishing, 2017. https://doi.org/10.1007/978-3-319-49736-5.

91 Падлецька Н. І., Дивак М. П.. Інформаційна технологія для ідентифікації зворотного гортанного нерва під час хірургічної операції на щитовидній залозі // Вимірювальна та обчислювальна техніка в технологічних процесах. - 2015. - № 1. - С. 151-157.

92 Зайченко, Ю. П., Мурга, Н. А. Застосування систем з нечіткою логікою до задачі медичної діагностики. Вісник НТУУ.quot; КПІ.quot;: Інформатика, управління та обчислювальна техніка, 2008(49).;

93 Tsai, Christina W., et al. ‘A Multiple-State Discrete-Time Markov Chain Model for Estimating Suspended Sediment Concentrations in Open Channel Flow’. Applied Mathematical Modelling, vol. 40, no. 23–24, Dec. 2016, pp. 10002–19. https://doi.org/10.1016/j.apm.2016.06.037.

94 Bevilacqua, Maurizio, et al. ‘Timed Coloured Petri Nets for Modelling and Managing Processes and Projects’. Procedia CIRP, vol. 67, Jan. 2018, pp. 58–62. ScienceDirect, https://doi.org/10.1016/j.procir.2017.12.176.;

95 Boubeta-Puig, Juan, et al. ‘MEdit4CEP-CPN: An Approach for Complex Event Processing Modeling by Prioritized Colored Petri Nets’. Information Systems, vol. 81, Mar. 2019, pp. 267–89. ScienceDirect, https://doi.org/10.1016/j.is.2017.11.005.

96 Kadri, Hela, et al. ‘Formal Approach to Control Design of Complex and Dynamical Systems’. Procedia Computer Science, vol. 108, Jan. 2017, pp. 2512–16. ScienceDirect, https://doi.org/10.1016/j.procs.2017.05.134.

97 Perumal, Varalakshmi, et al. ‘Detection of COVID-19 Using CXR and CT Images Using Transfer Learning and Haralick Features’. Applied Intelligence, vol. 51, no. 1, Jan. 2021, pp. 341–58. https://doi.org/10.1007/s10489-020-01831-z.

98 Chimmula, Vinay Kumar Reddy, and Lei Zhang. ‘Time Series Forecasting of COVID-19 Transmission in Canada Using LSTM Networks’. Chaos, Solitons & Fractals, vol. 135, June 2020, p. 109864. https://doi.org/10.1016/j.chaos.2020.109864.

99 Ramos-Rincón, Jose Manuel, et al. ‘Cardiometabolic Therapy and Mortality in Very Old Patients With Diabetes Hospitalized Due to COVID-19’. The Journals of Gerontology: Series A, edited by Lewis Lipsitz, vol. 76, no. 8, July 2021, pp. e102–

09 , https://doi.org/10.1093/gerona/glab124.

100 Maleki, Sepideh. ‘Personalizing Health CareHugh Kaul Personalized Medicine Institute, University of Alabama Birmingham’. XRDS: Crossroads, The ACM Magazine for Students, vol. 25, no. 2, Jan. 2019, pp. 54–55, https://doi.org/10.1145/3292418.

101 Djulbegovic, Benjamin, and Gordon H. Guyatt. ‘Progress in Evidence-Based Medicine: A Quarter Century On’. The Lancet, vol. 390, no. 10092, July 2017, pp. 415–23. , https://doi.org/10.1016/S0140-6736(16)31592-6.

102 Danhof, Meindert, et al. ‘The Future of Drug Development: The Paradigm Shift towards Systems Therapeutics’. Drug Discovery Today, vol. 23, no. 12, Dec. 2018, pp. 1990–95. , https://doi.org/10.1016/j.drudis.2018.09.002.

103 Alabadla, Mustafa, et al. ‘ExtraImpute: A Novel Machine Learning Method for Missing Data Imputation’. Journal of Advances in Information Technology, vol. 13, no. 5, 2022. pp. 470-476, https://doi.org/10.12720/jait.13.5.470-476.

104 Mishyna, M., et al. ‘Effects of Radiation Damage in Studies of Protein-DNA Complexes by Cryo-EM’. Micron, vol. 96, May 2017, pp. 57–64. , https://doi.org/10.1016/j.micron.2017.02.004.

105 Бідюк П. І. Методи прогнозування / П. І. Бідюк, О. С. Меняйленко, О. В. Половцев. – Луганськ: «Альма Матер», 2008 (у 2-х томах). – 605 с.

106 Perov, Y., Graham, L., et al. ‘Multiverse: causal reasoning using importance sampling in probabilistic programming’. In Symposium on advances in approximate bayesian inference. PMLR, 2020, https://doi.org/10.48550/arXiv.1910.08091.

107 Tang, Yan, et al. ‘Towards Big Data Bayesian Network Learning - An Ensemble Learning Based Approach’. 2014 IEEE International Congress on Big Data, IEEE, 2014, pp. 355–57., https://doi.org/10.1109/BigData.Congress.2014.58.

108 Lakho, Shamshad, et al. ‘Decision Support System for Hepatitis Disease Diagnosis Using Bayesian Network’. Sukkur IBA Journal of Computing and Mathematical Sciences, vol. 1, no. 2, Dec. 2017, pp. 11–19. https://doi.org/10.30537/sjcms.v1i2.51.

109 Seixas, Flávio L., et al. ‘A Bayesian network decision model for supporting

the diagnosis of dementia'. Alzheimer′ s disease and mild cognitive impairment. Computers in biology and medicine, no. 51, 2014, pp. 140-158.

110 Structure and stability of symptoms in first episode psychosis: a longitudinal network approach //SL Griffiths, SP Leighton, PK Mallikarjun, G Blake… - Translational Psychiatry, 2021. № 11 Issue 1. P. 567.

111 Bhatt, Chintan, et al., editors. 'Internet of Things and Big Data Technologies for Next Generation Healthcare'. Springer International Publishing, 2017 , doi:10.1007/978-3-319-49736-5.

112 Podletskaya N., and Divak M. 'Information technology for the identification of the reverse laryngeal nerve during thyroid surgery'. Measuring and computing technology in technological processes, vol.1, 2015, pp. 151-157.

113 Silva-Ramírez E. L., et. al. 'Single imputation with multilayer perceptron and multiple imputation combining multilayer perceptron and k-nearest neighbours for monotone patterns'. Applied Soft Computing, vol. 29, 2015, pp.65-74.

114 Subbotin, S., Oliinyk, A., Levashenko, V.,Zaitseva, E. Diagnostic rule mining based on artificial immune system for a case of uneven distribution of classes in sample. Komunikacie, 18(3), 3-11.

115 Tsai C.W., et. al. 'A multiple-state discrete-time Markov chain model for estimating suspended sediment concentrations in open channel flow'. Applied Mathematical Modelling, 2016, vol. 40, no. 23-24, pp. 10002-10019.

116 Kadri, H., et. al. 'S. Formal approach to control design of complex and dynamical systems. Procedia Computer Science 108C (2017) 2512-2516.

117 Masic, I., et. al. 'Evidence based medicine–new approaches and challenges'. Acta Informatica Medica, 2008. 16 (4): 219-225. doi: 10.5455/aim.2008.16.219-225.

118 Sobrinho, A., et. al. 'Towards medical device certification: A colored petri nets model of a surface electrocardiography device'. In IECON 2014-40th Annual Conference of the IEEE Industrial Electronics Society, 2014, pp. 2645-2651.

119 Boubeta-Puig, J., et. al. 'An approach for complex event processing modeling by prioritized colored Petri nets'. Information Systems, no. 81, 2019, pp. 267-289.

120 Anand, N., et. al. 'A. Feature selection on educational data using Boruta

algorithm’. International Journal of Computational Intelligence Studies, vol. 1, no. 10, 2021, pp. 27-35.

121 Pakhira, M. K. ‘Finding number of clusters before finding clusters’. Procedia Technology, vol. 4, 2012, pp. 27-37.

122 Li, Z., et. al. ‘Awareness of line-of-sight propagation for indoor localization using Hopkins statistic’. IEEE Sensors Journal, vol. 9, no. 18, 2018, pp. 3864-3874.

123 Khurana, K., and Sharma, S. ‘A comparative analysis of association rule mining algorithms’. International Journal of Scientific and Research Publications, vol. 5, no. 3, 2013, pp. 23-45.

124 HealthKit | Apple Developer Documentation [Електронний ресурс]. – 2022. – Режим доступу: https://developer.apple.com/documentation/healthkit

125 Newman, M. E. ‘Mixing patterns in networks’. Physical Review E, vol. 2, no. 67, 2015, pp. 126.

126 Bonchi, F., Castillo, C., Gionis, A., and Jaimes, A. “Social network analysis and mining for business applications”. ACM Transactions on Intelligent Systems and Technology (TIST), vol. 3, no. 2, 2011, pp. 22.

127 Nataliia Melnykova, Yurii Patereha / Imbalanced data: a comparative analysis of classification enhancements using augmented data / Computer science, cybernetics and automation, Intellectual capital is the foundation of innovative development ‘2024, pp. 54-72, DOI: 10.30890/2709-2313.2024-28-00-017

128 Comparative analysis of data augmentation methods for image modality / Andrii Bokhonko, Nataliia Melnykova, Yurii Patereha // Scientific Journal of TNTU. — Tern.: TNTU, 2024. — Vol 113. — No 1. — P. 16–26./ https://visnyk.tntu.edu.ua/index.php?art=762

129 Mykhaylo Melnyk, Yurii Patereha / Prediction of the occurrence of stroke based on machine learning models / Computer Design Systems. Theory and Practice, CDS. 2024; Volume 6, Number 1: 17- 27 pp.

130 Paterega, I. Main Strategies for Autonomous Robotic Controller Design / I. Paterega // Радіоелектроніка і інформатика : науч.-техн. журн. – Х. : Вид-во ХНУРЕ, 2011. – Вип. 4. – С. 36-41.,

https://openarchive.nure.ua/entities/publication/505fca9b-c6b2-445c-9c23-e4338981c56d

131 Особливості використання штучних нейронних осциляторів у робототехніці / Ю. І. Патерега // Наук. вісн. : зб. наук.-техн. пр. / Нац. лісотехн. ун-т України. – Л., 2010. – Вип. 20.13. – С. 322–331.

132 Штучні нейронні осцилятори / П. В. Тимощук, Ю. І. Патерега // Комп'ютерні системи проектування. Теорія і практика : [зб. наук. пр.] / відп. ред. М. В. Лобур . – Л. : Вид-во Нац. ун-ту "Львів. політехніка", 2009. – С. 40–45. – (Вісник / Нац. ун-т "Львів. політехніка" ; No 651).

133 Nykoniuk, Mariia; Melnykova, Nataliia; Patereha, Yurii; Sala, Dariusz; Cichoń, Dariusz/Classification of Patients with the Development of Alzheimer's Disease using an Ensemble of Machine Learning Models/ 2023/ CEUR Workshop Proceedings, pp. 198 – 216 / https://ceur-ws.org/Vol-3609/short4.pdf

134 Analysis of neural network controller for mobile robot navigation / Yu. I. Paterega // 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.

135 Artificial neural oscillators in robotics / Yu. Paterega // Перспективні технології і методи проектування МЕМС : матеріали шостої міжнар. конф. MEMSTECH 2010, 20–23 квіт. 2010, Поляна, Україна / Нац. ун-т «Львів. політехніка». – Л. : Вежа і Ко, 2010. – С. 123– 130.

136 Izhikelich's model of spiking neurons / Yu. Paterega // 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.

137 Mathematical models of spiking neurons / P. V. Tymoshchuk, Yu. I. Paterega // 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.

138 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;

139 Paterega I. Artificial evolution mechanisms in robot navigation. 2011 11th Intern. Conf. “The Experience of Designing and Application of CAD Systems in Microelectronics” CADSM 2011, 23-25 Febr. 2011. P. 281–286.

ДОДАТОК А

АКТИ ВПРОВАДЖЕННЯ

"Затверджую"
Перший проректор
Національного університету
"Львівська політехніка"
_____ проф. Олег МАТВІЙКІВ
"30" 04 2024 р.

АКТ
про впровадження в навчальний процес результатів
дисертаційної роботи
Патереги Юрія Ігоровича

Цей акт складено про те, що результати дисертаційної роботи Патереги Юрія Ігоровича впроваджено у навчальний процес кафедри "Систем автоматизованого проєктування" Національного університету "Львівська політехніка".

Впровадження результатів дисертаційної роботи полягає в їхньому використанні при викладанні навчальних дисциплін як окремих розділів лекційних курсів, так і в циклах лабораторних робіт.

Зокрема для викладання дисципліни "Методи багатокритеріальної оптимізації" для студентів освітньо-кваліфікаційного рівня "бакалавр", що навчаються за напрямом 122 "Комп'ютерні науки" використано такі результати:

- метод оптимізації процесу класифікації персоналізованих даних за рахунок введення етапу аугментації;
- метод оптимізації процесу опрацювання персоналізації даних особи за рахунок введення ансамблю моделей класифікації та ансамблевого голосування.

У лекційному курсі "Проектування інформаційних систем" для студентів освітньо-кваліфікаційного рівня "бакалавр", що навчаються за напрямом 122 "Комп'ютерні науки" використано такі результати:

- архітектура інформаційної системи опрацювання персоналізованих даних для аналізу стану особи;
- модель інформаційної технології опрацювання персоналізованих даних для аналізу стану особи.

Завідувач кафедри САП,
д.т.н., професор _____ Михайло ЛОБУР

Директор інституту ІКНІ,
д.т.н., професор _____ Микола МЕДИКОВСЬКИЙ

ЗАТВЕРДЖУЮ

Проректор з наукової роботи
Національного університету
«Львівська політехніка»
д.т.н., проф. _____ Іван ДЕМИДОВ
16 04 2024 р.

АКТ
про використання наукових результатів дисертаційної роботи Патереги Юрія Ігоровича на тему «Інформаційна технологія опрацювання персоналізованих даних для аналізу стану особи» представленої на здобуття наукового ступеня кандидата технічних наук, 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»

Комісія в складі голови - начальника НДЧ, д.т.н., Небесного Р.В. та членів: завідувача відділу науково-організаційного супроводу наукових досліджень, к.т.н. Лазько Г.В., в.о. заступника начальника планово-фінансового відділу Фаст І.І. та завідувача кафедри систем штучного інтелекту д.т.н., професора Шаховської Н.Б. цим актом підтверджують, що результати дисертаційної роботи Патереги Ю.І. використані 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 № 0120U100025). Зокрема, розроблено формалізовану модель інформаційної технології опрацювання персоналізованих даних для аналізу стану особи, що забезпечує пошук рішень щодо покращення процесу ідентифікації стадії захворювання.

Голова комісії:

Начальник НДЧ
д.т.н., ст. дослідник — Роман НЕБЕСНИЙ

Члени комісії:

завідувач відділу науково-організаційного супроводу наукових досліджень, к.т.н. — Галина ЛАЗЬКО

в.о. заст. нач. ПФВ — Ірина ФАСТ

зав. каф. СШІ
д.т.н., проф. — Наталя ШАХОВСЬКА

ЗАТВЕРДЖУЮ
Голова «Львівської асоціації
алергологів, імунологів,
імунореабілітологів»
ЧОПЯК_Валентина Володимирівна
«19» 01 2024 р.

АКТ

про впровадження результатів дисертаційної роботи дисертаційної роботи
на тему «Інформаційна технологія опрацювання персоналізованих даних для аналізу стану особи»
представленої на здобуття наукового ступеня кандидата технічних наук,
аспіранта кафедри «Системи автоматизованого проектування»
Національного університету «Львівська політехніка»
Патереги Юрія Ігоровича

Цей акт підтверджує, що результати дисертаційної роботи Патереги Ю.І. були використані для розроблення інформаційної системи підтримки прийняття рішень для лікування хворих з орфанними хворобами «Реєстру пацієнтів зі спадковим ангіоневротичним набряком (САН)» в рамках виконання Госпдоговору №01-2023 від 30.10.2023 р.

Впровадження дисертаційних досліджень Патереги Ю.І. полягає у наступному: створено комплекс методів, алгоритмів і програм, які покладені в основу функціонування інформаційної технології опрацювання персоналізованих даних для аналізу стану особи; розроблено алгоритм обробки персоналізованих даних особи для аналізу стану особи, що дає змогу формалізувати процес підготовки даних пацієнтів з різною патологією.

Експерт з питань імунології та алергології департаменту охорони здоров'я ЛОДА _____ ЛІЩУК-ЯКИМОВИЧ Х.О.

Завідувач поліклінічного відділення Регіонального центру алергології та клінічної імунології _____ БІЛЯНСЬКА Л.М.

Завідувач відділу лабораторної діагностики регіонального центру алергології та клінічної імунології _____ МАРІТЧАК Н.В.