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

Кваліфікаційна наукова
праця на правах рукопису

Мельникова Наталія Іванівна

УДК 004.89: 002.53

ДИСЕРТАЦІЯ

МОДЕЛІ ТА МЕТОДИ ПІДТРИМКИ ПЕРСОНАЛІЗОВАНИХ РІШЕНЬ У МЕДИЧНИХ СИСТЕМАХ

05.13.23. - системи та засоби штучного інтелекту

Подається на здобуття наукового ступеня
доктора технічних наук

Дисертація містить результати власних досліджень. Використання ідей, результатів і текстів інших авторів мають посилання на відповідне джерело _____ /Н.І. Мельникова/

Науковий консультант Шаховська Наталія Богданівна
професор, д.т.н.

Львів 2023

АНОТАЦІЯ

Мельникова Н.І. Моделі та методи підтримки персоналізованих рішень у медичних системах. – На правах рукопису.

Дисертаційна робота на здобуття наукового ступеня доктора технічних наук за спеціальністю 05.13.23 – системи штучного інтелекту. – Національний університет «Львівська політехніка» Міністерства освіти і науки України, Львів, 2023.

У дисертаційній роботі вирішено важливу науково-прикладну проблему розроблення та удосконалення моделей, методів і засобів машинного навчання в задачах класифікації, кластеризації, прогнозування та візуалізація результатів опрацювання персональних даних для адаптації медичних рішень до пацієнта.

Об’єктом дослідження є процес обробки та аналізу персоналізованих медичних даних.

Предметом дослідження є методи прийняття рішень, методи штучного інтелекту, зокрема машинного навчання для аналізу персоналізованих медичних даних.

Методи дослідження. Для досягнення поставленої мети використано: методи прийняття рішень – для забезпечення чіткості та напрамленості та у пошуку рішень стосовно вибору цільових схем лікування; теорію реляційних баз даних – для відображення простору умов щодо ідентифікації стану хворого та урахування додаткових вимірів та параметрів; систему операцій реляційної алгебри – для формалізації залежностей між множинами параметрів пацієнта та показників оцінки його стану; методи штучного інтелекту – для виявлення закономірностей та залежностей між цільовими змінними, кластеризації груп пацієнтів за відповідними ознаками, для усунення невизначеностей у просторі даних, для прогнозування цільових змінних; методи об’єктно-орієнтованого аналізу і проектування – для визначення семантичних зв’язків між джерелами даних; методи консолідації даних – для агрегації даних з різною структурою; методи моделювання інформаційних систем.

У роботі введено поняття персоналізації, персоналізованих рішень та

медичних даних пацієнта. Проаналізовано правові вимоги щодо використання персоналізованих медичних даних, обумовлених дотриманням набору правил міжнародних стандартів і внутрішніх законодавчих вимог, які захищають призначені для користувача дані та створюють їхню прозорість. Проаналізовано існуючі практичні рішення, що відображають основні підходи до використання медичних даних та урахування персоналізації за результатами аналізу існуючих систем штучного інтелекту у сфері медицини. Проведено порівняльний аналіз класичних методів та визначені обмеження, які характерні при опрацюванні медичних персоналізованих даних. Визначені найбільші обмеження під час опрацювання універсальними методами великих та малих мультимодальних медичних наборів даних, що дало змогу визначити актуальну науково-прикладну проблему розробки нових чи удосконалення існуючих методів штучного інтелекту, які підвищують точність обробки великих та малих мультимодальних медичних даних для пошуку персоналізованих рішень, та сформувати завдання наукового дослідження.

У межах дисертаційної роботи введено модель відображення стану пацієнта, яка спрощено відображає його структуру, подає інформацію про його стан і поведінку, та представлена як система, що консолідує різні елементи, представлені у вигляді множин, які взаємно залежні та залежні від середовища оцінки; визначено, що продукційні правила, які формулюють рішення щодо перегляду та зміни тактики лікування; представлено простір стану пацієнта, як евклідовий простір, що дозволило змоделювати інформаційну модель простору станів, як багатовимірну систему в тимчасовій області; формалізовано відображення фізичного стану пацієнта з урахуванням часовозалежних та часовонезалежних даних пацієнта, що дає можливість оцінити його в певний момент часу та приймати сталі показники стану; розроблена модель простору станів пацієнта представлена у вигляді куба як відображення функціонального відношення загального стану пацієнта. Удосконалено метод пошуку зміни стану пацієнта, шляхом використання простору умов та аналізу зміни приросту значень часовозалежних даних. На відміну від методу упорядкованого пошуку він надає чіткості та направленості у

пошуку рішень при виборі цільових схем лікування, це дає змогу зменшити ймовірність появи похибки (кількість ліжкоднів) при виборі схеми лікування, за рахунок агрегованого опрацювання даних з різних просторів моделі.

Проведено аналіз опрацювання медичних персоналізованих даних та досліджено процес прогнозування, визначено необхідність застосувати декілька кроків попередньої обробки отриманих даних. Досліджено особливості та принципи видобутку даних за допомогою інструментів їх аналізу, задля пошуку попередньо невідомої закономірності і зв'язків у даних, що можуть бути використані як валідна інформація; визначені типи класифікації завдань видобутку даних, що орієнтовані на передбачення даних та описові завдання. Охарактеризовано інструменти для очистки, препроцесингу, аналізу даних і виявлення взаємозалежностей між даними, що відіграють важливу роль при досягненні успіху в системах діагностування та прогнозування. Запропоновано варіанти використання методів машинного навчання для їхнього застосування при різній структурі медичних даних. Запропоновано твердження щодо вибору ефективних моделей ансамблів методів для пошуку залежностей між параметрами пацієнтів та факторами і показниками появи хвороби. З метою підвищення точності процесу узагальнення результатів на малих наборах даних на 4% у порівнянні з алгоритмами Random Forest та XGBoost, розроблено метод двоетапної обробки даних на основі ієрархічного предиктора. Розроблено метод групування моделей методів машинного навчання, за рахунок застосування стекінгової моделі для підвищення точності прогнозування даних і паралельної обробки даних як малої, так і великої розмірності, використовуючи Random Forest як метаалгоритм, що на відміну від подібних моделей, базується на деформації метаознак та повторному навчанні на розширеному наборі даних. Розроблено гібридний ансамбль моделей, що містить кілька селекторів за допомогою агрегування результатів, які будуть використовуватися на етапі попередньої обробки. Вбудовані алгоритми виконують вибір ознак під час процедури навчання класифікатора, і оптимізують набір ознак, що використовуються для досягнення вищої точності. Розроблено метод зменшення розмірності вхідних даних на основі ансамблю моделей слабких

предикторів для вибору найважливіших ознак, що базується на прирості інформації з урахуванням результатів застосування декількох селекторів та агрегації кінцевих результатів на підставі урахування індекса Жакара для оцінки подібності одержаних підмножин ознак та проведення мажоритарного голосування результатів. Розроблено модель тришарового групування ансамблю методів та описано етапи укладання, що забезпечують можливість об’єднати асоціативну класифікацію зі слабкими класифікаторами в ансамбль для узагальнення результатів. Модель класифікації ансамблю трирівневого стекування демонструє високу точність для інтелектуального аналізу коротких наборів медичних даних, асоціативні правила разом зі слабкими предикторами покращують якість класифікації. Розроблений ансамбль використовує модель Випадкового лісу як агрегатора для узагальнення результатів слабких регресорів; розроблена трирівнева класифікаційна модель ансамблю стекінгу з агрегатом логістичної регресії, яка має такі значення метрик оцінки продуктивності у вибраній підмножині ознак. Проведено порівняння точністі прогнозування для стандартних моделей зменшення розмірності під час наступних кроків, для цього був використований аналіз головних компонентів.

Запропоновано формулювання аномалій, де означено, що визначення аномалій залежить від контексту, де найважливішим значенням необхідно враховувати показник часу. Визначені типи аномалій та шляхи їхнього пошуку, що дають можливість диференціації підходів щодо аналізу персоналізованих медичних даних та визначені особливості їхнього застосування. Визначені основні проблеми, які виникають при обробці медичних персоналізованих даних, за рахунок чого постає проблема розробки ефективної стратегії аналізу даних, яку можна застосувати до розподілених баз даних різних доменів. Розроблено метод заповнення пропусків даних на основі ймовірних продукційних залежностей, що збільшує стійкість моделі до помилок у даних та забезпечення аналізу мультимодальних даних при паралельній реалізації в розподілених базах даних. Розроблено два алгоритми PPD майнінгу та заповнення відсутніх медичних персоналізованих даних, що забезпечують збереження характеристик стійкості до

помилок у даних, можливість паралельної реалізації в розподілених базах даних, автоматизація та виконання аналізу різних типів даних.

Розроблено архітектуру системи підтримки прийняття медичних рішень щодо прогнозування станів пацієнта на підставі опрацювання та аналізу персоналізованих медичних даних; подано опис імплементації рішень у системі підтримки прийняття рішень для лікування хворих на ПІД з порушеннями антитілоутвореннь. Використання цієї системи дає змогу вести облік хворих, контролювати та супроводжувати їхнє лікування, досліджувати динаміку перебігу хвороби та змін у стані пацієнта. Подано опис імплементації програмного модуля щодо прогнозування динаміки поширення COVID-19 за урахуванням процесу класифікації пацієнтів відповідно до стану, проведення оцінки помилок одержаних рішень. Представлено результати розробки програмного модулю щодо прогнозування динаміки поширення COVID-19, що дозволяє аналізувати дані пацієнтів і на їх основі визначати результат про наявність або відсутність COVID у пацієнта, а також представлено результати імплементації системи для виявлення помилок, що дає змогу експериментувати із різними параметрами для моделі, знову і знову на одних і тих же самих даних.

Практичне значення результатів дисертаційного дослідження дають змогу підвищити точність прогнозування цільових показників в підмножині простору умов на 5%, підвищити точність класифікації на 4% порівняно з результатами логістичної регресії як кращого класифікатора у порівнянні з існуючими для набору даних по COVID-19, підвищити точність прогнозування даних на 7-9% та забезпечити паралелізацію процесу обробки даних як малої, так і великої розмірності, забезпечити заповнення пропусків даних за рахунок одержання додаткових значень а також зменшити ймовірність появи похибки (кількість ліжкоднів) при пошуку шаблонів динаміки стану пацієнта шляхом використання простору умов.

Достовірність наукових та практичних результатів підтверджується за рахунок відповідних матеріалів про впровадження дисертаційних досліджень, а також за рахунок порівняння одержаних практичних результатів з результатами

застосування існуючих класичних методів та підходів щодо опрацювання персоналізованої інформації про пацієнта.

Дисертація виконана в межах науково-дослідних робіт у рамках проєкту НДДКР: «Інформаційна технологія опрацювання персоналізованої медичної інформації» № 0119U002257. 2019-2020; госпдоговору ТОВ «БІОФАРМА ПЛАЗМА», розробки інформаційної системи обліку пацієнтів з орфанними хворобами «Реєстр первинних імунодефіцитів», 2019 – 2021; проєкту НДФ щодо наукових досліджень і розробок «Система підтримки прийняття рішень моделювання поширення вірусних інфекцій» № 211/01.220 ННВЦ від 06.11.2020 р. 2020 – 2021; проєкту НДДКР «Технології та системи оброблення і зберігання персоналізованих військових медичних даних» № 0121U107809, 2021 – 2022; проєкту НДДКР «Розроблення інформаційної технології оцінювання та прогнозування надійності програмного забезпечення методами машинного навчання» № 0121U109527, 2021 – 2022.

Розроблені в дисертаційному дослідженні концепції, методи та ансамблі моделей прогнозування даних для оцінки стану хворого доведено до практичної реалізації інформаційної системи підтримки прийняття рішень для лікування хворих з орфанними хворобами використано ТОВ «БІОФАРМА ПЛАЗМА» та програмного модулю щодо прогнозування динаміки поширення COVID-19 у межах України, Білорусії, Німеччини використано в дослідженнях Львівського медичного університету ім. Д. Галицького, програмний модуль для розв’язання задачі класифікації щодо виявлення наявності COVID-19 у пацієнта, використано під час аналізу стану хворого для призначення чи корекції схеми лікування в Клінічній лікарні швидкої допомоги, Першому медичному об'єднанні міста Львова.

Результати роботи включені у навчальний процес студентів спеціальності 122 «Комп’ютерні науки», а саме в дисципліни «Дискретна математика», «Організація баз даних та знань» за рахунок використання навчальних посібників: Бойко Н. І. Людино-машинна взаємодія в системах штучного інтелекту: навч. посіб. / Н. І. Бойко, О. Б. Вовк, Н. Б. Шаховська, Н. І. Мельникова, Ю. П. Кривенчук. – Львів: Видавництво Тараса Сороки, 2018. – 248 с.; Журавчак Л.М., Мельникова

Н.І., Сердюк П.В. Практикум з комп’ютерної дискретної математики: навч. посібник / Л.М. Журавчак, Н.І. Мельникова, П. В. Сердюк. – Львів : Видавництво Львівської політехніки, 2019. – 279 с. ISBN 978-966-3466-85-9; Василюк А. С. Комп'ютерна графіка : навч. посіб. [для студентів напряму підгот. 6.040303 "Систем. аналіз"] / А. С. Василюк, Н. І. Мельникова ; М-во освіти і науки України, Нац. ун-т «Львів. політехніка». – Львів: Вид-во Львів. політехніки, 2016. – 308 с.

Ключові слова: персоналізовані медичні рішення, ієрархічний предиктор, залежності між цільовими змінними, стекінгова модель, заповнення відсутніх даних, пошук зміни стану, модель вибору ознак, великі набори даних, консолідація мультимодальних даних, інформаційна модель стану пацієнта.

ABSTRACT

Melnykova N.I. Models and methods and principles for supporting personalized solutions in medical systems. – On manuscript rights.

Dissertation for obtaining a technical science doctorate on specialty 05.13.23 – artificial intelligence systems. – Lviv Polytechnic National University, Ministry of Education and Science of Ukraine, Lviv, 2023.

The thesis solves the current scientific and applied problem of developing and improving models, methods and means of machine learning (classification, clustering, forecasting) and visualization of the results of personal data processing for adapting medical solutions to the patient.

The object of research is the process of processing and analyzing personalized medical data.

The subject of research is decision-making methods, artificial intelligence methods, in particular machine learning for analyzing personalized medical data.

Research methods. To achieve the goal, the following decision-making methods were used - to ensure clarity and directness and in the search for solutions regarding the selection of targeted treatment regimens; the theory of relational databases - to display the space of conditions for identifying the patient's condition and taking into account additional dimensions and parameters; the system of operations of relational algebra - for formalizing dependencies between sets of patient parameters and indicators of his condition assessment; methods of artificial intelligence - to identify regularities and dependencies between target variables, clustering groups of patients according to relevant characteristics, to eliminate uncertainties in the data space, to predict target variables; methods of object-oriented analysis and design - to determine semantic relationships between data sources; data consolidation methods – for aggregating data with different structures; methods of modeling information systems.

The work introduces the concepts of personalization, personalized solutions and patient medical data. The legal requirements for the use of personalized medical data are analyzed and are determined by compliance with a set of rules of international standards and domestic legal requirements that protect user data and create their transparency. The

existing practical solutions reflecting the main approaches to the use of medical data and taking into account personalization and based on the results of the analysis of existing artificial intelligence systems in the field of medicine were analyzed. A comparative analysis of classical methods was carried out and the limitations that are characteristic of the processing of personalized medical data were determined. The biggest limitations during the processing of large and small multimodal medical data sets by universal methods have been identified, which have been identified as an urgent scientific and applied problem in the development of new or improvement of existing artificial intelligence methods in order to increase the accuracy of the process of processing large and small multimodal medical data to find personalized solutions and to formulate tasks scientific research.

Within the scope of the dissertation, an information model of the patient's condition was introduced, which simply reflects its structure, provides information about its condition and behavior, and is presented as a system that consolidates various elements, presented in the form of sets that are mutually dependent and dependent on the assessment environment; it is determined that the production rules, which formulate decisions regarding the review and change of treatment tactics; the space of the patient's state is presented as a Euclidean space, which made it possible to simulate the information model of the space of states as a multidimensional system in the time domain; the display of the patient's physical condition is formalized, taking into account the time-dependent and time-independent data of the patient, which makes it possible to evaluate him at a certain point in time and take constant indicators of the condition; the proposed information model of the space of the patient's states is presented in the form of a cube, as a reflection of the functional relationship of the patient's general state; a pattern search method is proposed, which is based on a modification of the method of associative rules, which allows reducing the workload of the doctor and using parallel and distributed mode for calculation, and is an improvement of the ordered search method, and provides clarity and directness in the search for decisions regarding the selection of target treatment regimens, which allows to reduce the probability of an error when choosing a treatment scheme.

The analysis of the processing of medical personalized data was carried out and the forecasting process was investigated, and the need to apply several steps of preprocessing of the received data was determined; the features and principles of data extraction using data analysis tools were investigated in order to find previously unknown regularities and connections in the data that can be used as valid information, in addition to using the extracted information to build a predictive model; defined types of classification of data mining tasks focused on data prediction and descriptive tasks; characterized tools for cleaning, preprocessing, data analysis and detection of interdependencies between data play an important role in the success of diagnostic systems and forecasting; proposed options for using machine and deep learning methods for their use in different structures of the patient's medical data; statements regarding the selection of effective models of ensembles of methods for searching for dependencies between patient parameters and factors and indicators of the appearance of the disease are proposed; a hierarchical predictor is proposed, which includes two-stage processing of small data sets using object clustering methods and prediction for each resulting cluster, which improves the model's robustness to new input data and provides 4% higher accuracy compared to the Random Forest and XGBoost algorithms; proposed a stacking model based on machine learning algorithms, which uses Random Forest as a meta-algorithm, and which, unlike similar models, is based on deformation of meta-features and retraining on an extended data set, which provides increased accuracy of data prediction and parallel processing data of both small and large dimensions.

A hybrid ensemble feature selection model containing multiple selectors by aggregating the results to be used in the preprocessing step is proposed, the built-in algorithms perform feature selection during the classifier training procedure, and they clearly optimize the set of features used to achieve better accuracy; an ensemble of machine learning models was developed for the selection of priority features on large data sets, which consists of classifiers, associative rules and a generalized rank of features based on the Jacquard index, which allows avoiding the correlation of features and increases the generalization of the model; a model of three-layer stacking of an ensemble of methods is proposed and the stages of stacking are described, which provides an

opportunity to combine associative classification with weak classifiers into an ensemble to generalize results; a three-level stacking ensemble classification model demonstrates high accuracy for intelligent analysis of short medical data sets. Associative rules together with weak predictors improve the classification quality. The proposed ensemble uses a random forest model as an aggregator to generalize the results of weak repressors; a three-level classification model of the stacking ensemble with a logistic regression aggregate was developed, which has the following values of performance evaluation metrics in the selected subset of features; a comparison of the prediction accuracies of standard dimensionality reduction models in subsequent steps was performed using principal component analysis with eight components.

The formulation of anomalies is proposed, where it is specified that the definition of anomalies depends on the context, where the most important value must be the time indicator, because the context will depend on it; defined types of anomalies and ways of their search, which makes it possible to differentiate approaches to the analysis of personalized medical data and defined features of their application; the main problems that arise in the processing of personalized medical data are identified, due to which the problem of developing an effective data analysis strategy that can be applied to distributed databases of various domains arises; a method of imputation of missing data based on probable production dependencies is proposed, which increases the model's resistance to data errors and provides analysis of multimodal data during parallel implementation in distributed databases; two PPD algorithms for mining and imputation of missing medical personalized data are proposed, which ensure the preservation of the characteristics of resistance to errors in the data, the possibility of parallel implementation in distributed databases, automation and analysis of various types of data.

The architecture of the medical decision-making support system for predicting the patient's condition based on the processing and analysis of personalized medical data is proposed; a description of the implementation of decisions in the decision-making support system for the treatment of PID patients with antibody formation disorders is presented. The use of this system makes it possible to keep records of patients, monitor and accompany their treatment and study the dynamics of the course of the disease and

changes in the patient's condition; a description of the implementation of the software module for forecasting the dynamics of the spread of COVID-19, taking into account the process of classifying patients according to the condition, evaluating the errors of the received decisions, is presented; presented the results of the development of a software module for forecasting the dynamics of the spread of COVID-19, which allows analyzing patient data and based on them to give a result about the presence or absence of covid in the patient; presented the results of implementing an error detection system that allows you to experiment with different parameters for the model, over and over again on the same data. With this, the user will be able to choose the model that best describes the data and has the best results. That the shorter the time interval between the points, the more accurate the model. It can recognize the time when the anomaly occurred is more accurately, but high accuracy requires more calculations, and with the help of these graphs, the user will be able to determine the optimal option.

The dissertation work was carried out within the scope of research works within the: R&D Project: "Information technology for processing personalized medical information" No. 0119U002257. 2019-2020; Government contract of BIOPHARMA PLAZMA LLC, development of the information system for recording patients with orphan diseases "Register of primary immunodeficiencies". 2019-2021; NDF project on scientific research and development "Decision-making support system for modeling the spread of viral infections" No. 211/01.220 NNVC dated November 6, 2020, 2020-2021; R&D project: "Technologies and systems for processing and storing personalized military medical data" No. 0121U107809. 2021-2022; R&D project: "Development of information technology for evaluating and predicting software reliability using machine learning methods" No. 0121U109527. 2021-2022.

The concepts, methods and ensembles of data forecasting models developed in the dissertation study for assessing the patient's condition were brought to the practical implementation of the decision-making support information system for the treatment of patients with orphan diseases. BIOPHARMA PLAZMA LLC and, a software module for forecasting the dynamics of the spread of COVID-19 within the limits of Ukraine, Belarus, and Germany was used in the studies of the Lviv Medical University named after

D. Halytskyi, the software module for analyzing the presence of an ischemic stroke and determining the time from its appearance, the software module for solving the classification problem for detecting the presence of COVID-19 in a patient was used during the analysis of the patient's condition during treatment and correction of the prescription of the treatment regimen in the Clinical emergency hospital, the First Medical Association of the city of Lviv.

The results of the work were included in the educational process of students of the specialty 122 "Computer Science", namely in the disciplines "Discrete Mathematics", "Organization of Databases and Knowledge" through the use of educational aids: Boyko N. I. Human-machine interaction in systems artificial intelligence: education manual / N. I. Boyko, O. B. Vovk, N. B. Shakhovska, N. I. Melnikova, Yu. P. Krivenchuk. – Lviv: Taras Soroka Publishing House, 2018. – 248 c.; L.M. Zhuravchak, N.I. Melnikova, P.V. Serdyuk Workshop on computer discrete mathematics: teaching. manual / L.M. Zhuravchak, N.I. Melnikova, P. V. Serdyuk. – Lviv: Publishing House of Lviv Polytechnic, 2019. – 279 p. ISBN 978-966-3466-85-9; Vasylyuk A. S. Computer graphics: teaching. manual [for students of the training course. 6.040303 "System analysis"] / A. S. Vasyliuk, N. I. Melnikova; Ministry of Education and Science of Ukraine, National Lviv Polytechnic University. – Lviv: View of Lviv Polytechnics, 2016. – 308 p.

Keywords: personalized medical decisions, hierarchical predictor, dependencies between target variables, stacking model, missing data imputation, patient's state change search, feature selection model, big and small data sets, multimodal data consolidation, information model patient's state.

СПИСОК ПУБЛІКАЦІЙ ЗДОБУВАЧА

1. Melnykova, Nataliia, et al. ‘Data-Driven Analytics for Personalized Medical Decision Making’. Mathematics, vol. 8, no. 8, July 2020, p. 1211, https://doi.org/10.3390/math8081211. (індексована в наукометричній базі Scopus та Web of Science Core Collection, квартиль Q1 відповідно до класифікації SCImago Journal);
2. Melnykova, Nataliia, et al. ‘Personalized Data Analysis Approach for Assessing Necessary Hospital Bed-Days Built on Condition Space and Hierarchical Predictor’. Big Data and Cognitive Computing, vol. 5, no. 3, Aug. 2021, p. 37, https://doi.org/10.3390/bdcc5030037. (індексована в наукометричній базі Scopus та Web of Science Core Collection, квартиль Q1 відповідно до класифікації SCImago Journal);
3. Melnykova, Nataliia, et al. ‘The Hierarchical Classifier for COVID-19 Resistance Evaluation’. Data, vol. 6, no. 1, Jan. 2021, p. 6., https://doi.org/10.3390/data6010006. (індексована в наукометричній базі Scopus та Web of Science Core Collection, квартиль Q2 відповідно до класифікації SCImago Journal);
4. Melnykova, Nataliia, et al. ‘An Ensemble Methods for Medical Insurance Costs Prediction Task’. Computers, Materials and Continua, vol. 70, no. 2, 2022, pp. 3969 – 84, https://doi.org/10.32604/cmc.2022.019882. (індексована в наукометричній базі Scopus та Web of Science Core Collection, квартиль Q2 відповідно до класифікації SCImago Journal);
5. Melnykova, Nataliia, et al. ‘The Ensembles of Machine Learning Methods for Survival Predicting after Kidney Transplantation’. Applied Sciences, vol. 11, no. 21, Nov. 2021, p. 10380, https://doi.org/10.3390/app112110380. (індексована в наукометричній базі Scopus та Web of Science Core Collection, квартиль Q2 відповідно до класифікації SCImago Journal);
6. Мельникова, Н. І. ‘Особливості опрацювання медичної інформації для систем підтримки прийняття лікувальних рішень’. Вісник Національного університету Львівська політехніка. Інформаційні системи та мережі, no. 832, pp.190 – 204;

7. Шаховська Н. Б., Мельникова Н. І. ‘Методи побудови моделі поведінки користувачів’. Український журнал інформаційних технологій, 2020, no. 2, vol. 1, pp. 43–51;
8. Мельникова Н.І., та ін. ‘Розроблення інформаційної технології опрацювання персоналізованих медичних даних’. Вісник Національного університету «Львівська політехніка». Інформаційні системи та мережі: збірник наукових праць, 2015, no. 814, pp. 90 – 99.;
9. Shakhovska, N. B., and Melnykova. N. I. ‘Нові Методи Та Рішення Щодо Побудови Моделі Поведінки Користувачів’. Scientific Bulletin of UNFU, vol. 30, no. 5, Nov. 2020, pp. 76 – 83., https://doi.org/10.36930/40300513.;
10. Кривенчук Ю. П., Шаховська Н. Б., Вовк О. Б., Мельникова Н. І. ‘Комп’ютерне моделювання функцій перетворення оптичних схем засобу вимірювання температури, побудованого на ефекті рамана та структура алгоритму їх дослідження’. Радіоелектроніка, інформатика, управління, 2018, no. 346, pp. 25 – 33. DOI:10.15588/1607-3274-2018-3-3. (індексована в наукометричній базі Web of Science Core Collection);
11. Melnykova Nataliia, Basystiuk Oleh. ‘Multimodal Speech Recognition Based on Audio and Text Data’. Вісник ХНУ. Серія: технічні науки, 2022, no 3.;
12. Мельникова Н. І., Поберейко П. Б. ‘Дослідження методів пошуку ключових кадрів у відеопотоці з використанням нейронних мереж для систем пошуку’ Вісник Хмельницького національного університету. Серія: Технічні науки, 2022. vol. 3, no. 309. pp. 55 – 61.;
13. Melnykova, Nataliia, et al. Specifics personalized approach in the analysis of medical information. Econtechmod: an international quarterly journal on economics in technology, new technologies and modelling processes. Lublin. Rzeszow, 2016. vol. 5, no. 2. pp. 109 – 116.;
14. Kryvenchuk, Y., Shakhovska, N., Melnykova, N., and Boichuk, A. ‘Organization of the network connection in the Industry 4.0’. Econtechmod: An International Quarterly Journal on Economics of Technology and Modelling Processes, 8, 2019, pp. 39 – 45;

15. Melnykova N. ‘Using of personalized approach for assessment of the financial condition of the company’ Econtechmod: an international quarterly journal on economics in technology, new technologies and modelling processes. Lublin, Rzeszow, 2017, vol. 6, no 2, pp. 39 – 44;
16. Melnykova N. ‘Analysis of the Data Mining and Classification of Patients’ States’. Manažérska Informatika, 2020, no. 2 https://manazerskainformatika. sk/analysis-of-the-data-mining-and-classification-of-patients-states/;
17. Melnykova N. ‘Method for Clustering and Determining the Average Distance between Clusters’. Manažérska Informatika, 2021, no. 2, https://manazerskain formatika.sk/method-for-clustering-and-determining-the-average-distance-between-clusters/;
18. Melnykova N. ‘A New Approach to Modelling the Nature of Individual Morbidity Using Partial Functional Dependencies’. Manažérska Informatika, 2021, no. 1, https://manazerskainformatika.sk/a-new-approach-to-modelling-the-nature-of-individual-morbidity-using-partial-functional-dependencies/;
19. Melnykova N. ‘Model of States Warehouse of a State of the State of the Object and Personalized Decisions’. Manažérska Informatika, 2022, no. 1, https://manazerskainformatika.sk/model-of-states-warehouse-of-a-state-of-the-state-of-the-object-and-personalized-decisions/;
20. Мельникова Н. І., and Мельников В. А. ‘Персоналізований підхід до обробки та аналізу медичних даних пацієнтів’. Інформаційна безпека та інформаційні технології : монографія, Харків : Вид. Рожко С. Г., 2019., pp. 247 – 259;
21. Мельникова Н.І. ‘Методи оптимізації рішень щодо аналізу персоналізованих даних’. Кібербезпека та інформаційні технології: монографія. – Х.: ТОВ «ДІСА ПЛЮС», 2020, ISBN 978-617-7927-01-2, pp. 210 – 225;
22. Melnykova, Nataliia, et al. ‘Big Data Analysis in Development of Personalized Medical System’. Procedia Computer Science, vol. 160, 2019, pp. 229 – 34, https://doi.org/10.1016/j.procs.2019.09.461. (індексована в наукометричній базі Scopus та Web of Science Core Collection);

23. Melnykova, Nataliia, et al. ‘Using Big Data for Formalization the Patient’s Personalized Data’. Procedia Computer Science, vol. 155, 2019, pp. 624 – 29, https://doi.org/10.1016/j.procs.2019.08.088. (індексована в наукометричній базі Scopus та Web of Science Core Collection);
24. Melnykova, Nataliia, et al. ‘Anomalies Detecting in Medical Metrics Using Machine Learning Tools’. Procedia Computer Science, vol. 198, 2022, pp. 718 – 23, https://doi.org/10.1016/j.procs.2021.12.312. (індексована в наукометричній базі Scopus та Web of Science Core Collection);
25. Melnykova, Nataliia. ‘A Novel Approach for the Automatic Detection of COVID in a Patient by Using a Categorization Methods’. Procedia Computer Science, vol. 198, 2022, pp. 712 – 17, https://doi.org/10.1016/ j.procs.2021.12.311. (індексована в наукометричній базі Scopus та Web of Science Core Collection);
26. Melnykova, Nataliia. ‘Model of the system of personalized analysis of financial condition of the enterprise’. Advances in Intelligent Systems and Computing, 2018, no. 689, pp. 334-345, https://www.scopus.com/inward/record.uri?eid=2-s2.0-85036467510& doi=10.1007%2f978-3-319-70581-1_24&partnerID=40&md5= 909d5e2a66315 b464c19399e7482a86e. (індексована в наукометричній базі Scopus, квартиль Q3 відповідно до класифікації SCImago Journal);
27. Melnykova, Nataliia., et. al. ‘Smart Integrated Robotics System for SMEs Controlled by Internet of Things Based on Dynamic Manufacturing Processes’. Advances in Intelligent Systems and Computing, 2019, no. 871, pp. 535 – 549, https://doi.org/10.1007/978-3-030-01069-0_38. (індексована в наукометричній базі Scopus, квартиль Q3 відповідно до класифікації SCImago Journal);
28. Melnykova, Nataliia. ‘Semantic search personalized data as special method of processing medical information’. Advances in Intelligent Systems and Computing, 2017, no. 512, pp. 315 – 325, https://doi.org/10.1007/978-3-319-45991-2_22. (індексована в наукометричній базі Scopus та Web of Science Core Collection);
29. Melnykova, Nataliia., et al. ‘The special ways for processing personalized data during voting in elections’. Advances in Intelligent Systems and Computing, 1080 AISC,

2020, pp. 781 – 791, DOI: 10.1007/978-3-030-33695-0_52. (індексована в наукометричній базі Scopus);

30. Melnykova, Nataliia., et. al. ‘The personalized approach to the processing and analysis of patients’ medical data’. CEUR Workshop Proceedings, 2018, no. 2255, pp. 103 – 112, http://ceur-ws.org/Vol-2255/paper10.pdf. (індексована в наукометричній базі Scopus та Web of Science Core Collection);
31. Shakhovska, Nataliya. and Melnykova, Nataliia. ‘Feature Engineering and Missing Data Imputation Method of Medical Data Analysis’. CEUR Workshop Proceedings, 2022, no. 3137, pp. 48 – 57. http://ceur-ws.org/Vol-3137/paper4.pdf. (індексована в наукометричній базі Scopus та Web of Science Core Collection);
32. Melnykova, Nataliia., et. al. ‘The problem of analysing the relationships between individual characteristics of individuals with COVID`19’. CEUR Workshop Proceedings, 2020, no. 2753, pp. 473 – 482, https://pesquisa.bvsalud.org/global-literature-on-novel-coronavirus-2019-ncov/resource/pt/covidwho-984587. (індексована в наукометричній базі Scopus та Web of Science Core Collection);
33. Melnykova, Nataliia., et. al. ‘Advisory and accounting tool for safe and economically optimal choice of online self-education services’. CEUR Workshop Proceedings, 2019, 2588, pp. 290 – 300, http://ceur-ws.org/Vol-2588/paper24.pdf. (індексована в наукометричній базі Scopus та Web of Science Core Collection);
34. Melnykova, Nataliia., et. al. ‘Determination of characteristics discrete transfiguration for synthesized raster elements of non-regular structure’. CEUR Workshop Proceedings, 2019, no. 2533, pp. 249 – 258, http://ceur-ws.org/Vol-2533/paper23.pdf. (індексована в наукометричній базі Scopus та Web of Science Core Collection);
35. Melnykova, Nataliia., et. al. ‘Technologies of 3D-prototyping of Objects.’ CEUR Workshop Proceedings, 2019, no. 2533, pp. 271 – 281, http://ceur-ws.org/Vol-2533/paper25.pdf. (індексована в наукометричній базі Scopus та Web of Science Core Collection);
36. Melnykova, Nataliia. ‘Application of information technology for designing of treatment information systems’. 2015, Proceedings of 13th International Conference:

The Experience of Designing and Application of CAD Systems in Microelectronics, CADSM 2015, pp. 156 – 158, https://doi.org/10.1109/CADSM.2015.7230823. (індексована в наукометричній базі Scopus та Web of Science Core Collection);

37. Melnykova, Nataliia, and Oksana Markiv. ‘Semantic approach to personalization of medical data’. 2016 XIth International Scientific and Technical Conference Computer Sciences and Information Technologies (CSIT). IEEE, 2016., pp. 59 – 61, https://doi.org/10.1109/STC-CSIT.2016.7589868. (індексована в наукометричній базі Scopus та Web of Science Core Collection);
38. Melnykova, Nataliia. ‘The basic approaches to automation of management by enterprise finances’. 2017, Proceedings of the 12th International Scientific and Technical Conference on Computer Sciences and Information Technologies, CSIT 2017, pp. 288 – 291, https://doi.org/10.1109/STC-CSIT.2017.8098788. (індексована в наукометричній базі Scopus та Web of Science Core Collection);
39. Melnykova, Nataliia. et. al. ‘The personalized approach in a medical decentralized diagnostic and treatment’. 2017, 14th International Conference The Experience of Designing and Application of CAD Systems in Microelectronics, CADSM 2017, pp. 295 – 297, https://doi.org/10.1109/CADSM.2017.7916139. (індексована в наукометричній базі Scopus та Web of Science Core Collection);
40. Melnykova, Nataliia. et. al. ‘The new approaches of heterogeneous data consolidation’. 2018 International Scientific and Technical Conference on Computer Sciences and Information Technologies, 2018, pp. 408 – 411, https://doi.org/10.1109/STC-CSIT.2018.8526677. (індексована в наукометричній базі Scopus та Web of Science Core Collection);
41. Melnykova, Nataliia. et. al. ‘The special ways of application of neural networks for medical information processing, 2018 International Scientific and Technical Conference on Computer Sciences and Information Technologies, 2018, pp. 428 – 431, https://doi.org/10.1109/STC-CSIT.2018.8526708. (індексована в наукометричній базі Scopus та Web of Science Core Collection);
42. Melnykova, Nataliia. et. al. ‘Calculation of the Exact Value of the Fractal Dimension in the Time Series for the Box-Counting Method’. 2019 9th International

Conference on Advanced Computer Information Technologies, ACIT 2019, pp. 248 – 251, https://doi.org/10.1109/ACITT.2019.8780028. (індексована в наукометричній базі Scopus та Web of Science Core Collection);

43. Melnykova, Nataliia. et. al. ‘The Applying Processing Intelligence Methods for Classify Persons in Identify Personalized Medication Decisions’. 2020 10th International Conference on Advanced Computer Information Technologies, ACIT 2020, pp. 422 – 425, https://doi.org/10.1109/ACIT49673.2020.9208822. (індексована в наукометричній базі Scopus та Web of Science Core Collection);
44. Melnykova, Nataliia. et. al. ‘The Investigation of Artificial Intelligence Methods for Identifying States and Analyzing System Transitions between States’. 2020 International Scientific and Technical Conference on Computer Sciences and Information Technologies, 2020, vol. 1, pp. 41 – 75 https://doi.org/10.1109/CSIT49958.2020.9321841. (індексована в наукометричній базі Scopus та Web of Science Core Collection);
45. Melnykova, Nataliia. et. al. ‘The Scheme Application of the Medical Expert System’. XXIV Ukrainian-Polish Conference on CAD in Machinery Design. Implementation and Educational Issues, CADMD 2016, Lviv, 21.10-22.10.2016, pp. 65 – 66.;
46. Melnykova, Nataliia. et. al. ‘The Intelligent System Architecture of Personalized Management’. XXIV Ukrainian-Polish Conference on CAD in Machinery Design, Implementation and Educational Issues, CADMD 2016, Lviv, 21.10-22.10.2016, pp. 27 – 28.;
47. Мельникова Н.І., Жилко І.В. ‘Застосування хмарних обчислень для проектування систем підтримки лікарських рішень’. Міжнародна конференція «Інноваційні підходи і сучасна наука», Центр наукових публікацій, Київ, 30 квітня 2015 р., pp. 45– 46.;
48. Мельникова Н.І., Данилів В.М. ‘Інтелектуальна інформаційна система організації інтерактивних промоакцій’. Матеріалами міжнародної науково-практичної конференції: «IV осінні наукові читання», Київ: збірник статей (рівень стандарту, академічний рівень), Центр наукових публікацій, 2015. pp. 48 – 49.;

49. Мельникова Н.І., Копач М.І. ‘Методи оптимізації аналізу персоналізованих даних підприємства Математика’. Інформаційні технології. Освіта : тези доповідей VI Міжнародної науково-практичної конференції, Східноєвропейський національний університет ім. Лесі Українки, Луцьк-Світязь, 5 – 7 червня 2017, pp. 67 – 69.;
50. Мельникова Н.І., Мукалов П., Козій Д. ‘Особливості застосування нейронних мереж щодо опрацювання даних різного походження’. Матеріали VI Всеукраїнської науково-практичної конференції «Наукові дослідження: перспективи інновацій у суспільстві і розвитку технологій», Наукове партнерство «Центр наукових технологій», Харків, 24 – 25лютого 2017, pp. 104.;
51. Мельникова Н.І., Мельников В.А. ‘Персоналізований підхід до обробки та аналізу медичних даних пацієнтів’. Матеріали Міжнародної науково-практичної конференції «Інформаційна безпека та інформаційні технології», ХНЕУ імені Семена Кузнеця, 24 – 25 квітня 2019. pp. 56.;
52. Melnykova, Nataliia. ‘New approach to support of personalized decision making in medicine’. The 8 th International scientific and practical conference « Information, Its Impact on Social And Technical Processes, March Haifa, Israel,16 – 17, 2020, pp.261.;
53. Melnykova, Nataliia. et. al. ‘Semantic approach to personalization of medical data’. IX Міжнародна науково-практична конференція «Actual aspects of development in the context of globalization», Florecia, Italy, 23 – 24 March 2020, pp. 126 – 129.;
54. Мельникова Н.І. ‘Методи оптимізації рішень щодо аналізу персоналізованих даних’. Матеріали II Міжнародної науково-практичної конференції «Інформаційна безпека та інформаційні технології», Кропивницький: ЦНТУ, 2 – 3 квітня 2020, pp. 105.;
55. Melnykova, Nataliia., and Kolomyi, Anastasia. ‘Information System For Determination Of Early Symptoms Of Dementia Base On Mini-Cog Test And Mini-Mental State Examination Impact of modernity on science and practice’. XVIII International Scientific and Practical Conference, Boston, USA 2020. pp. 95 – 97.;

56. Melnykova, Nataliia. et. al. ‘Automatic audio to text conversion approaches along the radiology value chain’. XVII International Scientific and Practical Conference «Multidisciplinary academic notes. Theory, methodology and practice», Tokyo, Japan, 03 – 06 May 2022, pp. 994.

ЗМІСТ

ПЕРЕЛІК ПРИЙНЯТИХ СКОРОЧЕНЬ

ПМ – персоналізована медицина,

ПД – персоналізовані дані,

DDM – керування даними (data-driven medicine),

EMR/EHR – електронні медичні/охорони здоров’я записи (electrionic medical/healthcare records),

ШІ – штучний інтелект,

GDPR – міжнародне положення про захист даних,

ПІД – первинний імунодефіцит,

PPD – ймовірнісна продукційна залежність,

FD – функціональні залежності,

QMR – швидка медична довідка,

CD – нульова заміна,

AR – асоціативні правила.

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ

R – відношення, елементами якого є показники характеристики стану або продукційні правила рішень щодо стратегії пошуку оптимального стану;

A_{in} – множина часовонезалежних характеристик параметрів пацієнта або вхідних даних системи, що характеризують показники-фактори пацієнта, що отримують з динаміки змін його стану, які розраховуються на підставі інформації, яка міститься в історії хвороби;

A_t – множина часовозалежних параметрів пацієнта, що змінюються під впливом часу ;

a – параметри пацієнта;

Pd – множина, елементами якої є параметри стану пацієнта, а саме елементи множин часовонезалежних характеристик (A_{in}) та часовозалежних параметрів пацієнта (A_t);

ES – множина оцінок стану пацієнта, що залежить від коефіцієнтів оцінки;

k – коефіцієнт оцінки стану параметрів пацієнта;

K – множина коефіцієнтів оцінки стану пацієнта, що вказують на зміни результативного показника при зміні значення вхідної змінної;

e – оцінка стану параметрів пацієнта;

Fe – система оцінки стану пацієнта;

S – множина стратегічних рішень щодо оцінки стану пацієнта;

G – протокольні рішення;

P – стан пацієнта в просторі станів;

$P_o(t)$ – відображає часовий показник стану пацієнта на просторі станів;

$P_o\ (a)$ – параметричний показник стану пацієнта на просторі станів;

Ob – показник на окремі випадки, тобто множину індивідів (пацієнтів);

$\otimes$ – тензорний добуток просторів, який білінійно відображає вихідні простори станів;

ω – стан пацієнта.

ВСТУП

Актуальність теми. Персоналізована медицина з даними орієнтованими на пацієнта набуває все більшого значення з багатьох причин, серед яких: підвищення точності процесу обробки медичних даних для пошуку персоналізованих рішень, що впливає на забезпечення якісного та ефективного медичного обслуговування; можливість прогнозування зміни стану конкретного пацієнта тощо.

За результатами дослідження персональних даних пацієнта з точки зору протоколу лікування характеризуються викидами, шумами, пропусками даних, часовою залежністю та мультимодальністю. Важливим фактором є потреба опрацьовувати інформацію пацієнта, опираючись на протоколи хвороб, не лише однієї хвороби з якою звернувся хворий, але й усіх супутніх захворювань та індивідуальних особливостей його стану.

Існуючі прикладні системи, орієнтовані на забезпечення персоналізації рішень, в рамках виду робіт при опрацюванні індивідуальних медичних даних пацієнтів. Особливості таких систем характеризуються процесом прогнозування, яке спрямоване лише для одного пацієнта, а не для групи пацієнтів з однаковими параметрами, які у часі можуть змінюватися під впливом різних факторів, що значно знижує якість прийнятих персоналізованих медичних рішень.

На сьогодні існуючі підходи та рішення характеризуються стійкістю до шумів, розпаралельністю обробки даних, високою точністю і стабільністю, визначенням важливих ознак, знаходженням нелінійно роздільних кластерів, високою чутливістю та специфічністю у прогнозуванні, але це стосується лише класичних методів машинного навчання, які використовуються для вирішення такого типу задач, та які дають високі показники лише при розв’язанні окремих задач обробки персональних даних. Для комплексного підходу необхідно введення евристик, припущення про незалежність параметрів, робота не лише з бінарними ознаками об’єктів, пошук асоціативних залежностей з малою підтримкою, пошук оптимального рішення, підвищення точності прийняття лікарських рішень.

Наступна характерна особливість медичних даних – це або малі набори даних

(сотні екземплярів даних), наприклад, інформація про пацієнтів з орфанними хворобами, або великі набори даних (тисячі та сотні тисяч екземплярів даних), наприклад для пандемій, таких, як COVID-19. Іноді, залежно від типу захворювання та особливостей пацієнта кількість параметрів також може змінюватися від одиниць до сотень. При роботі з медичними даними, виникають проблеми як для малих наборів даних, так і для великих, а саме:

для малих – низька повторюваність результатів під час пошуку персоналізованих рішень;

для великих – необхідність відбору важливих ознак та висока обчислювальна складність.

Отже, опрацювання універсальними методами великих та малих наборів мультимодальних даних не дає змоги вирішити низку проблем, а саме:

- залежність якості моделі машинного навчання від результатів попередньої обробки даних для великих наборів даних;
- підвищення повторюваності результатів під час пошуку персоналізованих рішень на малих наборах даних;
- забезпечення генералізації при опрацюванні малої вибірки медичних даних слабкими предикторами.

Усе перераховане вище зумовлює потребу у розробленні нових методів пошуку персоналізованих рішень, які повинні враховувати наявність факту мультимодальності даних, необхідність заповнення даних, залежності між параметрами, забезпечувати точність прогнозування цільових даних. Існуючі методи штучного інтелекту вирішують лише окремі підзадачі опрацювання медичних даних.

Отже, проблема розроблення та удосконалення моделей, методів і засобів машинного навчання в задачах класифікації, кластеризації, прогнозування та візуалізація результатів опрацювання персональних даних для адаптації медичних рішень до пацієнта є актуальною науково-прикладною проблемою.

Зв'язок роботи з науковими програмами, планами та темами. Дисертаційна робота відповідає науковому напряму кафедри системи штучного інтелекту Національного університету, а саме: «Методи та засоби обробки, консолідації та аналізу персоналізованої медичної інформації», та є складовою частиною проєктів, які виконувалися в межах дербюджетних науково-дослідних робіт та грантових робіт: ДБ «Інформаційна технологія опрацювання персоналізованої медичної інформації» № 0119U002257 (керівник ДБ); ДБ «Технології та системи оброблення і зберігання персоналізованих військових медичних даних» № 0121U107809; ДБ «Розроблення інформаційної технології оцінювання та прогнозування надійності програмного забезпечення методами машинного навчання» № 0121U109527; Госпдоговір ТОВ «БІОФАРМА ПЛАЗМА», спрямований на розроблення інформаційної системи обліку та аналізу пацієнтів з орфанними хворобами «Реєстр первинних імунодефіцитів»; Проєкт Національного Фонду Досліджень України «Система підтримки прийняття рішень моделювання поширення вірусних інфекцій» № 211/01.220; Грант Central European Initiatives Extraordinary call «STOP COVID - 19»; Грант CELTIC EUROGIA PROPOSAL «Integrated care for next generation iCare4NextG».

Мета і завдання дослідження. Метою дисертаційної роботи є розроблення моделей, методів машинного навчання та засобів для підвищення прогнозованої точності та візуалізації результатів опрацювання персональних даних щодо оцінки стану хворого, що забезпечить якісний процес прийняття персоналізованих медичних рішень.

Мета дисертаційної роботи визначає необхідність вирішення таких задач для опрацювання персоналізованих медичних даних:

1. Провести аналіз проблеми опрацювання мультимодальних персоналізованих медичних даних і попередньої обробки даних та аналізу якості моделі даних;
2. Розробити інформаційну модель стану пацієнта у багатовимірному просторі умов, яка за рахунок додаткових вимірів та параметрів забезпечить прогнозування цільових змінних;
3. Розробити метод двоетапної обробки малих наборів даних, що забезпечуватиме

підвищення точності процесу узагальнення результатів на малих наборах даних;

4. Розробити метод групування моделей машинного навчання, який забезпечить прогнозування даних та паралелізацію процесу обробки даних при умові консолідації результатів;
5. Розробити метод зменшення розмірності вхідних даних, що забезпечить підвищення точності вибору пріоритетних ознак на великих наборах даних;
6. Розробити метод заповнення пропусків даних для підвищення стійкості моделі до помилок даних та паралелізації обчислень;
7. Розробити метод пошуку зміни стану пацієнта з урахуванням простору умов та аналізу характеристик його стану, що забезпечує підвищення точності підбору схеми лікування;
8. Розробити метод консолідації мультимодальних даних для забезпечення агрегації даних з різною структурою;
9. Розробити і апробувати програмні модулі щодо підтримки прийняття персоналізованих рішень на основі даних.

Об’єктом дослідження є процес обробки та аналізу персоналізованих медичних даних.

Предметом дослідження є методи прийняття рішень, методи штучного інтелекту, зокрема машинного навчання для аналізу персоналізованих медичних даних.

Методи дослідження. Для досягнення поставленої мети використано: методи прийняття рішень – для забезпечення чіткості та направленості та у пошуку рішень стосовно вибору цільових схем лікування; теорію реляційних баз даних – для відображення простору умов щодо ідентифікації стану хворого та урахування додаткових вимірів та параметрів; систему операцій реляційної алгебри – для формалізації залежностей між множинами параметрів пацієнта та показників оцінки його стану; методи штучного інтелекту – для виявлення закономірностей та залежностей між цільовими змінними, кластеризації груп пацієнтів за відповідними ознаками, для усунення невизначеностей у просторі даних, для прогнозування цільових змінних; методи об’єктно-орієнтованого аналізу і

проектування – для визначення семантичних зв’язків між джерелами даних; методи консолідації даних – для агрегації даних з різною структурою; методи моделювання інформаційних систем.

Наукова новизна роботи. У дисертаційній роботі вирішено важливу науково-прикладну проблему розроблення та удосконалення моделей, методів і засобів машинного навчання в задачах класифікації, кластеризації, прогнозування та візуалізації результатів опрацювання персональних даних для адаптації медичних рішень до пацієнта.

Отримано такі нові наукові результати:

вперше

- розроблено модель відображення стану пацієнта в багатовимірному просторі умов, що за рахунок додаткових вимірів та параметрів забезпечила підвищення точності прогнозування;
- розроблено метод двоетапної обробки малих наборів даних на основі ієрархічного предиктора, який за рахунок кластеризації та прогнозування забезпечує підвищення точності опрацювання нових вхідних наборів даних;
- розроблено метод групування моделей машинного навчання, який, враховуючи основні ознаки та повторне навчання, забезпечує підвищення точності прогнозування даних;
- розроблено метод зменшення розмірності вхідних даних, який завдяки класифікаторам, асоціативним правилам та узагальненому рангу ознак на основі індексу Жакара, забезпечує підвищення точності вибору пріоритетних ознак на великих наборах даних;

удосконалено:

- метод заповнення пропусків даних, який, використовуючи асоціативні та продукційні правила, забезпечує підвищення стійкості моделі до помилок даних;
- метод пошуку зміни стану пацієнта, який, використовуючи простір умов та аналіз характеристик стану, забезпечує підвищення точності підбору схеми

лікування;

– метод консолідації мультимодальних даних, який за рахунок попереднього визначення структури та семантики даних забезпечує агрегування даних з різною структурою;

набула подальшого розвитку:

– теорія підтримки прийняття персоналізованих рішень, яка ґрунтується на розробленні моделі відображення стану пацієнта в багатовимірному просторі умов та розроблених методах двоетапної обробки малих наборів даних, групуванні моделей машинного навчання, зменшенні розмірності вхідних даних, заповненні пропусків даних, пошуку зміни стану пацієнта та консолідації мультимодальних даних, що забезпечило адаптацію прийняття персоналізованих медичних рішень.

Практичне значення одержаних результатів та розроблених методів полягає у тому, що вони є структурною складовою процесу підтримки прийняття рішень з урахуванням персоналізованих медичних даних для підвищення точності прогнозування цільових показників стану пацієнта і приросту інформації під час процесів опрацювання, консолідації та аналізу медичних даних. Одержані результати дають змогу:

– підвищити точність прогнозування цільових показників в підмножині простору умов на 5%, що забезпечує індивідуальний підхід до моніторингу стану пацієнта на основі тривалого спостереження та контролю лікаря;

– підвищити точність класифікації на 4 % порівняно з результатами логістичної регресії, як одного з кращих класифікаторів для набору даних по COVID - 19 та на 6 %, порівняно з результатами методу опорних векторів на поліноміальному ядрі, як кращого класифікатора для набору даних по орфанних хворобах за рахунок ієрархічної класифікації та поєднання різних моделей машинного навчання;

– підвищити точність прогнозування даних на 7-9 % та забезпечити паралелізацію процесу обробки даних як малої, так і великої розмірності за рахунок розробленої стекінгової моделі на основі Випадкового лісу як метаалгоритму та

деформації метаознак під час повторного навчання на розширеному наборі даних;

- забезпечити заповнення даних за рахунок одержання додаткових значень, що керують доменом і функціональними залежностями, та включити їх у доступні навчальні дані. Правильність заповнених значень доводиться за допомогою предиктора, побудованого на вихідному наборі даних. Запропонований метод ймовірнісних продукційних залежностей на 12 % покращує результати у порівнянні з Випадковим лісом (Random Forest) та Очікування-максимізація (Expectation-Maximization) для 30 % відсутніх даних;
- зменшити ймовірність появи похибки (кількість ліжкоднів) при виборі схеми лікування за рахунок персоналізації стандартних схем лікування шляхом використання простору умов та аналізу динаміки приросту значень часовозалежних даних, що забезпечує чіткість та напрямленість у пошуку рішень щодо вибору цільових схем лікування;
- розробити інформаційну систему підтримки прийняття рішень у лікуванні орфанних захворювань;
- розробити програмний модуль для підтримки прийняття рішень у визначенні схеми лікування пацієнтів з постковідним синдромом.

Особистий внесок здобувача. Усі наукові результати, подані у дисертації, одержані здобувачем особисто. У друкованих працях, опублікованих у співавторстві ідеї та принципи, що використані в дисертаційному дослідженні, є результатом індивідуальної праці автора, а саме: [162, 163] – запропоновано принципи опрацювання персоналізованої інформації; [164, 176, 178] – розроблено модель стану пацієнта; [165, 172] – описано принципи опрацювання гетерогенних даних; [166] – аналіз характеристик мультимодальних даних; [167, 170] – розроблено метод обробки малих наборів даних; [167, 168] – розроблено метод прогнозування для окремих кластерів; [168, 169] – розроблено ієрархічний класифікатор хворих; [171, 173] – розроблено засоби прогнозування цільових змінних; [174, 175] – розроблено метод обробки персоналізованої інформації з урахуванням цільових змінних; [178,181] – розроблено алгоритм заповнення даних;

[179, 182] – принципи пошуку залежностей між цільовими змінними стану; [180, 183] – основи класифікації індивідуальних характеристик; [177, 184] – запропоновані критерії оцінювання стану пацієнта; [185, 186] – розроблено метод пошуку аномалій в оцінці стану хворого; [187, 188] – формалізовано процес пошуку пропусків даних; [189, 190] – запропоновані критерії ідентифікації даних; [191,192] – розроблено метод аналізу персоналізованої інформації; [193, 194, 195] – створено персоналізований підхід до обробки та аналізу медичних даних пацієнтів; [196, 198] – формалізовано базові операції щодо автоматизації обробки даних інформаційного об'єкта; [197, 199] – охарактеризовано принципи застосування методів машинного навчання щодо прогнозування цільових змінних інформаційних об'єктів; [200, 201] – описано особливості використання теорії множин при опрацюванні та формалізації даних інформаційних об'єктів; [202, 204] – окреслено особливості використання простору умов та аналізу зміни приросту значень часовозалежних даних; [202, 203] – формалізовано процес аналізу персоналізованих даних; [180, 205] – формалізовано етапи консолідації гетерогенних даних; [206] – описано особливості застосування нейронних мереж для опрацювання медичної інформації про пацієнта; [207] – описано підхід щодо опрацювання часовозалежних даних з кореляцією до термінів аналізу інформаційного об'єкта; [208, 209] – розроблено метод прийняття рішень щодо персоналізації стандартних схем лікування; [210, 211] – принципи проектування систем підтримки персоналізованих рішень; [212, 213] – спроектовано архітектуру системи управління персоналізованою інформацією; [214] – розроблено метод оптимізації аналізу персоналізованих даних інформаційного об'єкта; [215, 216] – розроблено методи опрацювання даних різного походження; [215, 217] – спроектовано етапи оптимізації опрацювання великих наборів персоналізованих даних; [216] – спроектована інформаційна система ідентифікації симптомів під час аналізу індивідуальних даних пацієнта; [217] – розглянуто особливості опрацювання та консолідації мультимодальних даних.

Апробація результатів дисертації. Основні теоретичні та практичні результати дисертаційної роботи були представлені та обговорені на науково-

практичних конференціях та семінарах: International Conference: The Experience of Designing and Application of CAD Systems in Microelectronics, CADSM 2015, 2016, 2017; International Scientific and Technical Conference on Computer Sciences and Information Technologies, CSIT 2016, 2017, 2018, 2020; International Conference on Advanced Computer Information Technologies, ACIT 2019, 2020; Міжнародна конференція «Інноваційні підходи і сучасна наука», м. Київ, 2015 р.; Міжнародна науково-практична конференція: «IV осінні наукові читання», м. Київ, 2015; VI Всеукраїнська науково-практична конференція «Наукові дослідження: перспективи інновацій у суспільстві і розвитку технологій», м. Харків, 2017; Міжнародна науково-практична конференція «Інформаційна безпека та інформаційні технології», ХНЕУ імені Семена Кузнеця, 2019; The 8 th International scientific and practical conference «Information, Its Impact on Social And Technical Processes», Haifa, Israel 2020; IX Міжнародна науково-практична конференція «Actual aspects of development in the context of globalization», 2020 р., Флоренція, Італія; II Міжнародна науково-практична конференція «Інформаційна безпека та інформаційні технології», 2020 р. – Кропивницький: ЦНТУ, 2020.; XVIII International Scientific and Practical Conference. Boston, USA 2020.

Публікації. Основні результати дисертаційного дослідження опубліковано у 56 наукових публікаціях, із них 21 стаття, з них: 2 статті опубліковано в журналах з Q1, 3 статті опубліковано в журналах з Q2, 7 статей у інших фахових іноземних виданнях, 7 статей у наукових фахових виданнях України, 2 монографії; 35 публікацій у матеріалах конференцій, 26 з яких індексуються в НБ Scopus. Загалом опубліковано 5 монографічних робіт, дві з яких – одноосібні, а також 3 навчальні посібники, 4 авторських свідоцтва на твір.

Структура та обсяг роботи. Дисертаційна робота складається зі вступу, шести розділів, висновків, списку використаних джерел із 302 найменувань та додатків. Обсяг дисертації становить 293 сторінки, у тому числі 239 сторінок основного тексту. Робота містить 116 рисунків та 32 таблиці.

РОЗДІЛ 1.

АНАЛІЗ ПРОБЛЕМИ ОПРАЦЮВАННЯ ТА ПОПЕРЕДНЬОЇ ОБРОБКИ МУЛЬТИМОДАЛЬНИХ ПЕРСОНАЛІЗОВАНИХ МЕДИЧНИХ ДАНИХ

У першому розділі введено поняття персоналізації, персоналізованих рішень та медичних даних пацієнта. Проаналізовано правові вимоги щодо використання персоналізованих медичних даних, обумовлено дотримання набору правил міжнародних стандартів та внутрішніх законодавчих вимог, які захищають призначені для користувача дані та створюють їхню прозорість. Проаналізовано існуючі практичні рішення, що відображають основні підходи до аналізу медичних даних з урахуванням персоналізації. Проведено порівняльний аналіз класичних методів та визначені обмеження, що характерні при опрацюванні медичних персоналізованих даних. Визначені обмеження під час опрацювання універсальними методами великих та малих мультимодальних медичних наборів даних, дали можливість визначити актуальну науково-прикладну проблему розробки нових чи удосконалення існуючих методів штучного інтелекту, а також сформували завдання наукового дослідження.

Матеріали розділу опубліковані у роботах автора [162, 168, 169, 176, 177, 179, 180, 195, 198, 204, 213, 214, 217, 218].

1.1 Визначення задачі персоналізації

1.1.1 Означення персоналізованої інформації

В епоху активного розвитку цифрових технологій, поняття персоналізованої медицини орієнтоване на опрацювання гетерогенних медичних даних, що дозволяє забезпечити керованість цьому процесу. Персоналізація набуває все більшого значення в науковому співтоваристві внаслідок багатьох причин серед яких, впровадження нових оптимізаційних методів та алгоритмів, що сприяють зменшенню медичних витрат, забезпеченню якісного та ефективного медичного обслуговування. Дане дослідження націлене на розв'язання проблеми формалізації профілю пацієнта та формалізації його станів під час лікування чи реабілітації, що дозволить оптимізувати процеси опрацювання, аналізу індивідуальних характеристик пацієнтів та прогнозування можливих персоналізованих рішень щодо надання медичної допомоги, орієнтуючись на оцінку стану здоров'я пацієнта.

Дані, орієнтовані на пацієнта – це дані, які пацієнт або його родичі повідомляють лікарю. Якість такого типу даних залежить від рівня грамотності пацієнта та можливості правильно передати покази, тоді як якість даних за фізичними показниками пацієнта залежить від якості та чутливості пристроїв та технологій, що беруть участь у зборі та опрацюванні поточних даних [1]. Для подальшого аналізу бралися до уваги дані за фізичними показами пацієнта і дані орієнтовані на пацієнта. Така інгетрація даних уможливлює керування ними, підтримку ітераційної перевірки медичних досліджень для зменшення частоти хибно/позитивних та помилково/ негативних результатів.

Поточна парадигма опрацювання медичних даних передбачає збір даних у рамках скринінгу для перевірки на наявність захворювань або діагностичних процесів. Результати скринінгу або діагностики повинні перевірятися (принципи ітераційної перевірки) медичними оглядами.

У роботі введено ряд означень поняття персоналізації та персоналізованої медицини:

Пацієнт – людина, яка проходить обстеження, лікування чи реабілітацію у медичних закладах охорони здоров’я.

Персоналізація – це процес пошуку рішень лікаря для розв’язання проблеми призначення або корекції лікування пацієнта, враховуючи індивідуальні особливості його стану та різні фактори впливу.

Персоналізація (широко відома як кастомізація в інших галузях) полягає у пристосуванні послуги чи товару для розміщення конкретних осіб, іноді пов’язаних із групами або сегментами осіб. Широкий спектр організацій використовує персоналізацію для покращення задоволеності клієнтів, цифрової конверсії продажів, результатів маркетингу, брендингу та покращення показників веб-сайтів, а також для реклами. Персоналізація є ключовим елементом у соціальних мережах та ректомендаційних системах [2].

Персоналізовані рішення – це рішення, які формуються на підставі індивідуальних показників пацієнта та сторонніх факторів, що визначають та впливають на його стан.

Медичні дані – індивідуальні дані стану пацієнта, що характеризуються викидами, шумами, пропусками, які опрацьовуються дослідниками під час процесу пошуку персоналізованих рішень.

Персоналізована медицина, прецизійна медицина – це медична модель, яка передбачає групування пацієнтів для визначення загальної схеми лікування та підбір/уточнення схеми для індивідуального пацієнта на основі прогнозованої реакції або ризику захворювання [1].

Отже, персоналізація виявляє моделі поведінки та кореляцію в поведінці реципієнтів [2].

Персоналізована медицина (ПМ) – це нова практика медицини, яка використовує генетичний профіль людини для управління рішеннями, що приймаються щодо профілактики, діагностики та лікування захворювань. Знання генетичного профілю пацієнта може допомогти лікарям підібрати належний препарат або терапію та вводити їх із відповідною дозою або режимом, як

зазначено в Національному інституті досліджень геному людини, Національному інституті охорони здоров'я, США [3].

Персоналізована медицина – це пристосування медичного лікування до індивідуальних особливостей кожного пацієнта. Підхід опирається на наукові обґрунтування щодо нашого розуміння унікальності індивідуума (людини), а генетичний профіль робить її сприйнятливою до певних захворювань. Це дозволить обґрунтовувати та прогнозувати, які медичні методи лікування будуть безпечнішими та ефективнішими безпосередньо для кожного пацієнта, а які не будуть [4].

Персоналізовану медицину можна вважати продовженням традиційних методів та підходів до розуміння та лікування хвороб. Оснащені спеціалізованими застосунками чи інструментами лікарі зможуть вибрати терапію або протокол лікування, заснований на молекулярному профілі пацієнта, який може не лише мінімізувати шкідливі побічні ефекти та забезпечити успішний результат, але допоможе стримати витрати у порівнянні зі "спробами та помилками" традиційного протоколу лікування захворювань [5].

Як зазначалося вище, персоналізована медицина в епоху цифрових технологій повинна базуватися на аналізі структурованих малих та / або великих даних. Також необхідні (електронні) системи медичних записів для збору, зберігання та обробки зібраних даних.

1.1.2 Актуалізація задачі персоналізації медичних даних

В останні роки були змінені парадигми та бачення медицини і охорони здоров'я. Так, політика охорони здоров'я західних країн зосереджена на профілактиці захворювань та поінформованості про них. Окрім того, Європейська Комісія розробляє та впроваджує різноманітні програми для сприяння широкому впровадженню комплексних цифрових рішень, які покращать якість життя громадян, демонструючи при цьому значне підвищення ефективності надання медичних послуг та догляду в Європі. Особливо важливими для реалізації цих програм серед інших заходів є підвищення рівня грамотності населення в галузі

охорони здоров'я, збір медичних, стаціонарних та амбулаторних спостережень, і принципово важливе значення – персоналізація в медицині.

За останні шість десятиліть з’явилося багато доказів, які свідчать про те, що значна частина мінливості реакції на ліки визначається генетичною, причому важливу роль відіграють вік, харчування, стан здоров’я, вплив навколишнього середовища, епігенетичні фактори та паралельна терапія. Для досягнення індивідуальної медикаментозної терапії з розумно передбачуваним результатом, необхідно додатково враховувати різні моделі реакції на них.

ПМ можна вважати продовженням традиційних підходів до розуміння та лікування захворювань, але з більшою точністю. Дані з профілю пацієнта впливають на вибір ліків або протоколів лікування, які мінімізують шкідливі побічні ефекти або забезпечують більш успішні результати. ПМ також може вказувати на сприйнятливість людини до певних захворювань до того, як вони проявляться, що дозволяє лікарям та пацієнтам скласти план моніторингу та профілактики.

Парадигма ПМ повинна розглядати можливість збору даних протягом життя, тобто від народження до моменту необхідності індивідуального або індивідуального лікування. Збір та автоматизований аналіз великих обсягів даних протягом життя суттєво сприяв би покращенню ПМ для пацієнта.

Окрім медичного процесу, медикаменти відіграють важливу роль в індивідуальній медицині. Тому багато фармакологічних напрямків досліджень, як фармакогеноміка, приділяють більше уваги окремим препаратам для індивідуального лікування. Хоча сьогодні ліки виробляються для груп людей (щодо їх генетики). Окрім цього, один підхід, придатний для всіх, призводить до невдалого лікування або токсичності препарату [218].

У Мюнхенському університеті Бундесверу проводилися дослідження впливу медичної грамотності на здоров’я пацієнта. Виявили, що пацієнти з високим рівнем обізнаності щодо захворювань (мають вищу медичну освіту) вміють краще запобігати захворюванням, ніж пацієнти з середнім або низьким рівнем медичної

освіти. Крім того, з’ясували, що дані, орієнтовані на пацієнта, краще допомагають при діагностиці. У цілому рівень обізнаності пацієнта з хворобою та якість його медичних даних можуть визначати результати його психологічного та фізичного здоров'я.

Зараз лікарі можуть вийти за рамки універсальної моделі призначення, щоб прийняти більш ефективні клінічні рішення для кожного пацієнта. ПМ пропонує структурну модель ефективного медичного обслуговування, що профілактично, скоординовано та доведено. ПМ краще працює з мережею електронних медичних записів, які пов'язують клінічну та молекулярну інформацію, щоб допомогти лікарям у прийнятті відповідних рішень щодо лікування. ПМ є учасником, залучаючи пацієнтів до вибору способу життя та активного підтримання здоров’я, щоб компенсувати генетичну сприйнятливість.

Одним із важливих факторів, що впливають на персоналізованість – це вік пацієнта. Виділяють п'ять вікових груп:

1. 0 – 14 років – діти;
2. 15 – 24 роки – підлітки;
3. 25 – 54 років – молоді;
4. 55 – 64 років – дорослі;
5. 65 років і старше — літні люди.

У [6] зазначено, що ПМ не буде повністю ефективною на будь-якому етапі життя, якщо відсутні відповідні медичні та індивідуальні дані попередніх періодів життя. Наприклад, для підтримки медичних рішень у дорослому віці додатково до даних, пов’язаних із «дорослим віком», необхідно мати дані «дитини, підліткового та молодого віку». Тому особливо важливо збирати достатньо даних і постійно вдосконалювати систему збору даних до результатів ПМ (більша ефективність та результативність). Для цього може бути застосована технологія Великих даних.

Великі дані передбачають збір даних з багатьох різних джерел. Важливо відзначити 3 основні V:

1. Velocity (швидкість зміни);
2. Veracity (достовірність, важливість даних), оскільки ПМ вимагає дуже точних та повних даних;
3. Volume (розмір).

Проблеми, які слід подолати щодо швидкості, полягають у з'ясуванні, які дані слід зберігати для подальшої обробки:

1. Зібрані вхідні дані або опрацьовані дані?
2. Як довго такі дані є дійсними? (важливо враховувати метрику валідності)

1.1.3 Правові аспекти консолідації персоналізованої інформації

Персональні дані – відомості чи сукупність відомостей за допомогою яких фізична особа може бути ідентифікована [10].

Персональні медичні дані пацієнта діляться на дві категорії: часовонезалежні та часовозалежні.

Отже, до категорії часовонезалежних можна віднести: прізвище та ім'я; дату та місце народження; стать, громадянство; сімейний стан; псевдонім; дані, записані в посвідченні водія; номер пенсійної справи; адресу місця проживання; антропометричні дані, що враховують параметри тіла, дані анамнезу хвороби, та інформацію про генетичні дані тощо.

Категорія часовозалежних охоплює інформацію про стан здоров'я, а саме: показники параметрів крові, артеріального тиску, усі характеристики відповідно до наявних симптомів захворювання, біометричні дані тощо.

Інформаційний бум призвів до тисячократного зростання даних, зібраних у різних сферах. До них належать комп'ютерні технології, економіка, соціологія, медицина, астрономія тощо. Збільшення обсягу зібраної інформації продовжує зростати в геометричній прогресії. Наприклад, на основі дослідження цифрового всесвіту, проведеного на замовлення EMC, загальний глобальний обсяг даних у 2005 році становив 130 екзабайт, до 2227 EB до 2015 року, а минулого року знову збільшився до 7 ZB (зетабайт). Передбачається, що до 2025 року обсяг цифрових

даних зросте до 10,9 ZB. Так само швидко зростає розмір окремих баз даних, які подолали петабайтний бар'єр. Більшість зібраних даних наразі не аналізується або є лише простим аналізом. Це важливо для медичної сфери, насамперед коли лікарні не підтримують медичні стандарти обміну даними (наприклад, HL7) [11].

Нормативно-правовий аспект є різним у різних країнах світу і зокрема в США. Роки широкомасштабних зловживань даними та широко розрекламованих скандалів, особливо фіаско Cambridge Analytica, змінили суспільне ставлення до приватного життя та підштовхнули законодавців до дії. Наприклад, в Каліфорнії, штаті, де працює багато провідних світових технологічних компаній, прийнято Закон про конфіденційність споживачів у Каліфорнії у червні 2018 року, який набув чинності з 1 січня 2020 року.

При розробці інформаційних систем, які опрацьовують дані пацієнтів, необхідно враховувати вимоги міжнародних протоколів, таких як GDPR та CCPA, що накладає додаткові обмеження у доступах даних відповідно до чинного законодавства.

Важливим фактором є аналіз та застосування протоколів відповідно до ідентифікованого захворювання у пацієнта. Протоколи лікування забезпечують чіткі вказівки для лікарів і покращують якість клінічних рішень навіть для фахівців, які звикли до традиційної медичної практики. Ще одна важлива перевага клінічних протоколів полягає в тому, що вони сприяють послідовному лікуванню пацієнтів на всіх рівнях.

Загальні Положення про захист даних (GDPR) описує певний набір правил, які захищають призначені для користувача дані і забезпечують їхню прозорість. Незважаючи на те, що GDPR є суворим, він дозволяє компаніям збирати анонімізовані дані без згоди, використовувати їх для будь-яких цілей і зберігати протягом невизначеного часу – доки компанії не видалять з даних усі ідентифікатори [12, 13].

1.1.4 Задачі анонімізації персоналізованої інформації

Анонімізація даних – це процес захисту приватної або конфіденційної інформації шляхом видалення або шифрування ідентифікаторів, які пов'язують особу зі збереженими даними [14]. Наприклад, особиста інформація (ідентифікаційна інформація), така як імена, номери соціального страхування та адреси, може бути зашифрована за допомогою процесу анонімізації даних, який зберігає дані та зберігає анонімність джерела (рисунок 1.1).

Рисунок 1.1. Типи анонімізації даних

Методи анонімізації даних.

Маскування даних – приховування даних із зміненими значеннями. Можна створити дзеркальну версію бази даних і застосувати методи модифікації, такі як перетасування символів, шифрування та заміна слів або символів. Наприклад, можна замінити символ значення таким символом, як «*» або «x». Маскування даних робить зворотне проєктування або виявлення неможливим [15].

Псевдонімізація – метод управління даними та деідентифікація, який замінює приватні ідентифікатори фальшивими ідентифікаторами або псевдонімами, наприклад, замінюючи ідентифікатор «Василь Кагуй» на «Петро Федоренко». Псевдонімізація зберігає статистичну точність та цілісність даних, дозволяючи

модифікованим даним використовуватись для навчання, розробки, тестування та аналітики, одночасно захищаючи конфіденційність даних [16].

Узагальнення – навмисне видаляє деякі дані, щоб зробити їх менш ідентифікованими. Дані можна модифікувати в набір діапазонів або широку область з відповідними межами. Ви можете видалити номер будинку за адресою, але переконайтесь, що не видаляєте назву дороги. Метою є усунення деяких ідентифікаторів при збереженні міри точності даних.

Обмін даними – техніка, що використовується для перестановки значень атрибутів набору даних, щоб вони не відповідали вихідним записам. Обмін атрибутами (стовпцями), які містять значення ідентифікаторів, таких як дата народження, наприклад, може мати більший вплив на анонімізацію, ніж кодування значення.

Змішування даних – дещо модифікує вихідний набір даних, застосовуючи методи округлення чисел та додавання випадкових шумів. Діапазон значень повинен бути пропорційним змішуванню. Мала база може призвести до слабкої анонімізації, тоді як велика база може зменшити корисність набору даних. Наприклад, можна використовувати базу 5 для округлення значень, таких як вік чи номер будинку, оскільки це пропорційно вихідному значенню. Можна помножити номер будинку на 15, і значення може зберегти свою достовірність. Однак використання даних, таких як 15, може зробити вікові показники фальшивими.

Синтетичні дані – алгоритмічно виготовлена інформація, яка не пов’язана з реальними подіямих [17]. Синтетичні дані використовуються для створення штучних наборів даних замість того, щоб змінювати оригінальний набір даних або використовувати його як є, ризикуючи конфіденційністю та безпекою. Процес включає створення статистичних моделей на основі шаблонів, знайдених у вихідному наборі даних. Можна використовувати стандартні відхилення, медіани, лінійну регресію або інші статистичні методи для створення синтетичних даних.

Недоліки анонімізації даних

GDPR передбачає, що веб-сайти повинні отримувати згоду користувачів на збір особистої інформації, наприклад IP-адрес, ідентифікатора пристрою та файлів cookie. Збір анонімних даних та видалення ідентифікаторів з бази даних обмежує здатність отримувати значення та розуміння даних. Наприклад, анонімізовані дані не можна використовувати для маркетингових зусиль або персоналізації користувацького досвіду [18, 19] .

Анонімізація або деідентифікація – актуальна, коли дані про стан здоров’я використовуються у другорядних цілях. Під вторинними цілями зазвичай розуміють цілі, які не пов'язані з наданням допомоги пацієнтам. Тому такі речі, як дослідження, охорона здоров’я, сертифікація чи акредитація та маркетинг, вважатимуться другорядними цілями. Більшість конфіденційності в законах у всьому світі ґрунтуються на згоді, якщо пацієнти дають свою згоду чи дозвіл, дані можуть потім використовуватися для цілей, які вони санкціонують. Однак, якщо дані анонімні, згода не потрібна. Взагалі, анонімізовані дані більше не вважаються особистою інформацією про здоров’я, і вони не підпадають під законодавство про конфіденційність.

Може здатися очевидним просто отримати згоду для початку. Але коли пацієнти звертаються до лікарні чи клініки для надання медичної допомоги, прохання у них про широку згоду на всі можливі подальші вторинні використання їхніх персональних даних, коли вони реєструються, як правило, розглядається як примусове, оскільки це насправді не буде інформованою згодою.

Більше того, у деяких юрисдикціях (наприклад, ЄС) особиста інформація повинна «збиратися для визначених, явних та законних цілей, а не оброблятися надалі способом, несумісним із цими цілями», відповідно до Директиви про захист даних 95/46 [21]. Згідно цієї директиви, мета обробки даних повинна бути визначена точно, і що подальше опарцювання не повинне бути несумісним з цілями, для яких персональні дані були зібрані спочатку [1]. Це піднімає питання про законність широкої згоди.

Звичайно, згода не завжди потрібна для передачі особистої інформації. Закон чи нормативний акт може передбачати розповсюдження особистої медичної інформації із правоохоронними органами за певних умов без згоди (наприклад, повідомлення про вогнепальні поранення) або повідомлення про випадки деяких інфекційних захворювань без згоди (наприклад, туберкульоз або ВІЛ).

Іноді медичний персонал має досить широкі повноваження щодо обміну інформацією з департаментами охорони здоров’я. Але не всі медичні працівники готові ділитися особистою інформацією про здоров'я своїх пацієнтів, і багато хто вирішує не робити цього, коли це вирішувати їм [23]. Отже, не слід сприймати як само собою зрозуміле, що зберігачі даних охоче передаватимуть особисту інформацію про здоров’я, навіть якщо їм це дозволено, і навіть якщо це буде для загального блага.

Анонімізація дозволяє обмінюватися інформацією про здоров’я, коли це не передбачено повноваженнями або практичним способом отримання згоди, а також коли обмін є дискреційним, а зберігач даних не хоче ділитися цими даними.

1.2 Аналіз існуючих рішень щодо опрацювання персоналізованих медичних даних

1.2.1 Практичні рішення щодо опрацювання медичних персоналізованих даних

Над створенням продуктів з використанням штучного інтелекту для установ охорони області працюють розробники великих компаній, зокрема Microsoft, Apple, Google і IBM. За підрахунками аналітиків, таких фірм у світі вже 800. Незважаючи на те, що інноваційні технології тільки почали впроваджуватися у сферу медицини, згідно з дослідженням Frost & Sullivan стверджують, що цей ринок зросте до 6,16 мільярда доларів при складеному річному темпі зростання (CAGR) 68,55% між 2018 і 2022 роками [6].

Ось, наприклад, повну історію хвороби, дані про аналізи за всі роки лікування, поточний стан організму – все це може зібрати і структурувати система штучного інтелекту. Дані, завантажені в базу, не вислизнуть від уваги електронного

мозку і будуть швидко оброблені. Це заощадить лікарям час, підвищить точність діагнозу і дозволить своєчасно призначити потрібне лікування.

Система від IBM під назвою Watson Health, здатна виявити потенційні проблеми з судинною системою, виявити рак і визначити, чи є у пацієнта схильність до утворення тромбів. IBM Watson може швидко реагувати, коли необхідно вивчити нову інформацію і зробити з неї висновки. Наприклад, за 10 хвилин штучний інтелект IBM проаналізував 20 мільйонів наукових статей про онкологію і на їх основі поставив пацієнту правильний діагноз. [24].

У кількох лікарнях Великобританії вже використовується схожа розробка від Google – DeepMind Health. Вона також допомагає обробити всю інформацію про здоров'я людини, ділиться своїми висновками з лікуючим лікарем, який в результаті ставить остаточний діагноз [25].

Безпосередньо спілкуватися з людиною і давати їй свої рекомендації можуть системи на зразок Ada – це сервіс, розроблений британською однойменною компанією. Медичний додаток спілкується з пацієнтом, розпитує про симптоми і скарги, а у відповідь дає рекомендації, в тому числі якого лікаря необхідно відвідати, пропонує зв'язатися з фахівцем для віддаленої консультації [26].

Програма Sense.ly на базі штучного інтелекту стежить за станом людей, які нещодавно пройшли тривале лікування або страждають на хронічні захворювання. Додаток створено для того, щоб структурувати дані про стан людини, відправляти їх спеціалісту та давати рекомендації. Також система здатна нагадувати, коли приймати ліки і коли звертатися до лікаря [8].

Схожим чином функціонують і системи генетичного аналізу, який допомагає зрозуміти первинну причину захворювання. Одна з платформ по перевірці генома людини – Sophia Genetics – виявляє схильність пацієнта до різних захворювань і звертає на це увагу лікаря [9].

За схожим принципом проводиться і підбір медикаменту для того чи іншого пацієнта, здійснюється аналіз впливу ліків на організм. Наприклад, система MedClueRx створена для того, щоб визначати, які препарати більше підійдуть при нервових розладах, захворюваннях шлунково-кишкового тракту, епілепсії [10].

Наступним класом систем є системи аналізу поширення та передбачення сили вірусів різної природи.

Аналізуючи повідомлення преси, платформи соціальних мереж і урядові документи, штучний інтелект може навчитися швидше виявляти осередки епідемії. Прикладом таких систем є канадська система-стартап BlueDot [11].

Програмне забезпечення цієї компанії призначене для захисту від ризику спалахів пандемії та захищає життя, зменшуючи вплив інфекційних захворювань, які загрожують здоров'ю людей. Розробники стверджують, що їх програмне забезпечення дозволяє донести всю інформацію про загрозу епідемії COVID-19 протягом декількох годин до Центрів з контролю і профілактики захворювань або Всесвітньої організації охорони здоров'я. Додаток також допомагає виявити потенційно заражених людей.

Китайська система спостереження використовувала технологію розпізнавання обличчя і програмне забезпечення від SenseTime для виявлення людей, у яких може бути жар. Уряд Китаю також розробив систему моніторингу під назвою «Кодекс здоров'я», яка використовує різні дані для виявлення є оцінки ризику кожної людини на основі її поведінки, а саме: історії пересування, часу, проведеного в гарячих точках, і потенційного контакту з тими, хто переносить вірус. Громадянам присвоюється кольоровий код (червоний, жовтий або зелений), доступ до якого вони можуть отримувати через популярні додатки WeChat або Alipay для перевірки [12].

Отже, існують різні типи систем для аналізу персоналізованих даних. Проте, усі системи призначені для аналізу даних та підтримки прийняття рішень з певної нозології.

1.2.2 Аналіз сучасних підходів щодо опрацювання медичних даних

Важливим фактором щодо обрання методів та підходів для опрацювання медичних даних є їхні характеристики. Зазвичай, медичними даними вважають тільки ті, що отримують при вимірюванні характеристик пацієнта.

Кількість характеристик пацієнта, хворої або здорової людини, чимала. Медичну інформацію, що збирається від пацієнтів можна класифікувати за характером збереження даних (Рисунок 1.2.), а саме :

- якісні ознаки (наявність болю, підвищеної температури, колір шкірних покривів, перкусійні та аускультативні феномени);
- одиничні числові дані (вага, артеріальний тиск, температура тіла, кількість лейкоцитів, ШОЕ);
- динамічні дані (електрограми – ЕКГ, ЕЕГ, ЕГГ; реограми РКГ, РЕГ, фонокардіограма);
- статичні картини (рентгенограма, авторадіограма);
- динамічні картини (поле біопотенціалів, електрокардіограма).

Рисунок1.2 Класифікація медичних даних за характером зберігання даних

Для медичних даних характерні такі особливості:

- нечітка, а часом і суперечлива термінологія;
- велика кількість якісних характеристик, які суб'єктивно оцінюють стан хворого;
- відсутність єдиних алгоритмів опису стану хворого щодо процесів діагностики і лікування;
- недостатній ступінь стандартизації медичної документації;
- велика варіабельність медичних даних, невеликі вибірки з невідомими законами розподілу, що дуже ускладнює статистичні розрахунки та проведення відповідних оцінок.

За результатами досліджень сформульовані основні етапи обробки даних пацієнта, що відображені на рисунку1.3.

Рисунок 1.3 Основні етапи обробки даних пацієнта

Деякі величини можуть набувати значень у певному інтервалі, інша частина може бути дискретною або аналоговою. Неперервними величинами є, наприклад, криві зміни маси тіла, температури, відстані тощо. Багато величин можуть набувати лише цілочислових значень. Приклади дискретних величин: частота пульсу, кількість хворих у відділенні, кількість ліжко-днів тощо.

Для медичних даних характерна наявність викидів, шумів, пропусків, надлишковості:

- викиди – елементи набору даних, подій або спостережень, що не відповідають очікуваній поведінці стану пацієнта (атипові дані, що виходять поза межі норми та не відповідають критеріям хвороби);
- шуми – це дані пацієнта, які є надлишковими чи додатковими, що не несуть інформаційного змісту для аналізу вибірки;
- пропуски – це відсутність даних при оцінці характеристик хвороби та стану пацієнта. При опрацюванні даних такі пацієнти зазвичай відсікаються.

Пошук персоналізованих рішень включає низку взаємопов'язаних процесів, а саме:

• збір даних (фізичних параметрів пацієнтів) із розумних датчиків, наприклад, електронних глюкометрів, автоматичних тонометрів, холтерів тощо;

• консолідація даних, орієнтована на визначення моделей даних, а також формування потоків даних відповідно до визначених завдань;

• обробка даних полягає у перевірці даних та зберіганні у базі даних, базі знань або сховищах даних, забезпечуючи при цьому захист даних;

• процес аналізу даних забезпечує сортування та групування даних за визначеними параметрами з використанням методів видобування даних, таких як методи кластерного аналізу, включаючи k-середні, k-медіану тощо; також використовуються методи візуалізації даних тощо,

• процес прогнозування передбачає побудову тестового інформаційного об'єкта майбутнього стану пацієнта на основі попередніх даних з використанням методів штучного інтелекту, а саме: штучні нейронні мережі, дерева рішень, Байєсівські мережі, лінійна регресія, кореляційно-регресійний аналіз, методи пошуку асоціативних правил, включаючи алгоритм Апріорі; остаточне проведення оцінки рішень, прийнятих на їх основі.

Схема процедури обробки медичних даних наведена на рисунку1.4.

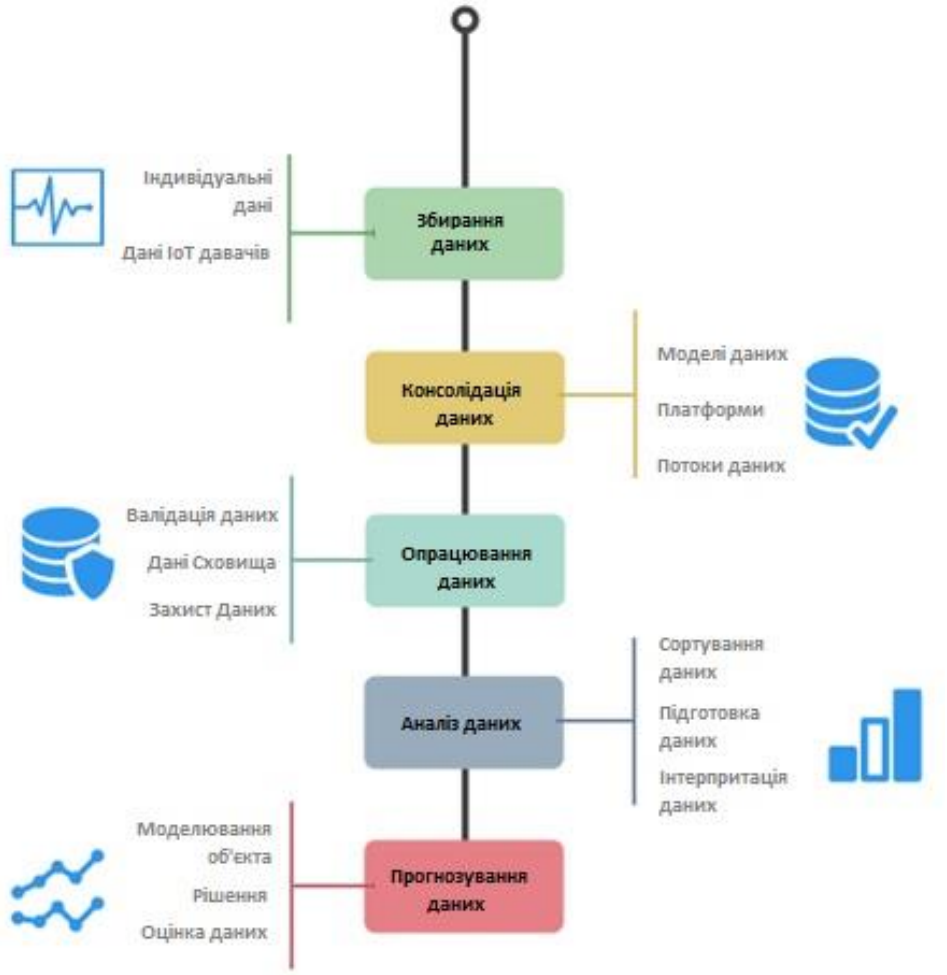


Рисунок 1.4. Залежність етапів дослідження медичних даних пацієнта для їх інтелектуалізації від персоналізованих рішень.

1.2.3 Аналіз методів штучного інтелекту для прикладних задач персоналізації

За результатами аналізу існуючих систем ШІ у сфері медицини для розв'язання проблем діагностики захворювань, дослідження геному, розробки ліків, медичної візуалізації та ін. виокремлено проблеми, що ще залишаються не розкритими, а саме за результатами консолідації даних про хворого, його індивідуальних особливостей, враховуючи міжнародні стандарти лікування та існуючу фармацевтичну продукцію, визначати схему персоналізованого лікування та спрогнозувати результати його застосування.

Різні пацієнти реагують на лікарські засоби та схеми лікування по-різному. Тому персоналізоване лікування має величезний потенціал для збільшення тривалості життя пацієнтів. Але визначити, які чинники повинні впливати на вибір лікування, дуже важко. Алгоритми ШІ можуть автоматизувати складну статистичну роботу і допомогти виявити, які характеристики свідчать про те, що пацієнт буде мати певну позитивну реакцію на певне лікування чи навпаки. Тому алгоритм може передбачити ймовірну реакцію пацієнта на певне лікування.

Інтелектуальним аналізом даних, в т.ч. і в медицині, займається низка українських та закордонних вчених. Так, Бідюк П.І. в [36] для аналізу пропусків даних та їх заповнення використав процедури обчислення на основі дерев рішення, Expectation-Maximization алгоритм та регресійний підхід до прогнозування відсутніх даних за допомогою функцій прогнозування. Подібні результати отримали Соколова О. у 2015 р і Sina Khanmohammadi у 2016 р. для асоціативної класифікації медичних даних та Anthony Costa Constantinou для комплексного анкетування та інтерв'ювання даних для інтелектуальних моделей Байєса для підтримки медичних рішень. Також застосовано мережу Байєса для системи, що використовується у Швидкій Медичній Довідці (QMR).

Проте з поширенням технологій Великих даних (Big data) байєсівські мережі виявилися повільним. Тому у роботі Y. Tang [16] розроблено метод розпаралелювання байєсівських мереж. Подібні результати отримав Anders L.Madse. Також баєсівські мережі використовуються для діагностування захворювань, наприклад у [35] та [36]. Проте, навіть за умов паралелізму, для багатопараметричних, великих за обсягом та з динамічним надходженням медичних даних байєсівські мережі доцільно застосовувати у комбінації з іншими методам машинного навчання.

Апарат штучних нейронних мереж, в тому числі із застосуванням нечіткої логіки, також активно застосовується для аналізу різноманітних медичних даних. Так, у роботах Зайченка Ю.П. [219, 220], Бодянського Є.В., Перової І.Г. [18], запропоновано систему швидкої медичної діагностики на основі автоасоціативної нейронечіткої пам’яті. Така система характеризується простотою як

архітектурного рішення, так і його програмної реалізації і забезпечує діагностування пацієнтів з невідомим діагнозом. Проте, однією із актуальних задач залишається підвищення точності результатів розв'язання задачі класифікації. Окрім цього, проблема швидкодії роботи подібних систем, які базуються на ітераційних алгоритмах навчання, незбалансованість вхідних даних, а також невеликі вибірки даних, зібрані медичним персоналом вручну, накладають ряд обмежень на застосування існуючих методів та засобів обчислювального інтелекту до розв'язання подібних задач.

Березький О.М. у [220] використовує нейронечіткий підхід для аналізу медичних даних, зокрема, зображень. Дивак М.П. [21] використовує інтервальні оцінки для оцінювання стану об’єкта в умовах обмеження часу на прийняття рішень. У Silva-Ramírez E. L [22] для розв’язання задачі класифікації та заповнення пропусків даних використано метод з використанням багатошарового персептрона, навчання якого проводиться за різними правилами, а також підхід до множинного підрахунку, який базується на поєднанні багатошарових персептронів та k-найближчих сусідів. У роботах Chiacchio F [23] та Tsai C.W [24] подано математичні методи перетворення динамічних дерев відмов у марковські моделі, Байєсові моделі або імітаційні моделі на основі методу Монте-Карло, пошук методів скорочення складності моделей та підвищення продуктивності обчислення. Проте, необхідним є розроблення нових евристичних підходів, які скорочують розмірність марковської моделі, а також обсяг операцій, необхідних для їх автоматизованого аналізу та обчислення. Актуальним питанням залишається застосування марковських моделей для обчислення параметра потоку відмов, ймовірності появи помилок першого та другого роду, аналізу причин непрацездатності системи.

Субботін С.О. у [219] використовує модифікований підхід дерев рішень для аналізу медичних даних. Перевагою цього є можливість візуалізації даних та пояснення результатів кінцевим користувачам.

Отже, методи інтелектуального аналізу даних використовуються для розв’язання багатьох задач з опрацювання та аналізу медичної інформації. Проте

відсутні комплексні дослідження, спрямовані на ідентифікацію стану пацієнта без специфікації виду анамнезу. Розв’язані окремі задачі цього напряму, проте у дослідженнях лише частково враховано феномени Великих даних та Інтернету речей, глибинного аналізу та візуалізації накопичених даних для підтримки прийняття рішень з персоналізованого лікування.

Математичний апарат мереж Петрі та його модифікації застосовано дослідниками для моделювання процесів різної природи. Наприклад, для моделювання процесів планування в режимі реального часу в середовищі з обмеженими ресурсами італійськими науковцями у [44] запропоновано використати апарат часових кольорових мереж Петрі (TCPN), що дало змогу аналізувати взаємозалежності, суперечливість пріоритетів та варіативність наявних ресурсів (на прикладі промислового проєкту). Іспанські дослідники у джерелі [45] використовують апарат кольорових мереж Петрі з пріоритетом для удосконалення процесу моделювання, аналізу та семантичної валідації складних системних подій (на прикладі технології опрацювання та корелювання великих потоків даних для процесу визначення рівня якості повітря). У спільній роботі 2016 р. у джерелі [46] дослідників з Тунісу та Франції розглянуто особливості процесу розроблення складної системи дискретних подій зі змінними структурами з використанням багатомоментного підходу, змодельованого засобами кольорових мереж Петрі.

Досліджуючи цю проблему, дослідники застосували модель на основі згорткової нейронної мережі для виявлення пацієнтів із COVID-19 за допомогою зображень CXR. Вони використали попередньо навчений ImageNet і навчили модель на основі даних із відкритим кодом рентгенівських зображень Chest (CXR) [47]. Застосування моделі LSTM для прогнозування специфічного для країни ризику зараження вірусом COVID-19, який спирається на дані про тенденції та метеорологічні дані певної країни для прогнозування ймовірного поширення захворювання COVID-19 [48]. Фахівці з штучного інтелекту застосовували методи машинного навчання для обробки інтернет-активності, новин, звітів організацій охорони здоров’я та діяльності засобів масової інформації, щоб передбачити поширення спалаху в Китаї. А також застосування байєсівського підходу для

прогнозування кількості смертей у майбутньому, використовуючи емпіричні дані [49].

Важливу роль відіграють проблема аналізу залежностей між індивідуальними характеристиками хворого та особливості поведінки чи динаміки змін його стану під час лікування та одужання. Так, застосування асоціативних правил та продукційних правил для виявлення прихованих залежностей у [221, 222] дозволяє скорочувати розмір вибірки, що є важливим для аналізу великих вибірок даних, а також підвищувати точність прогнозування наступного стану пацієнта. Тому асоціативні правила доцільно використовувати в задачах персоналізації.

Незважаючи на те, що алгоритми інтелектуального машинного навчання [75], нейронні мережі [86] та моделі типу SARIMA [97], які використовуються в дослідженнях, здатні визначати певні «шаблони» та тенденції в поведінці досліджуваних явищ, неможливо отримати прогноз високої точності за вищевказаних обставин [106]. Він може лише вказувати на визначальні тенденції та закономірності поширення хвороби.

Авторегресійне інтегроване середнє, або ARIMA, є одним із найбільш широко використовуваних методів прогнозування для однофакторного прогнозування даних часових рядів [113]. Хоча метод може обробляти трендові дані, він не підтримує часові ряди сезонних компонентів [114]. Розширення ARIMA, яке підтримує пряме моделювання сезонної складової ряду, називається SARIMA [115]. Проблема з ARIMA полягає в тому, що він не підтримує сезонні дані. Це часовий ряд із повторюваним циклом. ARIMA очікує дані, які не є сезонними або мають вилучений сезонний компонент, наприклад, сезонно скориговані з використанням таких методів, як сезонна дисперсія.

Сезонне авторегресійне інтегроване ковзне середнє, SARIMA або Seasonal ARIMA, є розширенням ARIMA, яке явно підтримує однофакторні часові ряди даних із сезонним компонентом [116]. Він додає три нові гіперпараметри для визначення авторегресії (AR), різниці (I) і ковзного середнього (MA) для сезонної складової ряду, а також додатковий параметр для періоду сезонності. Сезонна модель ARIMA формується шляхом включення додаткових сезонних умов в

ARIMA [117]. Сезонна частина моделі складається з термінів, які дуже схожі на несезонні компоненти моделі, але включають зворотні зрушення сезонного періоду [118].

Спеціальні методи моделювання вірусів проаналізовано в [129]. Найбільшою проблемою є невизначеність наявних офіційних даних, особливо щодо реальної початкової кількості інфікованих (випадків), що може призвести до неоднозначних результатів і неточних прогнозів щодо порядку, на що також вказували інші дослідники.

Тому в [120] агентний підхід поєднується з SEIR-моделлю. Використовуються такі агенти: людина, будинок, бізнес, уряд, система охорони здоров'я.

Завдання пошуку залежностей у даних вимагає аналізу залежностей між десятками параметрів досліджуваного процесу та сотнями можливих джерел впливу на цей процес. Залежності є недетермінованими, тому моделювання потребує використання статистичних методів для аналізу випадкових процесів [121]. Частина інформації часто прихована від спостереження або не контролюється. Тому в процесі аналізу зібраної інформації виникло багато труднощів.

Сьогодні розроблені методи статистичного аналізу дозволяють працювати з частково невизначеними або нечіткими процесами. Однак доступні методи мають суттєві обмеження щодо сфери дії та типів даних.

У прогностичній аналітиці часто згадуються методи машинного навчання, зокрема нейронні мережі [123]. Однак в цьому випадку ефективність їх застосування буде невеликою. Основна причина полягає в тому, що моделі машинного навчання корисні у випадку стаціонарних процесів. Передбачається, що майбутні прогнозні дані описуються тим же розподілом, що й навчальні дані. Однак зростання виявлених випадків коронавірусу є значно нестаціонарним процесом. Також для ідентифікації складних закономірностей методами машинного навчання необхідно мати достатньо великі навчальні вибірки з достатньою кількістю інформативних ознак, таких як стан пацієнтів, поведінка в

різних регіонах, відвідування різних закладів тощо. В даний час такі особливості аналізуються різними фахівцями, і коли такі дані стануть широко доступними, методи машинного навчання зможуть показати свою ефективність [74, 95, 99].

Також важливо оцінити невизначеність прогнозу, межі зміни прогнозних значень. Одним із вигідних підходів, на нашу думку, є використання байєсівського висновку, який базується на теоремі Байєса [124]. Метод найменших квадратів дозволяє знайти постійні коефіцієнти для моделей і, відповідно, деяке прогнозоване значення. За допомогою байєсівської регресії можна знайти розподіли для параметрів моделі та відповідно оцінити невизначеність прогнозування, що важливо для невеликої кількості даних.

1.2.4 Аналіз методів класифікації та кластеризації для попереднього опрацювання персоналізованих медичних даних

Для пошуку ключових характеристик даних, вибору основних об’єктів у вибірці даних здійснено аналіз методів зменшення розмірності, а також класифікації та кластеризації. Також задачею попередньої обробки даних є заповнення пропусків даних (імпутація).

Байєсові мережі, алгоритми ШНМ, метод кластеризації k-середніх використовуються в [54] для попереднього аналізу захворювань серця. Однак байєсівські мережі занадто повільні для обробки величезної кількості даних та для он-лайн діагностики. Однак з використанням паралелізації, доцільно використовувати байєсівські мережі в поєднанні з іншими методами машинного навчання для аналізу багатопараметричних, масштабних та динамічних потоків медичних даних.

Апарат штучних нейронних мереж з використанням нечіткої логіки також активно використовується для аналізу різних медичних даних. Так, у працях Є. Бодянського, І. Перової [56 – 59] було запропоновано систему швидкої медичної діагностики на основі багатоасоціативної нейронечіткої пам'яті. Однак одним із важливих завдань залишається підвищення точності результатів класифікації. Крім того, проблема дисбалансу вхідних даних, невелика кількість зразків даних,

зібраних вручну медичним персоналом, нав'язують ряд обмежень на використання існуючих методів та засобів обчислювального інтелекту для розв'язання такої проблеми [60].

Кластерний аналіз широко використовується для знаходження викидів. Викиди в медичній практиці означають різницю між оптимальними станами пацієнта на основі місцевого протоколу та індивідуальних особливостей. Одним з найпростіших алгоритмів кластеризації є клас методів, побудованих на основі розділення. Алгоритм K-means (k-середніх) будує k кластерів, розташованих на великій відстані один від одного. Основним типом задач, що вирішуються за допомогою алгоритму k-середніх, є припущення (гіпотеза) про кількість кластерів та різноманітність екземплярів у різних кластерах. Вибір числа k може базуватися на результатах попередніх досліджень та теоретичних міркуваннях [61].

Для алгоритму BIRCH (Збалансоване ітеративне скорочення та кластеризація з використанням ієрархій) швидкість кластеризації збільшується шляхом узагальненого вигляду кластерів. Цей алгоритм реалізує двоступеневий процес кластеризації. На першому етапі формується попередній набір кластерів. На другому кроці до виявлених кластерів застосовуються інші алгоритми кластеризації, придатні для оперативної пам'яті [62].

Алгоритм DBSCAN використовується для пошуку скупчень різних розмірів та форм. Якщо межа кластера містить більше точок, ніж мінімальна кількість об'єктів, створюється новий кластер з кореневим об'єктом. DBSCAN ітеративно збирає об'єкти безпосередньо близько до кореневих об'єктів, які можуть об'єднати кілька важкодоступних кластерів. DBSCAN не вимагає попереднього визначення кількості отриманих кластерів, на відміну від методів розділення. Хоча, гіперпарамерами цього алгоритму є значення радіусних параметрів будь-якого об'єкта та мінімальна кількість об'єктів, які безпосередньо впливають на результат кластеризації. Оптимальні значення цих параметрів важко визначити, особливо для багатовимірних просторів даних [63].

Використовуються три основні класи алгоритмів виділення важливих ознак ознак – фільтри, обгортки та вбудовані алгоритми [221].

Фільтри базуються на деяких показниках, які не залежать від методу класифікації, наприклад, співвідношення ознак з цільовим вектором і критеріями інформативності. Фільтри зазвичай застосовуються перед класифікацією. Найсуттєвішою перевагою фільтрації є те, що її можна використовувати як попередню обробку для зменшення розмірності простору ознак та подолання перенавчання моделі машинного навчання. Методи фільтрації, як правило, швидкі. Фільтри використовуються для вибору ознак у кластеризації або для побудови початкового наближення [222]. На жаль, такі методи не призначені для виявлення складних зв'язків між елементами і, як правило, недостатньо чутливі, щоб визначити всі залежності в даних.

Вбудовані алгоритми організовують виділення ознак під час навчання класифікатора, і саме вони явно оптимізують набір ознак, що використовуються для досягнення кращої точності [223]. Переваги вбудованих алгоритмів полягають у тому, що вони, як правило, швидко знаходять рішення, зменшують можливість перенавчання даних, усувають необхідність розділяти дані на навчальну та тестову підвибірки. Тим не менш, ці алгоритми зазвичай використовуються лише для певних наборів даних.

Обгортки покладаються на інформацію про значущість ознак із кількох класифікацій або регресійних моделей і таким чином можуть знаходити глибші шаблони в даних, ніж фільтри. Обгортки можуть використовувати будь-який класифікатор, який визначає ступінь значущості ознак [224].

Імпутація (заповнення) – це процедура оцінки невідомих або відсутніх значень на основі доступних даних, яка дозволяє сформувати повний набір даних із деякими правдоподібними оцінками.

Методи заповнення поділяються на немодельні та модельні. Існують підходи, засновані на методах одноразового та багаторазового заповнення з точки зору кількості значень, отриманих від методів імпутації [225]. Алгоритми одноразового заповнення формують єдиний повний набір даних, замінюючи кожен пропуск. Перевагами цього методу є використання методів комплексного аналізу даних на наступних етапах обробки. Алгоритми багаторазового заповнення

формують кілька повних наборів даних, які аналізуються окремо та пізніше об’єднуються відповідно до певних правил. Це мінімізує стандартні помилки на наступних етапах шляхом обробки повних наборів даних. Тим не менш, згадані методи вимагають наявності ресурсів для створення більшої кількості наборів даних, витрачають більше часу на виконання аналізу та вимагають більше пам’яті для зберігання результатів [226].

Метод, заснований на підстановці середнього, вирішує проблему неповних даних шляхом заміни кожної відсутньої змінної середнім значенням. Існують такі типи замін: медіанне значення, середнє значення для підгрупи [227], значення з найвищою частотою та заміна на мінімальне / максимальне значення. Цей метод може призвести до небажаних результатів [228], таких як зміна дисперсії, негативний кореляційний зсув і спотворення генеральної сукупності.

Підхід гарячої імпутації замінює кожен пропуск випадковим значенням, взятим із існуючого набору даних [229]. Його істотним недоліком є викривлення кореляцій і коваріат.

Підхід холодної імпутації (CD) реалізує заміну кожного розриву деяким постійним значенням із зовнішнього джерела [230]. Специфічний випадок CD – нульова заміна. Він має ті самі недоліки, що й гаряча імпутація.

Модель регресії використовується для заміни пропусків даних очікуваними значеннями, отриманими з рівняння регресії, побудованого з повного набору даних. Недоліки регресії: необхідність точного визначення регресійних моделей, збільшення кореляції та коваріації, ймовірність переходу до прогнозованих значень поза логічною послідовністю та потреба у великій кількості даних для отримання узгоджених оцінок.

Метод асоціативних правил (AR) використовує побудовані асоціативні правила для імпутації (заповнення) даних [231, 232]. Однак часова складність цього методу потребує вдосконалення [233].

Існуючі дослідження використовують стандартні методи машинного навчання для усунення пропусків, але також демонструють не дуже високу точність прогнозу. Так, у роботі [234] для визначення біомаркерів COVID

використовуються вибірка ознак, XGBoost та дерево рішень, причому показник F1 не піднімається вище 0,7. У роботі [235] маркери імунодефіциту CH4 і CH8 та їх асоціацію з коронавірусною інфекцією аналізували за допомогою статистичних моделей, у тому числі моделі Кокса. У роботі [236] для прогнозування епідемії COVID-19 використовується емпірична модальна декомпозиція (EEMD) і штучна нейронна мережа (ANN).

Отже, за результатами аналізу визначені задачі опрацювання персоналізованих медичних даних:

- анонімізація даних;
- попередня обробка даних (вибір важливих ознак, очистка даних, заповнення пропусків даних);
- консолідація даних;
- розроблення моделі даних;
- оцінка якості моделі даних.

У таблиці наведено результати порівняльного аналізу застосування класичних методів для опрацювання медичної інформації [249,266, 279].

Таблиця 1.1. Порівняльний аналіз застосування класичних методів

Метод	Для чого признач.	Дослідники	Переваги	Обмеження
Байєсовий підхід (Naïve Bayes)	Для класифікації ознак стану пацієнта, визначення ймовірності приналежності елемента вибірки до одного з класів при припущенні незалежності змінних [15,16,37,39,88]	Нільсон Н., Бідюк П. І. Wang, Tsai Ip RH, Ang LM, Seng KP, Tang Yan	Стійкий до шумів, може бути розпарарелеле-ний	Використання більше ніж 7 параметрів вимагає введення евристик, припущення про незалежність параметрів

Асоціативні правила	Знаходження асоціативних правил для визначення важливості параметрів стану пацієнта [58, 59, 90, 148, 231]	Арсеньєв Ю. Н., Дюк В. А., Субботін С.О., Hunyadi D., Шаховська Н., Runger, G. C.	Висока точність і стабільність	Працює лише з бінарними ознаками інформаційних об’єктів, «не знаходять» асоціативних залежностей з малою підтримкою
Класифікатори (Decision tree, Random Forest, SVM, gradient boosting)	Прогнозування цільових значень з високою точністю, щоб допомогти у пошуку нових показників стану пацієнта [62, 100, 145, 231, 233, 257]	Ramírez-Rubio, Nair LR, Shetty SD	Пояснює процес прийняття рішень, дерева рішень і Випадковий ліс може бути застосований для визначення важливих ознак	Наявність факту невизначеності ускладнює пошук оптимального рішення
Методи кластеризації (DBSCAN, k-means, hierarchical clustering)	Ідентифікація класів пацієнтів, для пошуку цільових рішень у відповідності до протоколів [23, 45, 220]	Zimek, Arthur; Campello Ricardo; Schubert Erich; Gribel, Daniel; Wang, Tsai, Березький О.М.	Створює ефект більш «уважного» відношення до предмету експертизи, знаходять нелінійно роздільні кластери	За неоднозначності результатів виникає багато альтернативних рішень, що ускладнює процес прийняття лікарських рішень
Нейронні мережі	Прогнозування ситуаційних рішень при існуючому стані пацієнта та для прогнозування станів пацієнта відповідно [18, 28, 76, 85, 219, 254, 302]	Кореневський А.Н., Мужичик А.В., Бодянський Є. В., Зайченко Ю.П., Ткаченко Р.О., Мисник А. В., Lei Zhang, Attallah O.	Навчена штучна нейронна мережа має високу чутливість та специфічність у прогнозуванні	Довгий час навчання

1.3 Формулювання проблеми дослідження

У дослідженні вирішується проблема пошуку персоналізованих рішень в контексті парадигм персоналізованої медицини, для оптимізації процесів оцінки та аналізу індивідуальних характеристик пацієнтів, а також прогнозування їхнього стану під час лікування чи реабілітації.

Окреслена проблема полягає у врахуванні класичного підходу щодо пошуку медичних рішень, який передбачає застосування протоколів лікування, що частково враховують індивідуальні особливості хворих. Водночас, виникає потреба у розробленні нових чи удосконаленні існуючих методів та проєктуванні програмних застосунків для підвищення якості медичних рішень за допомогою персоналізованої обробки та аналізу даних.

Наявний математичний апарат дає змогу опрацьовувати окремо часовозалежні або часовонезалежні дані, що унеможливлює визначення стану пацієнта з використанням методів класифікації і кластеризації з одного боку. З другого боку, прогнозування наступного стану можливо лише для одного пацієнта, а не для пацієнтів з подібними характеристиками, що значно знижує якість прийнятих медичних рішень. Тому проблема персоналізації медичних рішень для підвищення їх точності є актуальною.

В рамках дослідження повинні бути зібрані (або реконструйовані з медичних записів) наявні дані та інформація з дослідження захворювань. У разі відсутності відповідних даних будуть зібрані додаткові дані. Ці дані, включаючи лікування, діагностику, інформацію про результати здоров’я, спочатку мають бути проаналізовані для визначення деяких закономірностей. Для створення індивідуальних моделей прогнозування ризику будуть використовуватися машинне навчання, глибоке навчання або штучні нейромережеві алгоритми. Моделі повинні враховувати індивідуальний генетичний, соціальний зв’язок, навколишнє середовище та інші відповідні дані (які мають бути визначені).

Аналіз інформації про пацієнтів допоможе виявити загальні соціальні характеристики і, таким чином, сприятиме розробці рекомендацій щодо профілактики захворювань.

Усе подане вище дає змогу застосовувати індивідуальний підхід до лікування, аналізувати залежності між різними параметрами та запобігати захворюванню на основі застосування:

- статистичних моделей та методів ШІ для персоналізованого прогнозування ризику;
- методів запобігання захворювання, побудованих на модифікованих асоціативних правилах та ієрархічних правилах;
- ймовірнісних нейронних мереж та їх модифікації для он-лайн діагностики та цільового втручання.

Можна окреслити список проблем, що виникають під час персоналізованого підходу щодо надання лікування:

- індивідуалізація проблеми передбачає визначення вмісту та характеру даних для досягнення очікуваної мети та впливу;
- адаптація затверджених протоколів лікування для пошуку персоналізованих рішень;
- прогнозування результату медичних рішень;
- оцінка якості запропонованих медичних персоналізованих рішень;
- визначення обізнаності фахівців для збору точних та релевантних даних.

Проте, виникають обмеження застосування універсальних методів машинного навчання для великих та малих наборів мультимодальних медичних даних одночасно а саме:

- залежності якості моделі машинного навчання від результатів попередньої обробки даних;
- підвищення повторюваності результатів під час пошуку персоналізованих рішень;
- забезпечення генералізації при опрацюванні малої вибірки медичних даних слабкими предикторами.

Пошук персоналізованих рішень з використанням класичних методів оптимізації та зокрема класичних методів штучного інтелекту є недостатньо

ефективним для опрацювання індивідуальних показників пацієнта, що характеризуються викидами, шумами та пропусками даних.

При роботі з медичними даними властиво врахування існуючих медичних стандартів їхнього опрацювання та інформаційних процесів щодо підвищення точності опрацювання, зберігання та передачі персоналізованих даних, які різняться за структурою, джерелами надходження.

У зв'язку з цим виникає наукове протиріччя: необхідність опрацювання та аналізу великих та малих наборів мультимодальних медичних даних для підтримки прийняття рішень лікарем та обмеженість можливостей застосування існуючих моделей машинного навчання.

Методи в задачах кластеризації чи класифікації орієнтовані на опрацювання часовонезалежних даних. Але при опрацюванні даних пацієнта ми маємо вхідний вектор даних, який включає часовозалежні дані, що призводить до зміни стану хворого в часовому вимірі. В різні проміжки часу пацієнт може належати до різних кластерів.

Отже, процеси опрацювання даних з використання класичних методів машинного навчання, орієнтовані для прогнозування стану окремого хворого, а для групи пацієнтів з однаковими параметрами, що у часі можуть змінюватися під впливом різних факторів, знижується точність прийнятих персоналізованих медичних рішень.

Тому проблема розробки нових чи удосконалення існуючих методів штучного інтелекту щодо підвищення точності процесу обробки великих та малих мультимодальних медичних даних для пошуку персоналізованих рішень є актуальною науково-прикладною проблемою.

Отже, метою дисертаційного дослідження є розроблення моделей, методів машинного навчання та засобів для підвищення прогнозованої точності та візуалізації результатів опрацювання персональних даних щодо оцінки стану хворого, що забезпечить якісний процес прийняття персоналізованих медичних рішень.

Висновки до 1 розділу

1 Введено поняття персоналізації медичних даних та персоналізованої медицини, що характеризуються процесом пошуку рішень лікаря для розв’язання проблеми призначення або корекції лікування пацієнта, враховуючи індивідуальні особливості його стану та різні фактори впливу Визначені особливості застосування та опрацювання такого типу даних.

2 Важливим фактором опрацювання медичних даних є потреба опрацьовувати інформацію пацієнта опираючись на протоколи хвороб, але не лише цієї хвороби, з якою звернувся хворий, але й з багажем його супутніх хвороб та особливостей його стану з врахуванням протокольних вимог щодо зберігання та передачі персональних даних. Актуальність проблеми персоналізації зумовлена протиріччям, що лікар змушений дотримуватися протоколів хвороб та адаптовувати їх безпосередньо до індивідуальних характеристик хворого.

3 Аналіз міжнародних стандартів та внутрішніх законодавчих вимог показав, що є змога збирати анонімізовані дані без згоди пацієнта, використовувати їх для будь-яких цілей і зберігати протягом невизначеного часу – враховуючи те, що видаляють з даних усі ідентифікатори.

4 Проаналізовано існуючі практичні рішення, що відображають основні підходи до використання медичних даних та урахування персоналізації. За результатами аналізу існуючих систем штучного інтелекту у сфері медицини для вирішення проблем діагностики захворювань, дослідження геному, розробки ліків, медичної візуалізації та ін. виокремлено проблеми, що ще залишаються не розкритими, а саме: консолідація даних про хворого, його індивідуальних особливостей, враховуючи міжнародні стандарти лікування та існуючу фармацевтичну продукцію; визначення персоналізованого лікування та прогнозування результатів його застосування.

5 За результатами порівняльного аналізу класичних методів визначено обмеження, які характерні при опрацюванні медичних персоналізованих даних, а саме: використання більше ніж 7 параметрів вимагає введення евристик,

припущення про незалежність параметрів, опрацювання лише бінарних ознак інформаційних об’єктів, «не знаходять» асоціативних залежностей з малою підтримкою, наявність факту невизначеності ускладнює пошук оптимального рішення, за неоднозначності результатів виникає багато альтернативних рішень, що ускладнює процес прийняття лікарських рішень, довгий час навчання, чутливість до шуму.

6 Визначені обмеження існуючих методів опрацювання та аналізу великих та малих наборів мультимодальних медичних наборів даних дали змогу сформувати науково-прикладну проблему розроблення нових чи удосконалення існуючих методів штучного інтелекту для підвищення прогнозованої точності та візуалізації результатів опрацювання персональних даних щодо оцінки стану хворого.

РОЗДІЛ 2.

ФОРМАЛІЗАЦІЯ МОДЕЛІ ПРОСТОРУ СТАНУ ПАЦІЄНТА

У другому розділі побудовано модель стану пацієнта, представленої у вигляді сукупності множин, які взаємно залежні та залежні від середовища оцінки; визначено продукційні правила, які формулюють рішення щодо перегляду та зміни тактики лікування. Представлено простір стану пацієнта як евклідовий простір, що дозволило змоделювати інформаційну модель простору станів як багатовимірну систему з врахуванням фактору часу. Формалізовано відображення фізичного стану пацієнта з урахуванням часовозалежних та часовонезалежних даних пацієнта, що дає можливість оцінити його стан в певний момент часу. Розроблено інформаційну модель простору станів пацієнта та представлено у вигляді гіперкуба, як відображення функціонального відношення загального стану пацієнта. Розроболено метод пошуку шаблонів, який базується на модифікації методу асоціативних правил, що дозволяє зменшити час затрачений на процес формування рішення та використовувати паралельний та розподілений режим для розрахунку. Цей метод є удосконаленням методу упорядкованого пошуку та надає чіткості та направленості у пошуку рішень стосовно вибору цільових схем лікування, що дає змогу зменшити ймовірність появи похибки при виборі схеми лікування.

Матеріали розділу опубліковані у роботах автора [162, 163, 171, 175, 177, 178, 180, 181, 182, 188, 193, 196, 197, 203, 205, 209, 214, 215, 216, 217].

2.1 Формалізація моделі відображення стану пацієнта

Визначення індивідуальних характеристик, необхідних для вирішення задачі персоналізації, залежить від ключових факторів ідентифікації інформаційного об'єкта. Для формалізованого представлення пацієнта в медицині розглядаються основні параметри його загального стану з певними характеристиками.

Визначення необхідних індивідуальних характеристик для розв’язання проблеми персоналізації залежить від ключових факторів ідентифікації хворого. Для формалізованого представлення стану пацієнта беруть до уваги основні параметри його загального стану з визначеними його характеристиками.

Аналізуючи стан хворого під час лікування, експерти визнали, що важливим показником одужання є позитивна динаміка зміни основних показників загального стану у часі, а саме: результати мікробіологічних досліджень, температурні показники, аналіз поширення запальних процесів і т.д. Це свідчить про те, що важливим критерієм аналізу є значення основних показників, які змінюються в часі.

Для прикладу, на перебіг хвороб, викликаних інфекціями чи вірусами (навіть з відомими схемами запобігання на лікування) впливають різні фактори, а саме:

- варіабельність штамів,
- тип взаємодії,
- характеристика території поширення: кліматичні умови, розвиток інфраструктури та комунікацій, якість медичного обслуговування, хронічні захворювання, властиві цій території, політична ситуація тощо.

Саме тому розроблення імітаційних моделей поширення та протікання захворюваності різного роду інфекції та вірусів є складною науковою задачею. Ця задача характеризується:

- множинними критеріями;
- залежністю від часу;

- інтервалами моделювання;
- гетерогенністю вхідних даних.

Багатокритеріальність визначається під впливом типу поширення (епідемічне поширення, контрольоване поширення у легкій формі захворювання), початковими даними поширення, територією поширення.

Для об'єктивної оцінки стану пацієнта необхідно побудувати його формальну модель, що дозволило б представити його як деяку модель, що відображає його структуру, подає інформацію про його стан та поведінку.

Отже, модель стану пацієнта представлена як система, що консолідує різні елементи, подані у вигляді множин, які взаємнозалежні та залежні від середовища оцінки.

Формалізуємо окремі елементи інформаційної моделі стану пацієнта:

База знань представлена як множина правил R [28,35,80]. Припускається, що множина R – множина персоналізованих рішень, які має скінченний розмір *rank(R)*.

$$R = \{r_1, r_2, \dots, \}; \tag{2.1}$$

Pd – множина персоналізованих даних.

$$Pd = \{p_1, p_2, \dots, \}. \tag{2.2}$$

Для прийняття рішень використовуємо продукційні правила з множини R. При цьому встановлюється залежність між множиною персоналізованих даних та оцінкою стану пацієнта *ES*:

$$R: Pd \to ES. \tag{2.3}$$

Зазначимо, що елемент множини персоналізованих рішень подано як кортеж:

$$p_{\text{i}} = < A_{in}, A_t >, \tag{2.4}$$

де часовонезалежні характеристики подані як A_{in} та часовозалежні параметри пацієнта подані як A_t. Тоді *ES* – це множина оцінок стану пацієнта, що залежить від множини коефіцієнтів оцінки *K*.

Далі у продукційній моделі в опис продукції введено передумови та постумови у вигляді:

$$< G, A_t, Pd \rightarrow ES, K >, \quad (2.5)$$

де

– імплікація $Pd \rightarrow ES$ представляє власне правило;

– G – передумова вибору класу правил (протокольне рішення);

– A_t – передумова вибору правила у класі (у нашому випадку це часовозалежні параметри, які визначають особливості стану пацієнта);

– K – постумова правила, що визначає перехід на наступне правило (це множина коефіцієнтів оцінки стану пацієнта, що вказують на зміни результативного показника при зміні значення вхідної змінної).

Вибір правил здійснюється на основі багатокритеріального вибору Парето, де $ES = \{e_1, e_2, \dots, \}$ – векторна оцінка стану параметрів пацієнта для певних елементів множини параметрів:

$$S(ES) = \{x \in ES \mid \forall \mathrm{y} \in ES \quad [\forall \mathrm{i} \in \{1,\dots, \mathrm{m}\}[x_i \geq y_i]]\}. \quad (2.6)$$

Прикладом правил є пошук оптимальних станів та стратегічних рішень щодо лікування індивіда на основі вибраних характеристик та параметрів.

Склад множин A_{in}, A_t залежить від конкретної задачі.

У загальному випадку залежності даних можна розділити на лінійні та нелінійні. І якщо для лінійних залежностей інформативні фактори визначаються за відомими методами кореляційного та дисперсійного аналізу, то для нелінійних залежностей такі процедури найчастіше є емпіричними. У нашому випадку залежності прогнозуються як лінійні.

Параметри перед аналізом необхідно підготувати, а саме: необхідно визначити пріоритетні ознаки, що впливатимуть на результати. Важливість пошуку залежностей між даними оптимізує процес визначення наявного захворювання. Так, на рисунку 2.1 подано параметри особи для визначення наявності захворювання COVID-19.

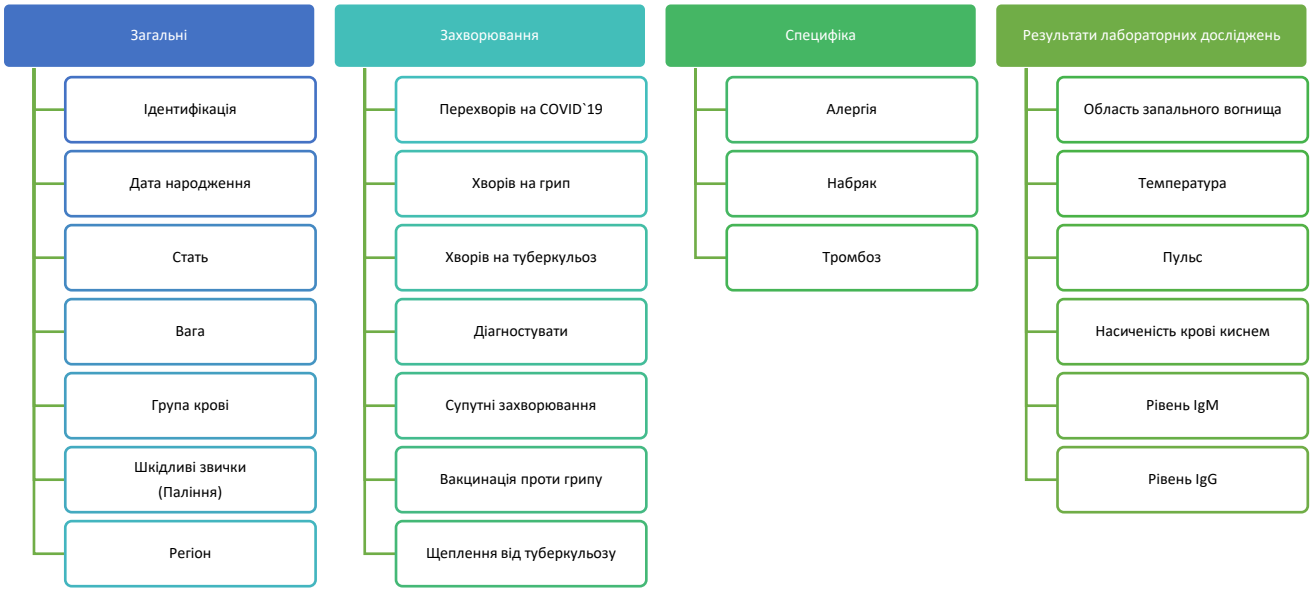


Рисунок 2.1. Параметри особи для визначення наявності захворювання COVID-19.

Отже, для заданого параметра пацієнта a визначається коефіцієнт оцінки стану параметрів пацієнта k, що вказує на зміни результативного показника при зміні значення вхідної змінної, на основі чого визначається оцінка стану параметрів пацієнта e для певних елементів множини параметрів пацієнта a та відповідного коефіцієнту оцінки k:

$$\forall aFK(a,k) \rightarrow \exists a \exists kFKS(a,k,e), \tag{2.7}$$

де FK – предикат для визначення коефіцієнтів оцінки k, релевантних до заданого параметру a; FKS – предикат для визначення параметрів оцінки стану пацієнта e на основі параметрів пацієнта a та коефіцієнт оцінки стану параметрів пацієнта k.

Виокремлення часовозалежних та часовонезалежних даних формує множину даних, які є необхідними для персоналізації даних хворого (вибір важливих ознак та вибір елементів з набору даних) у процесі пошуку індивідуального підходу щодо вибору лікарем стратегії лікування.

Слід зазначити, що часовозалежними та часовонезалежними даними можна вважати такі, що наведені в таблиці 2.1.

Таблиця 2.1. Часовозалежні та часовонезалежні параметри

Часовозалежні дані (A_t)	Часовонезалежні дані (A_{in})
діагноз, бактеріальний збудник, площа запального вогнища, супутні хвороби, попередня терапія, тривалість застосування, доза застосування, вага пацієнта, температурна карта, частота серцевих скорочень, алергічний анамнез, наявність набряків, лабораторні дослідження, та ін.	Прізвище Ім’я вік ріст професія анамнез хвороб спадковість шкідливі звички будова тіла конституція тіла та ін.

З використанням теорії функціонального аналізу часовозалежні дані можна подати як множину A_t, елементами якої є підмножини показників індивідуальних параметрів хворого $A_1, A_2, \ldots, A_n$:

$$A_t = \{A_1, A_2, \ldots, A_n\},$$

де

$$A_1 = \{a_{11}, a_{12}, \ldots, a_{1m}\},$$
$$A_2 = \{a_{21}, a_{22}, \ldots, a_{2m}\},$$
$$A_n = \{a_{n1}, a_{n2}, \ldots, a_{nm}\},$$

а саме:

$$A_t = \{(a_1, a_2, \ldots, a_n) | a_1 \in A_1, a_2 \in A_2 \ldots, a_n \in A_n\}. \quad (2.8)$$

Модель стану пацієнта, що забезпечує пошук оцінки його стану та пошук рішень щодо оптимізації процесу одужання, подана як:

$$Fe = \langle A_{in}, A_t, K, ES, S, G, R \rangle, \quad (2.9)$$

де Fe – система оцінки стану пацієнта, A_{in} – це множина часовонезалежних параметрів пацієнта або вхідних даних системи, що характеризують показники-фактори пацієнта, що отримують з динаміки змін його стану, які розраховуються на підставі інформації, що є в історії хвороби, K – це множина коефіцієнтів оцінки стану пацієнта, що вказують на зміни результативного показника при зміні значення вхідної змінної, S – множина стратегічних рішень щодо оцінки стану пацієнта, R – продукційні правила рішень щодо стратегії пошуку оптимального стану, ES – це множина оцінок стану пацієнта, що залежить від коефіцієнтів оцінки, A_t – характеристики пацієнта, що змінюються під впливом часу, G – протокольні рішення.

Для того, щоб врахувати особливості оцінки стану пацієнта, необхідно мати індивідуальний підхід до обробки його даних. Це допоможе визначити особливості пацієнта та особливості його індивідуальних характеристик.

Для представлення системи оцінки стану пацієнта та відповідної залежності між елементами використано формалізацію теорії множин та бінарних відношень (рисунок 2.2).

Рисунок 2.2 Відображення залежності множин даних системи оцінки стану пацієнта

При цьому залежності між сутностями системи можна представити бінарним відношенням на частковому прикладі, таблиця 2.2.

Таблиця 2.2**.** Частковий приклад бінарного відношення визначення коефіцієнтів оцінки, де k ∈ [0,1].

P	Показники	K	Коефіцієнти
a_1	Наявність супутніх хвороб	k_1	Коефіцієнт можливих ускладнень
a_2	Вік	k_2	Коефіцієнт надійності
a_3	Наявність алергії на медикаменти	k_1	Коефіцієнт можливих ускладнень
a_4	Протокол лікування	k_2	Коефіцієнт надійності
a_5	Вага	k_3	Коефіцієнт ризику
a_6	Температура	k_4	Коефіцієнт стану
a_1	Наявність супутніх хвороб	k_3	Коефіцієнт ризику
a_7	Лабораторні дослідження	k_4	Коефіцієнт стану
a_8	Площа запального вогнища	k_5	Коефіцієнт поширення
a_3	Наявність алергії на медикаменти	k_3	Коефіцієнт ризику

Значення отриманих коефіцієнтів визначають результат оцінки поточного стану об’єкта дослідження *ES*. Частковий приклад поданий у формулі 2.10.:

$$K_1(a_1,k_1) \vee K_2(a_2,k_2) \vee K_3(a_3,k_1) \vee K_4(a_4,k_2) \vee K_5(a_5,k_3) \vee K_6(a_6,k_4) \vee$$
$$\vee K_7(a_1,k_3) \vee K_8(a_7,k_4) \vee K_9(a_8,k_5) \vee K_{10}(a_3,k_3) \rightarrow \exists a \exists k ES(a,k,e). \quad (2.10)$$

Результатами оцінки визначають: оцінку поточного стану *ES*, стратегічні рішення *S*, що визначають тактику лікування, множину показників пацієнта *Pd*, протокольні рішення *G*.

Рішення щодо зміни стану пацієнта подано у вигляді n-ного відношення:

$$R = Pd \times K \times ES \times S \times G, \quad (2.11)$$

де × – декартів добуток.

Отримані правила визначають рішення щодо перегляду та зміни тактики лікування, а саме стратегічних рішень щодо зміни стану, зміни медикаментів, зміни

дозування, перелік лабораторних досліджень, фізіотерапевтичних заходів та ін. Водночас, значення параметрів-факторів визначають коефіцієнти оцінки стану.

На етапі аналізу результатів досліджень стану пацієнта, можемо сформувати властивість часовозалежних показників пацієнта, що часовозалежні дані у певний момент часу приймають сталі показники, які під впливом застосування персоналізованих рішень щодо лікування визначають зміну стану пацієнта.

Отже, завдання вибору лікувальної схеми зводиться до задачі розпізнавання та полягає в тому, щоб для кожного даного стану (пацієнта) ω, що є елементом множини станів P, за його описом часовозалежних ознак A_t і апріорною (навчальною) інформацією по часовонезалежних ознаках A_{in} обчислити значення предикатів:

$$FR_i(\omega) = (\omega \in P), \quad i = \overline{1,m}. \tag{2.12}$$

Таким чином, для розглянутого об'єкта (пацієнта) ω необхідно обчислити його інформаційний вектор:

$$\alpha(\omega) = \langle \alpha_1(\omega), \dots, \alpha_m(\omega) \rangle. \tag{2.13}$$

Процедура, що будує інформаційний вектор $\alpha(\omega)$ у цьому випадку виражає алгоритм ухвалення рішення про належність об'єкта до того або іншого класу й називається «вирішальною функцією».

За результатами фізичного стану хворого вирішальна функція надасть необхідні дані для виявлення аномалій лікування та виявлення тенденцій. Для цього ми агрегуємо різні параметри, які вказують на прогрес окремих пацієнтів: тривалість перебування в лікарні, динаміка змін стану здоров'я або одужання, рівень відновлення фізичної активності, повернення до роботи, смертність та ін. Враховуючи персоналізовані дані хворого під час процесу догляду за станом і за його межами, можемо змоделювати простір стану пацієнта в часі, як евклідовий простір, де кожній парі елементів a_1, a_2, поставлено у відповідність дійсне число (a_1, a_2), що задовільняє умови (аксіоми скалярного добутку):

$$(a_1, a_1) \geq 0, \text{при тому, що } (a_1, a_1) = 0, \text{коли } a_1 = 0,$$

$$(a_1, a_2) = (a_2, a_1),$$

$$(\lambda a_1, a_2) = \lambda(a_1, a_2),$$

$$(a_{11} + a_{12}, a_2) = (a_{11}, a_2) + (a_{12}, a_2)\,. \tag{2.14}$$

З урахуванням того, що простір стану пацієнта представлено, як евклідовий простір, можна змоделювати модель простору станів, як багатовимірну систему з врахуванням часу. Отже, за рахунок зміни індивідуальних характеристик пацієнта під час процесу лікування, модель простору його станів представлено рівнянням стану, що становить векторно-матричну форму запису системи диференціальних рівнянь першого порядку [191,195].

Рівняння стану має вигляд:

$$\frac{d}{dt}\overrightarrow{FS}(t) = A\overrightarrow{FS}(t) + B\vec{S}(t), \tag{2.15}$$

де $\overrightarrow{FS}(t)$ – вектор стану розмірності простору n, який включає параметри (характеристики) пацієнта, що однозначно визначають його стан:

$$\overrightarrow{FS}(t) = \begin{bmatrix} FS_1(t) \\ FS_2(t) \\ \dots \\ FS_n(t) \end{bmatrix}, \tag{2.16}$$

$\vec{S}(t)$ – вектор факторів впливу розмірності m, що впливають на систему ззовні, зарахунок пропонованих рішень щодо визначення терапевтичної схеми лікування:

$$\vec{S}(t) = \begin{bmatrix} S_1(t) \\ S_2(t) \\ \dots \\ S_m(t) \end{bmatrix}, \tag{2.17}$$

A, B – матриці параметрів, що включають в себе параметри системи, розмірність яких відповідно $n \times n$, $n \times m$:

$$A = \begin{bmatrix} a_{11}\, a_{12} \cdots a_{1n} \\ a_{21}\, a_{22} \cdots a_{2n} \\ \dots \quad \dots \quad \dots \quad \dots \\ a_{n1}\, a_{n2} \cdots a_{nn} \end{bmatrix}, B = \begin{bmatrix} b_{11}\, b_{12} \cdots b_{1m} \\ b_{21}\, b_{22} \cdots b_{2m} \\ \dots \quad \dots \quad \dots \quad \dots \\ b_{n1}\, b_{n2} \cdots b_{nm} \end{bmatrix}. \tag{2.18}$$

Рівняння стану можна записати в розгорнутій формі, формула 2.19 :

$$\frac{d}{dt}\begin{bmatrix} FS_1(t) \\ FS_2(t) \\ \dots \\ FS_n(t) \end{bmatrix} = \begin{bmatrix} a_{11}\, a_{12} \cdots a_{1n} \\ a_{21}\, a_{22} \cdots a_{2n} \\ \dots \quad \dots \quad \cdots \quad \dots \\ a_{n1} a_{n2} \cdots a_{nn} \end{bmatrix} \begin{bmatrix} FS_1(t) \\ FS_2(t) \\ \dots \\ FS_n(t) \end{bmatrix} + \begin{bmatrix} b_{11}\, b_{12} \cdots b_{1m} \\ b_{21}\, b_{22} \cdots b_{2m} \\ \dots \quad \dots \quad \cdots \quad \dots \\ b_{n1} b_{n2} \cdots b_{nm} \end{bmatrix} \begin{bmatrix} S_1(t) \\ S_2(t) \\ \dots \\ S_n(t) \end{bmatrix}. \quad (2.19)$$

Рівняння стану і структура описують вектор стану, що містить характеристики пацієнта, які описують його стан.

2.2 Побудова моделі простору станів пацієнта

Знаючи, що відображенням фізичного стану об'єкта є $FS(t): A_{in} \to A_t$, де $A_t = A_t(t)$ – функція зміни часовозалежних параметрів пацієнта, можемо представити з урахуванням усіх параметрів пацієнта його стан *P* у вигляді точки у просторі станів, як тривимірному просторі, де вісь *x* – відображає часовий показник $P_o(t)$, вісь *y* – параметричний показник $P_o(a)$, а вісь *z* – показник на окремі випадки, тобто множину об'єктів (пацієнтів) *Ob* (рисунок 2.3).

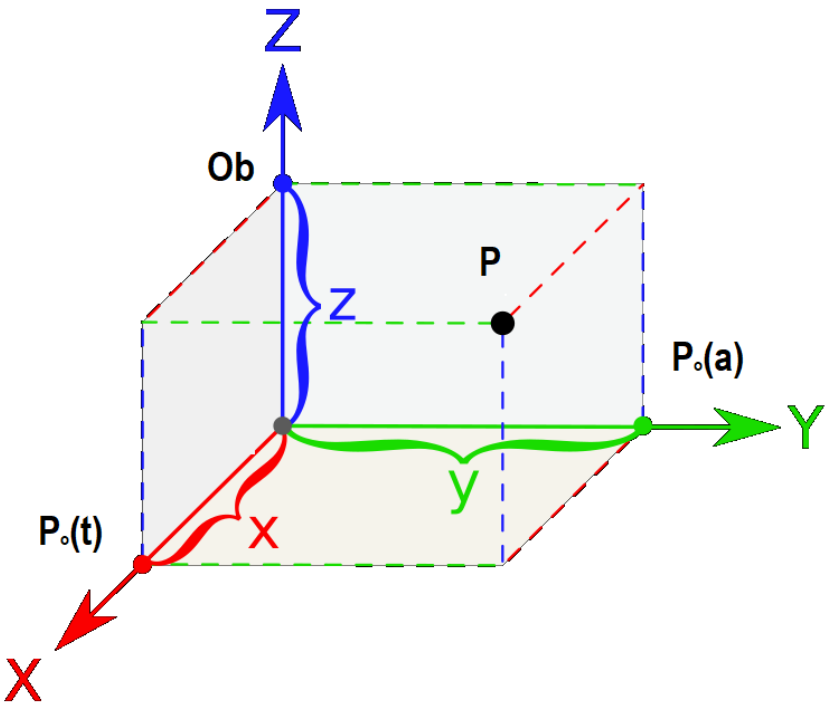


Рисунок 2.3. Простір станів пацієнта та його стан в певний проміжок часу

Таким чином, модель простору станів пацієнта представлено, як сукупність станів у вигляді гіперкуба, як відображення функціонального відношення загального стану пацієнта *GS*. Стан пацієнту відображений у вигляді точки *P* в декартовій системі просторів станів.

Отже, стан пацієнта представлено у вигляді точки *P(x,y,z)*.

Точка *P* має координати *час, параметр, пацієнт,* що записується, наприклад, як:

P(2022:20:10:14:20:56; 139/89 мм рт.ст.; Панасенко М.І.).

Будь-якій трійці показників *x, y, z* відповідає лише одна точка простору *P(x,y,z)*.

Визначаючи, що на одній часовій ітерації функціональне відношення фізичного стану пацієнта убезпечує залежність між часовим фактором та параметричним покажчиком, приймаємо:

$$FS_o(t) = FS_o(A_{in}, A_t(t)). \quad (2.20)$$

При умові аналізу фізичного стану пацієнта при наступній часовій ітерації, спостерігається залежність часових показників пацієнта від попередніх значень його фізичного стану:

$$A_t(t+1) = FS_o\big(A_{in}, A_t(t)\big), \quad (2.21)$$
$$FS_o(t+1) = FS_o(A_{in}, A_t(t+1)),$$
$$A_t(t+n) = FS_o\big(A_{in}, A_t(t+n-1)\big).$$

Тоді фізичний стан пацієнта з врахування часу представлено як:

$$FS_{o\,i} \subseteq \big(A_{in_i}, A_t(t)_{i1}\big) \otimes \big(A_{in_i}, A_t(t)_{i2}\big) \otimes \big(A_{in_i}, A_t(t)_{i3}\big) \otimes \ldots \otimes \big(A_{in_i}, A_t(t)_{in}\big), \quad (2.22)$$

де $\otimes$ – тензорний добуток просторів, який білінійно відображає вихідні простори станів.

На рисунку 2.4 зображено простір фізичних станів пацієнта.

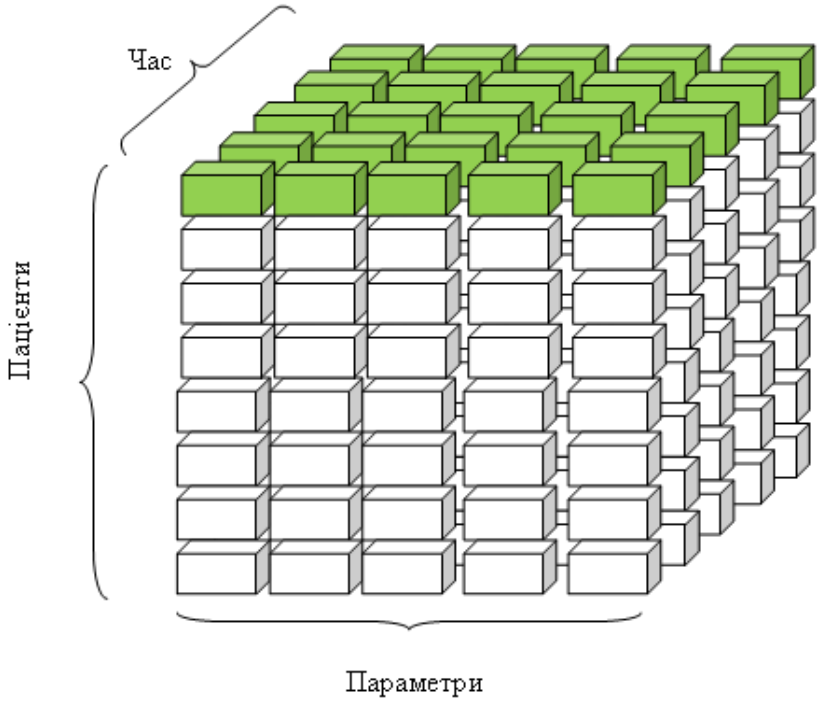


Рисунок 2.4. Простір фізичних станів пацієнта

Враховуючи, що FS_{o_i} це лише показник фізичного стану пацієнта під час процесу лікування чи реабілітації, то на основі (формула 2.22) можна проаналізувати поведінку стану пацієнта.

Важливим фактором є аналіз та застосування протоколів відповідно до ідентифікованого захворювання у пацієнта. Тому для формалізації просторів стану пацієнта, враховуючи при цьому його стан на відповідному етапі лікування, побудовано простір стану пацієнта з урахуванням протокольних рішень на різних етапах лікування:

$$PFS_{o_i} \subseteq FS_{o_{i1}}(t) \otimes FS_{o_{i2}}(t) \otimes \dots \otimes FS_{o_{in}}(t), \qquad (2.23)$$

$$PFS_{o_i} \subseteq FS_{o_{i1}}(A_{in_i}, A_t(t)_{i1}) \otimes FS_{o_{i2}}(A_{in_i}, A_t(t)_{i2}) \otimes \dots \otimes FS_{o_{in}}(A_{in_i}, A_t(t)_{in}),$$

$$PFS_o(t) = PFS_o(FS_o(t)),$$

$$PFS_o(t+1) = PFS_o(FS_o(t+1)).$$

На рисунку 2.5 зображено простір протокольних рішень для пацієнтів.

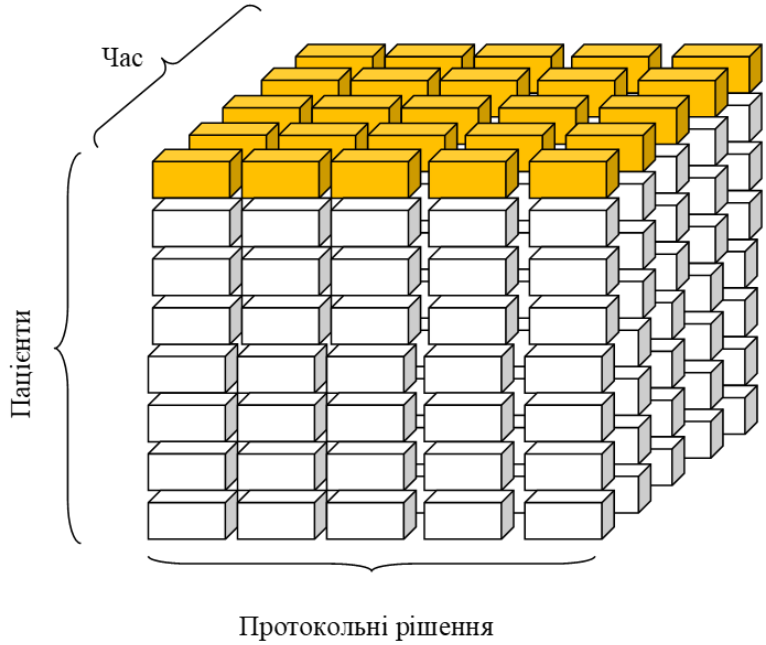


Рисунок 2.5. Простір протокольних рішень для пацієнтів

Відомий факт, що у багатьох хворих під час збору даних виявляється ще множина додаткових супутніх хвороб, на які обов’язково необхідно звернути увагу, через те, що кожна супутня хвороба потребує лікування з урахуванням діючих протоколів. Таким чином, необхідно врахувати при моделюванні простору станів конкретного пацієнта:

$$RPFS_{o_i} \subseteq PFS_{o_{i1}}(t) \otimes PFS_{o_{i2}}(t) \otimes \dots \otimes PFS_{o_{im}}(t), \quad (2.24)$$

$$RPFS_{o_i} \subseteq PFS_{o_{i1}}(FS_{o1}(t)) \otimes PFS_{o_{i2}}(FS_{o2}(t)) \otimes \dots \otimes PFS_{o_{im}}(FS_{on}(t)),$$

$$RPFS_o(t) = RPFS_o\big(PFS_o(t)\big),$$

$$RPFS_o(t+1) = RPFS_o(PFS_o(t+1)).$$

На рисунку 2.6 зображено простір протокольних рішень на основі простору умов та пов'язаних з ним захворювань.

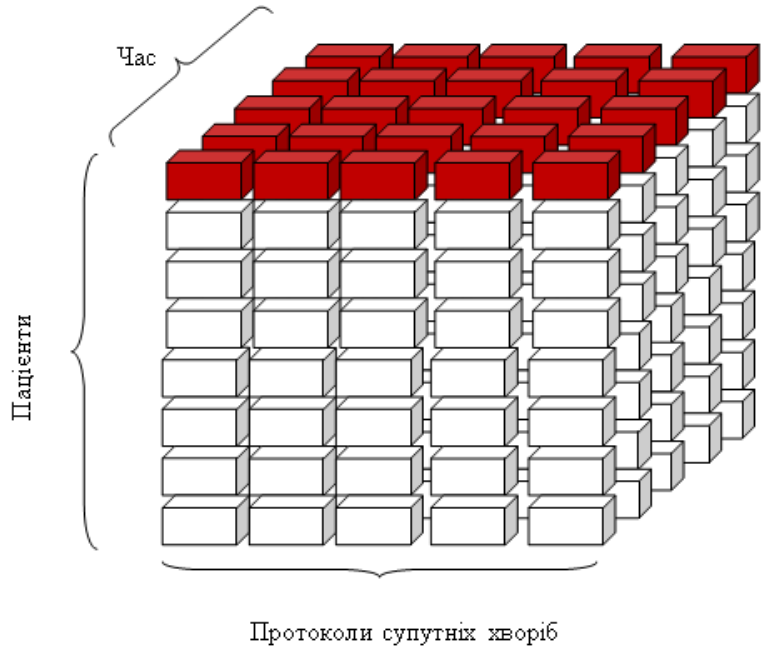


Рисунок 2.6. Простір протокольних рішень на основі простору умов та пов'язаних з ним захворювань

Як наслідок, результуючий стан пацієнта GSo – це результат застосування реляційної операції перетину між усіма просторами станів хворого в певний момент часу $RPFS_o(t)$, $PFS_o(t)$, $FS_o(t)$, що відображає необхідні персоналізовані рішення щодо лікування пацієнта:

$$GS_{oi} = RPFS_o(t) \cap PFS_o(t) \cap FS_o(t). \quad (2.25)$$

На рисунку 2.7 продемонстровано відображення залежності між персоналізованими даними та протокольними даними.

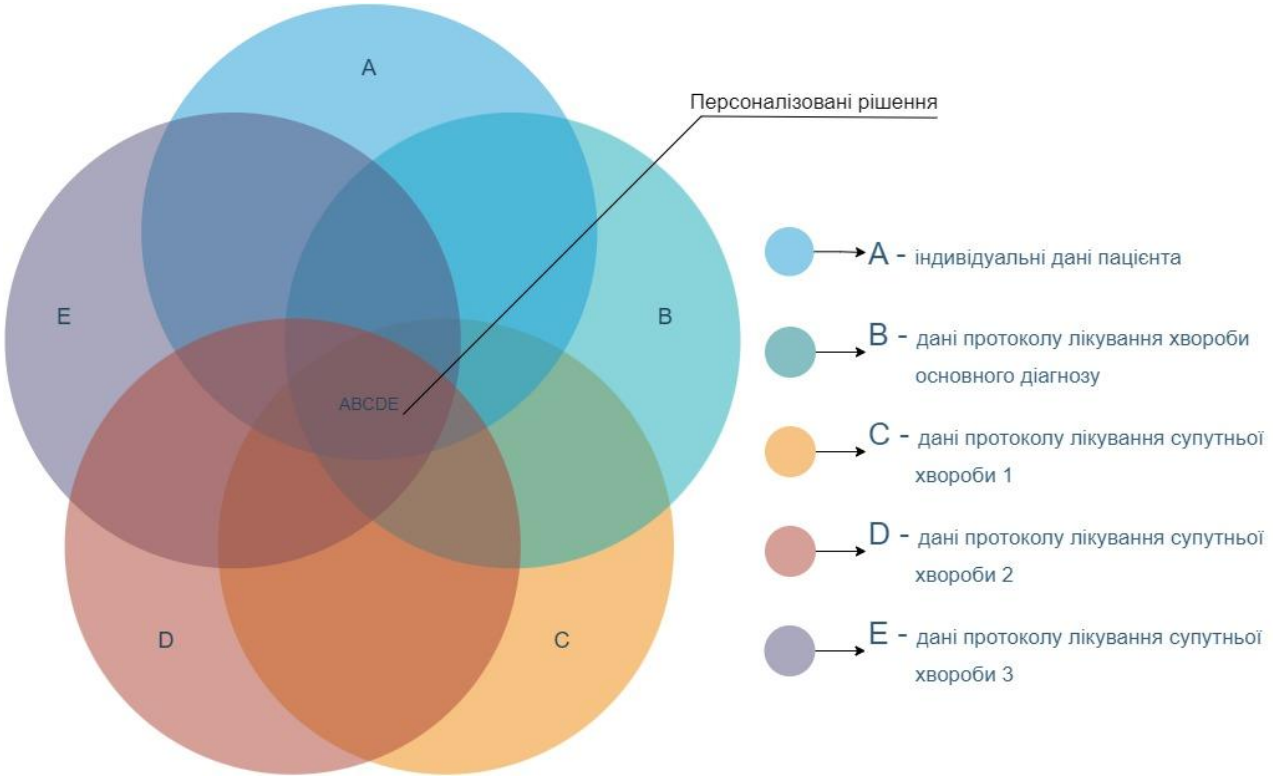


Рисунок 2.7 Відображення залежності між персоналізованими даними та протокольними даними

Схема залежності схем лікування від цільових параметрів індивіда, які враховують індивідуальні дані пацієнта та протокольні дані, подана на рисунку 2.8.

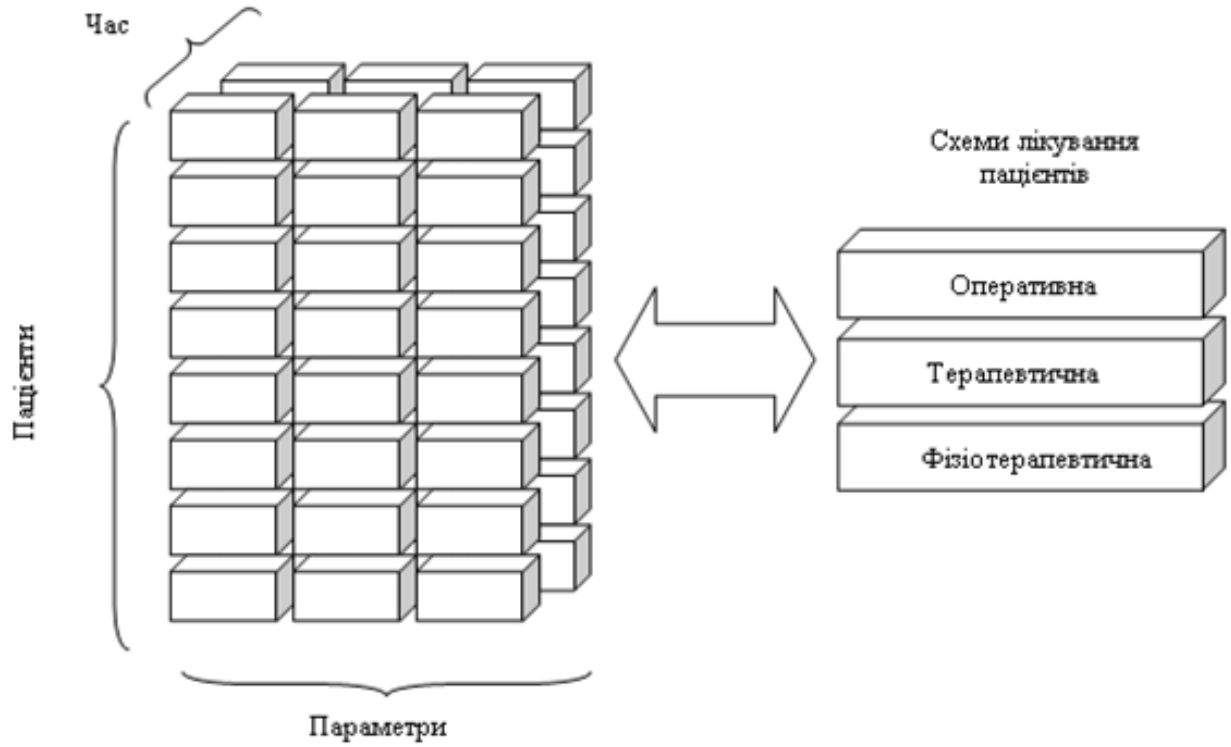


Рисунок 2.8. Схема залежності схем лікування від цільових параметрів індивіда

Наявність ефективного зберігання, доступу та модифікації інформації про стан об'єкта (пацієнта), а також об'єднання з фактичними даними потоку про досліджувану ситуаційну задачу, дозволить розробити структуру медичних даних. Для цього потрібно визначити структуру персоналізованих даних. Для вирішення такого класу задач необхідно формалізувати процес ідентифікації стану пацієнта шляхом побудови моделі простору його станів.

Отже, метою дослідження була побудова моделі консолідованих персоналізованих медичних даних пацієнта на підставі аналізу різнотипних його параметрів, а також побудови моделі простору станів досліджуваного хворого з урахуванням його індивідуальних характеристик, що враховуються під час процесу лікування.

2.3 Апробація моделі простору станів пацієнта для аналізу наборів даних

Схема дослідження подана таким чином:

- розвідковий аналіз даних (нормалізація ознак і кодування);

- побудова простору умов;
- розроблення та тестування моделей;
- валідація результатів.

Усі розрахунки зроблено за допомогою RStudio. Дані були передані через Data Sampler для збалансування. За результатами відбору об'єктів враховуються всі випадки. Зібраний набір даних є збалансованим.

Було опрацьовано персоналізовані дані пацієнтів із зібраного набору даних (https://doi.org/10.6084/m9.figshare.14865411.v1). Цей набір даних зібрано в хірургічному відділенні Львівської лікарні швидкої медичної допомоги (Україна). Хворі проходили клінічне лікування з приводу післяопераційних ускладнень на черевній порожнині.

Набір даних складається з таких характеристик:

- вік (залежний від часу параметр, показник ефективності A_{in}) – ціле;
- стать (параметр, що не залежить від часу) – логічний;
- вага (залежний від часу параметр, показник ефективності A_{in}) – категорійна змінна;
- дата поступлення – дата;
- діагноз (незалежний від часу параметр, використовується для протоколу $PFS_{o_i}\ choosing$) – категорійна змінна;
- пов'язаний діагноз (залежний від часу параметр, $RPFS_{o_i}$) – категорійна змінна,
- флора (залежний від часу параметр, $RPFS_{o_i}$) – категорійна змінна;
- медикамент (параметр, що залежить від часу, залежить від PFS_{o_i}) – категорійна змінна;
- діюча речовина (параметр, що залежить від часу, залежить від PFS_{o_i}) – категорійна змінна;
- час перебування в лікарні (параметр залежний від часу, ліжко-дні в лікарні, цільовий параметр) – ціле число;
- кожен екземпляр представляє результуючий стан пацієнта $GS_{o.}$

Завдання полягає в тому, щоб спрогнозувати кількість днів у стаціонарі (тривалість лікування) на основі медикаментозного лікування та особистісних параметрів пацієнтів.

Приклад 2.1.

Набір даних складається з 51 екземпляра та 10 параметрів. «Час перебування в лікарні» є цільовою змінною. Після етапу попередньої обробки та використання «гарячого кодування» для категоріальних змінних набір даних складається з 39 функцій. Пропущені значення містяться в атрибутах «флора» та «пов'язаний діагноз». Процедура заповнення відсутніх даних не використовується, оскільки природа відсутніх даних є дійсно порожнім значенням.

Для побудови простору умов виконано такі кроки:

1. Виділення найбільш істотних ознак;
2. Поділ на кластери за подібними залежними від часу та незалежними від часу параметрами.

Початковий вибір ознак виконується за допомогою кореляційної матриці, Boruta та дерева регресії. Жорстке голосування використовується для остаточного вибору функції.

Формулювання задачі кластерного аналізу, як класифікації багатомірних кількісних та якісних даних, поданих у просторі умов, полягає в наступному.

Нехай $Ob = \{o_1, o_2, \dots, o_n\}$ – множина пацієнтів, яку необхідно розбити на m підмножин (кластерів) Ob_j так, щоб кожен з об'єктів o_i належав лише одному кластеру.

Пацієнти, які належать до одного кластера мають спільні характеристики, а пацієнти, які належать до різних кластерів, навпаки, мають відмінні ознаки. Таке розбиття відповідає певним обмеженням, а також критерію оптимальності.

Для розв'язку цієї задачі розглядають набір параметрів (множину) $Pd = \{Pd_1, Pd_2, \dots, Pd_p\}$, ознак (параметрів, характеристик), якими володіють об'єкти множини Ob.

Характеристики пацієнтів можуть бути як кількісними так і якісними. За множиною ознак Pd, кожному хворому o_j ставиться у відповідність p-мірний

вектор (точка) $X_j = \langle x_{1j}, x_{2j}, \dots, x_{pj} \rangle$, де x_{ij} – значення i -ої характеристики пацієнта o_j.

Даному набору характеристик відповідає множина об'єктів (пацієнтів) Ob з деякою множиною $X = \{x_1, x_2, \dots, x_n\}$ точок (векторів) p -мірного простору. При цьому з рівності $X_i = X_j$ випливає, що відповідні об'єкти o_i і o_j або дійсно ідентичні, або ідентичні за даною множиною характеристик Pd.

Схожість пацієнтів у кластері $o_j \in Ob$ визначається за функцією $\mu(\cdot,\cdot)$, яку називають мірою схожості або подібності.

Для міри подібності використовують функцію $\mu(\cdot,\cdot)$, яка ставить у відповідність кожній парі об'єктів o_i і o_j невід'ємне число $\mu(o_i, o_j)$, що задовольняє умову:$0 \le \mu(o_i, o_j) \le 1$, де $\mu(o_i, o_j) = 1$ тоді і тільки тоді, коли o_i співпадає з o_j за даною множиною ознак, крім того має місце рівність:

$$\mu(o_i, o_j) = \mu(o_j, o_i).$$

Для визначення міри подібності насамперед вводять поняття відстані $\rho(o_i, o_j)$ між об'єктами o_i і o_j. Для цього обирають яку-небудь метрику в p-мірному просторі, тобто деяку невід'ємну функцію $\rho(x, y)$, яка задовільняє наступні умови:

$$\rho(x_i, x_j) = 0, x_i = x_j,$$
$$\rho(x_i, x_j) = \rho(x_j, x_i),$$
$$\rho(x_i, x_j) \le \rho(x_i, x_k) + \rho(x_k, x_j). \quad (2.26)$$

Відстань $\rho(o_i, o_j)$ між пацієнтами o_i і o_j визначають як $\rho(o_i, o_j) = \rho(x_j, x_i)$, де x_i і x_j точки p-мірного простору відповідно до об'єктів (пацієнтів) $o_i, o_j \in Ob$ з допомогою наборів ознак Pd.

Міру подібності між об'єктами $o_i, o_j \in Ob$ можна визначити наступним чином: $\mu(o_i, o_j) = \frac{D}{D+\rho(o_i,o_j)}, D > 0$. Оскільки, будь-яке розбиття множини Ob на кластери Ob_j зумовлює відповідне розбиття множини X на підмножини X_j (і навпаки), то відстань між кластерами $Ob_i, Ob_j \in Ob$, визначається:

$$\rho(Ob_i, Ob_j) = \min_{\substack{l,k \\ o_l \in Ob_i \\ o_k \in Ob_j}} \rho(o_l, o_k) = \min_{\substack{l,k \\ X_l \in Ob_i \\ X_k \in Ob_j}} \rho(X_l, X_k).$$

Діаметр кластера $Ob_j \in Ob$:

$$d(Ob_j) = \max_{\substack{l,k \\ o_l, o_k \in Ob_j}} \rho(o_l, o_k) = \max_{\substack{l,k \\ X_l, X_k \in X_j}} \rho(X_l, X_k).$$

Сукупність об'єктів $o_i \in Ob$, схожих до об'єкта $o_j \in Ob$, або множина точок $X_i \in X$, які близькі до точки $X_j \in X$, визначають як множину $\{o_i : \rho(o_i, o_j) < D\}$ або відповідно $\{X_i : \rho(X_i, X_j) < D\}$, де D – додатне число, яке називають порогом подібності.

Об'єкт (пацієнт) $o_i \in Ob$ вважають схожим з $o_j \in Ob$, якщо відстань $\rho(o_i, o_j)$ між цими об'єктами є меншою за поріг подібності D. Міру подібності та поріг подібності вибирають з міркувань та представлень про схожість пацієнтів множини Ob.

Використовуючи введені поняття, математичну модель задачі кластеризації можна записати в такий спосіб.

Розбити множину Ob на кластери Ob_j так, щоб

$$\cup_{j=1}^{m} Ob_j = Ob, Ob_j \cap Ob_k = \emptyset, j \neq k, d(Ob_j) < D, j = \overline{1, m}. \qquad (2.27)$$

Задача багаторівневої ієрархічної кластеризації полягає в наступному. Для кожного $h = \overline{1, H}$ (h – рівень ієрархії, H – кількість таких рівнів) множину Ob необхідно розбити на неперетинні підмножини (кластери) $Ob_j^h, j = \overline{1, m_h}$ таким чином, щоб діаметри кластерів $d(Ob_j^h)$ не перевищували заданих величин (порогів подібності) D^h і при цьому були досягнуті екстремуми деяких цільових функцій Φ^h.

Об'єкти кластеризації:

на першому рівні ієрархії – це кластери $o_j = Ob_j^0, j = \overline{1, n}$ вихідної множини Ob;

на другому рівні ієрархії – кластери $Ob_j^1, j = \overline{1, m_1}$ першого рівня;

на третьому – кластери $Ob_j^2, j = \overline{1, m_2}$ другого рівня і т.д.

Таким чином, кожен об’єкт кластеру h -го рівня є деякою множиною об’єктів кластеру $(h-1)$-го рівня, тобто $Ob_j^h = \bigcup_{i=J_{jk}} kOb_i^{h-1}$.

На кожному рівні ієрархії об’єкти Ob_j^h описують різними наборами ознак $Pd^h = \{Pd_1^h, Pd_2^h, \dots, Pd_{p_h}^h\}$ і схожість об’єктів визначають різними мірами подібності μ^h, які вибирають з представлень про схожість об’єктів даного рівня.

Математична модель задачі ієрархічної кластеризації:

$$\bigcup_{j=1}^{m_h} Ob_j^h = Ob; Ob_j^h = \bigcup_{j \in J_{jk}} Ob_j^{h-1}; Ob_j^h \cap Ob_i^h = \emptyset, i \neq j;$$

$$d\left(Ob_j^h\right) = D^h, j = \overline{1, m_h}; h = \overline{1, H}. \quad (2.28)$$

Розв’язок задачі кластеризації суттєво залежить від вибору мір подібності μ^h і порогу подібності D^h.

Приклад 2.2.

Для початку побудовано кореляційну матрицю, наведену на рисунку 2.9.

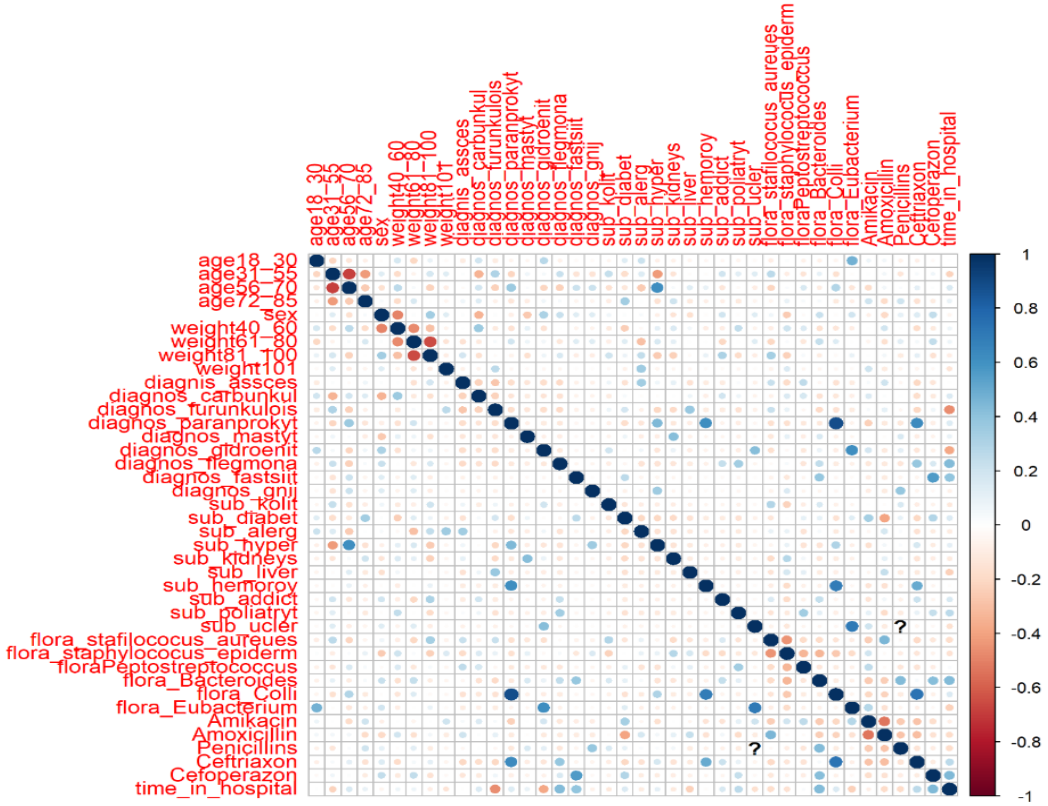


Рисунок 2.9 Кореляційна матриця

Отже, суттєва кореляція відсутня.

Для визначення важливих ознак використано алгоритм Boruta [21]. Результат подано на рисунку 2.10.

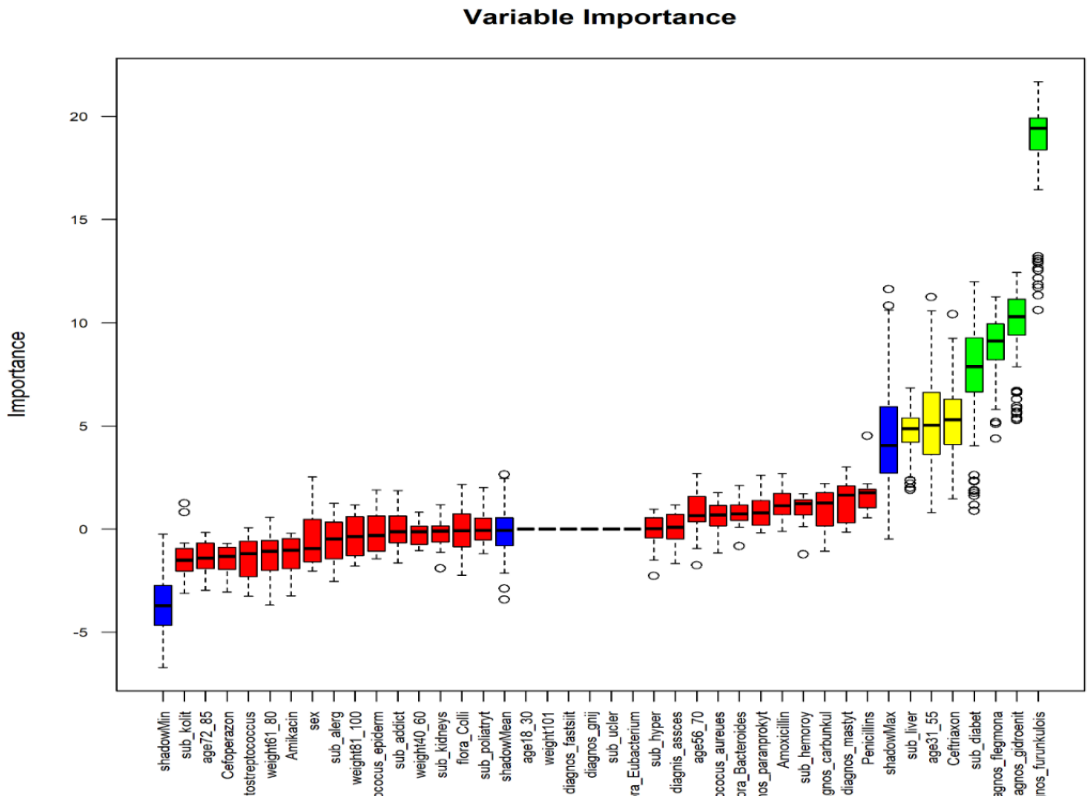


Рисунок 2.10 Результат алгоритму Boruta

Boruta – це оболонка, побудована навколо алгоритму класифікації випадкового лісу. На рисунку 2.10 показано важливі (зелений колір) та дотично-важливі (жовтий колір) ознаки. Сині прямокутні діаграми відповідають мінімальному, середньому та максимальному Z-балам тіньового атрибута.

Таблиця 2.3. Характеристика цільових змінних

Цільові змінні	meanImp (міра важливості)	рішення
діагноз_фурункульоз	18,631571	Підтверджено
діагноз_гідроденіт	9,901603	Підтверджено
діагноз_флегмона	8,915723	Підтверджено
супутній_діабет	7,557117	Підтверджено

Наступне дерево регресії також використовується для вибору ознак.

Дерево регресії виглядає наступним чином:

1) root 51 1981.92200 10.372550

2) діагноз_фурункульоз >=0.5 11 17.63636 4.818182 *

3) діагноз_фурункульоз < 0.5 40 1531.60000 11.900000

6) супутній_діабет < 0.5 33 1075.87900 11.060610

12) Цефтріаксон < 0.5 26 818.65380 10.115380

24) стать>=0.5 12 602.66670 8.666667 *

25) стать< 0.5 14 169.21430 11.357140 *

13) Цефтріаксон>=0.5 7 147.71430 14.571430 *

7) супутній_діабет >=0.5 7 322.85710 15.857140 *

За результатами дослідження спостерігається, що важливі ознаки для обох методів схожі.

Результат жорсткого голосування для селектора 3 функцій наведено нижче:

- діагноз_фурункульоз;
- супутній_діабет;
- діагноз_гідроденіт;
- Цефтріаксон;
- діагноз_флегмона;
- стать.

Лікування препаратом Цефтріаксон впливає на тривалість перебування в лікарні. Тому лише параметри, що залежать від часу, важливі для прогнозування кількості днів у лікарні.

Далі використовується кластеризація. Візуальна оцінка (кластерної) тенденції (VAT) використовується для аналізу можливості розбиття об'єктів. VAT демонструє погану тенденцію до кластеризації (рисунок 2.11). Невеликі відмінності представлені темними відтінками, а великі – світлими [22].

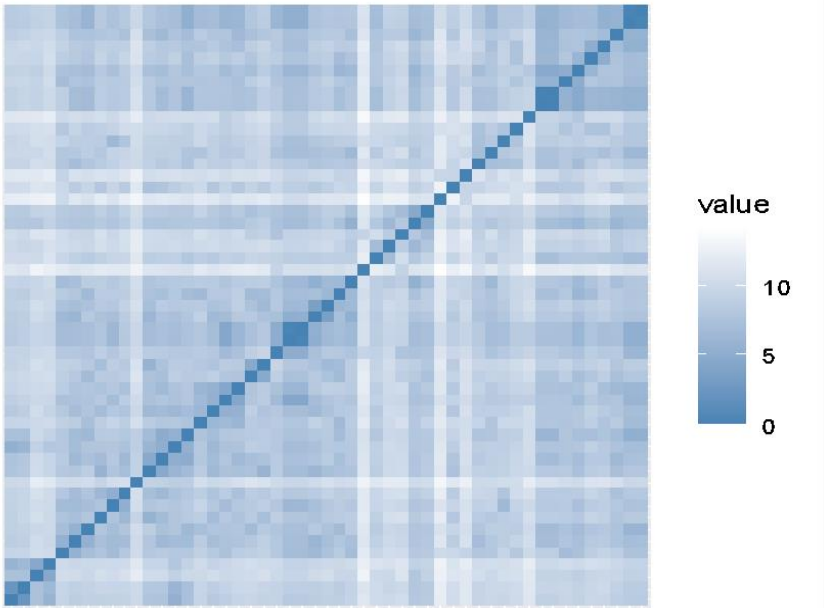


Рисунок 2.11. Результати застосування VAT

Те саме ми маємо з візуалізацією k-means (рисунок 2.12). Перекривання кластерів знайдено за допомогою k-means.

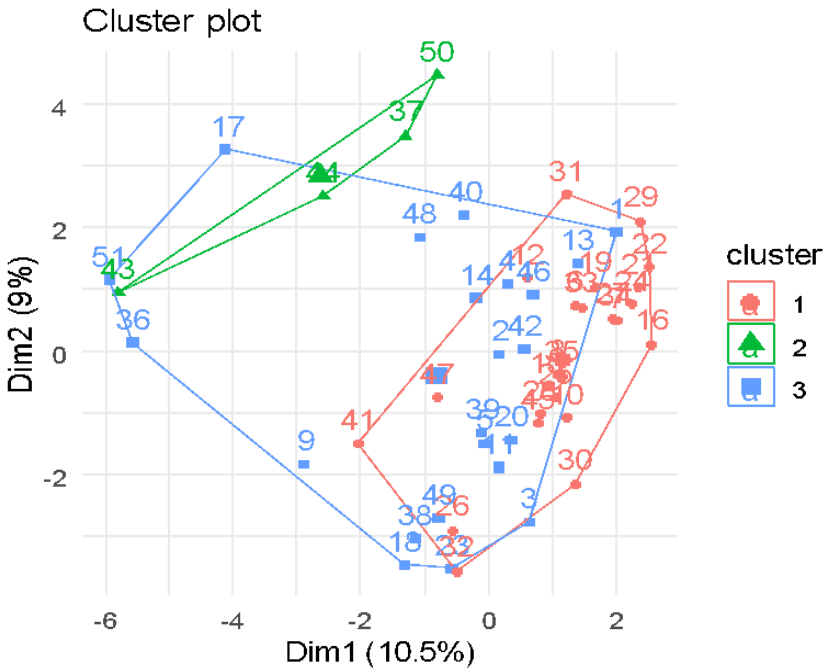


Рисунок 2.12 Результати кластеризації за допомогою k-means

Статистика Хопкінса [23] дорівнює 0,71. Це означає, що набір даних не є суттєво кластеризованим. Нечітке c-mean показує кращі результати (таблиця 2.4). Усі сильні екземпляри в кожному кластері позначені жирним шрифтом.

Таблиця 2.4. Результати нечіткого c-mean

Об’єкти	1	2	3	Об’єкти	1	2	3
[1,]	**0.89150163**	0.06212372	0.046374647	[26,]	0.31107702	0.65580938	0.033113601
[2,]	**0.77796788**	0.08358509	0.138447034	[27,]	0.15995697	**0.81908319**	0.020959847
[3,]	0.56614562	0.09556459	0.338289785	[28,]	0.07447833	**0.90846830**	0.017053375
[4,]	**0.86018278**	0.10377855	0.036038668	[29,]	0.07030249	**0.91359600**	0.016101503
[5,]	0.68808007	0.09069853	0.221221402	[30,]	0.05056537	**0.94033247**	0.009102165
[6,]	0.03321895	**0.96077754**	0.006003508	[31,]	0.07316832	**0.90999577**	0.016835911
[7,]	0.30215319	0.66503031	0.032816504	[32,]	0.07499601	**0.90979079**	0.015213199
[8,]	0.14472035	**0.83639243**	0.018887226	[33,]	0.05565317	**0.93171882**	0.012628002
[9,]	**0.81184745**	0.15376100	0.034391551	[34,]	0.47632097	0.48284576	0.040833272
[10,]	0.04098111	**0.95165817**	0.007360716	[35,]	0.03983866	**0.95303284**	0.007128497
[11,]	**0.91012330**	0.05144097	0.038435725	[36,]	**0.90736664**	0.06052699	0.032106373
[12,]	0.14787502	**0.83223998**	0.019885001	[37,]	0.08044003	0.03597534	**0.883584625**
[13,]	0.64440750	0.31189553	0.043696971	[38,]	**0.89430505**	0.07879489	0.026900054
[14,]	**0.89189168**	0.06118803	0.046920295	[39,]	0.55963468	0.09392417	0.346441148
[15,]	0.05546220	**0.93451510**	0.010022703	[40,]	**0.90959793**	0.05132587	0.039076209
[16,]	0.04969225	**0.94011428**	0.010193473	[41,]	0.18072499	**0.79517059**	0.024104421
[17,]	0.64818641	0.30658830	0.045225286	[42,]	0.78670219	0.08077068	0.132527121
[18,]	**0.89320798**	0.07949093	0.027301094	[43,]	0.09733532	0.02808194	**0.874582736**
[19,]	0.06356521	**0.92325893**	0.013175858	[44,]	0.06184420	0.02651523	**0.911640561**
[20,]	0.89348356	0.07989919	0.026617253	[45,]	0.48991432	0.46830619	0.041779495
[21,]	0.05293598	**0.93743658**	0.009627441	[46,]	**0.85462052**	0.06862292	0.076756557
[22,]	0.08066868	**0.90078226**	0.018549056	[47,]	0.31122970	0.65498914	0.033781156
[23,]	**0.88746299**	0.08389872	0.028638293	[48,]	**0.88266229**	0.08734761	0.029990095
[24,]	0.06510596	**0.92012271**	0.014771324	[49,]	**0.89659897**	0.05858958	0.044811455
[25,]	0.07236656	**0.91123006**	0.016403377	[50,]	0.17377591	0.04574853	0.780475563
				[51,]	0.66297817	0.29258686	0.044434975

Таким чином, набір даних розділений на 3 кластери.

Об’єкти 3, 5, 7, 13, 17, 20, 26, 34, 39, 42, 45, 47, 50, 51 слід аналізувати окремо як кластер 4.

Отже, простір умов включає обрані важливі ознаки та номери кластерів для кожного об’єкта. На етапі аналізу кожного окремого кластера використовуватимемо слабкі предиктори.

На основі простору умов розроблено ієрархічний предиктор, метою застосування якого є підвищення точності прогнозування для всього набору даних та обраних важливих ознак, що базується на дотриманні таких етапів:

- нечітке k-means ділить об’єкти на кластери (таблиця 2.4.);
- Випадковий ліс лінійної регресії, SVM з ядром Radial Basis і SVM з ядром Polynomial використовуються для кожного кластера окремо;
- на основі з отриманими результатами обирається середнє значення.

Прогностична точність ієрархічного предиктора представлена в таблиці 2.5.

Таблиця 2.5. Точність прогнозування для всього набору даних та обраних ознак

Модель на основі усіх змінних	RMSE	MAPE
Ієрархічний предиктор	1.401258	0.137792
Модель на основі обраних змінних	RMSE	MAPE
Ієрархічний предиктор	1.401257	0.102961

За результатами дослідження спостерігається, що точність прогнозування гірша для всього набору даних, ніж для набору з вибраних ознак, а також що RMSE для ієрархічного предиктора на всьому наборі даних не набагато вища, ніж для вибраних функцій. З іншого боку, точність прогнозування для розділених кластерів краща, ніж для одного з кращих слабких предикторів, а саме штучної нейронної мережі (ANN) з 12 нейронами в одному прихованому шарі.

Отримані результати дослідження, розроблених моделей в цілому, підтверджують гіпотезу про відмінності в точності прогнозування для всього набору даних і простору умов, побудованого на основі кластеризації аналізу та вибору ознак.

Проте слід зробити зауваження:

- кількість екземплярів у зібраному наборі даних замала. Дана гіпотеза повинна бути підтверджена великим набором даних;
- сформульоване припущення, що точність прогнозу буде сильно залежати від кількості порожніх значень.

Різниця між кращим слабким предиктором (персептрон) у 1,01 краща для обраних ознак. Якість розробленого ієрархічного предиктора для метрики RMSE на 1,47 краща, ніж для кращого слабкого предиктора. Дерево регресії показує

однаковий результат для всього набору даних і вибраних змінних. Лінійна регресія показує на 1,53 кращі показники RMSE для вибраних функцій у порівнянні з усім набором даних. Решта слабких предикторів показує кращі результати за вибраними функціями.

2.4. Побудова моделі поведінки стану пацієнта у просторі станів

Аналіз спорідненості є одним із широко розповсюджених методів аналізу даних. Метою цього методу є дослідження взаємозв'язку між подіями, що відбуваються разом. Фактично задача аналізу спорідненості – це різновид задачі аналізу споживчого кошика, ідентифікація асоціацій між різними подіями, послідовний пошук тощо. Отже, завдання аналізу спорідненості полягає у пошуку правил кількісного опису взаємозв'язку двох або більше подій. Такі правила називаються правилами асоціації [31]. Зазвичай такі асоціативні правила потребують визначення попередньої структури даних, наприклад, групування даних за певними критеріями [32].

Зважаючи на існуючі методи та алгоритми машинного навчання, та враховуючи їхні переваги та недоліки, внаслідок чого виникає потреба у розробленні процедури пошуку елементарних умовних асоціацій для побудови моделі поведінки стану об’єкта.

За допомогою кластеризації вирішено такі завдання аналізу даних, як:

- групування та розпізнавання об'єктів,
- пошук представників однорідних груп (зменшення розмірності даних),
- пошук найближчої групи для нового об’єкта,
- пошук незвичних об'єктів (виявлення викидів).

Основними методами кластеризації є: ієрархічні методи, секціонування, штучні нейронні мережі, штучні нейронні мережі на основі щільності або мережі [76].

Ієрархічна кластеризація створює ієрархію кластерів, або, іншими словами,

дерево кластерів, яке також називають дендрограмою. Кожен вузол кластера містить дочірні кластери; Нащадки кластера мають спільні вузли, що належать їх спільному предку. Такий підхід дозволяє досліджувати дані на різних рівнях деталізації. Недоліки пов'язані з тим, що більшість ієрархічних алгоритмів не використовують для вдосконалення вже побудовані (проміжні) кластери; перетин кластерів; проблеми з масштабованістю під час застосування великих обсягів даних.

Поступова побудова кластерів, на відміну від ієрархічного підходу, забезпечує одночасний процес дослідження сегментів методами секціонування [56]. Вони або намагаються ідентифікувати кластери шляхом ітеративного переміщення точок між підмножинами, або визначають кластери як області, щільно заповнені об'єктами. Алгоритми першого роду належать до кластеризації переміщення та поділяються на ймовірнісний, k-means та k-medoid методи.

Алгоритми секціонування другого типу належать до групи методів секціонування на основі щільності [44]. Вони намагаються виявити щільно пов'язані компоненти даних, гнучкі з точки зору їх форми. Методи засновані на групуванні сусідніх об'єктів у кластери на основі їх локальної компактності, а не близькості. Ці методи розглядають скупчення як ділянки густо розташованих об’єктів, які розділені на більш рідкісні регіони. Основними перевагами методів кластеризації на основі щільності є можливість виявлення кластерів вільної форми різного розміру та стійкості до шуму та викидів. До недоліків можна віднести високу чутливість до встановлення вхідних параметрів, поганий опис класів та непридатність для даних з великими розмірами.

Сіткові методи кластеризації працюють опосередковано, поділяючи простір елементів даних на кінцеву кількість клітинок і залишаючи ці клітинки з високою щільністю об’єктів для подальшої обробки, а ізольовані елементи ігноруються. Просторовий поділ базується на елементах сітки, зібраних із вхідних даних. Методи кластеризації на основі сіток мають наступні переваги: накопичення даних робить метод незалежним від їх порядку; відстань не розрахована; можливість обробки атрибутів різних типів; легко ідентифікувати сусідні кластери [35].

Недоліки стосуються визначення відповідного розміру решітчастої конструкції; Виявлення кластерів різної щільності та форми; вибір умов асоціації для формування ефективних кластерів.

Іншим способом виявлення прихованих залежностей даних є правила асоціації.

Правила асоціації (AR) – це набір спеціальних правил, які дозволяють знаходити та описувати відповідності у великих наборах даних [47].

Теорія асоціативних правил зосереджена на застосуванні предметного набору і транзакцій. Предметний набір – це непорожній набір елементів (станів), які можуть бути частиною транзакції:

$$I = \{i_1, i_2, \dots, i_k, \dots, i_n\}, \qquad (2.29)$$

де i_k – елементи, що входять до предметних наборів, $k = 1..n$, n – кількість елементів набору I.

Транзакція – це певний набір, який містить певні елементи набору I, що застосовуються разом. Транзакція також має унікальний ідентифікатор TID (Transaction ID).

База даних містить відповідний набір транзакцій:

$$T = \{t_1, t_2, \dots, t_i, \dots, t_m\}, \qquad (2.30)$$

де t_i – відповідна транзакція, m – загальна кількість транзакцій.

Поняття множини та асоціативного правила тісно пов'язані з іншою характеристикою асоціативного правила – довірою, яка обчислюється як відношення множини, що має як умову, так і наслідок (іншими словами, це підтримка асоціативного правила), щоб підтримати множину, яка має лише умову:

$$Conf(X \to Y) = \frac{Supp(X \to Y)}{Supp(X)} = \frac{\left(\frac{|X(t) \cap Y(t)|}{|T|}\right)}{\left(\frac{|X(t)|}{|T|}\right)} = \frac{|X(t) \cap Y(t)|}{|X(t)|}. \qquad (2.31)$$

Для визначення значущості правил використовуються порогові значення мінімальної підтримки та достовірності MinSupp та MinConf, які зазвичай визначаються експертами, де $X - умова, Y - наслідок, X \to Y - подія$, виходячи з

досвіду:

$$Supp(X \to Y) \geq MinSupp, \qquad (2.32)$$

$$Conf(X \to Y) \geq MinConf.$$

Методи пошуку асоціативних правил знаходять усі асоціації, які відповідають обмеженням підтримки та достовірності. Однак це призводить до необхідності переглянути досить велику кількість асоціативних правил; цю кількість необхідно скоротити таким чином, щоб проаналізувати лише найбільш значущі з них.

Серед основних алгоритмів генерації асоціативних правил виділяють AIS, SETM, Apriori, AprioriTid, AprioriHybrid [58]. Ефективність та доцільність використання кожного з них зумовлені структурою та обсягом набору даних, для яких здійснюється пошук асоціативних правил, оскільки основа цих методів лежить у різних принципах генерації та вибору предметних сукупностей – кандидатів.

AIS – це перший розроблений алгоритм пошуку асоціативних правил, який складається з двох етапів:

- перший крок реалізує процедуру генерації частих предметних наборів;
- другий – побудова частих правил із заданою впевненістю.

Недоліком цього алгоритму є те, що в процесі пошуку правил він неодноразово проходить через один набір даних.

Алгоритм SETM, як і AIS, складається з двох етапів і виконує формування предметних наборів кандидатів на льоту, використовуючи мову інструмента SQL. У ньому зберігається копія тематичного набору кандидатів разом із TID у спеціальній, послідовній структурі. Після проходження всього набору даних проводиться підрахунок підтримки кандидатів шляхом сортування та агрегування отриманої структури. Недоліками алгоритму SETM, як і AIS, є багаторазові проходження через набір даних та генерація надлишкових кандидатів, які в результаті не належать до частих предметних наборів.

Недоліки вищезазначених алгоритмів вирішуються алгоритмом Апріорі,

запропонованим Р. Агравалем та Р. Срікантом. На відміну від AIS та SETM, він усуває генерування та підрахунок надмірної кількості кандидатів завдяки використанню антимонотонних властивостей та дозволяє значно зменшити множину частих наборів предметів і тим самим зменшити простір пошуку асоціативних правил. Властивість різноманітності стверджує, що якщо набір станів P не часто зустрічається, то додавання якогось нового елемента Y до набору P не змінює його частоту (відповідно, якщо P не є частим набором станів, то і PY не є частим також). Модифікаціями класичного алгоритму Apriori є AprioriTid та AprioriHybrid.

За допомогою методу Апріорі реалізують пошук асоціативних правил. Оскільки розмір сучасних баз даних може досягати досить великих обсягів (гігабайти та терабайти), пошук асоціативних правил вимагає ефективних алгоритмів, які є масштабованими і дозволяють знайти рішення цього завдання в прийнятний час.

Алгоритм Апріорі використовує ітераційний підхід. На першому кроці алгоритму є одноелементні, часті набори даних, що позначаються набором L_1. На наступному кроці набір L_1 використовується для пошуку частих двоелементних наборів, з яких формується набір L_2, який, у свою чергу, використовується для пошуку триелементних наборів L_3, і так далі, поки всі можливі часті k-елементи знайдено множини L_k.

Модель AOG – це орієнтований ациклічний граф, де кожна вершина графа відповідає змінній із заданими параметрами. У байєсівських мережах параметри подаються як локальний умовний розподіл ймовірностей значень змінних $P\ (Xi \mid A_{in}(Xi))$. А в гауссових мережах – як коефіцієнти лінійних рівнянь (для ребер) і дисперсії відхилень (для вершин). Побудова AOG-моделей відповідає проблемі відтворення моделі зі статистичних даних. Сюди входять методи відновлення моделі AOG "Collifinder" та "Proliferator-C", узагальнюючи метод Chow & Liu. Застосування Collifinder та Proliferator-C дозволяє розпізнавати транзитивні, синергетичні та комбіновані асоціації, а отже, забезпечує надійний та ефективний метод відтворення структур однопоточних моделей залежностей без

тестів першого рівня.

Проблема вищеописаних методів полягає у необхідності задати елементарні умовні співвідношення для побудови графіка AOG-моделі. Пошук таких залежностей виходить за межі цих алгоритмів [59, 60, 61].

Алгоритми розпізнавання шаблонів з навчанням припускають наявність історичної інформації, що дозволяє будувати статистичні моделі зв'язків $x \rightarrow y$, де $y \in Y$, Y – значення спостереження за станом досліджуваного об'єкта (пацієнта), яке моделюється за допомогою $x \in X$, X – набір змінних (предиктори), за допомогою яких передбачається мінливість змінної y. Більшість моделей з викладачем розроблені таким чином, що їх можна записати як:

$$y = f(x, \beta) + \varepsilon, \qquad (2.33)$$

де f – функція, вибрана з довільного сімейства, β – вектор параметрів цієї функції, ε – помилки, які зазвичай породжують неупереджені, некорельовані випадкові процеси.

Під час побудови модель при фіксованих значеннях вибірки y мінімізують залишки деякої функції $Q(y, \beta)$. В результаті знайдено $\hat{\beta}$. Це вектор з оптимальними оцінками параметрів моделі. Змінюючи форму функцій f і Q, можна отримати різні моделі, з яких перевага віддається найбільш ефективній моделі. Ця модель забезпечує неупереджені, точні та надійні прогнози відповіді y.

2.4.1 Метод прогнозування зміни стану пацієнта

У роботі розроблено метод зміни стану пацієнта, що забезпечує пошук шаблонів поведінки стану пацієнта шляхом модифікації методу видобування асоціативних правил, що є удосконаленням методу упорядкованого пошуку та надає чіткості та направленості у пошуку рішень стосовно вибору цільових схем лікування, що дає змогу зменшити ймовірність появи похибки при виборі схеми лікування.

Для взаємозв'язку зі схемою пошуку продукційних правил

$R = \{Pd_i, dom(Pd_i)\},\ i = \overline{1,m},$ дозволяє знаходити статистично значущі правила, що відображають залежність атрибута Pd_m на атрибути $Pd_1, Pd_2, \dots, Pd_{m-1}$, тобто залежність виду $Pd_1, Pd_2, \dots, Pd_{m-1} \to Pd_m$. Як міра статистичної значущості використовується інформаційний показник Кульбака-Лейблера. Алгоритм дозволяє шукати лише залежності, визначені загальним набором вхідних даних; крім того, він має високу обчислювальну складність, якщо існує багато правил класифікації.

Для величезного набору даних з невідомою структурою асоціації з високою підтримкою окремих подій практично відсутні. Таким чином, такі асоціації, хоча вони можуть представляти інтерес, будуть виключені з розгляду, оскільки вони не будуть відповідати певному мінімальному порогу підтримки $Supp_{min}$.

Для вирішення цієї проблеми ми пропонуємо знайти асоціативні правила не лише для окремих *o*, а також ознак (параметрів) хворого й для їх ієрархії. Якщо на нижчих ієрархічних рівнях немає таких цікавих асоціацій, то вони можуть виникати на вищих рівнях. Іншими словами, підтримка окремого об'єкта завжди буде меншою, ніж підтримка групи, до якої він належить:

$$Supp(I) > Supp(i_j), \qquad (2.34)$$

де *I* – група знаходиться в ієрархії; i_j– елемент, включений до даної групи. Причини цього очевидні: загальна підтримка групи дорівнює сумі підтримки для включених до неї параметрів:

$$Supp(I) = \sum_{j=1}^{n} i_j, \qquad (2.35)$$

де *n* – кількість елементів у групі.

Асоціативні правила, знайдені для об’єктів чи подій, розташованих на різних ієрархічних рівнях, називаються багаторівневими правилами.

Спускаючись до нижчих рівнів абстракції, аналізуються нащадки лише тих категорій та підкатегорій, які є частими наборами, тобто існує принаймні заздалегідь визначена кількість разів, де *k* – кількість рівня.

Існує кілька підходів до пошуку ієрархічних асоціативних правил.

Повторювані методи часто використовуються, коли набори предметних предметів досліджуються на кожному ієрархічному рівні, від першого до рівня з найбільшою деталізацією. Простіше кажучи, як тільки виявляються всі часті набори предметів на першому рівні, починається пошук популярних наборів предметів на другому тощо. На кожному рівні для пошуку частих наборів можна використовувати будь-який алгоритм, такий як Apriori та його модифікації.

Відомо кілька стратегій проведення правил пошуку.

1. Використовуйте той самий мінімальний поріг для підтримки $Supp_{min}^{k} = const$ на всіх ієрархічних рівнях. При пошуку правил встановлюється один раз певний мінімальний поріг підтримки (наприклад, 5%), коли досягнення певного набору вважається частим, воно не входить до списку правил. Перевага підходу: висока швидкість аналізу предметної області зумовлена відмовою від оцінки часткових наборів, отриманих з тих, які недостатньо поширені. Відсутність – ризик передачі тонких асоціацій на нижчих рівнях ієрархії. Іноді цю ваду намагаються обійти, зменшивши мінімальний рівень підтримки. Як результат, поява десятків і сотень вільних правил з низькою підтримкою та ймовірністю.

2. Зниження порогу мінімальної підтримки при переході на нижчі рівні ієрархії. Він може бути реалізований індивідуально для кожної підгрупи. Функціональний тип порогового зменшення, як правило, пов'язаний з кількістю підкатегорій або серійним номером рівня.

Іншим варіантом функціонального підходу є встановлення порогових значень для мінімальної підтримки виключно залежно від рівня, незалежно від того, яка б підкатегорія не була $Supp_{min}^{k} = \frac{Supp_{min}^{1}}{k}$.

Перевага такого підходу: він дозволяє пройти набагато далі в ієрархії у пошуках асоціацій. Однак у випадку індивідуального завдання граничного рівня для кожної підкатегорії ця процедура відніме левову частку часу. У разі встановлення порогу, залежно від рівня або кількості підкатегорій, не враховуються індивідуальні переваги та недоліки окремих виробників та моделей.

3. Міжшарова фільтрація заснована на так званому проході рівня (рівень проходу). Відносно високий для верхніх рівнів ієрархії, граничний рівень підтримки залишається внизу під час першого проходження. Потім при кожному проходженні цей рівень знижується. Ці низькопрофільні предмети, які мають асоціації, таким чином ідентифікуються раніше, ніж їхні батьківські категорії, які можуть не мати необхідної підтримки. Для підключення рівнів довіри представників різних рівнів рекомендується здійснити три-чотири проходження бази даних транзакцій.

На основі проведеного аналізу даних, зібрано інформацію для формування портрету пацієнта, а саме:

- статуси – у якому стані перебуває пацієнт під час лікування;
- усі попередні стани пацієнта;
- час – коли пацієнт перебуває у певному стані;
- використання даних для організації планування стратегії щодо динаміки змін у стані пацієнта;
- параметри індивіда щодо діагнозу;
- рішення експертів щодо лікування пацієнта;
- інформація про лабораторні дослідження.

Для розпізнавання шаблонів використано триступеневий алгоритм:

1. Формування кластерів пацієнтів – для пошуку поведінки стану;
2. Побудова шаблону – для пошуку послідовності зміни стану;
3. Передбачення наступного стану хворого на основі побудованих на попередньому кроці шаблонів.

Статистика розривів може бути застосована до будь-якого методу кластеризації. Він порівнює загальну варіацію внутрішнього кластеру для різних значень k з їхніми очікуваними значеннями при нульовому еталонному розподілі

даних (тобто розподіл без явної кластеризації). Довідковий набір даних генерується за допомогою моделювання методом Монте-Карло процесу відбору проб.

2.4.1.1 Побудова шаблону поведінки стану пацієнта

Існують об’єктивні та суб’єктивні міри відповідності асоціативного правила. Часто вживаними є зазначена вище підтримка та достовірність. Суб'єктивними мірками значущості є підвищення інтересу (англ *Lift*) і важелі (*Lev*). Підвищення визначається відношенням збереження асоціативного правила до параметра підтримки стану та ефекту окремо:

$$Lift(X \to Y) = \frac{Supp(X \to Y)}{Supp(X) * Supp(Y)}. \quad (2.36)$$

Піднесення (*Lift*) – це так звана узагальнена міра зв'язку між двома предметними сукупностями. Його значення можна інтерпретувати так:

якщо $$Lift(X \to Y) = 1, \quad (2.37)$$

то $$Supp(X \to Y) = Supp(X) * Supp(Y),$$

тобто стан і наслідок не залежать один від одного;

якщо $$Lift(X \to Y) > 1, \quad (2.38)$$

то $$Supp(X \to Y) > Supp(X) * Supp(Y),$$

тобто наслідок позитивно залежить від стану;

якщо $$Lift(X \to Y) < 1, \quad (2.39)$$

то $$Supp(X \to Y) < Supp(X) * Supp(Y),$$

тобто наслідок негативно залежить від стану.

Приклад 2.3.

Результати аналізу даних наведені нижче (рисунок 2.13 та таблиця 2.6).

Набори параметрів (LHS – ліва частина)

[1] "{Вік}" "{Стать}" "{Регіон}" "{Перенесений COVID}"
[5] "{Паління}""{Рівень IgM}" "{Рівень IgG}"

Набори параметрів як наслідок (RHS – права частина)

[1] "{Вік}" "{Стать}" "{Регіон}" "{Перенесений COVID}" "{Рівень IgM}" "{Рівень IgG}"
[5] "{Паління}"

Рисунок 2.13. Структура асоціативних правил

Таблиця 2.6. Параметри асоціативних правил

LHS	RHS	підтримка	довіра	піднесення	номер
[1]	{Стать} => { Рівень IgG }	0.007692308	0.09090909	0.5371901	1
[2]	{ Паління } => { Рівень IgM}	0.007692308	0.04545455	0.5371901	1
[3]	{Вік} => { Перенесений COVID }	0.015384615	0.14285714	1.1607143	2
[4]	{Вік} => { Рівень IgG }	0.01538461	0.12500000	1.1607143	2
[5]	{ Перенесений COVID } => {Паління}	0.015384615	0.14285714	0.8441558	2
[6]	{Стать} => {Паління}	0.015384615	0.09090909	0.8441558	2
[7]	{Рівень IgM} => {Стать}	0.007692308	0.06250000	0.3693182	1
[8]	{Фінанси} => {Розваги}	0.007692308	0.04545455	0.3693182	1
[9]	{Рівень IgM} => {Рівень IgG}	0.046153846	0.46153846	1.8181818	6
[10]	{Перенесений COVID}	0.046153846	0.18181818	1.8181818	6

	=> {Рівень IgG}				
[11]	{Перенесений COVID} => {Стать}	0.015384615	0.11111111	0.6565657	2
[12]	{Паління} => {Рівень IgM}	0.015384615	0.09090909	0.6565657	2

Показники *Lift* правила 3, 4, 9, 10 є найвищі, отже важливі для аналізу.

В іншому випадку значення правил є важелем Lev, який запропонував Г. Пятецький-Шапіро [110]. Важіль – це різниця між частотою, з якою умова і наслідок з'являються разом, тобто підтримкою асоціативного правила, добутком підтримки умови та ефекту окремо:

$$Lev(X \to Y) = Supp(X \to Y) - Supp(X) * Supp(Y). \qquad (2.39)$$

Поліпшення – це відношення частоти спостережуваного виконання правила до добутку ймовірностей виникнення стану та ефекту окремо:

$$I(X \to Y) = \frac{P(X \to Y)}{P(X) * P(Y)}. \qquad (2.40)$$

2.4.1.2 Формування правил та шаблонів

Створений шаблон не відповідає на питання про те, як пов’язати асоціацію з станом, а також встановлює часові залежності, тобто не лише для того, щоб відповісти на питання, *Як* себе почуває пацієнт у певний момент, але *Який* стан протягом дня. А якщо є можливість – передбачити, який стан пацієнта буде на *Наступний День*.

Відповіді на це питання забезпечують використання послідовних шаблонів, заснованих на теорії асоціацій, обов’язковими полями яких є дата/час та ідентифікатор пацієнта.

При розгляді послідовностей транзакцій використовується одне припущення – один і той же клієнт не виконує дві різні транзакції одночасно.

Послідовність *S* називається максимальною, якщо вона не міститься в жодній іншій послідовності. Послідовність *S* називається клієнтом, якщо вона, крім набору об'єктів, дати та часу, містить також ідентифікатор досліджуваного об’єкта (пацієнта).

Послідовність S_1 міститься в послідовності S_2, якщо всі набори предметів S_1 містяться в наборах S_2 предметів.

Послідовність *S1* є послідовною, якщо всі набори предметів містяться в наборах предметів.

Наприклад, послідовність < (3); (4, 5); (8) > міститься в послідовності < (7); (3, 8); (9); (4, 5, 6); (8) >, оскільки (3) (3, 8), (4, 5) (4, 5, 6) та (8) (8).

Послідовність *S* називається підтримуваною, якщо вона міститься в її клієнтській послідовності. Підтримка послідовності визначається як кількість пацієнтів, які її підтримують, і зазвичай виражається у відсотках від загальної кількості пацієнтів. Таким чином, концепція підтримки послідовних шаблонів дещо відрізняється від аналогічного поняття про асоціативні правила.

Для бази даних транзакцій завдання пошуку послідовних шаблонів полягає у визначенні максимальної кількості послідовностей серед усіх, що мають підтримку вище заданого порогу. Кожна така максимальна послідовність і буде послідовним шаблоном. Послідовності, які задовільняють обмеженню мінімальної підтримки, далі ми будемо називати частими (за аналогією з частими множинами в теорії асоціативних правил).

2.4.1.3 Метод пошуку послідовних шаблонів

Далі в роботі розроблено метод пошуку послідовних шаблонів та визначено етапи пошуку:

1. Кластеризація пацієнтів методом k середніх із попереднім визначенням кількості кластерів методом ліктя – для пошуку поведінки стану,
2. Побудова ієрархії шаблонів в кожному кластері на основі модифікованого алгоритму Apriori та міри Кульбака-Лейблера – для пошуку послідовності

зміни стану,

3. Пошук найчастішого стану в кластері на різних рівнях ієрархії – для передбачення наступного стану хворого. Для кожного кластера:

3.1. Сортування. Транзакції вихідної бази T_{id} даних сортуються за кодом пацієнта – id, а транзакції T_o кожного пацієнта – за датою d та часом t. Результат – база даних послідовностей пацієнтів. (п. 2.4., формули 2.29, 2.30)

$$T_D = \{t_{id1}, t_{id2}, \dots, t_{idm}\},$$

Вхід: послідовність транзакцій вихідної бази $\langle t_{id1}, t_{id2}, \dots, t_{idm}\rangle$,

Вихід: перестановка $\langle t'_{id1}, t'_{id2}, \dots, t'_{idm}\rangle$,

Для вихідної послідовності, для усіх її членів виконується співвідношення

$$t'_{id1} \le t'_{id2} \le \cdots \le t'_{idm}.$$

3.2. Пошук частих станів FS'. Частим називаються стани пацієнта, які зустрічаються в більшості пацієнтів з однаковими параметрами $X_i = X_j$, і ступінь довіри до яких перевищує мінімально допустиме значення. (п. 2.4., формула 2.32., п. 2.4.1., формула 2.35). Вибраний набір частих станів трансформується у числове або символьне подання.

$$FS' = \{p \in P | X_i = X_j, Conf(X \to Y) \ge MinConf\}.$$

3.3. Трансформація. Необхідно визначити, які з найбільш частих послідовностей містяться в послідовності пацієнтів. Для цього кожна транзакція в послідовності замінюється множиною її частих станів. Якщо в транзакції немає частотного набору, він більше не розглядається. На додаток, якщо конкретний пацієнт у послідовності не має єдиного набору станів, він також виключається з розгляду. Після перетворення кожна послідовність пацієнтів є упорядкованим набором частих станів. (п. 2.4.1.1.)

3.4. Пошук частих послідовностей. Часті послідовності шукаються на множині частих наборів. Мінімальна частота – параметр алгоритму. (п.2.4.1.1)

3.5. Пошук максимуму послідовностей. Серед частих послідовностей є максимум. Іноді цей етап поєднують з попереднім, щоб скоротити час, витрачений на обчислення не максимальних послідовностей. (п.2.4.1.2)

Найбільш проблематичним кроком у пошуку послідовних шаблонів є ідентифікація частих послідовностей, оскільки аналіз даних стану пацієнта вимагає розгляду величезної кількості можливих комбінацій та декількох проходів через набір транзакцій. Кожен уривок починається з початкового набору послідовностей, які використовуються для генерації нових потенційних частих послідовностей, які називаються послідовностями кандидатів або просто кандидатами. Для цього вони обчислюють свою підтримку і, після завершення проходження, визначають, чи є вони часто виявленими кандидати. Виявлені часті послідовності будуть відправною точкою для нового проходу.

Приклад 2.4.

Для перевіреного набору даних після кроку 3, беручи до уваги виключення пацієнта, були отримані послідовності пацієнтів (Таблиця 2.7):

Таблиця 2.7. Послідовність пацієнтів

Пацієнт	Послідовність
1	< {1,5}; {2}; {3}; {4} >
3	< {1}; {3}; {4}; {3, 5} >
5	< {1}; {2}; {3}; {4} >
6	< {1}; {3}; {5}; {4} >
8	< {4}; {5} >

Пошук частих послідовностей відбувається починаючи з рівня 1 до максимально можливого. Результати послідовних передач наведені в таблиці 2.8. Наведено підтримку *Supp* кожного правила.

Таблиця 2.8. Матриця прогнозування

1- послідовність		2- послідовність		3- послідовність		4- послідовність	
F_1	*Supp*	F_2	*Supp*	F_3	*Supp*	F_4	*Supp*
1	4	1; 2	2	1;2;3	2	1;2;3;4	2
2	2	1; 3	4	1; 2:4	2		

3	4	1; 4	3	1; 3;4	3
4	4	1; 5	3	1; 3;5	2
5	4	2; 3	2	2; 3:4	2
		2; 4	2		
		3; 4	3		
		3; 5	2		
		4; 5	2		

Таким чином, частими є послідовності <1; 2; 3; 4>, <1; 3; 5> і <4; 5> оскільки вони не містяться в послідовностях більшої довжини. На них будуть шукати послідовні шаблони.

Отже, метод прогнозування зміни стану пацієнта складається з 3 етапів:

1. Створення кластера пацієнта;
2. Пошук частих шаблонів;
3. Прогнозування наступного стану пацієнта на основі частих шаблонів.

При оцінці параметра у моделі розраховується ймовірність, а не конкретне значення 0 або 1. Потрібно визначити поріг ймовірності, далі пацієнт може бути віднесений до групи 0 або 1. Порогове значення приймається рівним 0,09. Модель розпізнавання виглядає так:

якщо $y \leq$ поріг, тоді відповідь = 0;

якщо $y>$ поріг, то відповідь = 1.

Порівняємо результати прогнозної моделі з реальними даними. Матриця невідповідності (confusion matrix) фактичних та прогнозованих значень відповіді наведена в таблиці 2.9.

Таблиця 2.9. Матриця невідповідностей

Факт / Прогноз	0	1
0	18	2
1	2	37

Помилка прогнозування не є великою. Але ми можемо передбачити лише наступний стан, а не тривалість існуючого стану наступного періоду часу.

Порівняємо наші результати з відомими методами. У [112, 113] використовується двоступенева модель для розпізнавання образів людини. Перша частина моделі, заснована на розширенні ConvNets до 3D-випадку, автоматично вивчає просторово-часові особливості. Потім другий крок полягає у використанні цих вивчених особливостей для навчання рекурентної моделі нейронної мережі для класифікації всієї послідовності. Викиди оцінюються за набором даних. Матриця невідповідностей для цього набору даних з використанням двоступеневого алгоритму, наведеного в [112], показує середню точність 67,9 та 78,2 для розробленого алгоритму.

Продемонстровано результати розпізнавання шаблонів зміни стану. Для розв'язання проблеми використовується ансамбль моделей. Зокрема, для розв'язання поставлених задач використано кластеризацію, застосування асоціативних правил та послідовних правил. Алгоритм k-means використовується для групування пацієнта. Після цього будуються послідовні асоціативні правила для кожного кластера окремо. Вводиться поняття правил асоціації; розроблено метод пошуку залежностей. Оцінюється точність моделі. Матриця невідповідностей показує відповідні результати прогнозування. Рівень помилок менше 6,7%. Найбільша помилка трапляється для нульового класу.

Наступні дослідження орієнтовані на прогнозування *наступного стану*, що пов’язане з інформацією про передбачуваний *кластер стану*.

Приклад 2.5.

Дослідження проводилися на наборі даних, що збирався за допомогою форми Google https://docs.google.com/forms/d/1o8CMGVZv6BDkw-QIYg2F8VQzqcXxqklomRwXCLOZIcY/, фінансується Центральноєвропейською ініціативою та перевіряється Львівським регіональним центром COVID`19. Цей набір даних складається з таких характеристик:

- Вік (категорійний): 0-15, 16-22, 23-40, 41-65,> 66z;
- Стать (категорійнв): чоловік, жінка;
- Регіон (рядок): Львів (Україна), Чернівці (Україна), Білорусь, Німеччина інше;
- Ви курите? (логічне значення): так, ні;
- Чи був у вас COVID? (категоричний): так, ні, можливо;
- Рівень IgM (числовий): [0..0,9) (негативний), [0,9..1,1) (невизначений), > = 1,1 (позитивний);
- Рівень IgG (числовий): [0..0,9) (негативний), [0,9..1,1) (невизначений), > = 1,1 (позитивний);
- Група крові (числова);
- Щеплені від грипу? (категорично): так, ні, можливо;
- Щеплені від туберкульозу? (категорично): так, ні, можливо;
- Ви захворіли на грип цього року? (категорично): так, ні, можливо;
- Ви хворіли на туберкульоз цього року? (категорично): так, ні, можливо.

Опрацьовували набір даних, де представлено 2 279 відповідей.

Для аналізу даних було запропоновано:

- групування даних за ідентифікатором хворого;
- розділення факторів за ідентифікатором хворого;
- розділення на фактори за станом хворого;
- розділення факторів за схемою лікування.

Для розпізнавання шаблонів використано метод прогнозування зміни стану.

Створення кластера досліджуваних об’єктів (пацієнтів) виконано засобами RStudio, пакети factoextra та кластер (на рисунку 2.14 показані результати ієрархічної кластеризації):

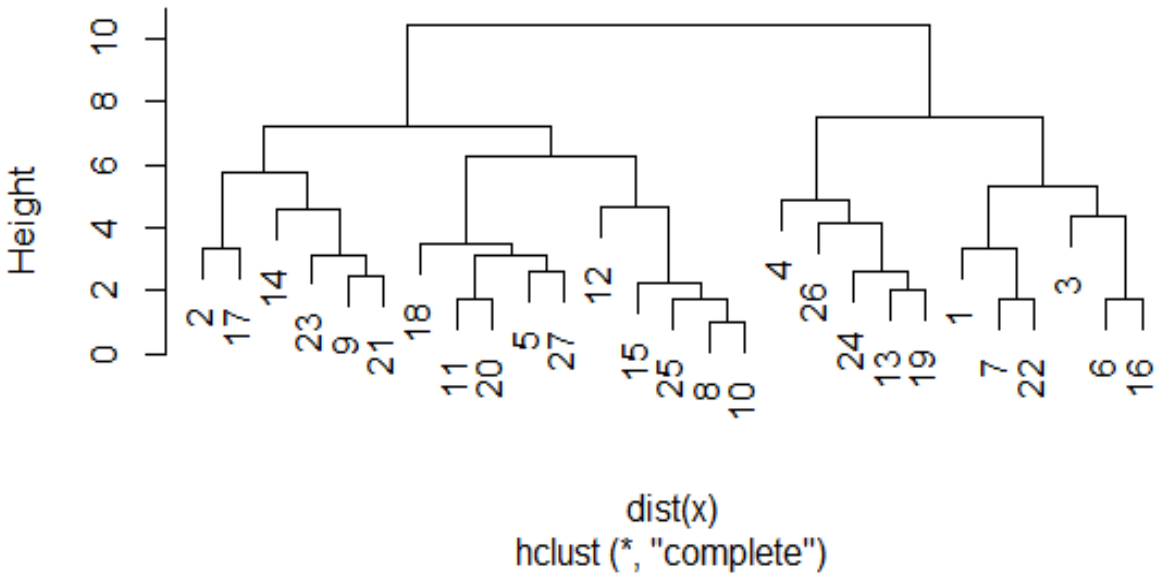


Рисунок 2.14 Дендрогорама досліджуваних об’єктів (пацієнтів)

У роботі використано k-means та k-medoid алгоритми кластеризації. По-перше, знайдено оптимально кількість кластерів за допомогою методу Elbow та статистики розривів. Статистика розривів може бути застосована до будь-якого методу кластеризації. Цей метод порівнює загальну варіацію внутрішнього кластера для різних значень k з їхніми очікуваними значеннями при нульовому еталонному розподілі даних (тобто розподіл без явної кластеризації). Навчальний набір даних генерується за допомогою моделювання методом Монте-Карло процесу відбору проб, рисунок 2.15.

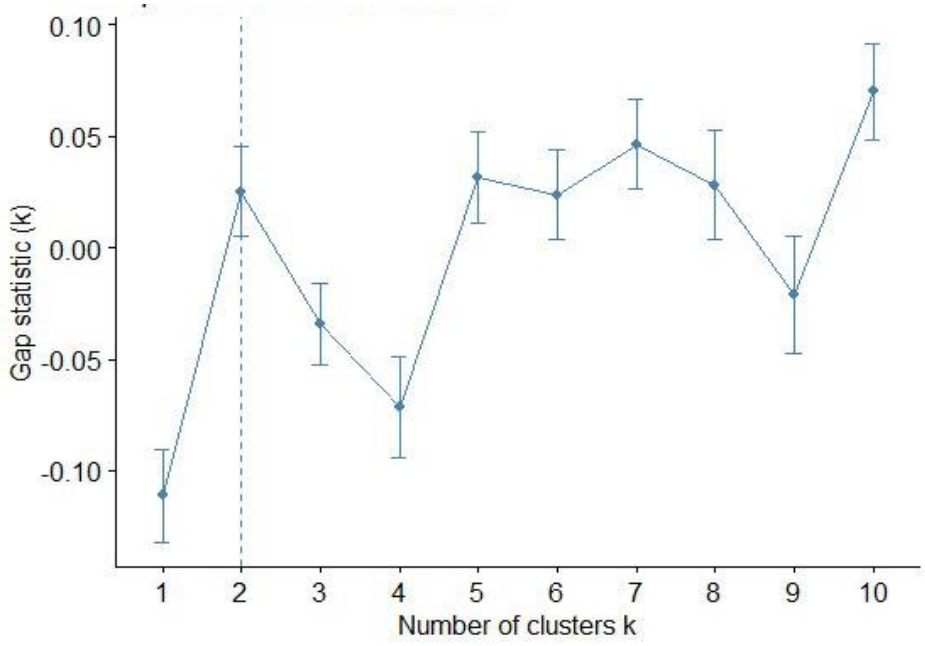


Рисунок 2.15 Оптимальна кількість кластерів

Для подальшого вивчення мір важливості змінних використано Випадковий ліс. На рисунку 2.16 визначено, які ознаки найчастіше з’являлися в коренях дерев, глибина дерев, помилка передбачення тощо.

	mean_min_depth	no_of_nodes	mse_increase	node_purity_increase	no_of_trees	times_a_root
1 Age	1.511688	3023	0.050938625	11.558790	499	112
2 Blood group	1.820000	3387	0.013201974	9.690663	500	53
3 Gender	1.723688	1484	0.052533296	6.322656	499	111
4 Had influenzas	1.727688	2225	0.034935393	7.862122	499	92
5 Smoke	2.030000	1507	0.005027793	4.138718	500	79
6 Vaccinated influenza	2.372752	1896	0.008977258	3.495168	496	25
7 Vaccinated tuberculosis	2.164000	2348	0.015751848	5.870700	500	28

Рисунок 2.16. Фрейм ознак вимірювання_значення.

Сформувавши набір найважливіших змінних, ми можемо дослідити взаємодії щодо них, тобто розбиття, що з'являються в максимальному піддереві щодо однієї із вибраних змінних. Щоб визначити 5 найважливіших змінних, аналізувались середня мінімальна глибина дерев, а також кількістю дерев, у яких з’явилася така змінна. Отримано такий результат:

[1] "Вік" "IgG" "Група крові" "грип" "IgM"

Naive Bayes показує щільність для кожного об’єкта в наборі даних (рисунок 2.17). Точність наївного Байєса набагато менше, ніж Випадковий ліс, і дорівнює 67%.

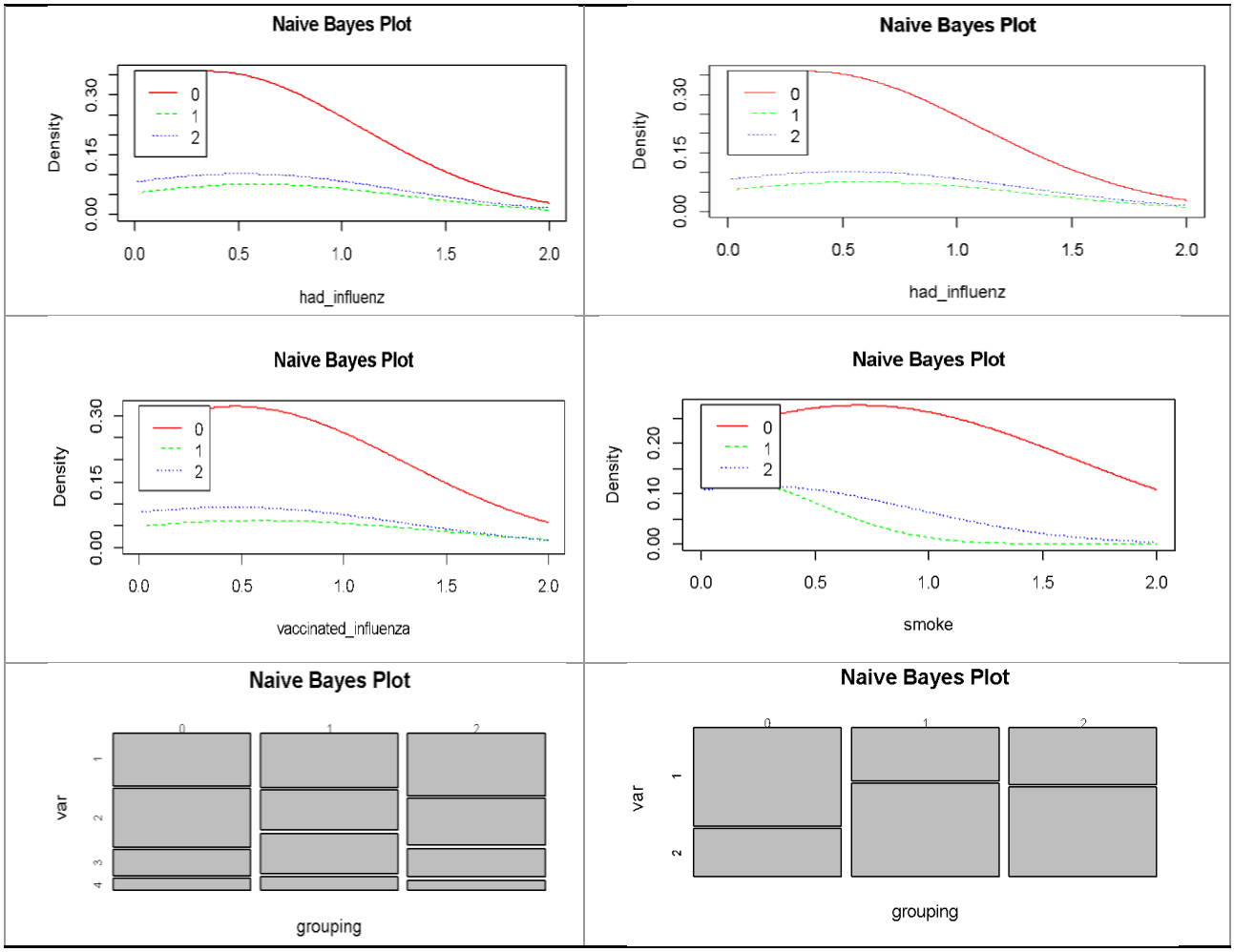


Рисунок 2.17 Наївний Баєс – діаграма щільності

Видно, що Випадковий ліс показує кращі результати. Побудовано 500 дерев. При цьому помилка першого роду становить 16,61%.

Найбільша помилка для класу 1 (covid – так). Це можна пояснити різницею у представленні IgG та IgM (розкид даних становить від 0,00 до 18,00) у різних країнах.

Мінімальні значення глибини для всіх дерев у випадковому лісі наведені на рисунку 2.18.

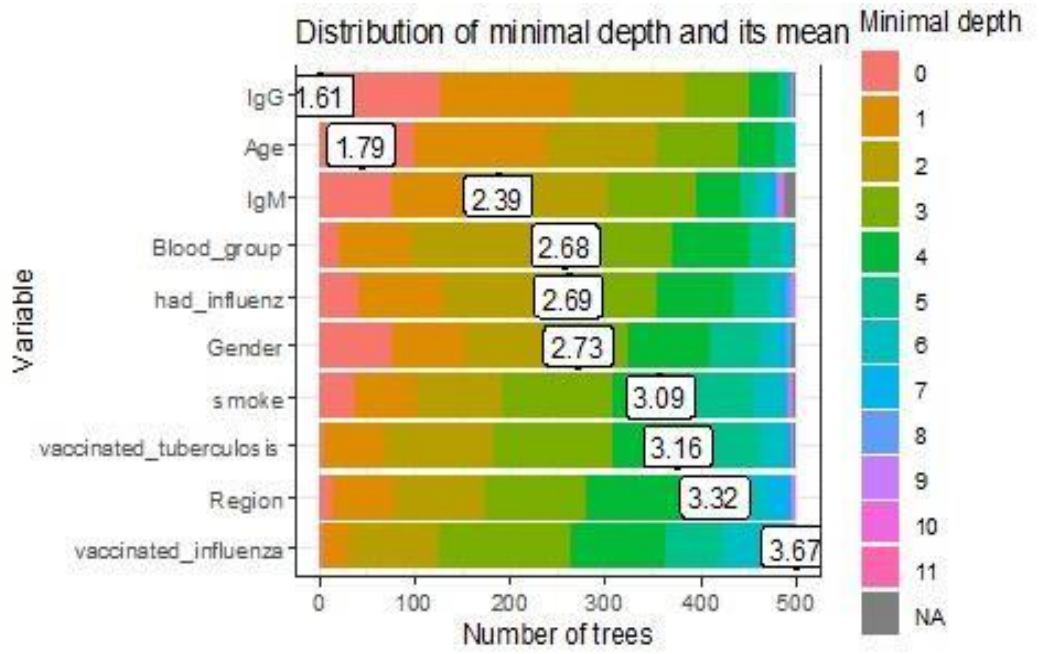


Рисунок 2.18 Розподіл мінімальної глибини розвинених дерев

Залежності з частковими функціональними залежностями мають низький рівень підтримки, що не дозволяє використовувати їх для подальшого аналізу даних, а часткові функціональні залежності є модифікованими асоціативними правилами, але вони виконуються лише для частини даних і залежать від фактора часу.

Досліджено результати розробленого методу пошуку шаблонів, який базується на модифікації методу асоціативних правил, що дозволяє зменшити трудомісткість і використовувати паралельний та розподілений режим для розрахунку.

2.5 Консолідація мультимодальних даних пацієнтів

Основні проблеми, що виникають при обробці даних різних типів (об'єкт, що вивчається – це числові дані, дані зображень, погано структуровані звіти тощо), спостерігається швидке збільшення обсягу зібраних даних, відсутність методів для їх ефективного аналізу, потреба у значних людських ресурсах для підтримки процесу аналізу даних, висока обчислювальна складність наявних алгоритмів

аналізу. Це призводить до збільшення часу, необхідного для аналізу, з урахуванням регулярного оновлення апаратного забезпечення, а також необхідності роботи з розподіленими базами даних. Існуючі методи аналізу даних не в повній мірі використовують принципи розподілу та зберігання даних. Отже, необхідно розробити ефективний метод аналізу даних, який можна застосовувати до розподілених баз даних різних предметних областей великого розміру. Тому для досліджуваного сховища даних доцільно розробити методи та засоби їх консолідації.

Таким чином, процес консолідації даних, а саме: даних різного походження великого розміру, щодо аналізу та прогнозування поведінки об'єкта на досліджуваній території вимагає вирішення низки проблем:

- підвищення ефективності отримання, аналізу та використання інформації, необхідної для підтримки прийняття рішень щодо визначення стану об'єкта;
- підвищення якості прогнозування прийняття рішень за рахунок використання інформації, отриманої з надійного джерела;
- ідентифікація нових аспектів стану пацієнта шляхом аналізу даних, які не передбачались і не враховувались при прийнятті рішень;
- усунення негативних тенденцій та небажаних наслідків для прогнозування змін у стані об'єкта, що досліджується, при своєчасному виявленні.

Існує потреба у побудові моделі зведених мультимодальних даних пацієнта, які в сукупності наділяються ознаками повноти, цілісності, послідовності та становлять адекватну об'єктну інформаційну модель досліджуваної території з метою її аналізу обробки та ефективне використання в процесах підтримки прийняття рішень.

Виникає потреба у побудові моделі консолідованих мультимодальних даних пацієнта, які в сукупності наділені ознаками повноти, цілісності, несуперечності та складають адекватну інформаційну модель об’єкта досліджуваної області, з метою її аналізу опрацювання та ефективного використання в процесах підтримки прийняття рішень. На рисунку 2.19 подано схему сховища консолідованих даних простору системи супроводу лікування хворого в стаціонарі.

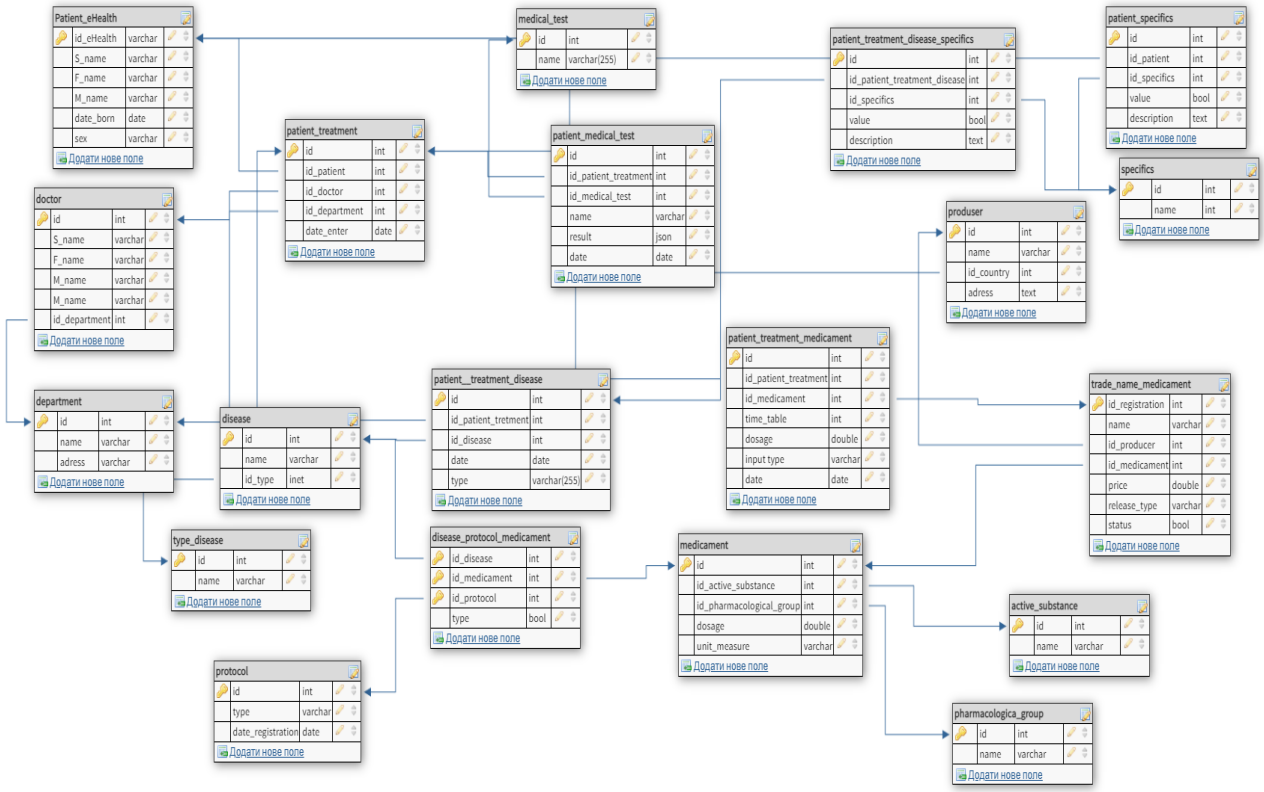


Рисунок 2.19 Сховище консолідованих даних (CDR)

Як видно з рисунку 2.20, джерела даних можуть мати різні моделі зберігання даних (платформи), а, отже, в них застосовуються різні методи опрацювання інформації.

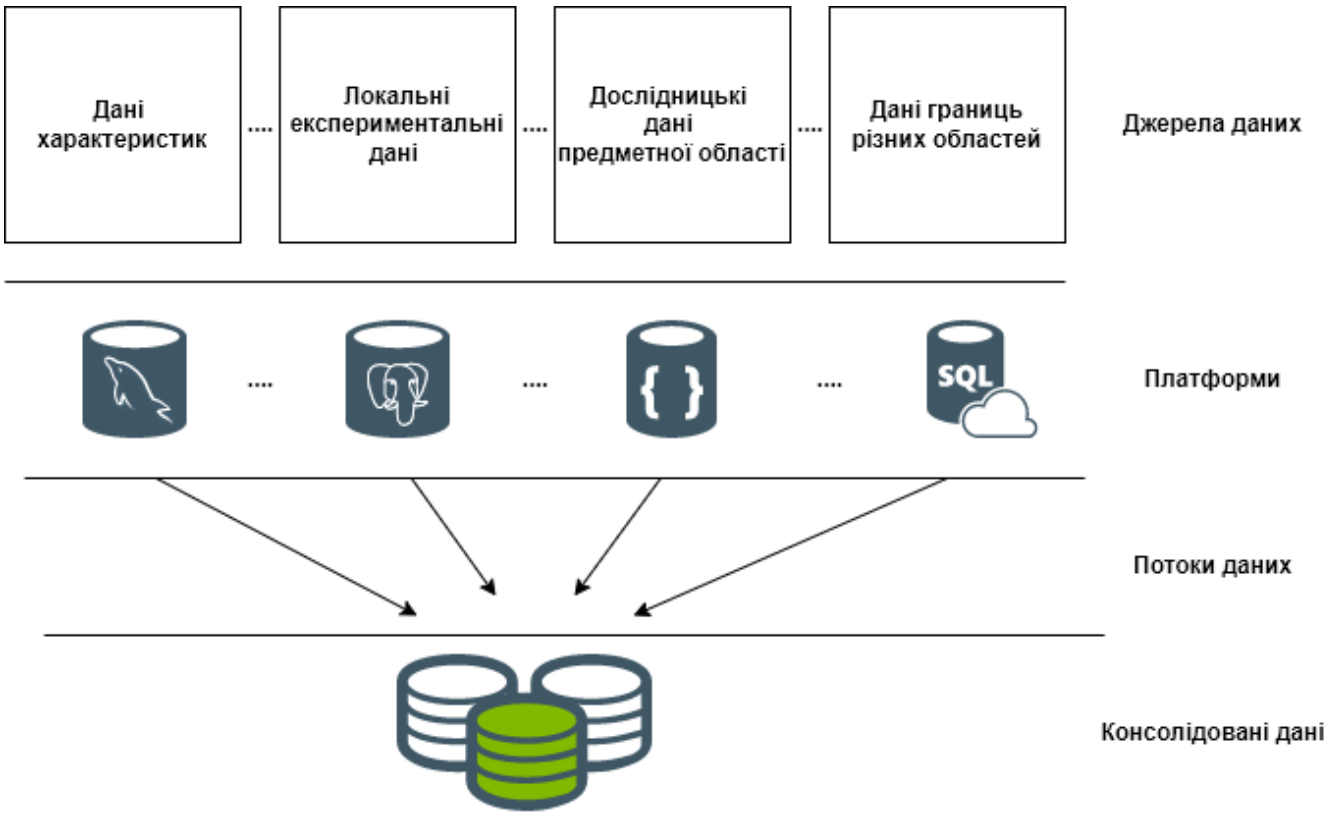


Рисунок 2.20 Схема консолідації мультимодальних даних

Основним завданням інтеграції даних є отримання даних з інформаційних продуктів простору даних з можливістю подальшого опрацювання. Проте, основною проблемою консолідації та використання існуючих методів та засобів інтеграції даних є неможливість попереднього визначення структур даних з метою їх узгодження, неможливість встановлення усіх протоколів обміну між джерелами даних, а також необхідність опрацювання неструктурованих даних.

Наступною проблемою є виразна потужність операцій пошуку для інформаційних продуктів різних типів (для текстових даних – лінійний пошук, для реляційних баз даних є операція селекції тощо). Для вирішення проблеми спершу побудуємо загальну схему консолідації даних з різних джерел. Консолідація даних полягає у завантаженні даних з заданих джерел у сховище консолідованих даних (CDR) рисунок 2.19.

2.5.1 Основні підходи консолідації

Особливості консолідації мультимодальних даних пацієнта, що складаються з наступних етапів:

1. Аналіз інформації, яку необхідно знайти;
2. Аналіз семантичних значень сутностей і атрибутів;
3. Уточнення семантичних відповідностей – за допомогою набору даних та сховища даних, а також визначаються відсутні зв'язки між концептами;
4. Побудова єдиної метамоделі – на основі визначених на попередніх етапах відповідностей та відмінностей структур даних формується структура сховища консолідованих даних, автоматично будується схема даних. Вказується, яка саме характеристика з якого джерела даних потраплятиме у кожен з вимірів простору станів;
5. Виведення результуючих відображень між сутностями й атрибутами – власне «інтеграція» – віртуальне перенесення даних з джерел у сховище консолідованих даних.

Семантичні зв’язки між джерелами даних наперед невідомі. Встановлення залежності здійснюється за допомогою детального опису структури джерела (метамоделі) та порівняння її з метамоделлю простору станів. Також використовується словник даних для визначення назв – синонімів характеристик об’єкта.

Приклад 2.6.

Семантичні зв’язки та їх відображення у каталозі та словнику простору станів.

Нехай є база даних показників пацієнта та перелік рішень оптимізації діяльності за близькими ознаками. Тоді каталог простору даних міститиме інформацію про відповідність кожного показника певному етапу стратегії, а словник синонімів – перелік усіх термінів для ідентифікації атрибутів бази даних.

Сховища даних дозволяють опрацьовувати інтегровані дані, що побудовані на основі наперед визначених моделей даних. У випадку роботи у всесвітній мережі

з величезною кількістю ресурсів (прикладами таких задач є туристичний бізнес – збирання інформації про місця відпочинку, її інтеграція та зберігання у внутрішніх базах даних, бізнес-аналітичні системи – інформація збирається на основі результатів функціонування підприємства та інформації показників його звітності для відображення ситуацій предметної області, а її збирання також проходить із джерел з наперед невідомими моделями даних) неможливо визначити, які саме моделі даних використовуватимуться. Тому лише за допомогою баз даних та сховищ даних не можна організувати ефективної взаємодії між усіма об’єктами у цих предметних областях.

Для створення системи інтеграції даних потрібна значна попередня робота, яка полягає у погодженні структур даних, встановленні зв’язків між характеристиками об’єкту, організації доступу до даних інформаційного об’єкту.

Залежно від усієї реалізації простору станів для каталогу даних Cg можна використовувати відношення реляційної моделі, XML–файли, програмні модулі тощо.

Над каталогом розміщене середовище керування моделями MM, яке дає змогу створювати нові зв'язки і маніпулювати наявними зв'язками (наприклад, об'єднувати або інвертувати відображення, зливати схеми і створювати єдині подання декількох джерел).

Для ідентифікації та роботи з неоднорідними колекціями в просторі станів доцільно використовувати глобальну схему імен (Uniform Resource Identifiers) як механізм посилань на глобальні константи, щодо яких є деяка угода між декількома джерелами даних.

Важливою компонентою простору станів є сховище консолідованих даних, яке служить для досягнення наступних цілей:

- створення асоціацій між об'єктами даних від різних учасників;
- вдосконалення доступу до джерел з обмеженими власними засобами доступу;
- забезпечення можливості виконання деяких запитів без доступу до реального джерела даних;

- консолідації даних як результату запиту;
- підтримання високого рівня доступності і відновлення.

Зв'язок між елементами простору даних GS_{oi}, визначеними в певному джерелі, та консолідованим сховищем даних CDR подано у вигляді відображення:

$$Mm(GS_{oi}) \Rightarrow CDR \tag{2.41}$$

Чим більше моделей здатне «розрізнити» середовище керування, тим точнішою буде інформація в CDR і ефективніше можна буде здійснювати процедури інтеграції, пошуку та опрацювання даних у просторі станів. У Клермонтському звіті про напрями дослідження надвеликих баз даних зазначено, що втрата інформації під час консолідації повинна становити не більше 25%. Проте не визначено, яка саме частина інформації може втрачатися.

2.5.2 Метод консолідації мультимодальних даних

Визначимо основні етапи процедури оцінювання якості консолідованих даних.

1. Складання системи характеристик якості консолідованих даних. Ця система має вигляд ієрархічної структури. Для різних методик характерна різна кількість рівнів ієрархії, а також різна кількість критеріїв кожного рівня ієрархії. Система критеріїв якості може включати як внутрішні, так і зовнішні характеристики даних. Однак перевагу слід віддавати зовнішніми характеристиками. Крім того, критерії можуть носити як кількісний, так і якісний характер. Перевагу варто віддавати кількісним характеристикам.

2. Визначення значень відносних вагових коефіцієнтів характеристик якості зі залученням думок експертів. Деякі методики базуються на допущенні про те, що всі критерії якості однаково важливі. Однак для отримання адекватної оцінки якості цей етап необхідний.

3. Оцінка значень одиничних показників якості за абсолютною шкалою. Інформація про їх значення може бути отримана за результатами випробувань, експертного чи соціологічного опитування. Найкращим є перше джерело, але у

випадку, якщо оцінка критеріїв цим методом неможлива, або надмірно трудомістка, то залучається експертна інформація.

4. Нормування значень одиничних показників якості. У різних методиках використовуються різні функції приведення.

5. Обчислення факторів якості на підставі розрахунку зваженої згортки значень одиничних показників якості. У різних методиках використовуються різні оператори згортки і різне число кінцевих показників якості.

Метод консолідації мультимодальних даних базується на оцінюванні їх якості та корисності:

1) визначення списку параметрів, необхідних для формування простору станів GS_{oi}.

2) визначення джерела даних для кожного параметру з GS_{oi}. Якщо коефіцієнт оцінки k більший за порогове значення, провести екстракцію значень.

3) пошук відхилень даних від нормованих показників по кожному типу захворювання;

4) формування результуючого набору даних.

Особливістю розробленого методу консолідації є попереднє визначення структур даних та узгодження семантики на рівні простору станів.

Розроблено методику керування структурою простору даних на основі визначеного значення функції якості даних.

Враховуючи особливості етапів процедури оцінювання якості консолідованих даних, розроблений метод консолідації даних з урахуванням якісного прийняття рішень, що передбачає оцінювання якості та корисності консолідованих даних інформаційного продукту.

Отже, на *першому етапі* необхідно визначити оцінку корисності даних на підставі визначення неперервної функції якості.

Одержана множина вхідних даних пацієнта $A = (a_1, a_2, a_3, .., a_n)$. Визначено цільову функцію якості Q, яка при обмеженнях даних n прямує до глобального максимуму:

$$Q(a_1,\dots,a_n)=\sum_{i=1}^{n}\left(a_i\sum_{k_i}d_{k_i}t_{k_i}\right)\rightarrow\max_{d_k\in[0;1],t_k\in[0;1](i=\overline{1,n},)}, \qquad (2.42)$$

де t_k – рівень довіри до інформаційного продукту для рішення, d_k– оцінка рішення.

На *другому етапі* наступним після фактичного оцінювання якості консолідованої інформації (2.42) необхідно провести оцінювання якості еталонного зразка (2.43), що відображає краще прийняте рішення.

Значення функції якості еталонного зразка визначається на основі даних не з консолідованого сховища даних інтелектуальної системи, а безпосередньо з джерел:

$$Q_i^e=\sum_i k_i a_i^e\,, Q_{const}^e=\sum_i k_i Q_i^e, \qquad (2.43)$$

На *третьому етапі* виконується нормування фактичної оцінки за формулою, де k_i – коефіцієнт оцінки, причому $k_i\in[0;1]$

$$Q_{const}=Q_{const}/Q_{const}^e. \qquad (2.44)$$

На *четвертому етапі* формування фактичної оцінки використано для прийняття рішень стосовно:

- зменшення невизначеності;
- визначення доцільності додавання інформаційного продукту системи прийняття рішень.

Отже, за результатами досліджень визначено підходи щодо процесу консолідації даних, а саме підвищення оперативності отримання, аналізу та використання інформації, підвищення якості прогнозування рішень, визначення нових аспектів діяльності об’єкта, усунення негативних тенденцій та небажаних наслідків

Основними проблемами консолідації та використання існуючих методів та засобів інтеграції даних є неможливість попереднього визначення структур даних з метою їх узгодження, неможливість встановлення усіх протоколів обміну між джерелами даних, необхідність опрацювання неструктурованих даних.

Для ефективної консолідації мультимодальності даних, їхнього пошуку та опрацювання у просторі станів необхідно забезпечити більше моделей, які середовище керування здатне було б «розрізнити».

Висновки до 2 розділу

1 Уведено модель стану пацієнта, яка спрощено відображає структуру параметрів пацієнта і зв'язків між ними, подає інформацію про його стан і поведінку та представлена як система, що консолідує різні елементи, що представлені у вигляді множин, які взаємно залежні та залежні від середовища оцінки.

2 Визначено, що продукційні правила, які формулюють рішення щодо перегляду та зміни тактики лікування, а саме стратегічних рішень щодо зміни стану, зміни медикаментів, зміни дозування, перелік лабораторних досліджень, фізіотерапевтичних заходів та ін., представлені у вигляді п'ятірки, враховуючи параметри пацієнта a, оцінки стану параметрів пацієнта e, коефіцієнту оцінки k, стратегічним рішенням s та протокольним рішенням g.

3 Простір стану пацієнта представлено, як евклідовий простір, що дозволило змоделювати інформаційну модель простору станів, як багатовимірну систему.

4 Формалізовано відображення фізичного стану пацієнта з урахуванням часовозалежних та часовонезалежних даних пацієнта, що дає можливість оцінити показники в певний момент часу, що уможливлює застосування персоналізованих рішень щодо лікування, а також аналізування зміни стану.

5 Розроблена модель простору станів пацієнта представлено у вигляді гіперкуба, як відображення функціонального відношення загального стану пацієнта GS, та його стан P у вигляді точки у просторі станів, як тривимірний простір, де вісь x – відображає часовий показник $P_o(t)$, вісь y – параметричний показник P_o (a), а вісь z – показник на окремі випадки, тобто множину індивідів (пацієнтів) Ob.

6 Розроблено метод пошуку шаблонів, який базується на модифікації методу асоціативних правил, що дозволяє зменшити трудомісткість і використовувати паралельний та розподілений режим для розрахунку, що є удосконаленням методу упорядкованого пошуку та надає чіткості та направленості у пошуку рішень стосовно вибору цільових схем лікування, що дає змогу зменшити ймовірність появи похибки при виборі схеми лікування.

7 Розроблено метод консолідації мультимодальних даних за рахунок попереднього визначення структур даних та узгодження семантики, що на відміну від методів консолідації даних, на рівні сховища даних, дало змогу агрегувати дані з різною структурою та підвищити точність прийнятих рішень.

8 Проведено аналіз проблеми опрацювання мультимодальних персоналізованих медичних даних та попередньої обробки даних та аналізу якості моделі даних. Обґрунтовано актуальність розв’язання цієї проблеми на основі введення нової абстракції керування даними – простору умов для прогнозування цільових змінних, що дало змогу підвищити точність прогнозування цільових показників в підмножині простору умов на 5 %, що забезпечило індивідуальний підхід до моніторингу стану пацієнта на основі тривалого спостереження та контролю лікаря.

РОЗДІЛ 3.
РОЗРОБЛЕННЯ МЕТОДІВ ОПРАЦЮВАННЯ ВЕЛИКИХ ТА МАЛИХ НАБОРІВ ПЕРСОНАЛІЗОВАНИХ ДАНИХ

У розділі проведено аналіз опрацювання медичних персоналізованих даних та досліджено процес прогнозування, визначено необхідність застосування препроцесингу даних. Досліджено особливості та принципи видобування даних за допомогою інструментів аналізу даних, для пошуку попередньо невідомої закономірності і зв'язків у даних, які можуть бути використані для побудови прогнозної моделі. Визначені класи завдань видобутку даних, що орієнтовані на передбачення даних та описові завдання. Охарактеризовано інструменти для очистки, препроцесингу, аналізу даних та виявлення взаємозалежностей між даними, що є важливим для досягнення успіху в системах діагностування та прогнозування. Запропоновані варіанти використання методів машинного та глибинного навчання для різнотипних медичних даних пацієнта. Здійснено вибір ефективних моделей ансамблів методів для пошуку залежностей між параметрами пацієнтів та факторами і показниками появи хвороби. Розроблено ієрархічний предиктор, який включає двоетапну обробку малих наборів даних методами кластеризації об'єктів та прогнозування для кожного одержаного кластера, що забезпечує покращення стійкості моделі до нових вхідних даних та забезпечує вищу точність на 4% у порівнянні з алгоритмами Random Forest та XGBoost. Розроблено стекінгову модель на основі алгоритмів машинного навчання, яка використовує Random Forest як метаалгоритм, і яка, на відміну від подібних моделей, базується на деформації метаознак та повторному навчанні на розширеному наборі даних, що забезпечує підвищення точності прогнозування даних та паралельної обробки даних як малої, так і великої розмірності.

Матеріали розділу опубліковані у роботах автора [163, 167, 172, 174, 177, 178, 181, 184, 188, 191, 193, 197, 201, 202, 207, 208, 211, 212, 216, 217].

3.1 Аналіз етапів опрацювання медичних персоналізованих даних пацієнта

Великі дані та машинне навчання мають великий потенціал для постачальників медичних послуг. Медичні дані – це одні із найбільш корисних, але в деякій мірі одній з найскладніших даних для аналізу. Аналітика даних використовується для виявлення закономірностей даних шляхом аналізу великої кількості неструктурованих, неоднорідних, нестандартних та неповних даних.

Особливо доцільною є прогностична аналітика у сфері прогнозування того чи іншого захворювання людини. Досить часто аналіз результатів лабораторних досліджень для прогнозування стану пацієнта займає досить багато часу, зусиль та коштів, проте не завжди вистачає необхідних ресурсів і як результат це може призводити до летальних випадків. Саме тому машинне навчання є досить хорошим помічником у цьому напрямку і може допомогти зберегти велику кількість людських життів.

Проте для того, щоб правильно аналізувати дані та виконати прогнозування, потрібно зробити декілька кроків препроцесингу отриманих даних:

- зібрати необхідні дані з різних джерел;
- очистити, опрацювати пропущені дані, виявити викиди та різного роду аномалії;
- виявити взаємозв’язки між досліджуваними даними та робити різного роду передбачення.

Загалом є два різних типи класифікації завдань із пошуку даних. Ці категорії відображено на рисунку 3.1 – це схема передбачення даних та описові завдання.

Рисунок 3.1. Завдання видобування даних

Основним завданням прогнозування даних є створення деяких моделей із наданого набору даних для того, щоб надати корисне та правильне прогнозування майбутніх чи невідомих значень іншого набору даних. Оскільки система передбачена для аналізу певних даних та прогнозування загального стану хворого під час його лікування, то вона попадає під категорію задачі класифікації.

У проаналізованих джерелах запропоновано алгоритм машинного навчання для прогнозування майбутньої інтубації серед пацієнтів з діагнозом або підозрою на COVID-19. Це ретроспективне когортне дослідження пацієнтів, яким поставлено діагноз або, які перебувають під дослідженням на перенесення COVID-19. Алгоритм машинного навчання був навчений прогнозувати наявність інтубації в майбутньому на основі попередніх показників, лабораторних та демографічних показників. Розроблений алгоритм може бути використаний для виявлення пацієнтів з високим ризиком можливості захворіти для надання допомоги у клінічній допомозі, що в певній мірі є дотичним до задач дослідження [138].

У джерелах [139, 155] описані параметри коагуляційної функції хворих на COVID-19 та виявлення факторів ризику розвитку важкої хвороби. Проведено багатовимірний аналіз регресії Кокса для виявлення потенційних біомаркерів для прогнозування прогресування захворювання. За результатами цього аналізу, побудовано номограму та оцінено точність прогнозування за допомогою калібрувальної кривої, кривої прийняття рішення, кривої клінічного впливу та аналізу Kaplan–Meier. Досліджено задачу виявлення залежності між певними показниками хворого та перенесення COVID-19.

Серед існуючих методів широко використовується також відрегульована модель випадкового лісу, підсилена алгоритмом AdaBoost. Модель використовує географічні, туристичні, медичні та демографічні характеристики пацієнта для прогнозування важкості перенесення та можливого результату – одужання або смерті. Модель має точність 94% та оцінку F1 0,86, на використаному наборі даних, аналіз виявляє позитивну кореляцію між статтю пацієнтів та смертністю, а також вказує на те, що більшість пацієнтів мають вік від 20 до 70 років [140].

Відображений процес очищення даних – це очищення відсутніх значень із використанням інтерполяції сплайнів та ентропії-кореляції. Очищення даних потім піддається процесу вилучення ознак за допомогою Principle Component аналізу. Для вибору оптимальних функцій представлений алгоритм Dragon Fly, а результуючий вектор функцій подається до мережі Deep Belief [141].

Проводилися дослідження щодо пошуку залежності між раковим та коронавірусним захворюваннями, де досліджувався вплив певного захворювання в людини на ймовірність появи захворювання COVID-19 [142].

Використано декілька нейронних моделей, а саме ResNet18, ResNet50, SqueezeNet та DenseNet-121, для ідентифікації COVID-19 на аналізованих рентгенівських знімках грудної клітки. Проведений аналіз дає можливість детальніше розібратись із даними моделями та використати їх у даному дослідженні [143].

У джерелах поставлено за мету виявити взаємозв'язк між клінічними показниками (такими, як лейкоцити, тромбоцити, вік, стать, наявність суміжних захворювань і т.д.) і тяжкістю коронавірусної хвороби COVID-19 та дослідити їхню роль у прогнозуванні тяжкості захворювання COVID-19. Пошук взаємозв’язку був здійснений за допомогою багатовимірної логістичної регресії. Це приклад щодо реалізації аналізу взаємозалежності багатьох показників стану пацієнта та захворюваністю на COVID-19 [144].

У [145] описано проведення ретроспективного дослідження у 70 безсимптомно інфікованих пацієнтів, підтверджених тестами на нуклеїнові кислоти в провінції Хунань, Китай, з 28 січня 2020 року по 18 лютого 2020 року.

Для оцінки потенційних предикторів появи симптомів приведена модель регресії Кокса. Як результат, виявлено те, що пацієнти, які курять чи мають легеневу хворобу, були схильні до безсимптомного перенесення хвороби, потребують особливої уваги. Отже, як бачимо ця робота є дуже дотичною до даного дослідження [145].

На основі ретроспективного аналізу [146] виявлено, що показник FIB підвищений у важких пацієнтів і є кращим, ніж кількість лімфоцитів та міоглобіну, для розрізнення загальних та важких пацієнтів. У межах дослідження припустили, що гормональне лікування не має суттєвого впливу на COVID-19 [146].

У [147] досліджено, чи комп'ютерна томографія (КТ) точно виявляє важкість перенесення COVID-19. Зображення КТ отримували за допомогою LK2.1. Для виявлення суттєвих особливостей використано двовибірковий t-тест або U-тест Манна-Уїтні. Метод мінімальної надмірності та максимальної релевантності (MRMR) проведений, щоб знайти характеристики з максимальною кореляцією та мінімальною надмірністю. Потім ці особливості використовувались для виявлення важких пацієнтів за допомогою багатовимірного методу логістичної регресії. Крім того, побудована клінічна модель. Для оцінки ефективності двох моделей проведені аналізи ROC. Співвідношення клінічних особливостей та особливостей текстури КТ аналізували за допомогою кореляційного аналізу Spearman [147, 148].

Отже, на основі проведеного аналізу літературних джерел розроблено схему послідовності виконання процесів опрацювання медичних персоналізованих даних (рисунок 3.2).

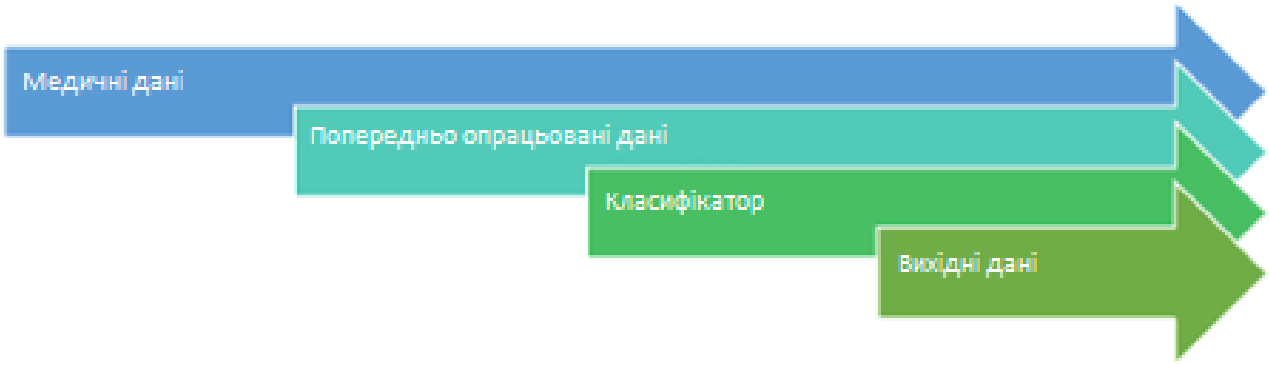


Рисунок 3.2. Схема послідовності виконання процесів опрацювання медичних персоналізованих даних

Послідовність виконання процесів опрацювання медичних персоналізованих даних визначено на рисунку 3.2. Характерним для інформаційних систем діагностики, що залежать від медичних числових чи категоріальних даних, є те, що інструменти для очистки, препроцесингу, аналізу даних та виявлення взаємозалежностей між даними відіграють важливу роль в успіху такого роду систем. Фаза попередньої очистки даних забезпечує опрацювання нульових значень, відкидання викидів та різного роду аномалій, відкинути дані, які несуть найменшу кількість важливої інформації. Фаза аналізу даних дозволяє нам знаходити важливі ознаки та взаємозв'язки між досліджуваними даними.

Етапом класифікації є етап аналізу даних для того, щоб вивчити набір даних та подальша їхня класифікація на ряди категорій. Кожна із категорій може володіти певними власними характеристиками і даними, які відносяться до цієї категорії та мають ті ж властивості. Існує великий ряд класифікаторів, які досить добре працюють із такими проблемами, як виявлення наявного захворювання в пацієнта чи ні. Найпоширенішими серед них є нейронні мережі, Байєсові мережі, класифікатор Байєса, Дерева рішень та ін.

Під час етапу тестування здійснюється оцінка продуктивності та точності розробленої системи. Для того, щоб забезпечити достатньо високий рівень точності розробленої системи, переважно використовується техніка перехресної перевірки (cross-validation).

У задачах машинного чи глибинного навчання для оцінки якості роботи моделі використовуються спеціальні метрики для порівняння точності різних видів алгоритмів.

3.1.1 Аналіз методів прогнозування ризиків захворювання

Методи машинного навчання використовують як для структурованих, так і для неструктурованих даних. Неструктуровані дані – це нестиснуте чи стиснуте аудіо, текст чи картинка. А структуровані дані – це, наприклад, бази даних, в яких містяться особисті дані пацієнтів та дані щодо їхніх аналізів, їхнього стану здоров’я і т.д.

Традиційно для прогнозування ризику захворювання використовувались стандартні статистичні методи та інтуїція лікаря, знання та досвід. Така практика часто призводить до небажаних упереджень, помилок та великих витрат і негативно впливає на якість послуг, що надаються пацієнтам [152]. Зі збільшенням доступності електронних даних про здоров'я, обчислювальні підходи машинного навчання, стали більш практичними для застосування та дослідження в області прогнозування захворювань. У літературі більшість відповідних досліджень використовували один або кілька алгоритмів машинного навчання для прогнозування певної хвороби. З цієї причини порівняння ефективності різних керованих алгоритмів машинного навчання для прогнозування захворювань є основним напрямком цього дослідження.

У навчанні з вчителем ми маємо вхідні дані *x*, і прагнемо навчити функцію відображати опрацьовані вихідні дані. У цьому випадку використовували багатокласову класифікацію. Багатокласова класифікація – це завдання класифікації машинного навчання, яке складається з більш ніж двох класів або результатів. Існують сотні моделей для класифікації. Насправді часто можна взяти модель, яка працює для регресії, і перетворити її на модель класифікації. В основному так працює логістична регресія. Моделювання лінійної регресії WX + b на вхід, перетворює її у значення ймовірності від 0 до 1, подаючи реакцію у

сигмоїдальну функцію. Потім ми прогнозуємо, що вхідні дані належать до класу 0, якщо модель видає ймовірність більше 0,5, а в іншому випадку – до класу 1.

Процес прикладного машинного навчання складається з послідовності етапів.

Ми можемо переходити між кроками для будь-якого проєкту, але всі проєкти мають однакові загальні кроки:

Крок 1: ідентифікація проблеми;

Крок 2: попередня підготовка даних;

Крок 3: оцінка моделі;

Крок 4: фіналізація моделі.

Одним з дуже важливих етапів є етап підготовки даних (крок 2), і є загальні або стандартні завдання, які можна використовувати або дослідити під час етапу підготовки даних у проєкті машинного навчання.

Отже, перед поданням вхідних даних класифікатору, дані повинні бути опрацьованими. Типи підготовки даних залежать від первинних даних. Однак, більшість проєктів прогнозування вимагають однакових завдань з підготовки даних знову і знову (рисунок 3.3).

Рисунок 3.3 Схема опрацювання даних для пошуку персоналізованих рішень

Це забезпечує приблизну структуру, яку ми можемо використовувати для обмірковування та навігації різними алгоритмами підготовки даних, які проаналізовано у цьому розділі.

3.1.2 Розроблення рішення з багатомірного статистичного аналізу персоналізованих даних

Виділено 4 генеральні ідеї багатомірного статистичного аналізу на яких базуються всі основні розділи і підходи математичного апарату класифікації та зменшення розмірності.

1. Ефект суттєвої багатомірності є принципом, суть якого полягає в тому, що висновки, які отримують в результаті аналізу та класифікації множини статистично досліджуваних (за низкою властивостей) об'єктів, повинні опиратися одночасно на сукупність цих взаємозв'язаних властивостей з обов'язковим врахуванням структури і характеру їх зав'язків.

2. Можливість лаконічно пояснити природу багатовимірних структур, які підлягають даному аналізу. Суть цього принципу полягає в наступному.

Визначимо поняття багатовимірної структури. Нехай $\{Ob_1, Ob_2, \ldots, Ob_n\}$ – множина статистично досліджуваних об'єктів (пацієнтів). Результати досліджень можуть бути представлені у двох формах.

Найбільш поширеною формою є таблиці (матриці) «об'єкт – властивість», в якій кожен пацієнт представлений вектором значень $X_i = \left(X_i^{(1)}, X_i^{(2)}, \ldots, X_i^{(p)}\right)^T$, врахованих ознак (характеристик) $X_i^{(1)}, X_i^{(2)}, \ldots, X_i^{(p)}$, зареєстрованих при аналізі i-го об'єкта.

Другою формою є матриці (таблиці) – «попарних порівнянь» наступного виду

$$\overrightarrow{Ob} = \begin{pmatrix} o_{11} & \cdots & o_{1n} \\ o_{21} & \cdots & o_{2n} \\ \vdots & \vdots & \vdots \\ o_{n1} & \cdots & o_{nn} \end{pmatrix},$$

де елементи o_{ij} – є результатом зіставлення об'єктів Ob_i і Ob_j в значенні деякого

заданого відношення. Величина a_{ij} може виступати як міра подібності або відмінності об'єктів; міра зв'язку або взаємозв'язку між об'єктами в будь-якому процесі; геометрична відстань між двома конкретними об'єктами, відношення надання переваги, наприклад: $o_{ij} = 1$, якщо $Ob_i > Ob_j$ і $o_{ij} = 0$, якщо $Ob_i < Ob_j$, тощо. Під лаконічним поясненням природи розуміють апріорне припущення того, що існує порівняно невелике число визначальних факторів, з допомогою яких можна досить точно описати, не лише спостережувані характеристики $X_i^{(k)}$ пацієнтів (всі елементи $X_i^{(k)}$ та елементи o_{ij} в матрицях попарних порівнянь) і характер зав'язків між ними, але також і шукану класифікацію самих об'єктів.

3. Максимальне використання навчання при налаштовуванні математичних моделей класифікації та зменшення їх розрядності. Пояснюють цей принцип з допомогою схеми «на вході задачі – на виході задачі».

Якщо дослідник володіє і «входами» і «виходами» задачі, то початкову інформацію називають навчальною і метою дослідження є опис процедур, з допомогою яких при поступленні лише вхідних даних стосовно нового пацієнта, його можна було б з найбільшою (в певному сенсі) точністю віднести до одного з класів (в задачі класифікації) або наділити його значеннями визначальних факторів (в задачі зменшення розмірності). До таких ситуацій відносять задачі медичної діагностики: «входи» – результати обстежень, «виходи» – діагнози.

Метою діагнозу є використання «навчання» для вибору з множини результатів невеликого числа найбільш інформативних показників і побудови на їх основі формального діагностуючого правила.

4. Оптимізаційне формулювання задач класифікації та зменшення розмірності. Суть цього принципу полягає в тому, щоб серед множини можливих методів, які реалізують поставлену мету статистичної обробки даних – розбиття сукупності статистично досліджуваних об'єктів (пацієнтів) на однорідні класи, перехід від заданого широкого набору ознак (характеристик) $X_i^{(1)}, \ldots, X_i^{(P)}$ до невеликого числа визначальних факторів – вміти знайти кращий метод з допомогою оптимізації деякого заданого критерію (функціоналу) якості методу.

Виділено класичні етапи аналізу багатовимірних даних:

- етап формування даних, їх аналіз та опрацювання;
- етап тренування класифікатора на тренувальних зразках даних.

Етап препросесингу включає в себе етап пошуку нульових значень у наборі даних та їхнє опрацювання. Є два основні способи обробки відсутніх значень при попередній обробці даних.

Перший метод попередньої обробки даних зазвичай використовується для обробки нульових значень. Видаляється певний рядок, якщо він має нульове значення для певної функції, і певний стовпець, якщо він має більше 75% відсутніх значень.

Цей метод рекомендується застосовувати лише тоді, коли в наборі даних достатньо зразків. Потрібно переконатися, що після того, як ми видалили дані, не буде додано упереджень. Видалення даних призведе до втрати інформації, що не дасть очікуваних результатів під час прогнозування результату.

Також можна обчислити середнє, медіану або максимальне чи мінімальне та замінити її відсутніми значеннями.

Це наближення, яке може додати дисперсію до набору даних. Але втрата даних може бути знижена з використанням цього методу, він дає кращі результати порівняно з видаленням рядків і стовпців.

Далі здійснено кодування категорійних даних у числові дані.

Переважна більшість методів класифікації та регресії сформульовані в термінах евклідових або метричних просторів, тобто мають на увазі представлення даних у вигляді дійсних векторів однакової розмірності. У реальних даних, однак, не такі рідкісні категоріальні ознаки, що приймають дискретні значення, такі як так/ні або січень/лютий/.../грудень.

Для перетворення таких даних використано LabelEncoder. LabelEncoder – це об’єкт, який використовують і який допомагає нам при передачі категоріальних даних у числові дані. Вектор на основі одного гарячого кодування — це двійковий вектор, який встановлює одне значення відповідного індексу розмічених даних на 1, а інші — на 0.

Після опрацювання даних класифікатор може вчитись прогнозувати результат на тренувальній вибірці даних.

Досягнувши достатньої точності на тренувальній вибірці, система може класифікувати раніше невідомі їй дані за допомогою натренованого класифікатора. Після тренування, класифікатор зможе вказати приналежність невідомого йому вектору ознак.

На рисунку 3.4 наведено діаграму варіантів використання, на якій актором є власне система тому, що розроблювана система не потребує користувацьких чи адміністраторських функцій.

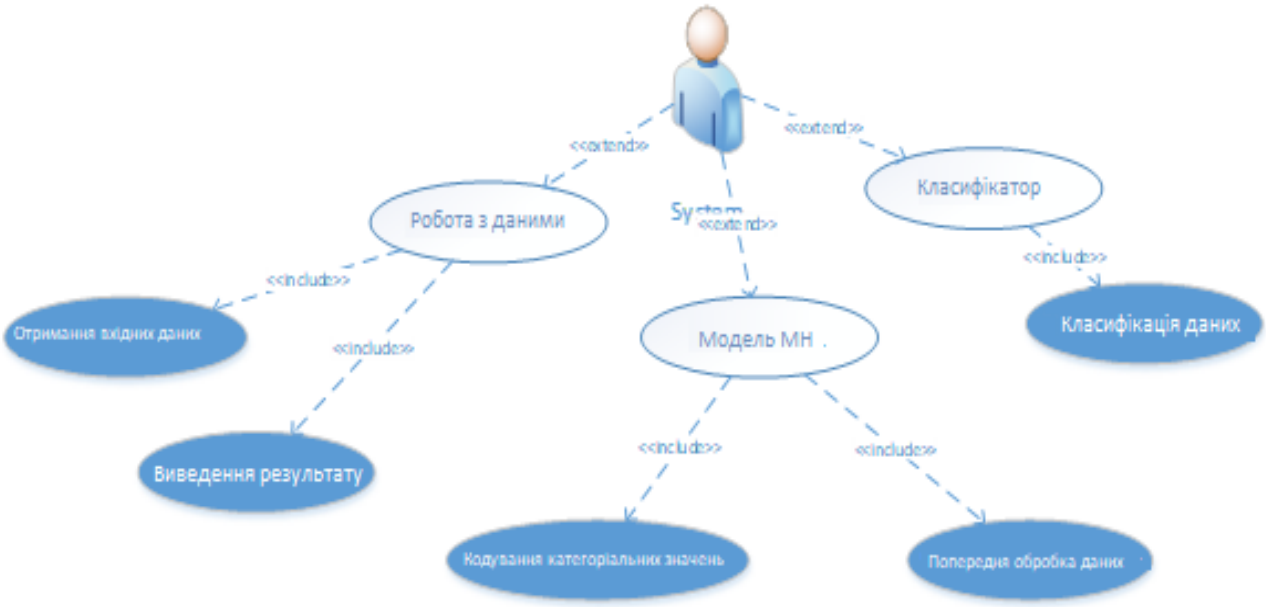


Рисунок 3.4. Діаграма варіантів використання системи

Використано ймовірнісну модель методу головних компонент Principal component analysis (PCA) для пониження розмірності таким чином [246, 247]:

$$X = W_t + \mu + \varepsilon. \tag{3.1}$$

Тут $W_t + \mu$ задає точку на гіперплощині, а $\varepsilon \approx N(\varepsilon|0, \sigma^2 I)$ – нормально розподілена шумова компонента (відхилення) з однаковою дисперсією σ^2 в усіх напрямах у просторі R^D. Символом $N(x|\mu, \sum)$ тут і далі буде позначатися щільність багатовимірного нормального розподілу:

$$N(x|\mu, \sum) = \frac{1}{\sqrt{2\pi}^D \sqrt{det\,\Sigma}} exp(-\frac{1}{2}(x-\mu)^T \sum -1(x-\mu)). \tag{3.2}$$

В якості апріорного розподілу на значення координат об'єкту t в базисі гіперплощини вибрано:

$$p(t) = N(t|0, I). \quad (3.3)$$

Повний спільний розподіл у ймовірнісній моделі PCA задається наступним чином:

$$p(X, T|W, \mu, \sigma) - \prod_{n=1}^{N} p(x_n, t_n|W, \mu, \sigma) - \prod_{n=1}^{N} N(x_n|Wt_n + \mu, \sigma^2 I)N(t_n|0, I).$$

Тут X – набір спостережуваних змінних, T – набір прихованих змінних і (W, μ, σ) – набір параметрів моделі.

Для пошуку значень параметрів моделі скористаємося методом максимальної правдоподібності:

$$p(X, T|W, \mu, \sigma) - \prod_{n=1}^{N} p(x_n, t_n|W, \mu, \sigma) \to max_{W,\mu,\sigma} . \quad (3.4)$$

Маргінальний розподіл $p(x_n)$ у ймовірнісній моделі PCA обчислюється як

$$p(x_n|W, \mu, \sigma) = \int p(x_n|t_n, W, \mu, \sigma)\, p(t_n) dt_n.$$

Останній інтеграл є згорткою двох нормальних розподілів і може бути обчислений аналітично [248]:

$$p(x_n|W, \mu, \sigma) = \int N(x_n|Wt_n + \mu, \sigma^2 I)N\,(t_n|0, I)dt_n = N(x_n|\mu, \sigma^2 I + WW^T).$$

Таким чином, ймовірнісна модель PCA є нормальним розподілом, в якому матриця коваріацій задається спеціальним чином

$$C = WW^T + \sigma^2 I \quad (3.5)$$

Зауважимо, що так само як і класична модель PCA, ймовірнісна модель PCA інваріантна щодо вибору базису в гіперплощині. Нехай $R \in R^d$ – довільна ортогональна матриця, що задає поворот базису гіперплощини. Це відповідає використанню матриці $\underline{W} = WR^C$. Тоді матриця коваріацій дорівнює [246]

$$C = \underline{WW^T} + \sigma^2 I = WRR^T W + \sigma^2 I = WW^T + \sigma^2 I\, . \quad (3.6)$$

Таким чином, матриця коваріацій не залежить від R.

Повернемося тепер до задачі оптимізації (формула 3.7). Цю задачу можна еквівалентно переписати таким чином [247,248]:

$$log\, p\,(X|W, \mu, \sigma) = \sum_{n=1}^{N} log\, N\,(x_n|\mu, \sigma^2 I + WW^T) = -\frac{ND}{2} log\, 2\,\pi -$$

$$\frac{N}{2} det(\,\sigma^{2I} + WW^T) - \frac{1}{2}\sum_{n=1}^{N}(x_n - \mu)^T(\sigma^2 I + WW^T)^{-1}(x_n - \mu) \to max_{W,\mu,\sigma} . \quad (3.7)$$

Можна показати, що зазначена задача оптимізації має аналітичне вирішення:

$$\mu = \frac{1}{N}\sum_{n=1}^{N} x_n,$$

$$W = Q(\wedge - \sigma^2 I)^{1/2} R\ , \qquad (3.8)$$

$$\sigma^2 = \frac{1}{D-d}\sum_{i=d+1}^{D} \lambda_i.$$

Тут $Q = (q_1|\ldots|q_d) \in R^{D\times d}, q_1, \ldots, q_d -$ власні вектори вибіркової матриці коваріації, що відповідають найбільшим власним значенням $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_d$, $\wedge = diag(\lambda_1, \lambda_2, ..., \lambda_d), R$ – довільна ортогональна матриця розміру $d \times d$.

3.2 Розроблення ієрархічного предиктора для оцінки резистентності пацієнта до хвороби

Отже, для ієрархічного предиктора здійснено поєднання етапів попереднього аналізу та аналізу даних. Для другого етапу використано методологію ансамблювання моделей.

Знаходження залежностей у даних вимагає аналізу зв'язків між десятками параметрів досліджуваного процесу та сотнями можливих джерел впливу на цей процес. Залежності є недетермінованими, тому моделювання потребує використання статистичних методів для аналізу випадкових процесів. Частина інформації часто прихована від спостереження або не контролюється. Тому в процесі аналізу зібраної інформації виникло багато труднощів. Новизна полягає в тому, що ієрархічна архітектура включає керовані та неконтрольовані методи. Це дозволяє розробити ансамбль методів, заснованих на кластеризації та класифікації. Попередньо необхідно визначити класифікатори, які будуть об’єднані в ансамбль моделей. Додатково слід зазначити про необхідність часовозалежних персональних даних, що також впливає на вибір моделі.

Останнім кроком у будь-якій системі, заснованій на ансамблі, є механізм, який використовується для об’єднання окремих класифікаторів. Стратегія, що використовується на цьому етапі, частково залежить від типу класифікаторів, які

використовуються як члени ансамблю. Наприклад, деякі класифікатори, такі як машина опорних векторів, надають вихідні дані лише з дискретними значеннями. Найпоширенішим правилом комбінування для таких класифікаторів є (просте або зважене) голосування більшістю, за яким слідує підрахунок Борда. Інші класифікатори, такі як багатошаровий персептрон або (наївний) класифікатор Байєса, надають безперервні результати, специфічні для класу, які інтерпретуються як підтримка, яку надає класифікатор кожному класу. Для таких класифікаторів, на додаток до підходів, що ґрунтуються на голосуванні, доступний ширший набір опцій, наприклад арифметичні об'єднувачі (сума, добуток, середнє значення тощо) або більш складні шаблони рішень. Багато з цих об'єднувачів можна використовувати одразу після завершення навчання, тоді як для більш складних алгоритмів поєднання може знадобитися додатковий крок навчання (як це використовується в узагальненні з накопиченням або ієрархічному класифікаторі).

Отже, припустимо, що лише мітки класів доступні з виходів класифікатора, і визначимо рішення t-го класифікатора як $d_t \in [0,1]$, t=1..T, c=1..C, де T – кількість класифікаторів, а C – кількість класів. Якщо t-й класифікатор (або гіпотеза) h_t вибирає клас ω_c, то $d_{t,c}$=1 і 0, у іншому випадку. Зазначимо, що виходи безперервного значення можна легко перетворити на виходи міток (шляхом призначення $d_{t,c}$=1 для класу з найвищим виходом), але не навпаки. Тому правила комбінування, описані в цьому розділі, також можуть використовуватися класифікаторами, що надають підтримку певного класу.

Мажоритарне голосування має три варіанти, залежно від того, чи приймається рішення ансамблю:

- це клас, щодо якого згодні всі класифікатори (одностайне голосування);
- підтримано принаймні одним більше половини кількості класифікаторів (проста більшість);
- клас, який отримує найбільшу кількість голосів, незалежно від того, перевищує чи ні сума цих голосів 50% (голосування за системою більшості).

Якщо не зазначено інше, більшість голосів зазвичай належить до множинного голосування, яке може бути математично визначено таким чином: виберіть клас ω_{c*}, якщо

$$\sum_{t=1}^{T} d_{t,c*} = max \sum_{t=1}^{T} d_{t,c} . \quad (3.9)$$

Якщо результати класифікатора є незалежними, тоді можна показати, що більшість голосування – це правило оптимальної комбінації. Щоб підтвердити це, розглянемо непарне число T класифікаторів, причому кожен класифікатор має ймовірність правильної класифікації p. Тоді ансамбль приймає правильне рішення, якщо принаймні $\lfloor T/2 \rfloor + 1$ із них класифікатори вибирають правильну мітку. Тут функція повертає найбільше ціле число, менше або дорівнює його аргументу. Точність ансамблю регулюється за біноміальним розподілом; ймовірність мати $k \geq T/2 + 1$ класифікаторів, що повертають правильний клас. Оскільки кожен класифікатор має коефіцієнт успіху p, тоді ймовірність успіху ансамблю:

$$P_{ens} = \sum_{k=T/2+1}^{T} \binom{T}{k} p^k (1-p)^{T-k}. \quad (3.10)$$

Зазначимо, що P_{ens} наближається до 1 як $T \to \infty$, якщо $p > 0{,}5$; і наближається до 0, якщо $p < 0{,}5$.

Формула 3.10 стверджує – якщо ймовірність членського класифікатора дає правильну відповідь вищу за 0.5, що насправді є найменшою мірою, яку ми можемо очікувати від класифікатора у задачі бінарного класу, тоді ймовірність успіху наближається до 1 дуже швидко. Якщо ми маємо проблему з кількома класами, та сама концепція актуальна доти, поки кожен класифікатор має кращу ймовірність успіху, ніж випадкове вгадування (тобто p > ¼ для задачі чотирьох класів).

Якщо у нас є підстави вважати, що деякі з класифікаторів, швидше за все, правильніші ніж інші, зважування рішень цих класифікаторів може покращити загальну продуктивність порівняно з мажоритарним голосуванням. Припустимо, що ми маємо механізм передбачення (майбутнього) наближеного узагальнення продуктивність кожного класифікатора. Потім ми можемо призначити вагу W_t класифікатору h_t як частку його оціненої продуктивності узагальнення. Ансамбль,

комбінований відповідно до голосування зваженою більшістю, тоді обирає клас c*, якщо:

$$\sum_{t=1}^{T} w_t d_{t,c^*} = max_c \sum_{t=1}^{T} w_t d_{t,c} \, . \qquad (3.11)$$

Тобто якщо загальна зважена кількість голосів, отриманих класом c*, вища за загальну кількість голосів, отриману будь-яким іншим класом. Загалом ваги голосування нормалізуються таким чином, що їх сума рівна 1.

Якби ми апріорі знали, які класифікатори працювали б краще, ми будемо використовували б лише ці класифікатори. За відсутності такої інформації, вірогідною та загальновживаною стратегією є використання продуктивності класифікатора на окремий набір даних перевірки (або навіть навчання) як оцінка цього класифікатора продуктивність узагальнення.

Ансамблі моделей доцільно ви користовувати, враховуючи обмеження при роботі з великими та малими медичними персоналізованими даними: мала кількість об’єктів (<=100), велика кількість параметрів (>150).

Недолік існуючих методів:

- важливість ознак різна для різних методів. Це означає, що залежність між параметрами підтримується лише для частини набору даних;
- відсутність генералізації (здатності моделі належним чином адаптуватися до нових, раніше небачених даних, отриманих з того самого розподілу, що використовувався для створення моделі);
- необхідність регуляризації;

Аналіз поширення епідемії COVID-19 у різних країнах свідчить про різний характер ураження вірусом [125].

Застосування ансамблів здійснювалось для набору даних [111], зібраного за допомогою гуглформи, як результат проекту, який фінансується Центрально-Європейською Ініціативою. Набір даних перевірено Львівським регіональним центром резистентності до COVID`19. Проект Stop COVID-19 [122] має приклад використання, реалізований в Україні та Білорусі. Партнери з Німеччини теж

поділилися гугл -формою та допомогли зі збором даних. Набір даних – це дані, зібрані за період з 01 вересня по 29 жовтня. Набір даних містить результати про непідтверджені та підтверджені випадки COVID-19 [128].

Характеристики IgG та IgM є результат швидких тестів і наборів анти-SARS-CoV-2 IgG та IgM. Кількість антитіл IgG і IgM різна в різні терміни після інфікування. Тому враховується не лише категоріальне значення (позитивне, невизначене чи негативне), а й точні значення цих атрибутів.

3.2.1 Попередня обробка даних

По-перше, передбачена попередня обробка даних. Основне припущення аналізу: всі, хто заповнив анкету, були або хворі, або мали симптоми. Проаналізовано розподіл даних.

RStudio використовується для аналізу даних. За допомогою пакетів factoextra, cluster, corrplot і caret реалізовано більшість методів.

Вибір екземплярів базується на розподілі даних.

Розподіл характеристик набору даних наведено в таблиці 3.1. Частота віку <15 менше 0,013. Тому 4 рядки видаляються. Розподіл за статтю відносно однаковий.

Таблиця 3.1. Розподіл за віком, статтю, регіоном та COVID.

#	Вік	n
1	23-40	124
2	40-65	84
3	16-22	82
4	>65	19
5	<15	4

#	Стать	n
1	Чоловік	178
2	Жінка	135

#	COVID	n
1	так	105
2	ні	100
3	може бути	78

#	Регіон	n
1	Україна, Львів	159
2	Україна, Чернівці	67
3	Білорусь	56
4	Німеччина	27
5	інші	4

Розподіл за групами крові представлений на рисунку 3.5. Розподіл підтверджених випадків за групами крові такий: 1 група – 58, 2 група -76, 3 група – 18, 4 група – 15.

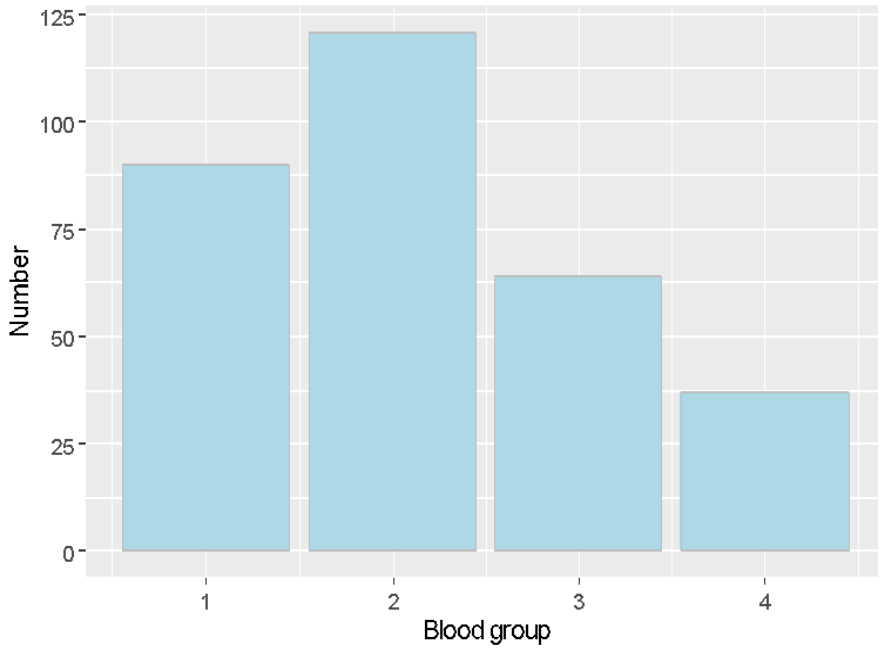


Рисунок 3.5. Розподіл крові по групах

Наступним припущенням є кореляція між ознаками (рисунок 3.6) обраних осіб та осіб з невідомим діагнозом.

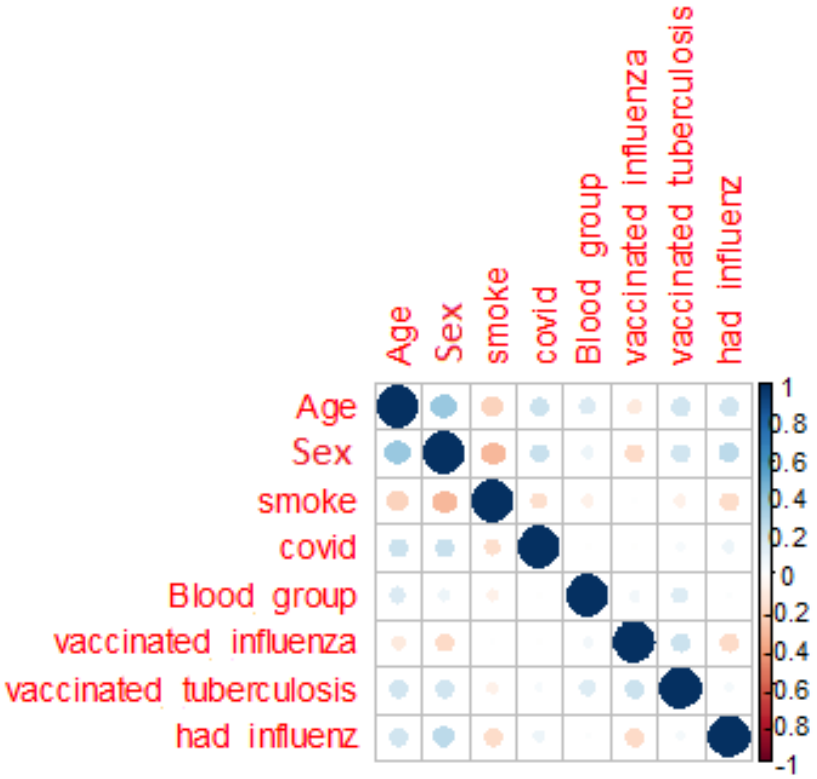


Рисунок 3.6. Кореляційна матриця

Представлена кореляційна матриця показує відсутність залежних параметрів для всього набору даних. Цільовий атрибут covid явно не визначений ознаками.

Спектральна декомпозиція, яка вивчає кореляцію між змінними, розроблена за допомогою аналізу основних компонентів (PCA). Залежність між змінними наведена на рисунку 3.7. Позитивно корельовані змінні вказують на одну сторону графіка. Негативно корельовані змінні вказують на протилежні сторони графіка. Тому представлено співвідношення між COVID та віком, статтю, групою крові, щепленим туберкульозом, хворим на грип.

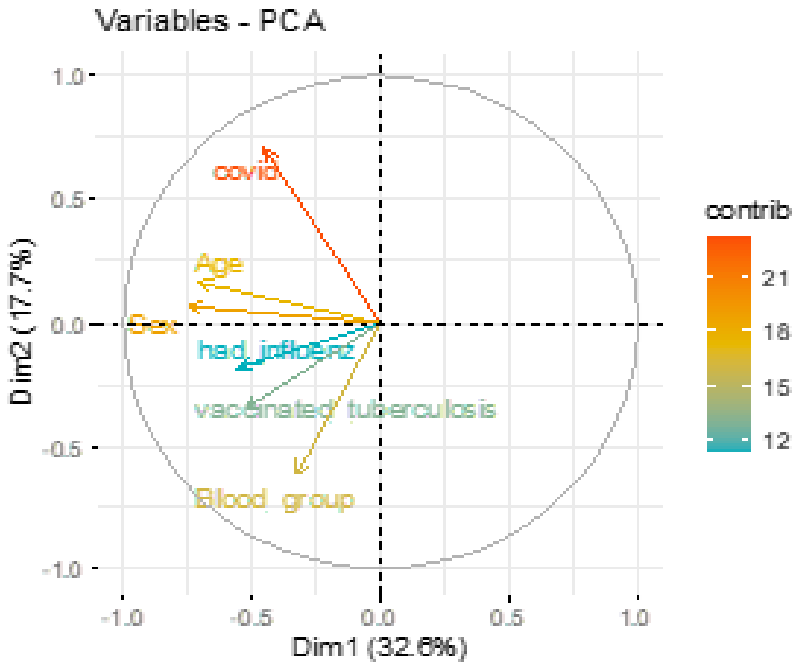


Рисунок 3.7. Залежність між змінними

Наступним кроком є кластеризація та аналіз даних усередині кластерів.

3.3. Вибір предикторів

На наступному етапі ми намагаємося побудувати декілька предикторів. До уваги беруться два показники: середньоквадратична помилка (RMSE) і середня абсолютна відсоткова помилка (MAPE). Результати наведено в таблиці 3.2.

Таблиця 3.2. Результати предикторів

Модель	RMSE	MAPE
Лінійна регресія	6.109753	0.4946111
Дерево регресій	4.970676	0.4763212
Випадковий ліс (500 дерев)	3.560431	0.3753441
k-найближчого сусіда	3.360431	0.3753441
МОВ з лінійним ядром	3.194611	0.2923931
МОВ з поліноміальним ядром	2.262171	0.2670496
ШНМ з 1 прихованим шаром, 12 нейронами	2.0972937	0.2025539

Аналіз вибраних змінних наведено нижче (Таблиця 3.3).

Таблиця 3.3. Точність прогнозування для вибраних функцій

Модель на основі обраних змінних	RMSE	MAPE
Лінійна регресія	3.972548	0.3240777
Дерево регресій	4.970676	0.4763212
Випадковий ліс (500 дерев)	3.546447	0.2730968
k-найближчого сусіда	3.346447	0.2730968
МОВ з лінійним ядром	3.095239	0.2386008
МОВ з поліноміальним ядром	2.226906	0.1766799
ШНМ з 1 прихованим шаром, 12 нейронами	2.059849	0.1513523

3.3.1 Класифікація

На етапі класифікації виконується побудова класифікатора для всього набору даних.

Цільова змінна буде «Чи був у вас COVID?», решта змінних – характеристики. Будується дерево рішень (рисунок 3.8).

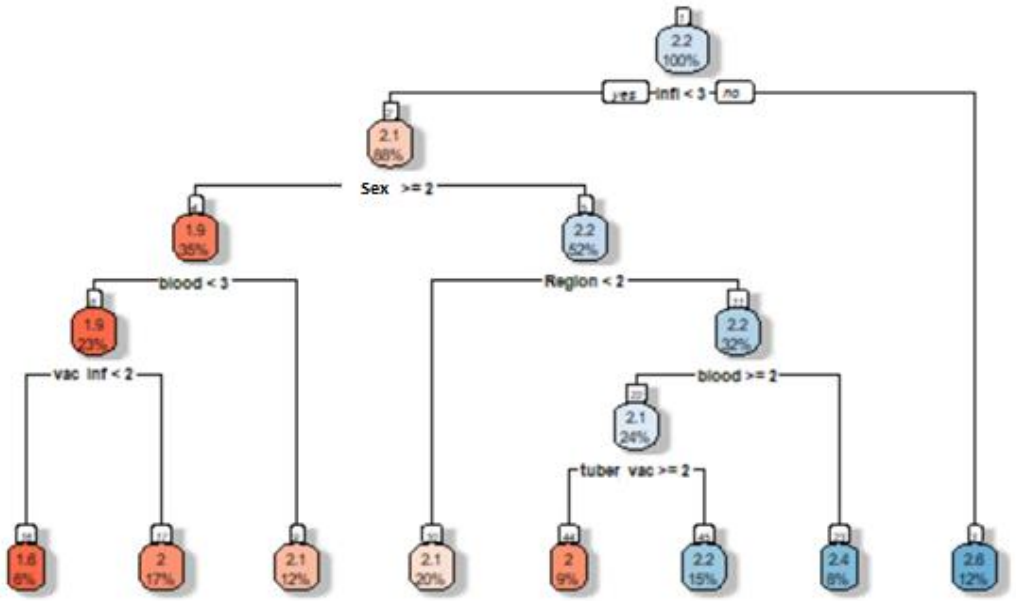


Рисунок 3.8. Візуалізація дерева рішень

Точність дорівнює 0,5135. Але ця модель дозволяє вибрати основні ознаки, такі як «Чи хворіли ви на грип цього року?», стать, група крові, регіон.

Окрім того, що вибір функцій на основі PCA та дерева рішень показує інший результат (рисунок 3.5 та рисунок 3.8), модель випадкового лісу розроблено на основі всіх ознак та ознак групування (рисунок 3.5). З трьох дерев будуються змінні, перевірені під час кожного розбиття. Середні квадрати залишків (MSR) враховують дисперсію фактичного значення цільової змінної та розрахункового значення цільової змінної, отриманого за допомогою лінійної регресії (таким чином, враховуючи середнє значення цільової змінної). MSR для всього набору даних і вибраних функцій дорівнює 0,5067292 і 0,5736409 відповідно. Таким чином, усі особливості враховуються для подальшого аналізу.

Out-of-bag measuring (OOB) — це помилка передбачення Випадкових лісів, розширених дерев рішень та інших моделей машинного навчання, які використовують пакетування для підвибірки зразків даних, які використовуються для навчання. OOB — це середня помилка передбачення для кожної навчальної вибірки x_i, використовуючи лише дерева, які не мали x_i у своїй початковій вибірці. Значення OOB дорівнює 16,61%. Матриця невідповідності наведена в таблиці 3.4.

Таблиця 3.4. Матриця невідповідності

	0	1	2	Помилка класу
0	153	9	17	0,14525139
1	8	88	12	0,18518518
2	1	3	20	0,16666666

Найбільша похибка – для 1 класу (COVID – так). Це можна пояснити різницею у представленості IgG та IgM (розкид даних від 0,00 до 18,00) у різних країнах.

Мінімальні значення глибини для всіх дерев у випадковому лісі наведено на рисунку 3.9.

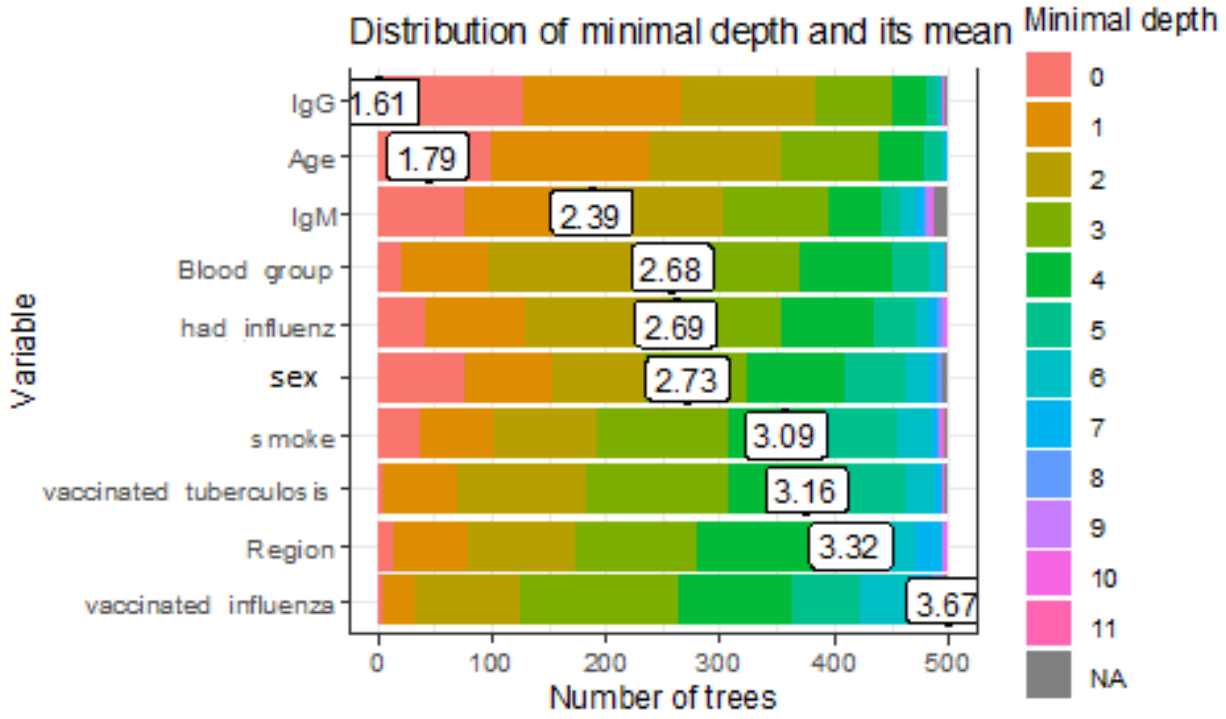


Рисунок 3.9. Розподіл мінімальної глибини розвинених дерев

Вісь X коливається від нуля дерев до максимальної кількості дерев. У кожному дереві будь-яка змінна використовувалася для розбиття на 500. Тому максимальна глибина створених дерев призначена для щепленого грипу. Перший рівень у більшості «поганих класифікаторів » представлений IgG.

Для подальшого вивчення мір змінної важливості ми передаємо ліс для вимірювання функції важливості та отримуємо наступний кадр даних (Таблиця 3.5). Вік і група крові є найчастішими коренями.

Таблиця 3.5. Вимірювання функції важливості

#	змінна	середня_мінімальна глибина	times_a_root
1	Вік	1,511688	112
2	Група крові	1,620000	111
3	стать	1,723688	53
4	Хворів на грип	1,727688	92
5	палить	2,030000	79
6	Щеплення від грипу	2,372752	25
7	Вакцинований туберкульоз	2,164000	28

На рисунку 3.10 представлено графік вибраних показників важливості змінних у лісі.

Кореляція між mean_min_depth і times_a_root знайдено. З цього факту ми робимо висновок, що атрибути Вік і Група крові є найбільш впливовими при аналізі захворюваності на COVID-19.

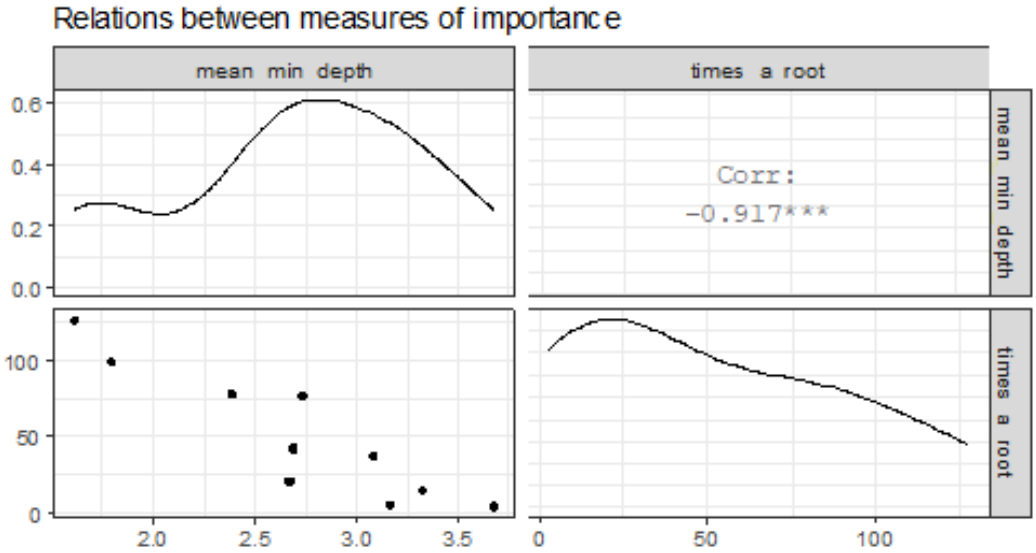


Рисунок 3.10. Співвідношення між мірами

Після вибору набору найважливіших змінних (таблиця 3.6) можна дослідити взаємодії, тобто розбиття, що з'являються в максимальних піддеревах відповідно до вибраних змінних. Щоб отримати назви 5 найважливіших змінних відповідно до середньої мінімальної глибини та кількості дерев, у яких з'явилася змінна, ми маємо наступний результат:

Таблиця 3.6. Цільові змінні.

#	Регіон
1	Вік
2	IgG
3	Група крові
4	Хворів на грип
5	IgM

Наівний Байєс показує щільність для кожної функції в наборі даних (рисунок 3.11). Точність наївного Байєса набагато менша за Випадкового лісу і дорівнює 67%. 9 візуалізує граничні ймовірності прогностичних змінних у даному класі (рисунок 3.11).

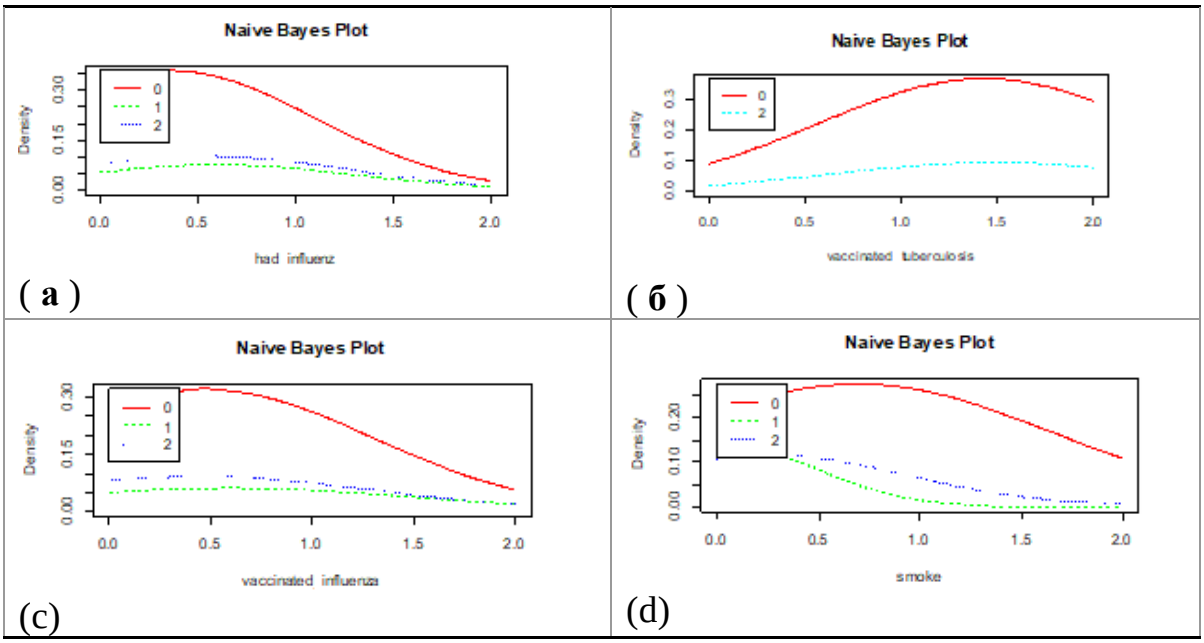


Рисунок 3.11. Графік щільності за параметрами: (a) хворіли на грип; (b) вакцинований туберкульоз; (c) щеплений грип; (г) паління

Наступним кроком є класифікація нейронної мережі. Архітектура нейронної мережі така:

- 1 прихований шар і 7 нейронів у прихованому шарі;
- зворотне поширення як алгоритм навчання;
- логістична активація.

Точність дорівнює 82%.

Наступні класифікатори також використовуються для класифікації COVID-19.

- Машина опорних векторів з лінійним ядром показує точність, рівну 60,5%;
- Логістична регресія для числових даних показує інформаційний критерій Akaike (AIC): 37,471. Точність дорівнює 55,3%.

На наступному етапі аналізу оцінюється кожен класифікатор для

- усього набору даних;
- набору даних по країнах;
- вибраних ознак;
- кожного кластеру окремо.

3.3.2 Кластерний аналіз

Методи кластеризації вимагають визначення відстані між примірниками. Ось чому one-hot-encoding використовується для категоріального перетворення даних у числові дані для кластеризації.

Запропоновано знайти кластери та використовувати їх для подальшого аналізу. Перший метод — це алгоритм k-середніх з 4 кластерами, оціненими за статистикою розривів [126]. Візуалізація k-середніх показує перетин між кластерами (рисунок 3.12). Це потребує подальшого аналізу.

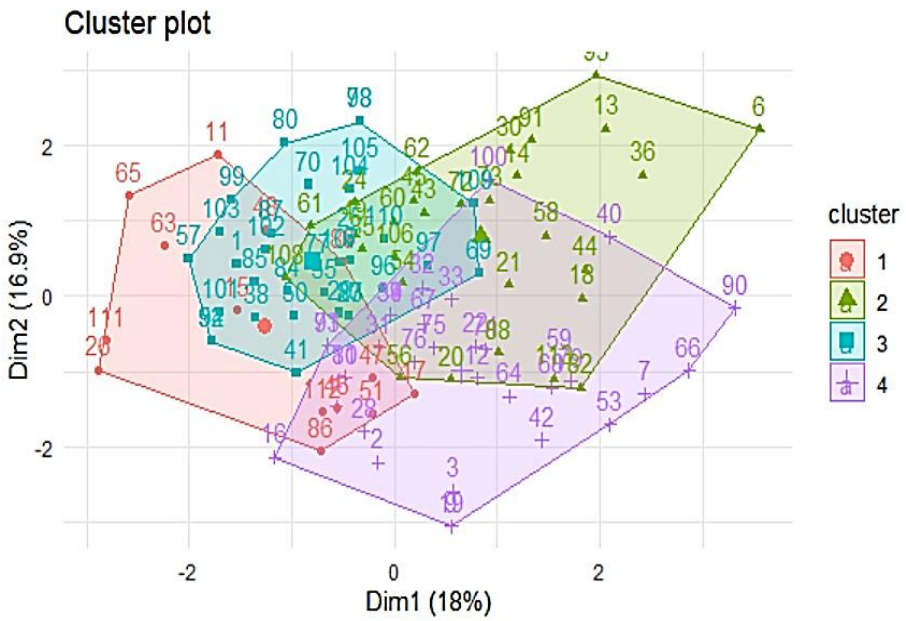
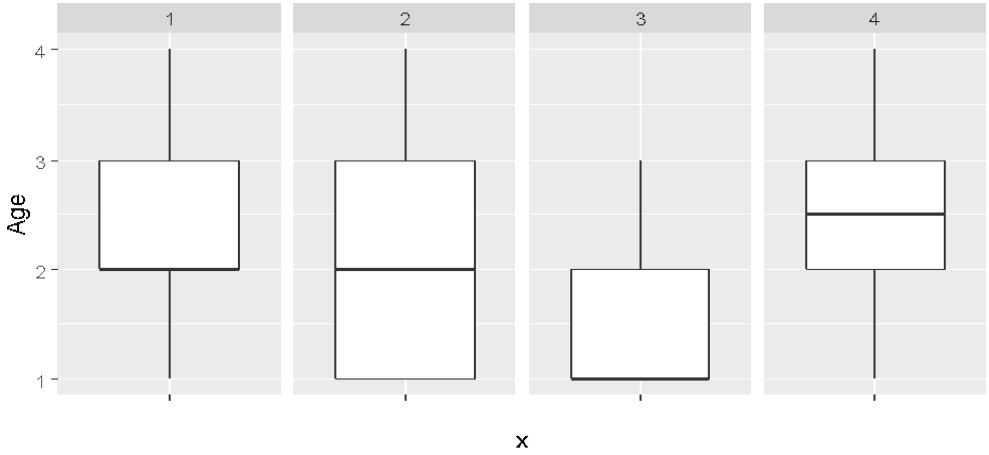


Рисунок 3.12. Візуалізація k-середніх

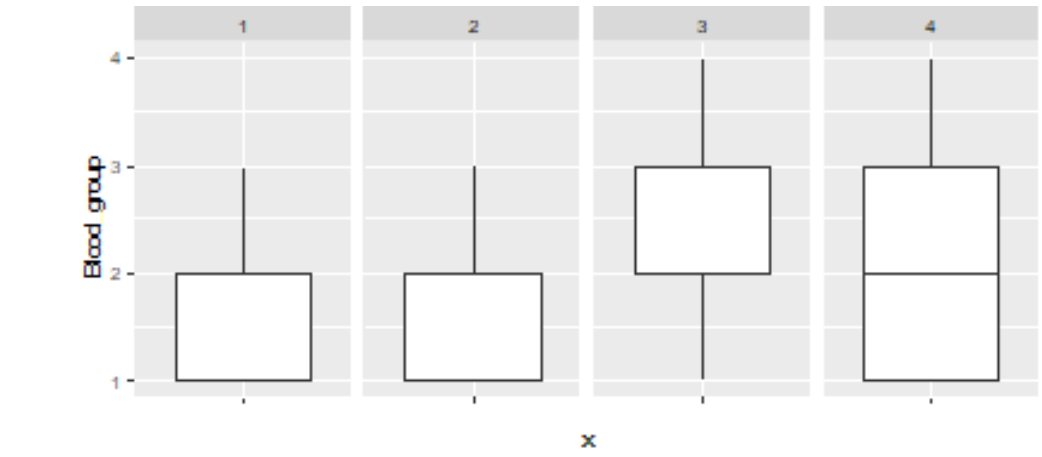
Проаналізовано тенденцію кластеризації. Статистика Хопкінса (H) [127] показує, що розподіл даних нерівномірний. Це означає, що дані підходять для кластеризації: H=0,5940309.

3.3.3 Аналіз кожного кластера

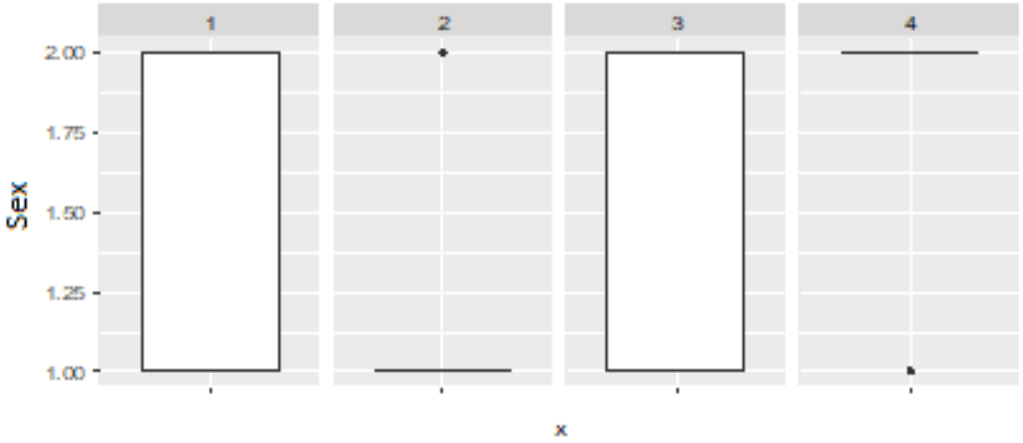
Наступним кроком є аналіз кожного кластера окремо (рисунок 3.13).



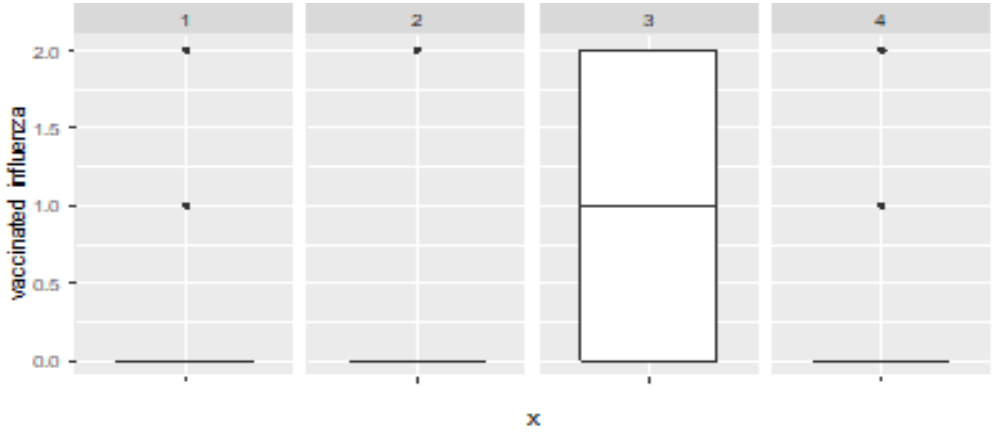
(a)



(б)



(г)



(д)

Рисунок 3.13. Розподіл об’єктів кластера: а – за віком; б) за групою крові; (г) за статтю ; (д) вакцинованим проти грипу.

Як бачимо, розподіли в різних кластерах повністю відрізняються один від одного: відрізняються не тільки медіани, але й розкид значень. Однак варто зазначити, що «діаграми коробок і вусів» найбільш інформативні, коли розподіл даних нормальний або близький до нормального. Кластер 2 складається лише з

чоловіків, а кластер 4 складається лише з жінок. Особи, вакциновані проти грипу, представлені лише в кластері 3.

Наведено розподіл об’єктів кластера за параметрами (рисунок 3.14). Кластер 3 має найбільшу кількість підтверджених випадків. Найпоширенішою є група крові 2. Цей факт підтверджує гіпотезу про те, що пацієнти з групою крові II більш вразливі до COVID-19 за згаданим набором даних. Найменша кількість підтверджених випадків наведена в кластері 4.

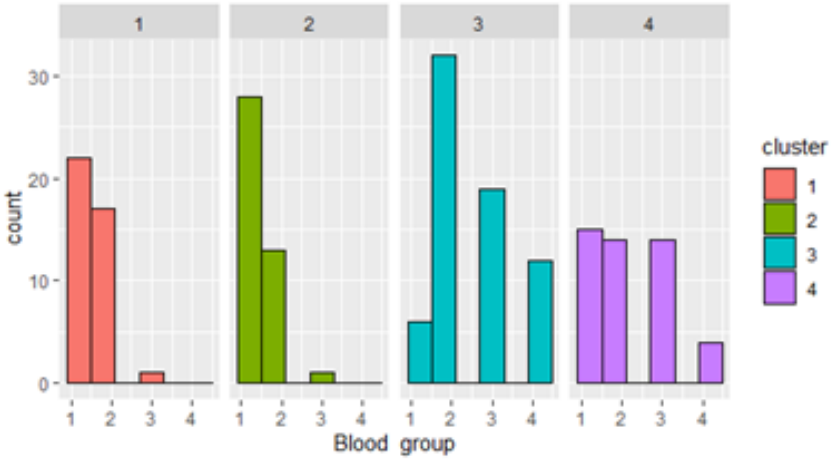


Рисунок 3.14. Розподіл об’єктів кластера за групами крові.

Візуалізація розподілу кластерів за групами крові показує викиди в кластерах 1 і 2 (група крові 3), у кластері 3 (група крові 1) і рівномірний розподіл осіб із групою крові 1-3 у кластері 4.

3.3.4 Розроблення методу двоетапної обробки даних

Важливість змінних різна для різних методів (таблиця 3.6). Це означає, що залежність між параметрами підтримується лише для частини набору даних. Тому ми пропонуємо знайти залежність для окремих кластерів і використовувати цю залежність для класифікації.

Результати точності моделей наведені в таблицях 3.7 і 3.8.

Таблиця 3.7. Точність моделей для цілих ознак

Модель	Повний набір даних	Відфільтровано Україною	Відфільтровано Білорусією	Відфільтровано Німеччиною	Кластер 1	Кластер 2	Кластер 3	Кластер 4
Логістична регресія	0,553	0,572	0,534	0,544	0,601	0,592	0,610	0,589
Машина опорних векторів	0,605	0,6327	0,570	0,584	0,621	0,694	0,635	0,637
Наївний Байєс	0,670	0,693	0,655	0,655	0,674	0,693	0,672	0,692
XGBoost	0,898	0,932	0,860	0,942	0,941	0,945	0,899	0,957
Випадковий Ліс	0,897	0,924	0,859	0,940	0,932	0,944	0,961	0,925
Нейронна мережа	0,820	0,849	0,828	0,79	0,830	0,849	0,8204	0,849
Дерево рішень	0,513	0,542	0,517	0,492	0,553	0,631	0,612	0,642

Таблиця 3.8. Точність моделей для обраних ознак

Модель*	Вік, IgG, Група_крові, перенесений_грип, IgM	Вік, стать, група_крові, перенесений_грип
Логістична регресія	0,633	0,671
Машина опорних векторів	0,671	0,722
Наївний Байєс	0,674	0,732
XGBoost	0,935	0,945
RandomForest	0,945	0,934
Нейронна мережа	0,832	0,845
Дерево рішень	0,553	0,631

Розроблено метод двоетапної обробки малих наборів даних, що базується на основі ієрархічного предиктора для підвищення точності процесу прогнозування та узагальнення результатів.

Перший етап — кластеризація;

Другий етап – побудова моделі класифікації для кожного відокремленого кластера.

Крім того, ієрархічний предиктор, побудований на основі ансамблю k-means і XGBoost, демонструє кращу точність для кластерів 1, 2 і 4. k-середніх разом із випадковим лісом не домінують над іншими моделями в кластері 3. Усі «погані» класифікатори показують кращу точність для окремих кластерів, ніж для всього набору даних.

Ієрархічний предиктор будується за такими етапами:

1. Кластеризація

1.1. За допомогою статистики розривів шукається кількість кластерів [224]:

$$Gap_n(k) = W_n(x_k) - x_k,\ W_n(x_k) = \frac{1}{B}\sum_b x_{kb}, \qquad (3.12)$$

де W_n – це центроїд, що є основою простої процедури визначення оптимальної кількості кластерів, B – довідкові набори даних, x – значення показника кластеру;

1.2. Застосування алгоритму k-means для знайденої кількості кластерів та пошуку найменшого сумарного квадратичного відхилення елементів кластерів від центрів цих кластерів, формула 3.13:

$$\min\left[\sum_{j=1}^{k}\sum_{i=1}^{N}\left\|x_i - cx_j\right\|^2\right],$$

$$Cost = \sum_{i=1}^{k}\sum_{j=1}^{N_i} d_{ij}d_N\left(x_j, cx_i\right), \qquad (3.13)$$

де k – число кластерів, cx – центри мас векторів.

2. Побудова моделі прогнозування/класифікації для кожного відокремленого кластера:

2.1. Пошук слабких предикторів:

$$\begin{cases} \sum_j w_{i,j} x_{i,j,n} = y_{i,n} \\ \sum_j w_{i,j} = 1 \end{cases}, \quad (3.14)$$

де w – вага оцінки слабкого предиктора, y – результат прогнозування.

2.2. Узагальнення результатів за допомогою зваженого жорсткого голосування:

$$S_i = w_{i,LR} p_{i,LR} + w_{i,SVM} p_{i,SVM} + w_{i,RF} p_{i,RF} + w_{i,XGBoost} p_{i,XGBoost} + w_{i,NN} p_{i,NN}. \quad (3.15)$$

Точність ієрархічного предиктора наведена в таблиці 3.9.

Таблиця 3.9. Точність ієрархічного класифікатора

Модель	Кластер 1	Кластер 2	Кластер 3	Кластер 4
Ієрархічний класифікатор	0,941	0,945	0,961	0,957

XGBoost і Випадковий ліс алгоритми також дають високу точність для моделі, заснованої на обраних ознаках, але меншу в порівнянні з ієрархічним класифікатором.

Таким чином, розроблений підхід полягає в ієрархічному класифікаторі та пошуку залежності між резистентністю до COVID-19 та оцінкою особливостей пацієнта. Набір даних був зібраний у різних країнах і станом на 29.10.2020 містить 313 спостережень. Одержано такі результати:

- зібрано набір даних із трьох країн (Україна, Німеччина та Білорусь), що дозволило провести більш глибокий аналіз та узагальнення;
- гіпотеза про те, що пацієнти з II групою крові більш вразливі до COVID-19;
- респонденти на які вплинули випадки COVID -19, були обрані на основі алгоритмів машинного навчання та порівняння їх результатів;
- розроблений ієрархічний класифікатор на основі комбінованого використання алгоритмів неконтрольованого та контрольованого машинного навчання забезпечує вищу точність на 4% у порівнянні з алгоритмами Random Forest та XGBoost;

- дозволило знайти поширену закономірність стійкості до COVID-19.

Новизна полягає в ієрархічному предикторі, заснованому на комбінованому використанні неконтрольованих і керованих алгоритмів машинного навчання. Оцінюються «слабкі» класифікатори на основі результатів k-середніх. Ієрархічний класифікатор побудований на основі k-середніх, випадкового лісу з 500 деревами та XGBoost. Класифікація для окремих кластерів дає нам на 4% вищу точність порівняно з аналізом набору даних. Розроблений підхід може бути використаний також для персоналізованої підтримки прийняття рішень в медицині в інших областях.

Гіпотеза про те, що пацієнти з групою крові II більш уразливі до COVID-19, підтверджується на основі зібраного набору даних. Цей факт може бути використаний у подальших дослідженнях.

Вибір характеристик дає нам змогу проаналізувати наступні ознаки, які найбільше впливають на COVID-19 : вік, стать, група крові, перенесений грип.

Розроблений шаблон резистентності пацієнта до COVID-19, що дозволяє точніше оцінювати нові випадки на основі традиційних моделей, таких як SSIR, SEIR, SARIMA тощо.

Розроблений підхід може бути використаний також для персоналізованої підтримки прийняття рішень в медицині в інших областях.

3.4 Стекінгова модель опрацювання великих наборів медичних даних

Розроблено метод групування моделей машинного навчання, як стекінгова модель опрацювання великих наборів медичних даних S_k на основі алгоритмів машинного навчання, які використовують Random Forest як метаалгоритм.

Математичне формулювання розробленої моделі наступне.

1. Випадково згенеровано із початкового набору даних K перехресних згорток, формула 3.16:

$$\{z_1^1, \dots z_B^1\}, \{z_1^2, \dots z_B^2\} \dots, \{z_1^K, \dots z_B^K\}, \tag{3.16}$$

де K – кількість піднаборів, B – розмір піднабору, z_b^l– спостереження l -го зразка.

2. Завдання полягає в тому, щоб навчити K незалежні слабкі предиктори

$$w_1(.), w_2(.), \dots, w_K(.),$$

та поєднати результати навчання за допомогою метамоделі mw, формула 3.17:

$$s_K(.) = mw\big(w_1(.) \times w_2(.), w_1(.) \times w_3(.), \dots, w_{K-1}(.) \times w_K(.)\big), \qquad (3.17)$$

де $w_i(.) \times w_j(.)$– результат попарного множення слабких предикторів.

3. Спотворені функції використовуються разом із навчальним набором даних у метамоделі. Ця комбінація дозволяє уникнути кореляції результатів слабких предикторів і збільшує узагальнення моделі.

Основним недоліком стекінгової моделі є те, що метаатрибути навчання та тестування різні. Метаатрибут у навчальній вибірці не є відповідями конкретного предиктора; він складається з частин, які є відповідями різних регресій (з різними коефіцієнтами). А метаатрибут на контрольній вибірці, взагалі, є відповіддю на зовсім іншу регресію, налаштовану на повне навчання. У класичному стекуванні можуть виникати ситуації, коли метаатрибут містить мало унікальних значень, але багато з цих значень не перетинаються під час навчання та тестування.

Розроблена модель стекінгу також поєднує лінійну регресію, k-найближчих сусідів, машина опорних векторів з радіальною базисною функцією, машина опорних векторів з лінійною функцією, як слабкі предиктори. Крім того, метаознаки деформуються за результатами попарного множення. Метаознаки є результатом навчання слабких предикторів.

Загальна схема розробленої нової моделі стекінгу наведена на рисунку 3.15

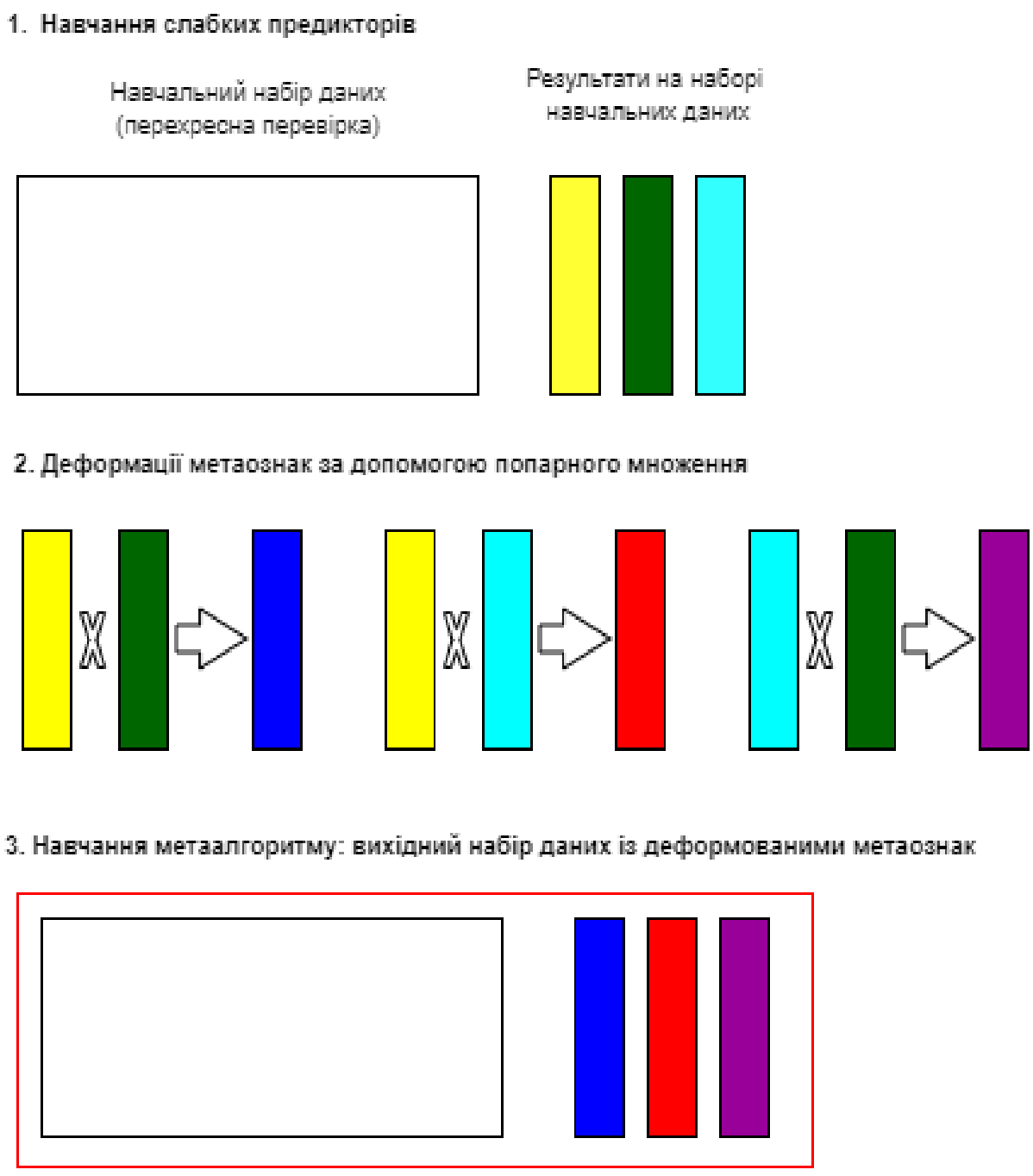


Рисунок 3.15 Схема стекування.

Висновки до 3 розділу

1. Проведено аналіз опрацювання медичних персоналізованих даних та досліджено процес прогнозування, та визначені етапи препроцесингу отриманих даних.
2. Досліджено особливості та принципи видобування даних задля пошуку попередньо невідомої закономірності і залежності між даними пацієнта, які можуть бути використані як валідна інформація, які можуть бути використані для побудови прогнозної моделі.
3. Здійснено класифікацію завдань видобутку даних, що орієнтовані на передбачення даних та формулювання рішень. Визначені інструменти для очистки, попередньої обробки, аналізу даних та виявлення взаємозалежностей між ними, що мають великий вплив на процеси опрацювання даних в системах діагностування та прогнозування станів пацієнта.
4. Запропоновані різні шляхи використання методів машинного та глибинного навчання для аналізу медичних даних пацієнта різної структури. Запропоновано рішення щодо вибору ефективних моделей ансамблів методів для пошуку залежностей між параметрами пацієнтів та факторами і показниками появи хвороби.
5. Розроблено метод двоетапної розробки на основі ієрархічного предиктора, що забезпечує підвищення точності процесу узагальнення результатів на малих наборах даних на 4 % порівняно з результатами логістичної регресії як кращого класифікатора для набору даних по COVID 19 та на 6% порівняно з результатами поліноміального методу опорних векторів як кращого класифікатора для набору даних по орфанних хворобах за рахунок ієрархічної класифікації та поєднання різних моделей машинного навчання.

6. Розроблено метод групування моделей машинного навчання, як стекінгову модель, для забезпечення точності прогнозування даних на 7-9 % та паралелізації процесу обробки даних, що на відміну від існуючих моделей убезпечує універсальність пошуку рішень щодо малих та великих наборів персоналізованих даних.

РОЗДІЛ 4.

АНАЛІЗ ТА ВИБІР ПРІОРІТЕТНИХ ОЗНАК ДЛЯ ВЕЛИКИХ НАБОРІВ ДАНИХ

У розділі розроблено модель вибору функцій гібридного ансамблю, що містить кілька селекторів за допомогою агрегування результатів, які будуть використовуватися на етапі попередньої обробки, вбудовані алгоритми виконують вибір ознак під час процедури навчання класифікатора, і вони оптимізують набір ознак, що використовуються для досягнення кращої точності. Розроблено ансамбль моделей машинного навчання для вибору пріорітетних ознак на великих наборах даних, який складається з класифікаторів, асоціативних правил та узагальненого рангу ознак на основі індексу Жакара, що дозволяє уникнути кореляції ознак та збільшує узагальнення моделі. Розроблено модель тришарового стекування ансамблю методів та описані етапи укладання, що забезпечує можливість об'єднати асоціативну класифікацію зі слабкими класифікаторами в ансамбль для узагальнення результатів. Модель класифікації ансамблю трирівневого стекування демонструє високу точність для інтелектуального аналізу коротких наборів медичних даних. Асоціативні правила разом зі слабкими предикторами покращують якість класифікації. Розроблений ансамбль використовує модель випадкового лісу, як агрегатор для узагальнення результатів слабких репресорів. Розроблена трирівнева класифікаційна модель ансамблю стекінгу з агрегатом логістичної регресії, яка має такі значення метрик оцінки продуктивності у вибраній підмножині ознак. Проведено порівняння точності прогнозування для стандартних моделей зменшення розмірності під час наступних кроків, для цього використано аналіз головних компонентів із вісьмома компонентами.

Матеріали розділу опубліковані у роботах автора [166, 176, 177, 178, 182, 183, 187, 191, 192, 198, 201, 203, 206, 209, 211, 212, 217].

4.1 Розробка моделі вибору ознак гібридного ансамблю

Методи виявлення залежностей у даних існують вже понад століття. Перші з них виникли як методи математико-статистичного аналізу (кореляційний і факторний аналіз). Комп’ютеризація процесів аналізу даних різко збільшила кількість аналізованих даних, якість аналізу та обсяг методів вилучення знань.

Вибір важливих ознак (наприклад, визначення генів, відповідних для певного типу раку) може допомогти розшифрувати механізми, що лежать в основі проблеми добробуту для дослідження. Повний перелік ознак можна здійснити методом селекції, тобто, перевіривши всі можливі набори, вибравши ті ознаки, для яких похибка мінімальна. Цей метод простий у реалізації, але він абсолютно неефективний для великих даних. Тому в даному випадку найчастіше використовуються інші алгоритми.

Використовуються три основні класи алгоритмів визначення ознак [136]:

- фільтри;
- обгортки;
- вбудовані алгоритми.

Фільтри базуються на деяких показниках, які не залежать від методу класифікації. Наприклад, співвідношення ознак з цільовим вектором і критеріями інформативності. Вони застосовуються перед класифікацією. Однією з переваг фільтрації є те, що її можна використовувати як попередню обробку, щоб зменшити розмірність простору та подолати надмірне оснащення. Методи фільтрації, як правило, швидкі. Фільтри використовуються для вибору ознак у кластеризації для побудови початкового наближення [139]. На жаль, такі методи не призначені для виявлення складних зв'язків між ознаками і як правило, недостатньо чутливі, щоб ідентифікувати всі залежності в даних.

Вбудовані алгоритми виконують вибір ознак під час процедури навчання класифікатора і оптимізують набір ознак, що використовуються для досягнення кращої точності [140]. Основна перевага вбудованих алгоритмів полягає в тому, що вони зазвичай знаходять рішення швидше, уникаючи повторного навчання даних з

нуля, усуваючи необхідність розділяти дані на навчальну та тестову підвибірки. Однак ці алгоритми не є універсальними.

Обгортки покладаються на інформацію про важливість функції з деяких методів класифікації або регресії, тому можуть знаходити глибші закономірності в даних, ніж фільтри. Обгортки можуть використовувати будь-який класифікатор, який визначає ступінь важливості ознак.

Базова лінія моделі вибору функцій гібридного ансамблю – це кілька селекторів за допомогою агрегування результатів. Кілька алгоритмів оболонки будуть використані на етапі попередньої обробки для першого етапу.

Кореляційна матриця показує числове значення коефіцієнта кореляції для всіх можливих комбінацій змінних. Він використовується в основному для з'ясування зв'язку між більш ніж двома змінними.

Дерево рішень повертає вагу функції як критерій для оцінки ознак. Це дозволяє побудувати ранжований список вибраних функцій з використанням різних показників. CART [123] використовувався для вибору ознак з індексом Джіні як мірою в нашому випадку.

Випадковий ліс [127] — це сукупність численних чутливих до навчання алгоритмів (дерев рішень). Ці алгоритми мають невелике зміщення. Зміщення методу навчання – це відхилення середнього відгуку навченого алгоритму від відгуку ідеального алгоритму. Кожен із цих класифікаторів побудовано на випадковій підмножині об’єктів і випадковій підмножині ознак.

Boruta — це евристичний алгоритм для вибору значущих ознак на основі використання випадкового лісу [141]. На кожній ітерації видаляються ті функції, для яких Z-міра менша за максимальну Z-міру серед доданих функцій. Щоб отримати Z-міру ознаки, необхідно обчислити її важливість, отриману за допомогою вбудованого алгоритму Випадкового лісу, і розділити її не стандартне відхилення важливості ознаки. Додані функції отримуються наступним чином: характеристики, наявні у вибраному, копіюються, а потім кожен новий атрибут заповнюється шляхом перемішування його значень. Ця процедура повторюється

кілька разів, щоб отримати статистично значущі результати, а змінні генеруються незалежно на кожній ітерації.

Індекс Жаккара вимірює подібність підмножин функцій, вибраних окремими селекторами функцій (кожен селектор організований як окрема ітерація):

$$(Sl_1, \dots, Sl_{1n}) = \frac{|Sl_1 \cap \dots \cap Sl_n|}{|Sl_1 \cup \dots \cup Sl_n|}, \quad (4.1)$$

де S_i є підмножиною ознак на i -й ітерації, для i =1,…, n. Значення індексу Жаккара змінюється від 0 до 1, де 1 означає абсолютну подібність підмножин.

Схема вибору ознак гібридного ансамблю наведена на рисунку 4.1

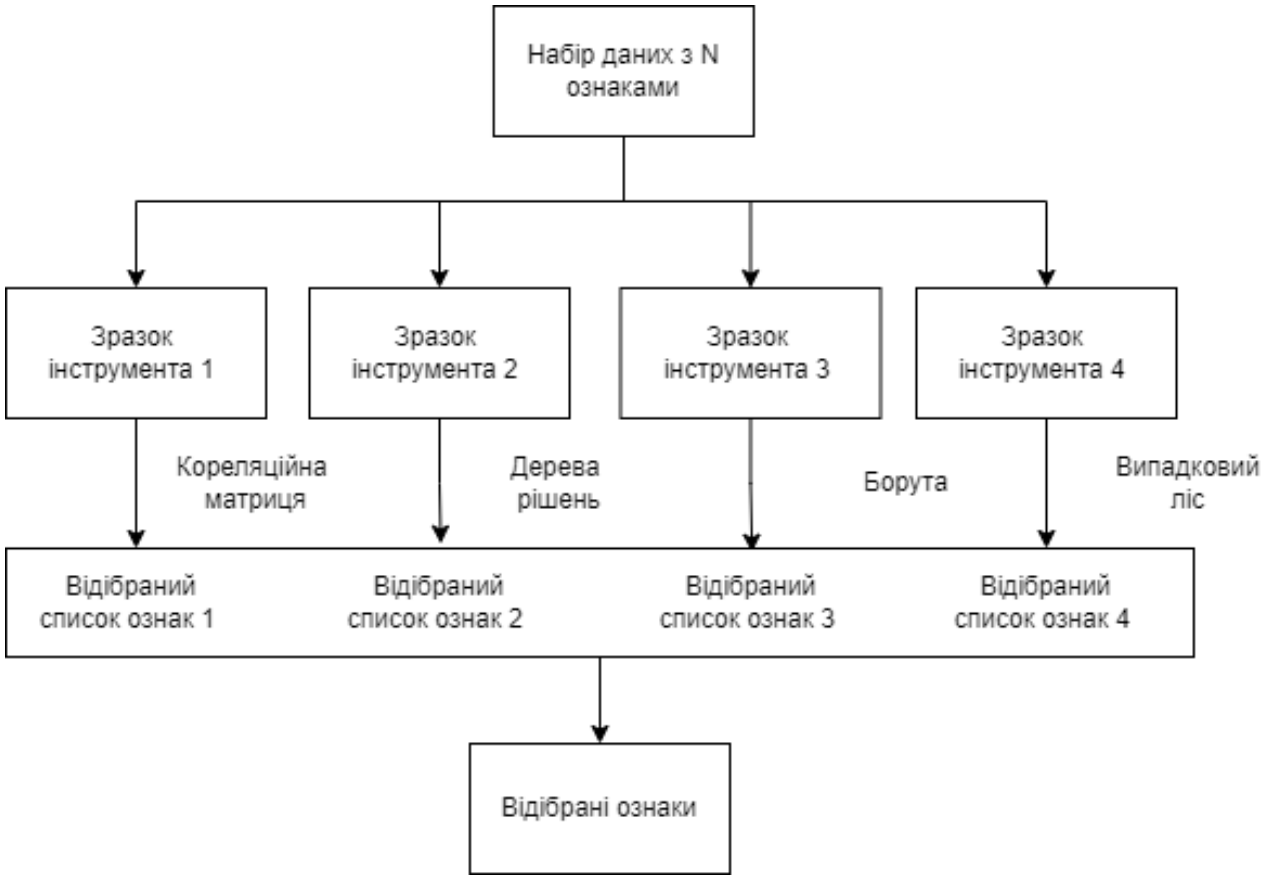


Рисунок 4.1. Модель вибору ознак гібридного ансамблю.

4.2 Розробка методу зменшення розмірності вхідних даних

У класифікаційних завданнях часто використовуються асоціативні правила [142,143]. Асоціативний інтелектуальний аналіз даних – це підхід до інтелектуального аналізу даних, який використовує методи виявлення правил асоціації для побудови систем класифікації. По-перше, правила асоціації генеруються з навчального набору даних із заданими пороговими значеннями підтримки та достовірності. Далі робиться прогноз для тестового набору даних і вимірюється точність класифікатора. Однак точність асоціативної класифікації багато в чому залежить від правил, які ми маємо перед класифікацією [144].

Розроблено метод зменшення розмірності вхідних даних, який за рахунок класифікаторів, асоціативних правил та узагальненого рангу ознак на основі індексу Жакара забезпечує підвищення точності вибору пріоритетних ознак на великих наборах даних, що передбачає тришарове стекування ансамблю моделей методів і дає можливість об'єднати асоціативну класифікацію зі слабкими класифікаторами в ансамбль для узагальнення результатів.

Базовий рівень розроблених трирівневих моделей класифікації складається з наступних кроків:

1. На першому рівні будуються асоціативні правила для видобутку прихованих залежностей. Використовується весь набір параметрів (ознак) хворого $X = \{X_1, X_2, \dots, X_p\}$, п. 2.3 .

   На цьому наборі використовуються правила мінімальної підтримки та достовірності MinSupp та MinConf, формула 2.32;
2. На другому рівні вибираються слабкі класифікатори для набору даних, що складається з важливих ознак $X_i = \langle x_{1i}, x_{2i}, \dots, x_{pi} \rangle$, п. 2.3:

$$C1: X -> X_1, C2: X -> X_2, \dots . Cn: X -> X_p \ ; \qquad (4.2)$$

3. На третьому етапі застосовується узагальнення рангу ознак на основі індексу Жакара, формула 4.1.

Схема класифікації ансамблю тришарового стекування наведена на рисунку 4.2

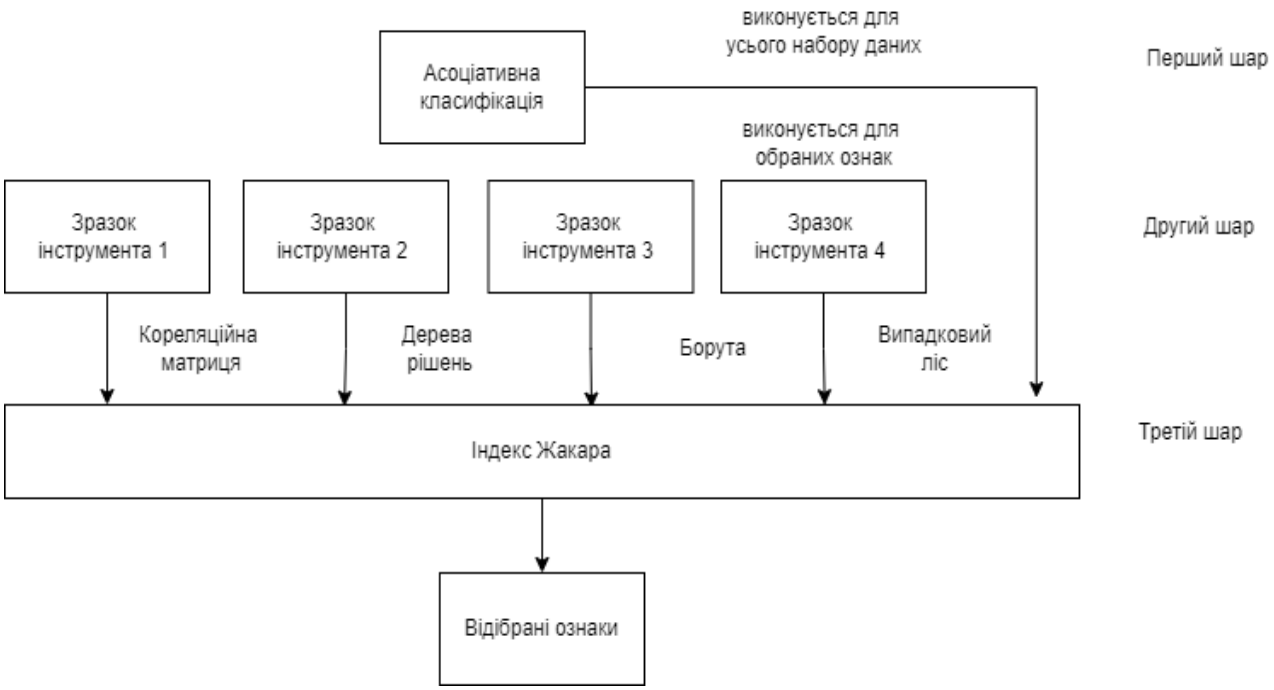


Рисунок 4.2. Тришаровий класифікатор ансамблю стекування.

Приклад 4.1.

Ми реалізували наш підхід у RStudio [145]. Основними пакетами, які ми використовували, були Caret, Rpart, Metrics, Boruta, RandomForest, Rules і Ggplot2 для візуалізації. Для створення асоціативних правил ми встановили мінімальний поріг підтримки в апріорному алгоритмі рівним 0,0001. Ми відфільтрували всі правила з достовірністю нижче 0,001.

По-перше, розроблена нова модель гібридного ансамблевого вибору функцій для системи прогнозування тривалості реабілітації після COVID-19 на основі машинного навчання реалізована як автоматичний ідентифікатор рангу відсікання функцій. Використовуються кореляційна матриця, дерево рішень, Випадковий ліс і Boruta. Результати селекторів функцій зібрані в таблиці 4.1.

Таблиця 4.1. Підсумок вибору ознак різними методами

Селектор функцій	Список функцій	Зважений список
Характеристики без високої кореляції	Вік ІМТ Пульс 6 хвилин тестової ходьби Sa02% Шкала Борга Силова ємність легень Обсяг форсованого видиху Об'єм пікового потоку при 25% (Vpeak25) Об'єм пікового потоку при 50% (Vpeak50) Об'єм пікового потоку при 75% (Vpeak75) CD16 Іл-8 ІЛ-10 CD4/CD8	ні
Дерево рішень (CART)	Силова ємність легень Обсяг форсованого видиху Vpeak25 Vpeak50 Vpeak75 CD16 Іл-8 CD4/CD8	так
Випадковий ліс	Силова ємність легень	так

(RandomForest)	Обсяг форсованого видиху Vpeak25 Vpeak50 CD16 CD4/CD8 Vpeak75	
Борута (Boruta)	Силова ємність легень Обсяг форсованого видиху Vpeak25 Vpeak50 Vpeak75 0-лімфоцити Іл-8	так

Для агрегації результатів використовується індекс Жаккара. На основі результатів, отриманих за допомогою різних селекторів функцій, створюється список важливих характеристик (Vpeak25 + Vpeak50 + Vpeak75 + Об'єм сили видиху + 0-лімфоцити + IL8 + CD4/CD8).

Далі для дискретизації даних використовується кластеризація. Визначено кількість кластерів за допомогою статистики розривів (рисунок 4.3).

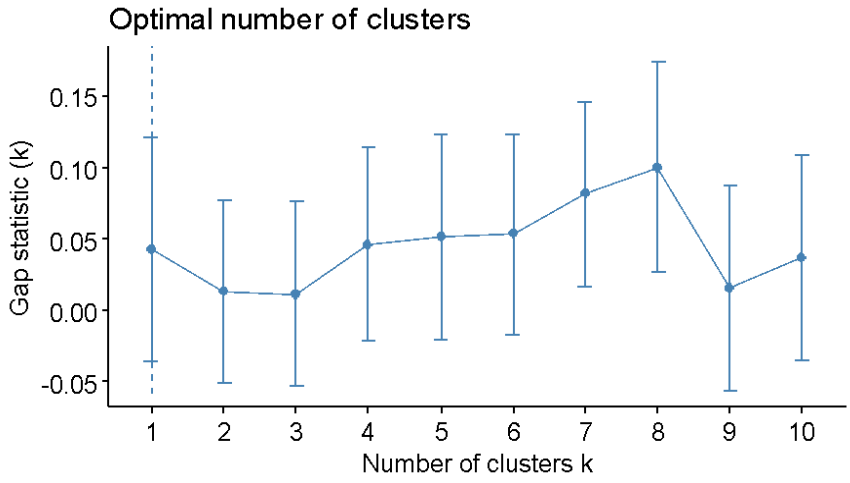


Рисунок 4.3 Оптимальна кількість кластерів за статистикою розривів.

Відповідно до рисунку 4.3 використовуються два кластери. Далі застосовується алгоритм k-середніх із двома кластерами. Результати показують, що майже всі об’єкти належать до одного класу. Такий же ефект досягається за допомогою самоорганізуючих карт (рисунок 4.4) [146]. Таким чином, прогнозні моделі були використані для всіх зразків.

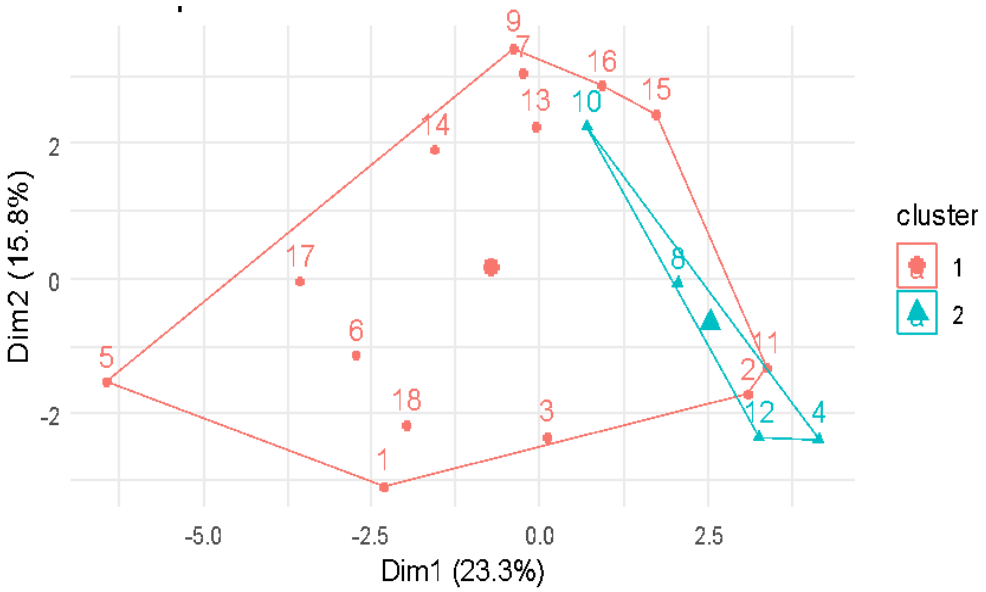


Рисунок 4.4. Кластерний графік для 2-х кластерного алгоритму k-середніх.

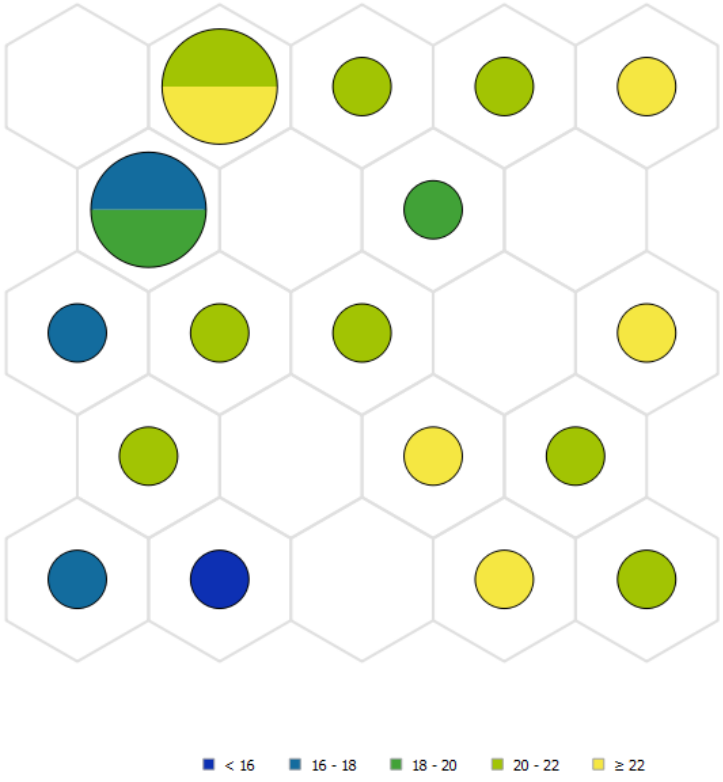


Рисунок 4.5. Теплова карта для самоорганізованої кластеризації карт.

По-друге. Виконується нормалізація за допомогою функції журналу. Основною метою нормалізації є приведення різноманітних даних у найрізноманітніших одиницях і діапазонах значень до єдиного вигляду, що дозволяє порівнювати їх між собою або використовувати для обчислення подібності об’єктів. Для прогнозування тривалості лікування використовуються різні моделі машинного навчання (використовується безперервна змінна). Випадкова вибірка; розмір навчального тесту 80%; повторний тренінг/тест=10. Використано десятикратну перехресну перевірку. Результати наведено в таблиці 4.2.

Таблиця 4.2. Прогноз тривалості реабілітації після COVID з використанням різних моделей ML

Предиктор	MAPE	RMSE
Лінійна регресія	0,0629	2,259
Дерево регресії	0,1220	0,155
Випадковий ліс	0,0191	0,071
k-найближчого сусіда	0,0189	0,071
Лінійне ядро МОВ	0,0163	0,076
Поліноміальне ядро МОВ	0,0054	0,016
ШНМ з 1 прихованим шаром, 12 нейронами, сигмовидною функцією активації	0,0096	0,034

Крім того, використовувалася п'ятикратна перехресна перевірка. Результати наведено в таблиці 4.3. З таблиць видно, що різниця між «ванільними моделями» та перехресною перевіркою відносно невелика. Це можна пояснити обмеженим розміром набору даних.

Таблиця 4.3. Прогноз тривалості реабілітації після COVID-19 з використанням різних моделей ML і перехресної перевірки

Предиктор	**MAPE**	**RMSE**
Лінійна регресія	0,0625	2,252
Дерево регресії	0,1218	0,152
Випадковий ліс	0,0185	0,064
k-найближчого сусіда	0,0184	0,064
Лінійне ядро МОВ	0,0159	0,072
Поліноміальне ядро МОВ	0,0050	0,012
ШНМ з 1 прихованим шаром, 12 нейронами, сигмовидною функцією активації	0,0093	0,030

Ті самі моделі використовуються в процесі прогнозування для вибраних функцій. Результати цього експерименту наведено в таблиці 4.4.

Таблиця 4.4. Прогноз тривалості реабілітації після COVID з використанням різних моделей ML з використанням лише вибраних функцій

Предиктор	**MAPE**	**RMSE**
Лінійна регресія	0,0276	0,112
Дерево регресії	0,0426	0,134
Випадковий ліс	0,0192	0,070
k- найближчого сусіда	0,0192	0,070
МОВ з лінійним ядром	0,0221	0,102
МОВ з поліноміальним ядром	0,0255	0,117
ШНМ з 1 прихованим шаром, 12 нейронами, сигмовидною функцією активації	0,0280	0,110

Як видно з таблиць 4.2 і 4.4, у деяких випадках ми отримали кращі результати, зокрема для лінійної регресії, дерева регресії, випадкового лісу та k-NN. Моделі SVM, і ANN дали гірші результати для вибраних функцій.

Біомаркери, що вказують на старіння, були обрані як підмножина ознак, і регресійний аналіз був застосований до цієї моделі. Біомаркери старіння включають ознаки CD3, CD22, CD4 і CD8. Результати прогнозування тривалості COVID-19 з використанням того самого набору моделей ML представлені в таблиці 4.5.

Таблиця 4.5. Прогноз тривалості реабілітації після COVID з використанням різних моделей ML на основі предикторів біомаркерів старіння

Предиктор	**MAPE**	**RMSE**
Лінійна регресія	0,1121	2,649
Дерево регресії	0,1279	2,967
Випадковий ліс	0,0823	1,822
k-найближчого сусіда	0,0823	1,822
МОВ з лінійним ядром	0,0787	2,402
МОВ з поліноміальним ядром	0,1006	2,739
ШНМ з 1 прихованим шаром, 12 нейронами, сигмовидною функцією активації	0,0958	2,336

Використаний аналіз головних компонент із вісьмома складовими для порівняння точності прогнозування зменшення розмірності під час наступних кроків. Пояснена дисперсія в цьому випадку становить 86 %. Результати прогнозування тривалості COVID-19 за допомогою моделей ML для всіх функцій і вибраних PCA наведено в таблиці 4.6.

Таблиця 4.6. Порівняння результатів прогнозування моделей ML з використанням усього набору функцій і восьми підмножин вибраних функцій PCA

Модель	Для всього набору даних			На вісім компонентів		
	MSE	MAE	R2	MSE	MAE	R2
k-НС	7,029	2,155	-0,162	5,781	1,825	0,044
МОВ	5,753	1,828	0,049	5,821	1,830	0,038
СГС	16,007	3,280	-1,647	12,328	2,603	-1,039
Лінійна регресія	17,826	3,571	-1,948	14,836	2,760	-1,453
БШП НМ	5,717	1,777	0,055	6,469	2,034	-0,070

З таблиці 4.6 можна побачити, що помилка класифікації зменшилася після застосування аналізу PCA. У цьому експерименті використовується багатошаровий персептрон (MLP NN) з такими параметрами: вісім нейронів прихованого шару, функція активації ReLu, вирішувач стохастичного градієнтного спуску, значення регуляризації (альфа) – 65.

Отже, обидві нейронні мережі з наведеною вище конфігурацією (12 нейронів + сигмоподібна функція та вісім нейронів + функція ReLu) дають другий за точністю результат і навіть кращий прогноз для повного набору даних. У нашому випадку вони менш чутливі до розмірності простору ознак. Але кількість нейронів у прихованому шарі дорівнює або перевищує 8, тобто кількість значущих компонентів, вибраних PCA, тому можливо, що мережа може передбачити весь набір даних з досить високою точністю.

По-третє. На наступному етапі розв’язуємо класифікаційну задачу. Для цього цільовий атрибут «Тривалість» перетворений у категоріальну змінну. Представлено задачу бінарної класифікації – короткі та тривалі реабілітаційні заняття. Набір даних є незбалансованим (90 довгих і 32 коротких), тому використовується техніка балансування. Набір даних був збалансований двома методами: випадковим відбором даних з більшого класу в кількості, що дорівнює кількості вибірок, які належать до меншого класу, і стратегією SMOTE, яка синтетично збільшує кількість вибірок меншого класу. Після балансування набір

даних складається з 62 екземплярів для довгого класу та 60 екземплярів для короткого класу.

Тепер для перетвореного набору даних використовується модель класифікації ансамблю трирівневого стекування. Спочатку використовуються асоціативні правила з алгоритмом Апріорі. Видобуті правила представлені в таблиці 4.7.

Таблиця 4.7. Асоціативні правила для класифікації тривалості реабілітації після COVID-19 згідно з алгоритмом Apriori

n/n	Набори	опорне значення
1	{CD4=[26,28)}	0,2222222
2	{ Vpeak25 =[94,100)}	0,2222222
3	{SaO2=[95,96)}	0,2222222
4	{Вік=[30,54)}	0,2777778
5	{Вік=[54,61)}	0,2777778
6	{Зріст=[161,168)}	0,2777778
7	{CD4=[26,28), CD4/CD8=[0,81,1,06)}	0,1111111
8	{6min_test_walk=[365,420), CD4=[26,28)}	0,1666667
9	{CD4=[26,28), CD8=[21,25)}	0,1111111
10	{Force_exhalation_volume =[100,105),CD4=[26,28)}	0,1111111
11	{CD4=[26,28), TNF-α =[11,7,27,3]}	0,1111111
12	{CD4=[26,28), IL-10=[3,7,7,83)}	0,1111111
13	{CD4=[26,28), IL-8=[43,8,98,1]}	0,1111111
14	{Вага=[59,75,7), CD4=[26,28)}	0,1111111

На другому шарі у розробленому ансамблі використовуються п'ять класифікаторів:

1. Дерево – алгоритм дерева з прямим скороченням;
2. Наївний класифікатор Байєса;
3. SVM з ядром RBF;
4. Логістична регресія;
5. Відкалібрований учень.

Відкалібрований учень створює модель, яка калібрує розподіл ймовірностей класу та оптимізує поріг прийняття рішення. Сигмоїдна функція була використана для калібрування ймовірності, тоді як оптимізація порогового значення рішення була застосована для оптимізації точності класифікації.

У таблиці 4.8 представлено середнє значення ефективності прогнозування за класами з використанням показників AUC, CA, F1, Precision і Recall. Навпаки, таблиця 10 містить ті самі показники для моделі класифікації на основі восьми ознак, вибраних PCA.

Таблиця 4.8. Ефективність категоріальної класифікації тривалості COVID-19 за п’ятьма класифікаторами ML і моделлю класифікації ансамблю трирівневого стекування, застосованої до всього набору ознак

Модель	AUC	ACC	F1	Точність	Відкликання
Дерево рішень	0,854	0,760	0,762	0,766	0,760
МОВ	0,988	0,910	0,921	0,924	0,920
Наївний Байєс	0,957	0,860	0,861	0,869	0,860
Відкалібрований учень	0,917	0,920	0,921	0,933	0,920
Логістична регресія	0,898	0,800	0,800	0,867	0,800
Модель класифікації ансамблю тришарового стекування з агрегатом випадкового лісу	0,992	0,930	0,960	0,964	0,960

Таблиці 4.7 і 4.8 показують деяке зниження точності прогнозу у випадку моделі класифікації на основі восьми обраних ознак. Цей результат потребує подальшого дослідження, і поки що його можна пояснити невеликим розміром набору даних, складним впливом біомаркерів на тривалість COVID-19 (що підтверджується низькою ефективністю класифікації наївного класифікатора Байєса), наявністю деяких особливостей/ супутні захворювання, не включені в поточний набір даних, або чутливість класифікаторів до розмірності простору ознак.

Таблиця 4.9. Ефективність категоріальної класифікації тривалості COVID-19 за п’ятьма класифікаторами ML і моделлю класифікації ансамблю трирівневого стекування, застосованої до вибраної підмножини ознак

Модель	AUC	ACC	F1	Точність	Відкликання
Дерево рішень	0,781	0,720	0,723	0,735	0,720
МОВ	0,908	0,840	0,842	0,847	0,840
Наївний Байєс	0,883	0,860	0,861	0,869	0,860
Відкалібрований учень	0,888	0,860	0,861	0,896	0,860
Логістична регресія	0,880	0,840	0,841	0,886	0,840
Модель класифікації ансамблю тришарового стекування з агрегатом випадкового лісу	0,978	0,920	0,921	0,924	0,920

У таблиці 4.10 наведено точність прогнозування цільового класу «Довгий» для згаданих п’яти класифікаторів ML і трирівневого класифікатора ансамблю стекування з агрегатом Випадкового лісу для моделі класифікації на основі восьми вибраних функцій PCA.

Таблиця 4.10. Ефективність класифікації цільового класу COVID-19 «Long» за п’ятьма класифікаторами ML і моделлю класифікації ансамблю трирівневого стекування, застосованої до вибраної підмножини функцій

Модель	AUC	ACC	F1	Точність	Відкликання
Дерево рішень	0,792	0,720	0,750	0,808	0,700
МОВ	0,922	0,860	0,846	0,958	0,821
Наївний Байєс	0,883	0,860	0,877	0,926	0,833
Відкалібрований учень	0,867	0,860	0,868	0,999	0,767
Логістична регресія	0,867	0,840	0,846	0,999	0,733
Модель класифікації ансамблю тришарового стекування з агрегатом випадкового лісу	0,967	0,900	0,909	0,999	0,833

На рисунку 4.6 показано, що в цілому, використовуючи існуючий набір даних, ефективність прогнозування цільового класу «Довгий» дещо вища, ніж для середнього за класами. З іншого боку, з таблиці можна помітити, що метрика Precision (Точність) є вищою для всіх використаних класифікаторів, тоді як Recall (Відкликання) має нижчі значення. Це зумовлено тим, що така модель більш точна в класифікації позитивних елементів. У той же час це гірше для класифікації справжніх позитивних результатів серед усіх позитивних випадків у наборі даних. З іншого боку, зважене гармонійне середнє значення точності та запам’ятовування (метрика F1) є трохи вищим у випадку класифікації цільового класу «Long». Отже, ми можемо зробити висновок, що вища точність компенсує нижчу цінність запам’ятовування для цього експерименту.

Модель класифікації ансамблю трирівневого стекування з агрегатом випадкового лісу показує кращий результат у порівнянні з деревом рішень, SVM та іншими класифікаторами.

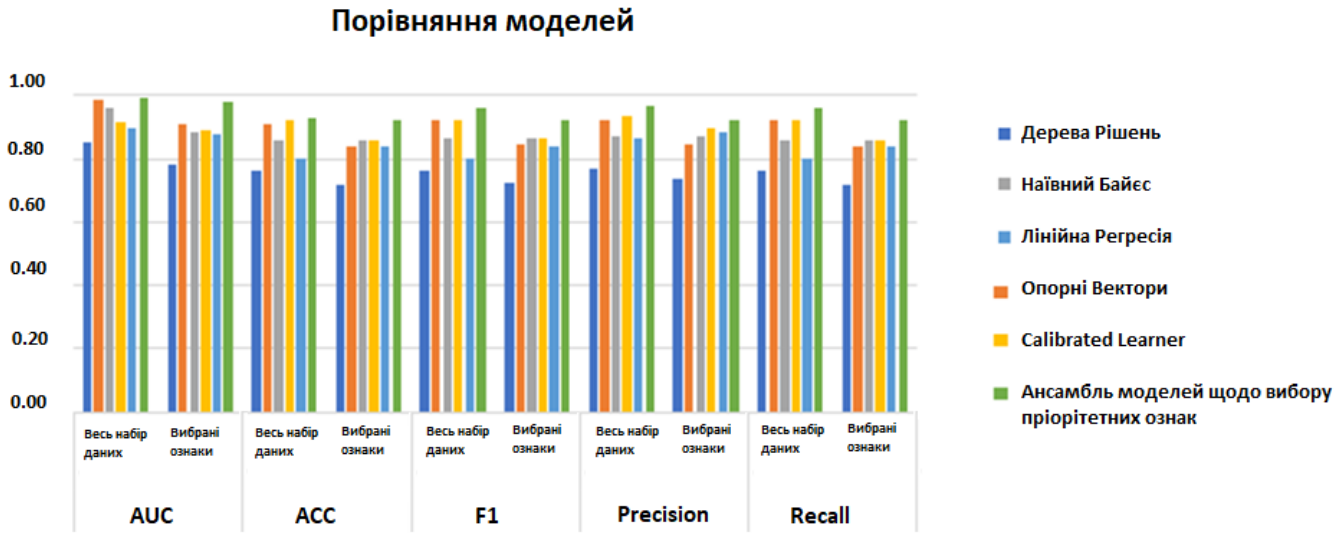


Рисунок 4.6. Порівняння моделей.

Нова модель гібридного ансамблевого вибору функцій для системи прогнозування після COVID-19 на основі машинного навчання пропонується як автоматичний ідентифікатор рангу відсікання функцій.

Знайдено асоціативні правила разом із використанням слабких предикторів для покращення якості класифікації.

За останні два роки хвороба SARS-CoV-2 стала одним із найбільших тягарів для глобальної системи охорони здоров'я, спричиняючи летальні випадки та значно виснажуючи ресурси охорони здоров'я. Захворювання може протікати у різних пацієнтів по-різному: від безсимптомного перебігу до госпіталізації у відділення інтенсивної терапії або навіть летального результату. Питання про те, від чого залежить тяжкість захворювання, цікавить дослідників усього світу. Однак остаточної відповіді на це питання досі немає. Загальновизнано, що тяжкість захворювання SARS-CoV-2 залежить як від супутньої патології, так і від стану імунної системи пацієнта, що відображається в кількох біомаркерах, які можна отримати за допомогою біохімічних лабораторних досліджень.

Методи штучного інтелекту, зокрема машинне навчання, можуть допомогти в проведенні таких клінічних випробувань, оскільки вони можуть працювати з меншими вибірками даних при використанні відповідних підходів. Дослідження з меншою вибіркою даних у цьому випадку дає змогу скоротити час збору набору

даних, що зменшить кількість клінічних пацієнтів, дозволить швидше використовувати прогнозні результати, і проводити профілактичні заходи для пацієнтів із потенційно тяжкими захворюваннями. Це, у свою чергу, може зменшити навантаження на систему охорони здоров'я та підвищити ефективність лікування та реабілітації хворих із важкими захворюваннями.

На основі набору даних, зібраного у Львівському обласному реабілітаційному центрі, що містить анонімну інформацію щодо імунного профілю та інших важливих діагностичних показників, у тому числі функції зовнішнього дихання та насичення киснем, побудовано модель класифікації тривалості реабілітації після COVID з використанням ансамблю методів машинного навчання. Розроблено нову модель вибору ознак гібридного ансамблю та модель класифікації ансамблю тришарового стекування. Розроблена гібридна ансамблева модель вибору функцій для системи прогнозування після COVID-19 на основі машинного навчання може використовуватися як автоматичний ідентифікатор рангу відсікання функцій. Модель класифікації ансамблю трирівневого стекування демонструє високу точність для інтелектуального аналізу коротких наборів медичних даних. Асоціативні правила разом зі слабкими предикторами покращують якість класифікації. Розроблений ансамбль використовує модель випадкового лісу як агрегатор для узагальнення результатів слабких регресорів.

Розроблена трирівнева класифікаційна модель ансамблю стекінгу з агрегатом логістичної регресії має такі значення метрик оцінки продуктивності у вибраній підмножині ознак: площа під кривою ROC (AUC) – 0,908; точність класифікації (ACC) – 0,840; оцінка F1 – 0,842; точність (Precision) – 0,867; відкликання (Recall) – 0,840. Отже, розроблена модель досягла хорошого прогнозу тяжкості захворювання SARS-CoV-2 на основі цитокінів і фізіологічних біомаркерів. Результати вказують на те, що зміни в досліджуваних біомаркерах, пов'язаних із тяжкістю захворювання, можуть бути використані для моніторингу тяжкості та прогнозування тривалості реабілітації.

Основні результати практичної імплементації:

- тривалість лікування COVID-19, як у днях, так і в категоріях (тобто довгих і коротких), прогнозується на основі важливих біомаркерів, отриманих шляхом профілювання цитокінів крові за допомогою підходу машинного навчання;
- нова модель гібридного ансамблевого вибору функцій для системи прогнозування після COVID-19 на основі машинного навчання пропонується як автоматичний ідентифікатор рангу відсікання функцій;
- асоціативні правила разом зі слабкими предикторами покращують якість класифікації;
- розроблена модель класифікації ансамблю тришарового стекування використовує модель випадкового лісу як агрегатор для слабкого узагальнення результатів регресора;
- біомаркери, пов'язані зі старінням, а саме: $CD3^{+,}$ $CD4^{+,}$ $CD8^{+,}$ $CD22^{+}$ досліджені для прогнозування тривалості реабілітації після COVID;
- дослідження пропонує прогностичні атрибути, які можна використовувати для моніторингу тяжкості захворювання та прогнозування тривалості реабілітації.

Подальші дослідження зосереджені на збільшенні вибірки даних шляхом додавання інформації про нових пацієнтів, вдосконалення моделі розширеного набору даних і застосування підходів до аналізу малих даних, таких як подвоєння вхідних даних за допомогою нелінійних ядер [147] або подвоєння вхідних даних на основі RBF [148] для покращення точність і надійність прогнозу.

Висновки до 4 розділу

1. Розроблено модель вибору функцій гібридного ансамблю, що містить кілька селекторів за допомогою агрегування результатів, які будуть використовуватися на етапі попередньої обробки, в якій вбудовані алгоритми виконують вибір ознак під час процедури навчання класифікатора, і вони оптимізують набір ознак, що використовуються для досягнення кращої точності.
2. Розроблено ансамбль моделей машинного навчання для вибору пріорітетних ознак на великих наборах даних, який складається з класифікаторів, асоціативних правил та узагальненого рангу ознак на основі індексу Жакара, що дозволяє уникнути кореляції ознак та збільшує узагальнення моделі.
3. Розроблено модель тришарового стекування ансамблю методів, що дає можливість об’єднати асоціативну класифікацію зі слабкими класифікаторами в ансамбль для узагальнення результатів; описані етапи укладання.
4. Модель класифікації ансамблю трирівневого стекування демонструє високу точність для інтелектуального аналізу коротких наборів медичних даних. Асоціативні правила разом зі слабкими предикторами покращують якість класифікації. Розроблений ансамбль використовує модель випадкового лісу як агрегатор для узагальнення результатів слабких репресорів.
5. Розроблена трирівнева класифікаційна модель ансамблю стекінгу з агрегатом логістичної регресії має такі значення метрик оцінки продуктивності у вибраній підмножині ознак.
6. Проведено порівняння точності прогнозування для стандартних моделей зменшення розмірності під час наступних кроків, для цього використовано аналіз головних компонентів із вісьмома компонентами.
7. Розроблено метод зменшення розмірності вхідних даних на основі ансамблю моделей слабких предикторів для вибору найважливіших ознак, що базується на прирості інформації з урахуванням результатів застосування декількох

селекторів та агрегації кінцевих результатів на підставі урахування індекса Жакара для оцінки подібності одержаних підмножин ознак та проведення мажоритарного голосування результатів, що дає змогу уникнути кореляції результатів слабких предикторів і збільшує узагальнення моделі.

РОЗДІЛ 5.

АНАЛІЗ АНОМАЛІЙ ПРИ ОПРАЦЮВАННІ ПЕРСОНАЛІЗОВАНИХ МЕДИЧНИХ ДАНИХ

У розділі запропоновано формулювання аномалій, де означено, що визначення аномалій залежить від контексту, де найважливішим значенням необхідно враховувати показник часу, тому що саме від нього і буде залежати контекст; визначені типи аномалій та шляхи їхнього пошуку, що дає можливість диференціації підходів щодо аналізу персоналізованих медичних даних та визначені особливості їхнього застосування; визначені основні проблеми, які виникають при обробці медичних персоналізованих даних, за рахунок чого постає проблема розробки ефективної стратегії аналізу даних, яка може бути застосована до розосереджених баз даних різних доменів; розроблено метод заповнення відсутніх даних на основі ймовірних продукційних залежностей, що збільшує стійкість моделі до помилок у даних та забезпечення аналізу мультимодальних даних при паралельній реалізації в розподілених базах даних; розроблено два алгоритми PPD майнінгу та заповнення відсутніх медичних персоналізованих даних, що забезпечують збереження характеристик стійкості до помилок у даних, можливість паралельної реалізації в розподілених базах даних, автоматизація та виконання аналізу різних типів даних.

Матеріали розділу опубліковані у роботах автора [167, 179, 180, 182, 184, 188, 189, 190, 195, 196, 204, 205, 206, 210, 213, 214, 215].

5.1 Виявлення аномалій у поточних медичних персоналізованих даних

Аномалія – це невідповідність прогнозованій поведінці системи. Це означає, що контекст вирішує, чи є те чи інше значення ненормальним, чи ні. З цієї причини показник часу необхідно розглядати як найважливіше значення, оскільки від нього залежить контекст. Не плутати зі звичайним виявленням аномалії, де порядок значень не має значення.

Коли системи моніторингу містять занадто багато показників, відстеження аномалій стає складним або навіть неможливим. У таких випадках зазвичай перевіряється, чи досягає показник певного критичного значення. Але така перевірка підходить не для всіх типів даних.

Існують різні види аномалій, а саме:

1. Викид – цей тип аномалії представляє значення, яке є занадто великим або занадто малим порівняно з нормальним розподілом даних у послідовності. Ці типи аномалій можна легко ідентифікувати за допомогою правила трьох сигм або міжквартального коливання, машинне навчання не потрібне, лише проста статистика;
2. Зсув – ця аномалія є різкий стрибок вгору або вниз середнього значення послідовності без зміни моделі поведінки;
3. Зміна шаблону – цей тип аномалії характеризується тим, що регулярна сезонність процесу порушується, і тепер часові послідовності, які ми спостерігали досі, не можуть надати суттєвої інформації про те, як процес поводитиметься в майбутньому, що призведе до цього може вказувати необхідність перенавчання моделі;
4. Відхилення моделі – цей тип аномалії не призводить до зміни сезонності, на відміну від зміни моделі, але є розрив у порядку, коли значення не повторюють модель, що спостерігалася в минулому;
5. Зміна характеру розподілу – ця аномалія означає, що дані показали більшу або меншу мінливість;

6. Поширені аномалії – цей тип аномалії виникає, коли значення двох спостережуваних показників знаходиться в межах нормального діапазону, але їх одночасна поява є ознакою ненормальної поведінки. Цю задачу відмінно вирішує кластерний аналіз, де точки, що не входять у жоден з кластерів, визначаються або як аномалія, або як шум.

Система виявлення аномалій повинна мати можливість виявляти кожен із заданих типів аномалій та їх комбінацію. Крім того, вона повинна покращуватися під час зміни, коли поведінка змінюється за часовим показником, і для цього необхідно як продовжувати вивчати нові послідовності та забувати старі.

Результатів може бути декілька, які залежатимуть один від одного. Така поведінка одного конкурента може визначити, як поводитиметься інший. Без цієї інформації неможливо точно сказати, аномалія це чи ні, і для цієї системи необхідно вводити не один, а багато наборів на вході та впорядковувати закономірності часу і між ними. Для вирішення всіх цих проблем однієї статистики недостатньо, тому пропонуються існуючі підходи для вирішення проблеми аномалій:

- статистичні методи;
- однокласовий SVM;
- кластеризація (k-середні);
- автокодери.

Не всі алгоритми були перераховані, але усі їх комбіновані рішення побудовані на основі нейронних мереж майже у всіх задачах.

5.1.1 Визначення аномалій за рахунок аналізу вхідних медичних персоналізованих даних

Виявлення аномалій буде відбуватись у метричних даних, зібраних обчислювальною системою, такі як завантаженість центрального процесора, оперативної пам’яті, мережі, операції запису та зчитування з диску і т.д. Ці дані – це потік метрик, що залежить від часу, тобто для центрального процесора у нас буде: [значення загруженості – точна в часі] і ці значення будуть утворювати

послідовність, де на осі x буде час, а на осі y – значення. Ці два значення і будуть входом для моделі, оскільки вони обоє важливі для виявлення аномалій.

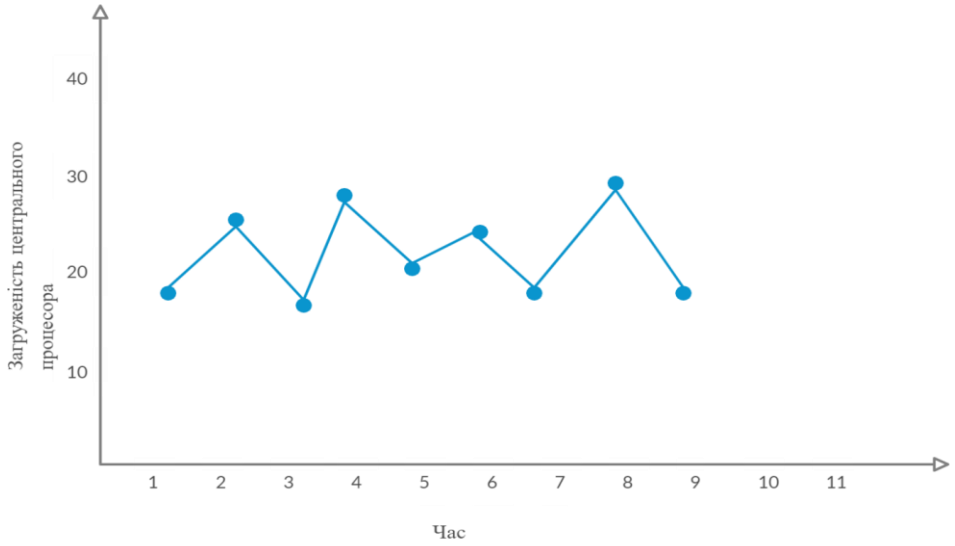


Рисунок 5.1 Формат часових рядів

5.1.2 Визначення аномалій за рахунок аналізу вихідних цільових даних

Розроблена модель для визначення аномалій повинна отримати на вхід часову послідовність. Часова послідовність складатиметься з точок і для кожної із цих точок буде робитись передбачення наступної точки, яка буде йти після цієї. Це значення, що модель передбачила буде потім порівнюватись із фактичним значенням з ціллю визначення наскільки точною є модель. Також ці значення будуть використовуватись для обчислення значення аномальності, воно розміщене в межах від 0 до 1. Чим більш відмінне фактичне значення від прогнозованого, тим вище значення аномалії та набору.

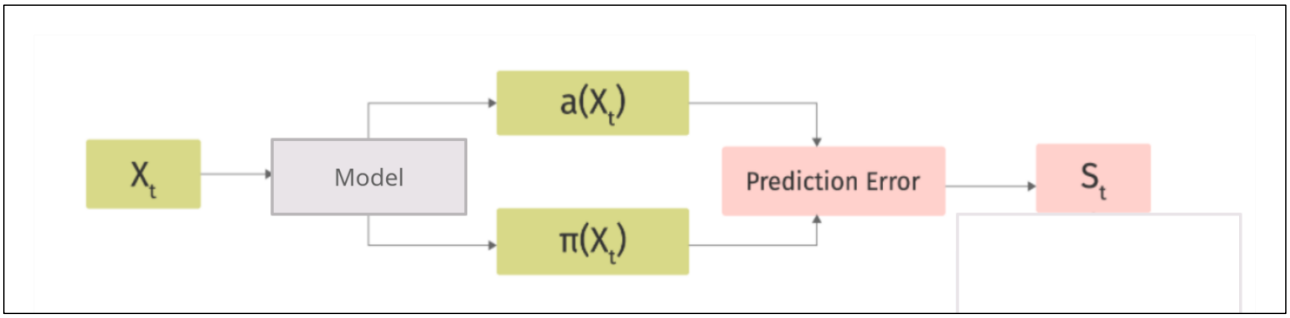


Рисунок 5.2 Схема отримання ймовірності аномалій

5.2 Метод заповнення відсутніх даних розроблений на основі інтеграції рішень

Основними проблемами, що виникають при обробці медичних персоналізованих даних, є:

- відсутність аналітичних методів, придатних для використання в кількох предметних областях;
- потреба в значних людських ресурсах для підтримки процесу дослідження даних;
- висока обчислювальна складність існуючих алгоритмів дослідження та швидке зростання складених даних.

До постійного збільшення часу дослідження, навіть при частому оновленні серверного обладнання та необхідності роботи з розподіленими базами даних, потужності яких більшість існуючих методів аналізу даних використовуються неефективно. Таким чином, постає проблема розробки ефективної стратегії аналізу даних, яка може бути застосована до розосереджених баз даних різних доменів.

Методи аналізу даних ефективно використовуються на чітких і попередньо оброблених даних. Тому робота спрямована на аналіз та розробку методів попередньої обробки даних. Зокрема, взяли до уваги вибір функцій (розробку функцій) та заповнення відсутніх даних.

Вибір відповідних функцій може бути більш важливим завданням, ніж скорочення часу обчислень або підвищення класифікації чи точності прогнозування. Наприклад, у медицині [16] пошук оптимального набору оптимальних ознак для класифікаційної або прогнозованої проблеми може допомогти розробити діагностичний тест.

Перед застосуванням етапу заповнення даних необхідно реалізувати аналіз цільових ознак наборів даних. Ця частина дослідження відображена у результатах застосування моделі вибору функцій гібридного ансамблю розглянута у розділі 4.1.

Надалі розробляється метод заповнення пропусків даних. Цей метод базується на класичних функціональних залежностях у реляційних базах даних і правилах асоціації з нереляційних баз даних. Він складається з двох частин:

1. Видобуток ймовірнісних виробничих залежностей;
2. Використання ймовірнісних виробничих залежностей для заповнення відсутніх даних.

Дослідження великих даних вимагає визначення кластерів атрибутів, які утворюють функціональні залежності (FD). Однак у реальному світі набори даних широко стандартизовані, з основними залежностями, визначеними лише на підмножині значень важливих груп атрибутів. Крім того, залежність може виникати не лише для кортежів у реляційних джерелах даних, але й між підмножинами значень у різних кортежах. Ми назвемо їх ймовірнісною продукційною залежністю (PPD).

Ймовірнісна продукційна залежність — це залежність, схожа на асоціативне правило у виборі первинного співвідношення, що є правильним для багатьох об'єктів [147], формула 5.1. Поріг значущості повинен бути визначений експертно або на основі оцінки ймовірності помилкового вибору цієї залежності. Основна відмінність між асоціативними правилами та PPD полягає в тому, що PPD буде створено з існуючих функціональних залежностей (FD) у наборі даних [59].

$$F_I: K = \{a_i\}, a_i \in A, D = \{a_j\}, \qquad (5.1)$$
$$a_j \in A, : P(k \in K \to d \in D) = p,$$

де k і d кортежі груп атрибутів K і D відповідно.

Прогалини та відсутні дані, представлені серед значень атрибута Y відношення r, класифікуються за допомогою PPD.

Пропонується наступний алгоритм PPD-видобування.

Алгоритм 1. Алгоритм PPD видобування (mining):

1. Сутності з однаковими X-значеннями будуть згруповані;
2. Вибрати атрибути з FD з тим самим X і додати їх до Y;
3. Щоб обчислити підтримку та впевненість, заповнення кортежу, вибраного на кроці 2;

4. Визначити кортежі з найвищим значенням Достовірності;
5. Щоб додати X→Y до PPDset;
6. Для заповнення відсутніх даних необхідно побудувати PPD.

Також пропонується новий алгоритм заповнення відсутніх даних.

Алгоритм 2. Алгоритм заповнення даних

1. Повнота=0;
2. Поки Повнота/100< Заповненняе;
3. Упорядкуйте всі атрибути з PPDset за рівнем надійності;
4. Для кожної групи:
    4.1. Якщо відсоток непорожній Значення Y вище або дорівнює підтримці заповніть порожні значення за допомогою PPDset:
        4.1.1. Повнота++;
    4.2. Інакше:
        4.2.1. Об’єднати PPD за допомогою правил Армстронга.

Приклад 5.1.

Для перевірки розроблених методів використовується набір даних із медичними даними.

Набір даних складається з 35 ознак та 122 екземплярів, зібраних у Львівському обласному реабілітаційному центрі для пост-COVID-пацієнтів із коротко- та довготривалим (понад 20 днів) лікуванням та реабілітацією. Особисті дані було видалено з набору даних і замінено унікальними випадковими ідентифікаторами. Наступна ознака, стать, обробляється за допомогою техніки одноразового кодування та в остаточному наборі даних представлена у двох компонентах – жіночому та чоловічому. Такі показники, як вік, вага, зріст, ІМТ, САТ, пульс, функція зовнішнього дихання, приймаються як фізіологічні параметри, виміряні перед стаціонарним лікуванням.

Ми реалізували наш підхід у Rstudio. Основними пакетами, які ми використовували, були caret, rpart, Metrics, Boruta, Randomforest, rules і ggplot2 для

візуалізації. Для генерації PPD мінімальний поріг підтримки дорівнює 0,0001 в апріорному алгоритмі вибрано. Крім того, відфільтровуються всі правила з достовірністю нижче 0,001.

Спочатку було проаналізовано прогнозну точність для всього набору даних і вибраних функцій (таблиця 5.2).

Таблиця 5.2. Порівняння результатів прогнозування моделей ML з використанням усього набору функцій і підмножини вибраних функцій

Модель	Для всього набору даних			Для вибраних функцій		
	MSE	MAE	R2	MSE	MAE	R2
k-НС	7,029	2,155	-0,162	5,781	1,825	0,044
МОВ	5,753	1,828	0,049	5,821	1,830	0,038
СГС	16,007	3,280	-1,647	12,328	2,603	-1,039
Лінійна регресія	17,826	3,571	-1,948	14,836	2,760	-1,453
БШП НМ	5,717	1,777	0,055	6,469	2,034	-0,070

Велика точність досягається для цілих вимірювань (середня квадратична помилка – MSE, середня абсолютна помилка MAE, R2).

Далі використовується метод заповнення відсутніх даних. Розроблений метод порівняли з існуючими: асоціативними правилами (AR), випадковим лісом (RF), машиною опорних векторів (SVM), багатошаровим перцептроном (MLP), очікуванням – максимізацією (EM) і k-найближчим сусідом (KNN) (рисунок 5.3). Помилка відновлення представлена за допомогою нормалізованої середньоквадратичної помилки (NRMSE).

Розроблений метод заповнення відсутніх даних створює додаткові значення даних, керуючи базованим доменом і функціональними залежностями, і включає ці значення до доступних навчальних даних. Правильність заповнених значень перевіряється на предикторі, побудованому на вихідному наборі даних. Розроблений метод PPD проводить на 12% краще, ніж моделі RF та EM при 30% відсутніх даних. Метод EM виглядає кращим для додаткових відсутніх даних

(діапазон приблизно 40–50% відсутніх даних), а PPD дає еквівалентні результати з SVM (машиною опорних векторів).

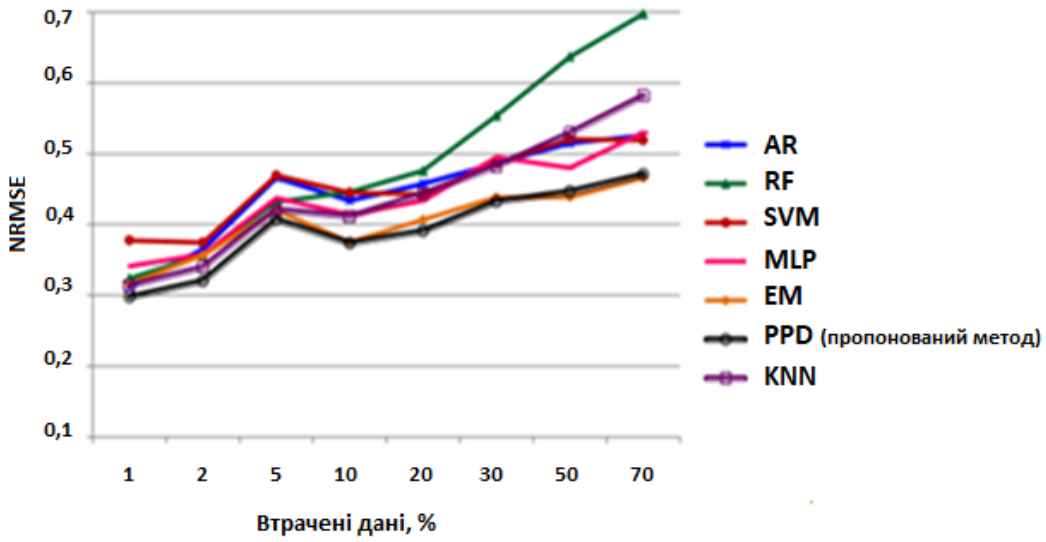


Рисунок 5.3. Помилка відновлення NRMSE.

Проаналізовано сучасні тенденції розвитку інформаційних технологій та баз даних. У результаті дослідження виявлені невирішені проблеми в області пошуку залежностей у великих базах даних кількох предметних галузей, зокрема, в медицині. Проведено аналіз існуючих методів і засобів виявлення залежностей у даних. Це дало змогу ідентифікувати новий підклас залежностей – Probabilistic Production dependencies. Розроблено метод отримання PPD в реляційних базах даних.

Порівняння підходів до дослідження та моделювання статистичних процесів за якісними критеріями підтвердило наступні переваги розробленого методу:

1. Збереження характеристик стійкості до помилок у даних;
2. Можливість паралельної реалізації в розподілених базах даних;
3. Автоматизація та виконання аналізу різних типів даних.

Асимптотична оцінка часу виконання алгоритму на системі із n комп'ютерів складає [228], формула 5.2:

$$t_n = O\left(minSupport^{-log_{avgD}(m)}\right). \quad (5.2)$$

Отримана оцінка часу виконання алгоритму є субполіноміальною, а отже, розроблений алгоритм є ефективним паралельним алгоритмом.

Отримана оцінка часу виконання алгоритму є субполіноміальною, а отже, розроблений алгоритм є ефективним паралельним алгоритмом.

Часова складність визначена так [228]:

$$t_n = O\left(minSupport \cdot \left(\frac{n}{minSupport}\right)^{1+log_{avgD}(m)} + \frac{Z_{ma}^2}{m \cdot D(A)} + \frac{Z_{aggr}^2 \cdot \log\left(sz_{aggr}\right)}{m \cdot D(A)}\right) =$$
$$= O\left(minSupport \cdot \left(\frac{n}{minSupport}\right)^{1+log_{avgD}(m)} + \frac{Z_{aggr}^2 \cdot \log\left(sz_{aggr}\right)}{m \cdot D(A)}\right). \quad (5.3)$$

Ємнісна складність визначена як [278]:

$$M = O\left(M_{stat} + M_{aggr} + M_{ma}\right) = \quad (5.4)$$
$$= O\left(\left(\frac{n}{minSupport}\right)^{1+log_{avgD}(m)} + Z_{aggr} \cdot sz_{aggr} + Z_{ma} \cdot sz_{ma}\right) =$$
$$= O\left(\left(\frac{n}{minSupport}\right)^{1+log_{avgD}(m)} + Z_{ma} \cdot sz_{ma}\right).$$

Розроблений алгоритм дає змогу стверджувати, що задача виявлення асоціативних залежностей у розподілених базах даних належить до класу P-задач. Отже, алгоритм пошуку асоціативних залежностей добре вирішувати з допомогою MapReduce.

Крім кількох послідовних реалізацій, паралельні реалії для роботи з великими даними не є широко доступними. Одним із прикладів серійної реалізації добре відомий статистичний: обчислення з пакетом R, що називається "arules".

Паралельна реалізація програми FP-Groth доступна в бібліотеці для вивчення комп'ютера з відкритим вихідним кодом (MLlib) Apache Spark та Apache Mahout.

Висновки до 5 розділу

1. Запропоноване формулювання аномалій, де означено, що визначення аномалій залежить від контексту, де найважливішим значенням необхідно враховувати показник часу, бо саме від нього буде залежати контекст.
2. Визначені типи аномалій та шляхи їхнього пошуку, що дає можливість диференціації підходів щодо аналізу персоналізованих медичних даних та визначені особливості їхнього застосування.
3. Визначені основні проблеми, що виникають при обробці медичних персоналізованих даних, за рахунок чого постає проблема розробки ефективної стратегії аналізу даних, яка може бути застосована до розосереджених баз даних різних доменів.
4. Розроблено та апробовано метод, заснований на ймовірнісній залежності. Відсоток відновлення даних становить 1,2 % порівняно з асоціативними правилами.
5. Розроблено метод щодо заповнення відсутніх даних, який грунтується на пошуку подібності даних, що можуть впливати на доменні значення та функціональні залежності між ними і забезпечує їхнє включення у навчальні дані, що дає змогу забезпечити стійкість до помилок даних, паралелізації обчислень та аналізу різнотипних даних. Розроблений метод заповнення даних на 12 % покращує результати краще, ніж моделі Random Forest та Expectation-Maximization для 30 % відсутніх даних.
6. Розроблено два алгоритми PPD майнінгу та заповнення відсутніх медичних персоналізованих даних, що забезпечують збереження характеристик стійкості до помилок у даних, можливість паралельної реалізації в розподілених базах даних, автоматизація та виконання аналізу різних типів даних.

РОЗДІЛ 6.

РОЗРОБКА ПРОГРАМНИХ МОДУЛІВ ДЛЯ ЕКСПЕРЕМЕНТАЛЬНИХ ДОСЛІДЖЕНЬ

У розділі розроблено архітектуру системи підтримки прийняття медичних рішень щодо прогнозування станів пацієнта на підставі опрацювання та аналізу персоналізованих медичних даних; подано опис імплементації рішень у системі підтримки прийняття рішень для лікування хворих на ПІД з порушеннями антитілоутвореннями. Використання цієї системи дає змогу вести облік хворих, контролювати та супроводжувати їхнє лікування та досліджувати динаміку перебігу хвороби та змін у стані пацієнта; подано опис імплементації програмного модуля щодо прогнозування динаміки поширення COVID-19 за урахуванням процесу класифікації пацієнтів відповідно до стану, проведення оцінки помилок одержаних рішень; подано результати розробки програмного модулю щодо прогнозування динаміки поширення COVID-19, що дозволяє аналізувати дані пацієнтів і на їх основі давати результат про наявність або відсутність COVID у пацієнта; подано результати імплементації системи для виявлення помилок, що дає змогу експерементувати із різними параметрами для моделі знову і знову, на одних і тих же даних. За допомогою цього користувач зможе вибрати модель, яка краще описує дані і має кращі результати, що чим менший часовий інтервал між точками, тим точніша модель, яка може визначити час, коли виникла аномалія, проте велика точність потребує значної кількості обчислень, і за допомогою даних графіків користувач зможе визначити оптимальний варіант.

Матеріали розділу опубліковані у роботах автора [162, 165, 169, 170, 174, 179, 183, 187, 188, 189, 196, 197, 199, 202, 206, 207].

6.1. Архітектура системи підтримки прийняття рішень

У будь-якій системі, що орієнтована на застосуваннння підходів машинного навчання, після визначення бізнес цінності та метрик успіху, процес надсилання моделей у використання складається з наступних етапів.

Пропонуємо архітектуру системи підтримки прийняття медичних рішень щодо прогнозування станів пацієнта на підставі опрацювання та аналізу персоналізованих медичних даних, де основними компонентами та етапами є:

1. Data Extraction – обираються та збираються релевантні дані з різних джерел;
2. Data Analysis – виконується підготовка exploratory data analysis (вивчення даних), щоб зрозуміти і виявити потрібні дані для побудови моделей машинного навчання. Цей процес веде до наступного:
   - розуміння схем даних та характеристик, на які очікує модель;
   - визначення підготовки даних та проектування функцій, необхідних для моделі;
3. Data Preparation – розподіл на навчальний, тестовий та валідаційний набори даних, застосування перетворень, аугментацій тощо. Результатом цього кроку є дані, отримані у підготовленому форматі;
4. Model Training – спеціаліст по даних (Data scientist) реалізує різні підходи до навчання з підготовленими на попередньому етапі даними, щоб отримати кілька моделей. Також він може перебирати різні гіперпараметри для отримання кращої моделі. Результатом цього кроку є натренована модель;
5. Model Evaluation – процес підрахунку метрик на тестовому наборі даних. Результатом цього кроку є набір метрик, що описує якість роботи моделі;
6. Model Validation – процес прийняття моделі для розгортання, порівняння її з попередньою моделлю, аналіз продуктивності;
7. Model Serving – модель впроваджується в обране оточення, що пройшла етап валідації, деплоймент може бути:
   - мікросервіси з REST API для виконання передбачень;
   - модель вбудована в якийсь пристрій, наприклад, мобільний;
   - частина системи прогнозування;

8. Model Monitoring – відстежуються продуктивність та деякі метрики якості, щоб можна було ініціалізувати новий етап у процесі машинного навчання.

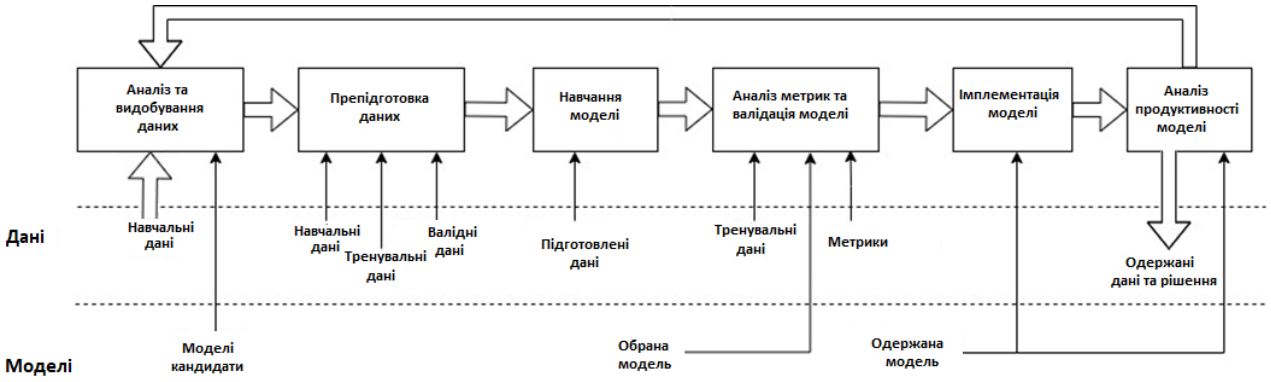


Рисунок 6.1 Архітектура системи підтримки прийняття медичних рішень

6.2. Імплементація рішень у прикладні медичні програмні модулі для медичних систем підтримки прийняття рішень

6.2.1 Інформаційна система підтримки прийняття рішень для лікування хворих з орфанними хворобами

Розроблена система складається з основних модулів:

1. Модуль авторизації користувача;
2. Модуль заповнення загальної інформації про пацієнта;
3. Модуль внесення даних для визначення шляху до діагнозу відповідно до анамнезу та лабораторних показників пацієнта;
4. Модуль встановлення діагнозу, враховуючи особливості історії хвороби та результатів лабораторних досліджень;
5. Модуль обліку проведення трансплантації стовбурових клітин чи проведена генна терапія;
6. Модуль визначення замісної терапії даному пацієнту з врахуванням цільових змінних стану пацієнта та особливостей застосування медикаментів;
7. Модуль моніторингу та обліку хворих;
8. Модуль проведення аналітичних досліджень;
9. Модуль супроводу операцій друку, фільтрації, допомоги користувачам.

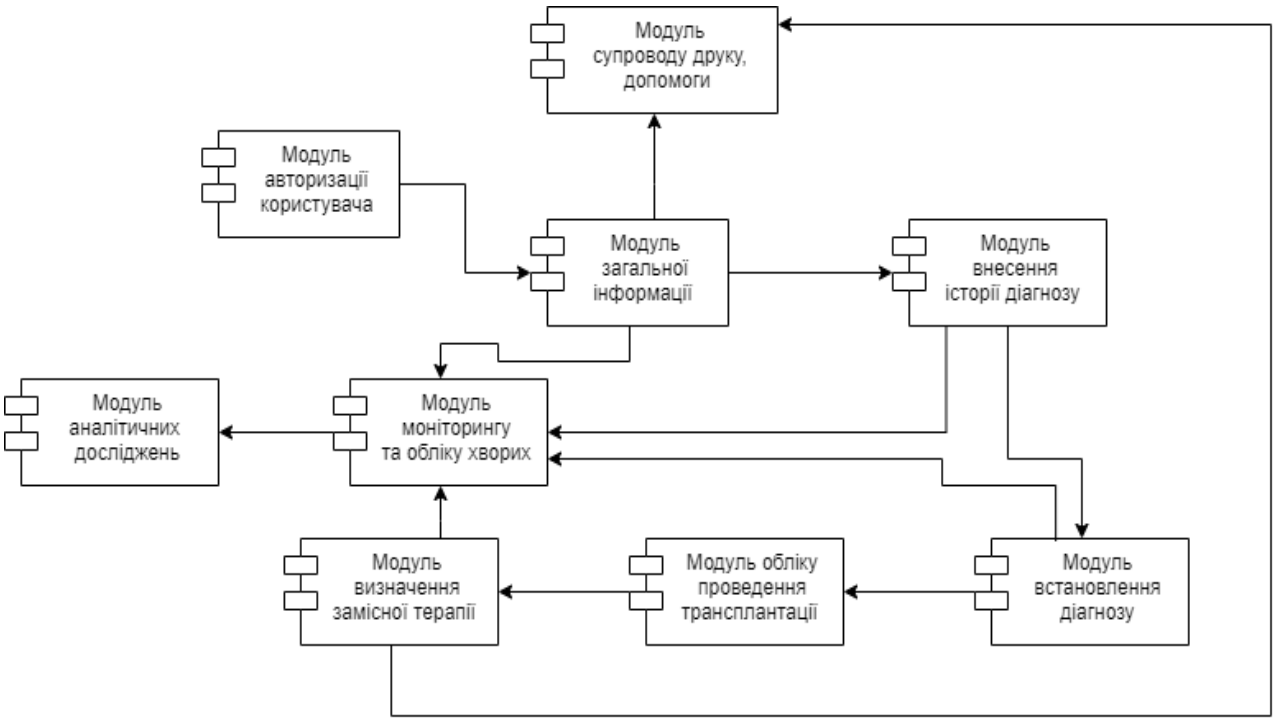


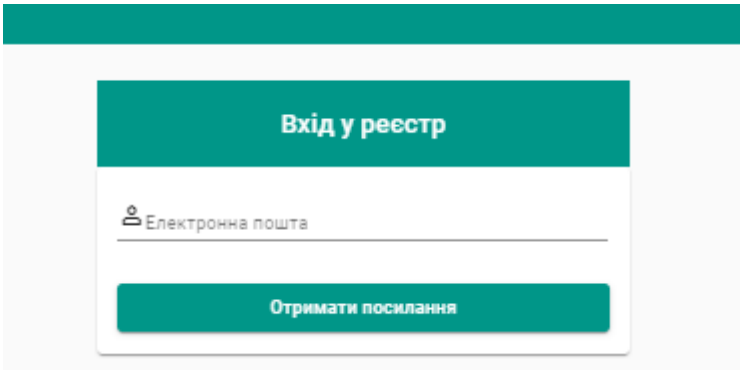
Рисунок 6.2 Структура інформаційної системи підтримки прийняття рішень

Детальний огляд реалізації інформаційної системи підтримки прийняття рішень

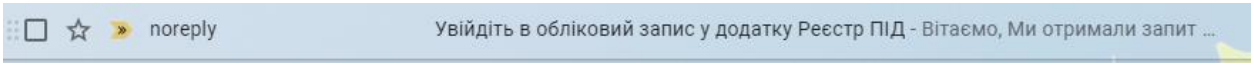
I. Інформаційна система підтримки прийняття рішень для лікування хворих на ПІД з порушеннями антитілоутвореннями доступна у браузерах Chrome, Mozilla, Opera, а також у мобільній версії.

II. Вхід в інформаційну систему у «Реєстр ПІД з порушеннями антитілоутворень»:

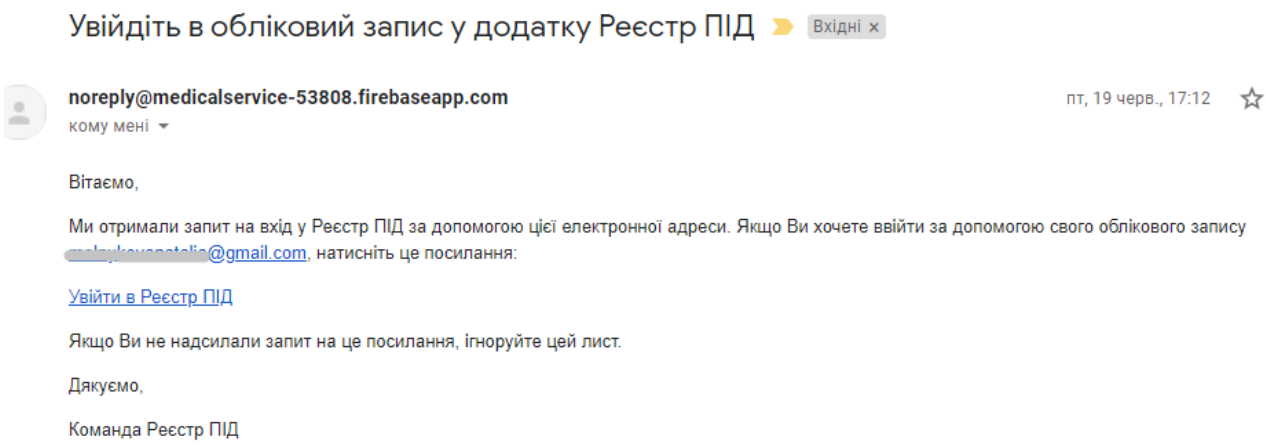
1. Перейти у реєстр за посиланням https://register-pid-ua.web.app/login;
2. Ввести свою електронну пошту у вікно реєстрації;



3. Перейти на свою пошту та перевірити наявність листа від системи. Користувач отримує підтвердження на поштову скриньку;



4. Текст листа такого змісту:



5. Перейдіть за вказаним посиланням «Увійдіть в Реєстр ПІД»;
6. Перейшовши за посиланням, користувач буде у власному профілі реєстру (рисунок 6.3);

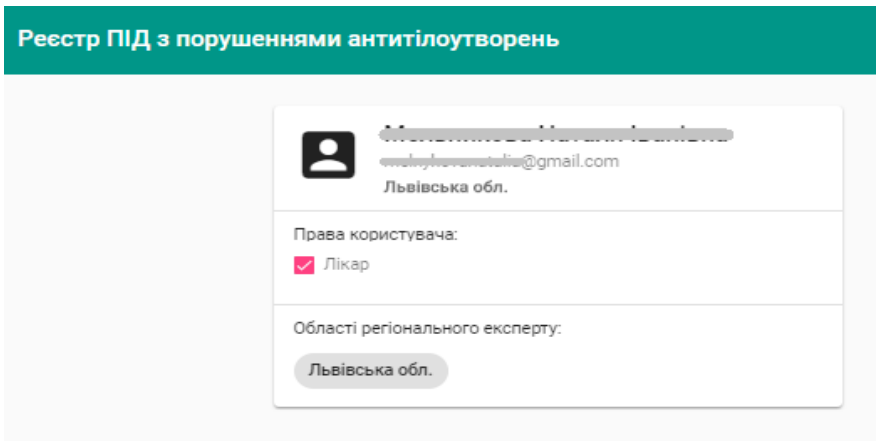


Рисунок 6.3 Інформація про профіль користувача

III. Внесення даних про пацієнта:

1. Внесення даних про пацієнта у форму «Загальна інформація»;

1.1. На стрічці меню системи обрати пункт «Додати пацієнта» (рисунок 6.4);

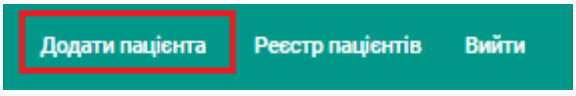


Рисунок 6.4 Меню «Реєстр ПІД з порушеннями антитілоутворень»

1.2. У формі «Загальна реєстрація» (рисунок 6.3) необхідно ввести дані про хворого, а саме:

1.2.1. Прізвище, ім’я, по батькові;

1.2.2. Стать (обрати один варіант відповіді із запропонованих Чоловіча/Жіноча);

1.2.3. Використання даних (обрати один варіант відповіді із запропонованих Повне/Науковий аналіз/Не використовувати);

1.2.4. Завантажити документ про згоду (скористатися кнопкою внизу панелі зліва «Завантажити файл згоди» (рисунок 6.5) для завантаження документу (після завантаження роздрукувати);

1.2.5. Регіон народження;

1.2.6. Місто народження;

1.2.7. Дата народження (обрати у запропонованому календарі)

1.2.8. Регіон проживання;

1.2.9. Місто проживання;

1.2.10. Дата смерті (обрати у запропонованому календарі, якщо відома);

1.2.11. Сімейний анамнез (обрати один варіант відповіді із запропонованих Так/Ні/Невідомо);

1.2.11.1. Внесення даних про тип родинного зв’язку та ідентифікатор особи у «Реєстр ПІД з порушеннями антитілоутворень» (рисунок 6.5);

1.2.11.2. Для внесення ще одного типу зв’зку необхідно скористатися кнопкою додати (рисунок 6.5);

1.2.11.3. Якщо помилково внесені дані про тип з’язку, скористайтеся кнопкою знищення (рисунок 6.5);

1.3. Перейти на наступну форму «Шлях до діагнозу», натиснувши кнопку внизу панелі справа «Наступна форма» (рисунок 6.5);

1.4. Перейти на іншу форму, можна клацнувши на закладку іншої форми реєстру.

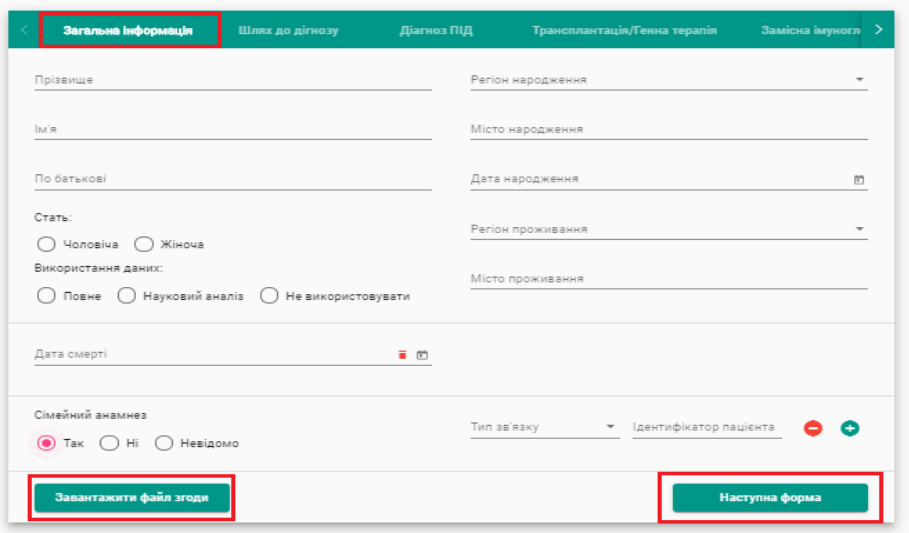


Рисунок 6.5 Форма «Загальна інформація» для заповнення загальної інформації про пацієнта

2. Внесення інформації у форму «Шлях до діагнозу» (рисунок 6.6):

2.1. Ввести дату перше встановленого клінічного діагнозу (обрати у запропонованому календарі);

2.2. Обрати одну або декілька відповідей із запропонованих варіантів інформації про первинний імунодефіцит, що підтверджено лабораторними дослідженнями. Допускається введенні додаткової інформації вручну (рисунок 6.6);

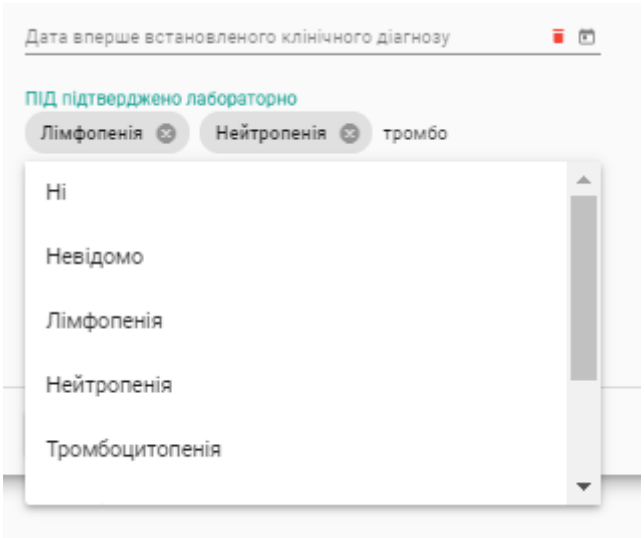


Рисунок 6.6 Поле «ПІД підтверджено лабораторно» для внесення інформації про результати лабораторних досліджень

2.3. Ввести інформацію про ПІД-асоційовані прояви (обрати один варіант відповіді із запропонованих Відомі/Невідомі/Відсутні):

2.3.1. Якщо відомі, потрібно обрати симптом або симптоми із запропонованого списку (рисунок 6.7):

2.3.1.1. Для внесення ще одного прояву необхідно скористатися кнопкою додати (рисунок 6.7);

2.3.1.2. Якщо помилково внесені дані про прояв, скористайтеся кнопкою знищення (рисунок 6.7);

2.4. Ввести дату кожного прояву (обрати у запропонованому календарі);

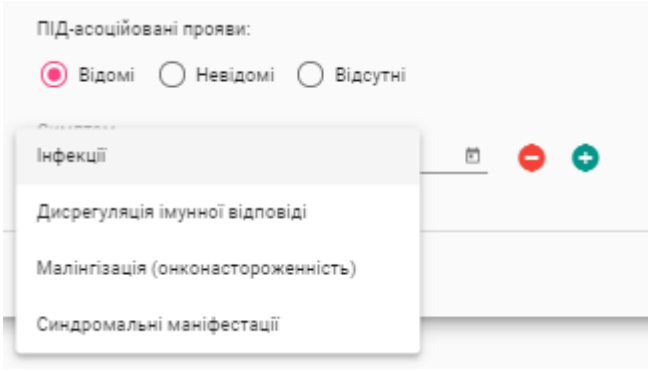


Рисунок 6.7 Поле «ПІД-асоційовані прояви» для внесення інформації прояви ПІД

2.5. Ввести лабораторні імунологічні показники (рисунок 6.8):

2.5.1. IgM (г/л);

2.5.2. IgG (г/л);

2.5.3. IgA (г/л);

2.5.4. IgE (м.о.);

2.5.5. Якщо показники виходять за межі норми під записом з'явиться коментар (рисунок 6.8);

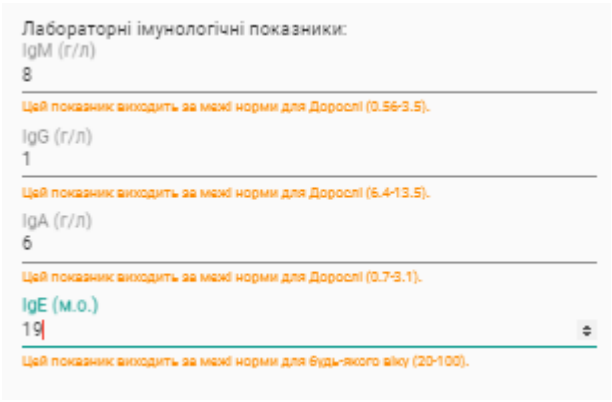


Рисунок 6.8 Поле «Лабораторні імунологічні показники» з коментарями щодо невідповідності показників нормам

2.6. Перейти на наступну форму «Діагноз ПІД», натиснувши кнопку внизу панелі справа «Наступна форма» (рисунок 6.9);

2.7. Перейти на іншу форму, можна клацнувши на закладку іншої форми реєстру;

2.8. Перейти на попередню форму «Загальна інформація», натиснувши кнопку внизу панелі справа «Попередня форма» (рисунок 6.9), або клацнувши на закладку «Загальна інформація».

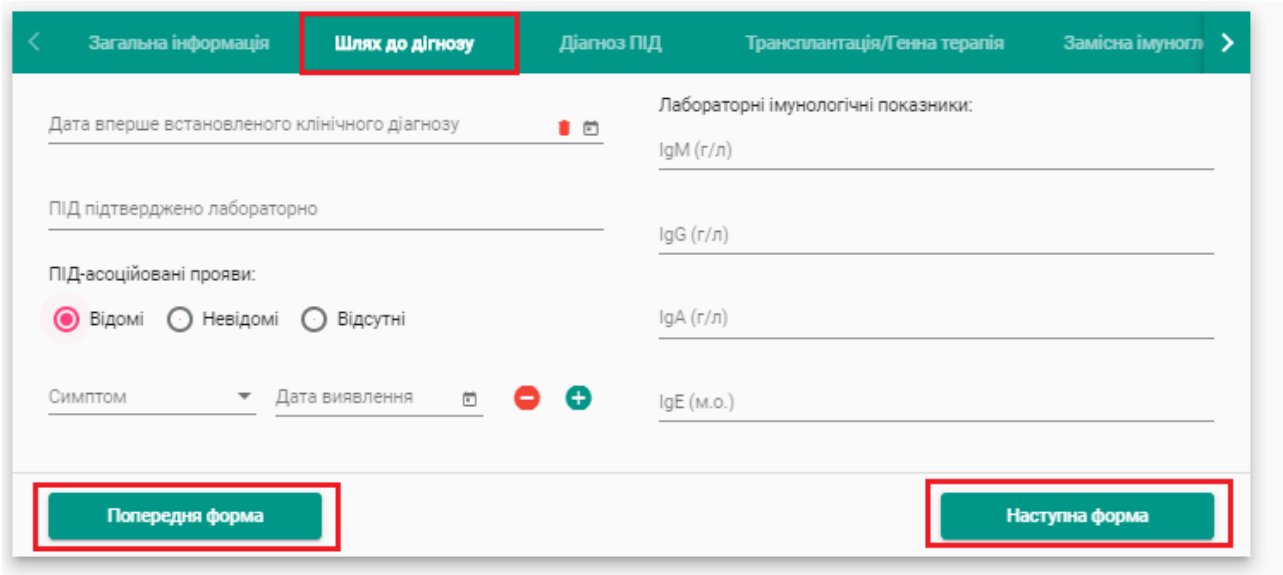


Рисунок 6.9 Форма «Шлях до діанозу» внесення даних щодо шляху до діагнозу.

3. Внесення даних у форму «Діагноз ПІД»:

3.1. Обрати категорію ПІД. Активна лише категорія «Дефіцити антитілоутворення» (рисунок 6.10);

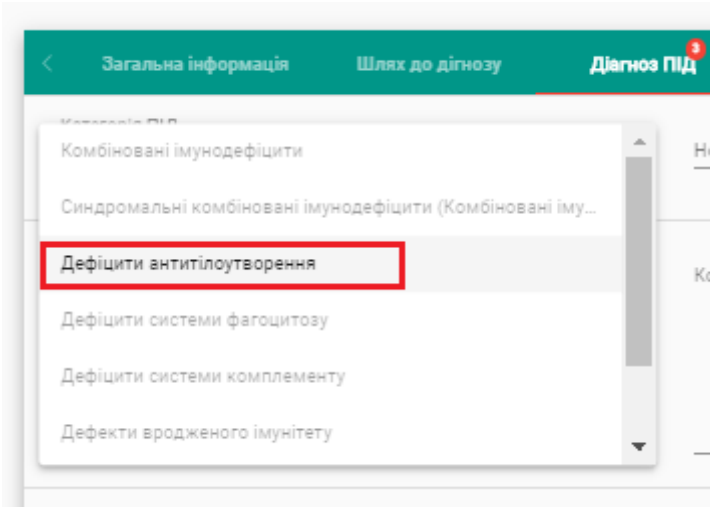


Рисунок 6.10. Поле вибору категорії ПІД

3.2. Обрати нозологію відповідної категорії з випадаючого списку (рисунок 6.11);

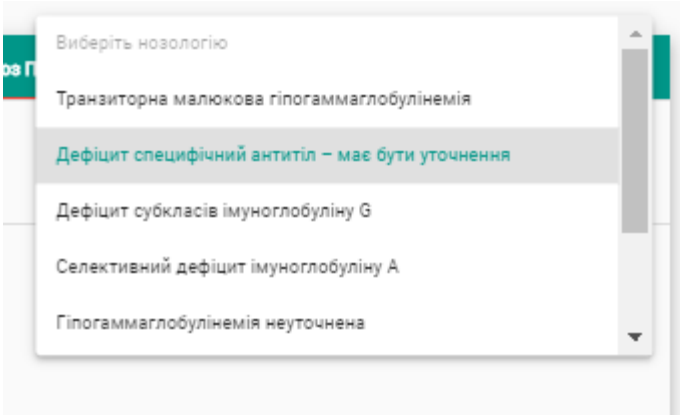


Рисунок 6.11 Поле вибору нозології ПІД

3.3. Обрати відповідну інформацію щодо історії генетичних досліджень (обрати один варіант відповіді із запропонованих):

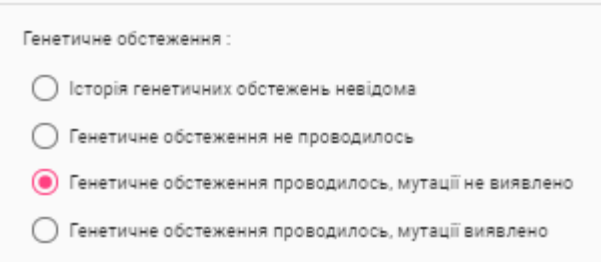


Рисунок 6.12. Поле вибору історії генетичних досліджень

3.3.1. Якщо обираєте варіанти «Генетичне обстеження проводилось мутації не виявлено» та «Генетичне обстеження проводилось мутації виявлено» необхідно внести:

3.3.1.1. Дату генетичного обстеження (обрати із запропонованого календаря);

3.3.1.2. Лабораторію обстеження (обрати із запропонованого списку)

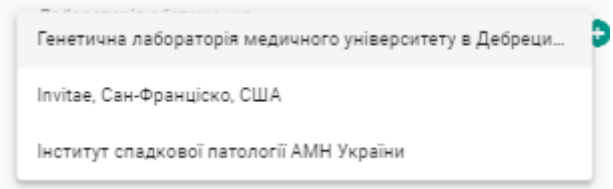


Рисунок 6.13 Перелік лабораторій

або внести іншу, скориставшись кнопкою Додати , ввівши дані у вікно внесення нової лабораторії;

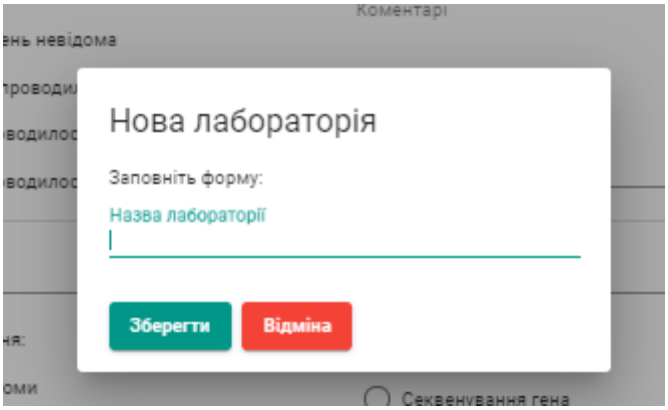


Рисунок 6.14. Вікно внесення інформації про нову лабораторію

3.3.2. Обрати один варіант відповіді щодо методу секвенування (рисунок 6.15):

3.3.2.1. Якщо обирають варіант панельне секвенування, тоді необхідно вказати кількість генів (рисунок 6.15);

Рисунок 6.15 Вікно внесення кількості генів

3.4. У поле коментарі (рисунок 6.16) можна вписати необхідну додаткову інформацію щодо визначення діагнозу ПІД;

3.5. Перейти на наступну форму «Трансплантація/ Генна терапія», натиснувши кнопку внизу панелі справа «Наступна форма» (рисунок 6.16);

3.6. Перейти на іншу форму можна, клацнувши на закладку іншої форми реєстру;

3.7. Перейти на попередню форму «Шлях до діагнозу», натиснувши кнопку внизу панелі справа «Попередня форма» (рисунок 6.16), або клацнувши на закладку «Шлях до діагнозу».

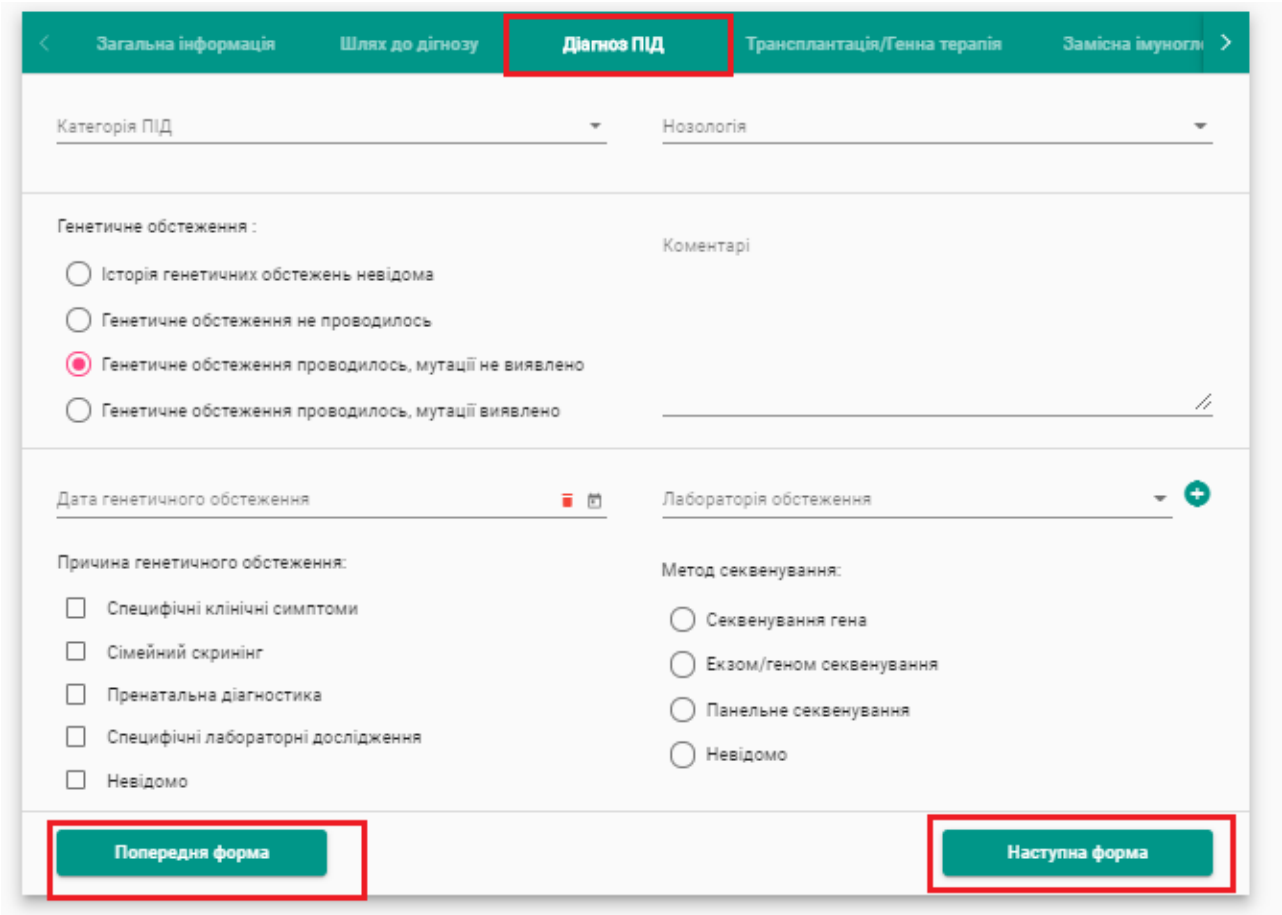


Рисунок 6.16. Форма «Діагноз ПІД» внесення даних щодо визначення діагнозу.

4. Внесення даних у форму «Трансплантація/Генна терапія»:

4.1. Обрати один із варіантів відповідей поля «Чи була трансплантація стовбурових клітин» – Так/Ні/Невідомо (рисунок 6.21);

4.2. Якщо обрано варіант «так»:

4.2.1. Необхідно ввести дату трансплантації (обрати із запропоновано календаря), (рисунок 6.17);

4.2.2. Необхідно обрати інформацію про джерело CD34 стовбурових клітин (обрати один із варіантів Кістковий мозок/Периферична кров/Пуповинна кров/Невідомо);

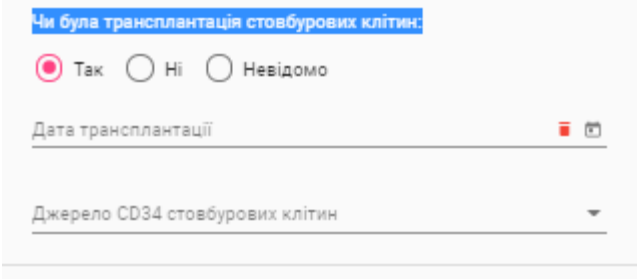


Рисунок 6.17. Внесення даних про наявність трансплантації

4.3. Обрати один із варіантів відповідей поля «Чи була генна терапія» – Так/Ні/Невідомо (рисунок 6.21):

4.3.1. Якщо обрано варіант «так»:

4.3.1.1. необхідно ввести дату проведення генної терапії (обрати із запропоновано календаря), (рисунок 6.18);

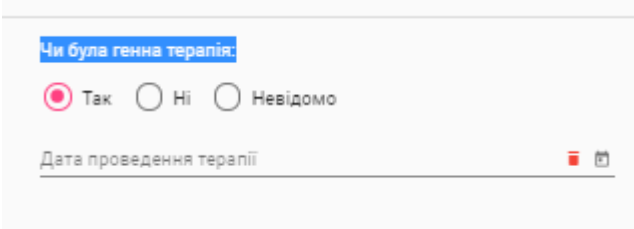
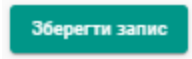
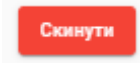


Рисунок6.18 Внесення даних про проведену генну терапію

4.3.2. У поле коментар (рисунок 6.21) за потребою внести додаткову інформацію про особливості проведення трансплантації та генної терапії;

4.3.3. Для збереження внесеної інформації про проведення трансплантації та генної терапії необхідно натиснути на кнопку внизу форми справа 

або якщо потрібно відмінити, тоді натиснути кнопку 

(рисунок 6.21);

4.3.4. Якщо необхідно видалити запис інформації, що була збережена, потрібно натиснути на кнопку знищити у полі таблиці «Дії» (рисунок 6.19);

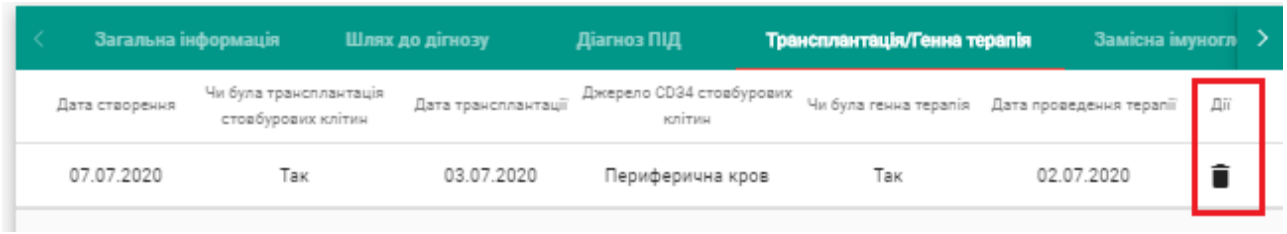


Рисунок 6.19. Керування даними після внесення

4.4. Після чого з’явиться вікно про підтвердження знищення інформації, (рисунок 6.20.);

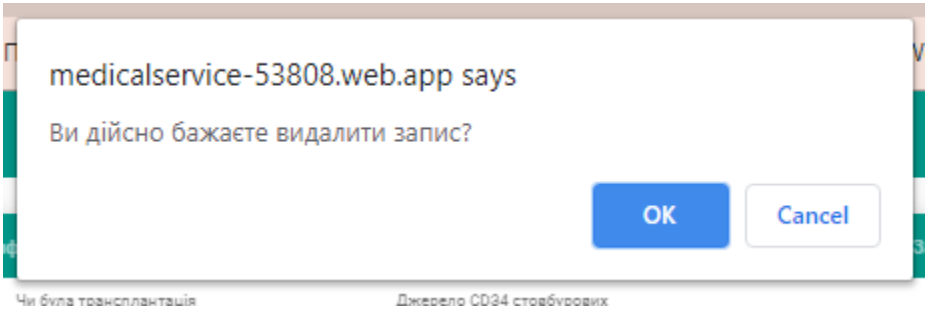


Рисунок 6.20. Вікно підтвердження знищення запису, що містить інформацію про проведену генну терапію та трансплантацію

4.5. Перейти на наступну форму «Замісна імуноглобулінотерапія», натиснувши кнопку внизу панелі справа «Наступна форма» (рисунок 6.21.);

4.6. Перейти на іншу форму, можна клацнувши на закладку іншої форми реєстру;

4.7. Перейти на попередню форму «Діагноз ПІД», натиснувши кнопку внизу панелі справа «Попередня форма» (рисунок 6.21), або клацнувши на закладку «Діагноз ПІД».

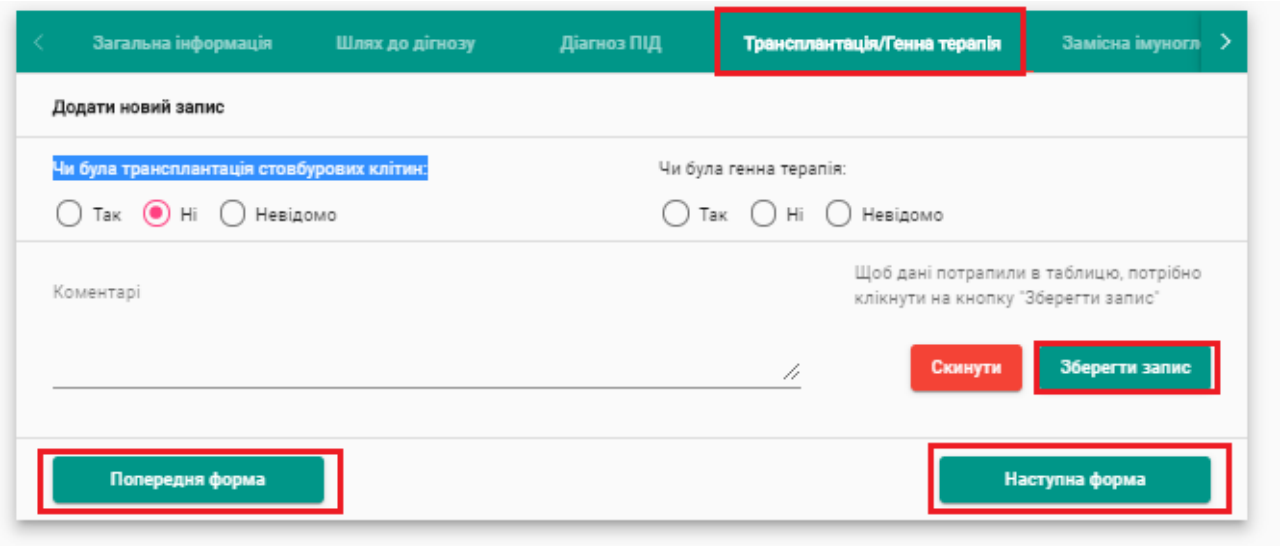


Рисунок 6.21. Форма «Трансплантація/Генна терапія» внесення даних щодо проведення у пацієнта трансплантації чи генної терапії.

5. Внесення даних у форму «Замісна імуноглобулінотерапія»:

5.1. Для формування історії призначень щодо замісної імуноглобулінотерапії необхідно додати запис, що передбачає вибір одного із варіантів відповідей поля «Чи потребує пацієнт на теперішній час замісну терапію?» – Так/Не регулярно/Ні/Невідомо (рисунок 6.22);

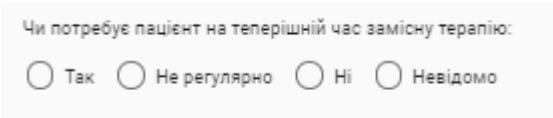


Рисунок 6.22 Поле вибору чи потребує пацієнт на теперішній час замісну терапію

5.2. Якщо обрано варіант «Так» чи «Нерегулярно» необхідно внести дані у наступні поля (рисунок 6.32):

5.2.1. Кінцеву дату введення імуноглобуліну (рисунок 6.30) (обрати із запропонованого календаря);

5.2.2. Вказати ефективність попередньої терапії (рисунок 6.23) (обрати один із варіантів Добра/Часткова/Відсутня);

Рисунок 6.23 Поле вибору ефективності попередньої імуноглобулінотерапії

5.2.3. Ввести рекомендовану дозу (г/кг) (рисунок 6.32);

5.2.4. Ввести поточну вагу пацієнта на день огляду (кг) (рисунок 6.32);

5.2.5. Виконається підрахунок місячної дози (рисунок 6.32);

5.2.6. Обрати виробника з існуючих (Невідомо/Біофарма/ Октафарма) (рисунок 6.24) або додати іншого, натиснувши кнопку додати (рисунок 6.25);

Рисунок 6.24 Поле вибору виробника імуноглобулінів

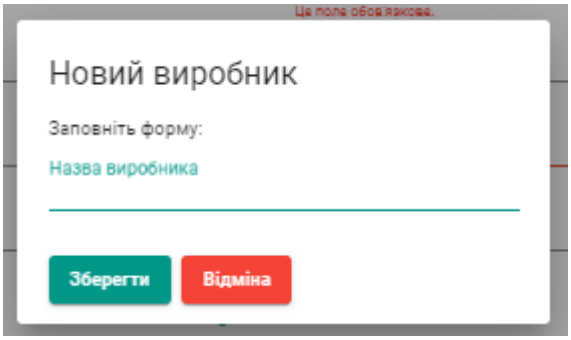


Рисунок 6.25 Поле внесення нового виробника імуноглобулінів

5.2.7. Ввести поточний рівень IgG у (г/л) (рисунок 6.32);

5.2.8. Обрати із запропонованих варіантів локації введення (Вдома/Стаціонарно/ Амбулаторно/ Обидві локації/ Невідомо) (рисунок 6.26);

Рисунок 6.26 Поле вибору варіантів локації введення

5.2.9. Обрати із запропонованих інтервалів введення (Кожні 7 днів/ Кожні 14 днів/ Кожні 21 днів/ Кожні 28 днів/ Невідомо) або ввести самостійно (рисунок 6.27);

Рисунок 6.27 Поле вибору варіантів інтервалів введення

5.2.10. Обрати із запропонованих шляхів введення (Внутрішньовенно/ Підшкірно/ Внутрішньом’язево), (рисунок 6.28);

Рисунок 6.28 Поле вибору варіантів шляхів введення

5.2.11. Ввести побічні явища під час проведення замісної імуноглобулінотерапії (Гарячка/ Анемія/ Лейкопенія/ Анафілактична реакція/ Біль голови/ Біль у м’язах/ Біль у суглобах/ Локальні/ Інші) або внести самостійно та натиснути клавішу Enter, (рисунок 6.29);

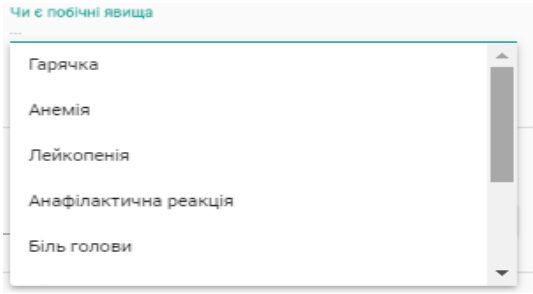
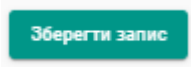


Рисунок 6.29 Поле вибору побічних явищ під час проведення замісної імуноглобуліно терапії

5.3. Внесену інформацію потрібно зберегти, натиснувши на кнопку внизу форми справа 

, або якщо необхідно відмінити, тоді натиснути кнопку 

(рисунок 6.30);

5.4. Якщо необхідно видалити запис інформації, яка була збережена потрібно натиснути на кнопку видалити  у полі таблиці «дії» (рисунок 6.30):

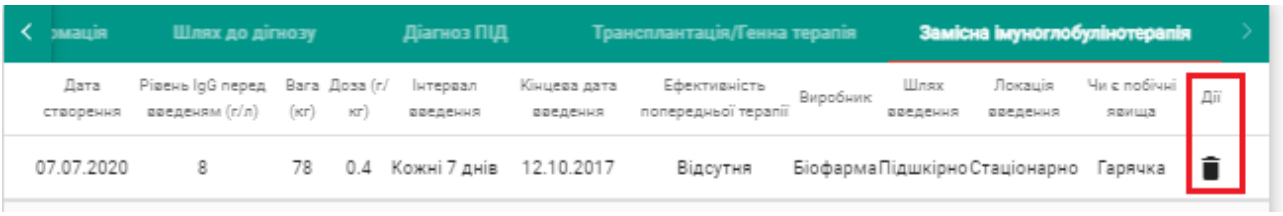


Рисунок 6.30 Керування даними після внесення

5.4.1. Після чого з’явиться вікно про підтвердження знищення інформації. (рисунок 6.31);

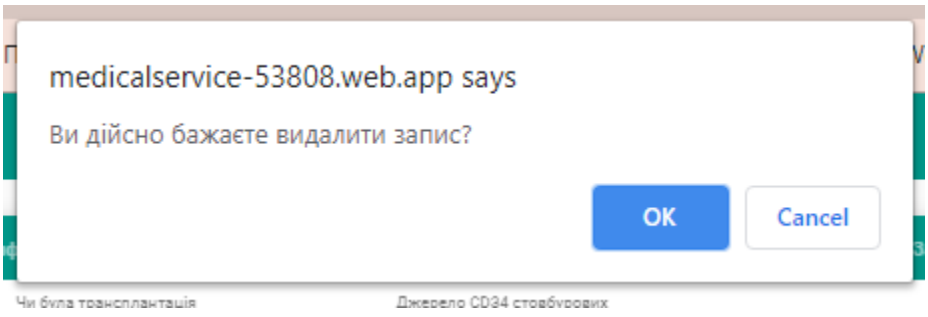


Рисунок 6.31 Вікно підтвердження знищення запису що містить інформацію про проведену замісну імуноглобулінотерапію

5.5. Перейти на іншу форму, можна клацнувши на закладку іншої форми реєстру;

5.6. Перейти на попередню форму «Трансплантація/Генна терапія», натиснувши кнопку внизу панелі справа «Попередня форма» (рисунок 6.32), або клацнувши на закладку «Трансплантація/Генна терапія»;

5.7. Зберегти дані про пацієнта, натиснувши на кнопку в самому низу справа форми «Замісна імуноглобулінотерапія» (рисунок 6.32). Після натискання з'явиться підтвердження внесених даних (рисунок 6.33). Якщо підтвердження з помилкою (рисунок 6.34), необхідно повернутися до даних форм та все ретельно перевірити.

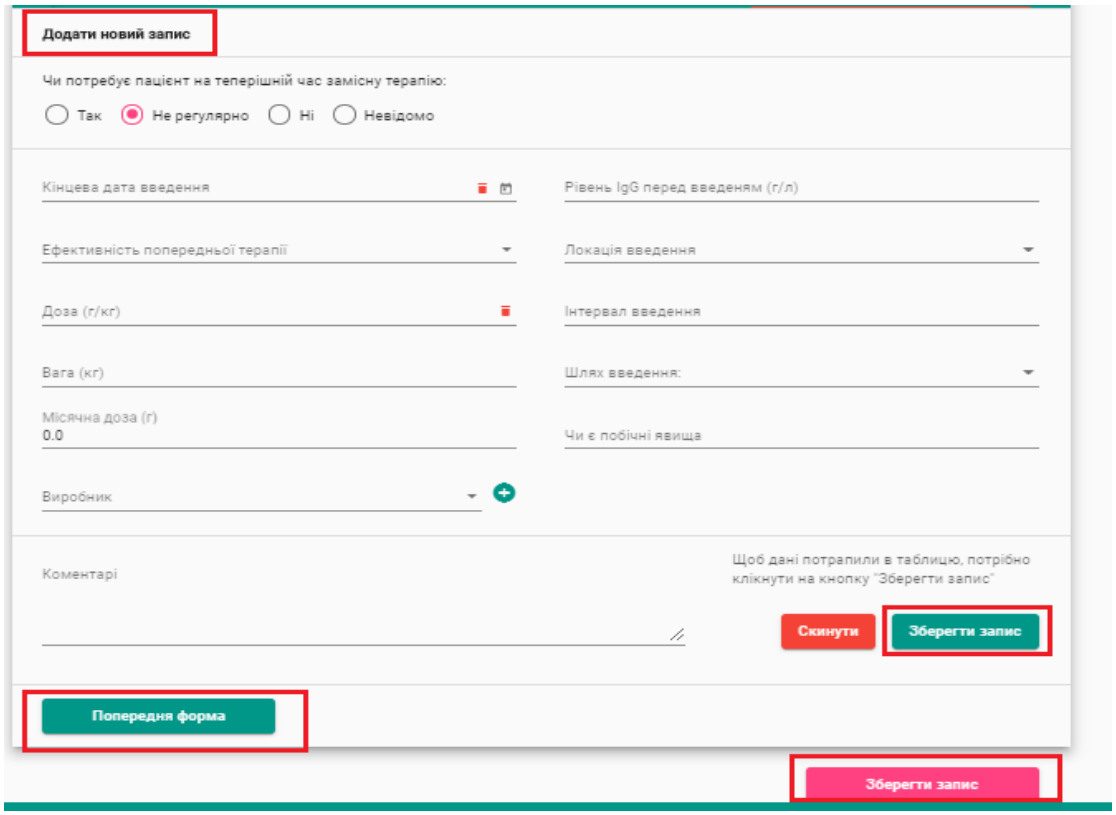


Рисунок 6.32 Форма «Замісна імуноглобулінотерапія» для внесення даних щодо проведення у пацієнта замісної імуноглобулінотерапії.

Рисунок 6.33 Вікно підтвердження про внесені усі дані про хворого

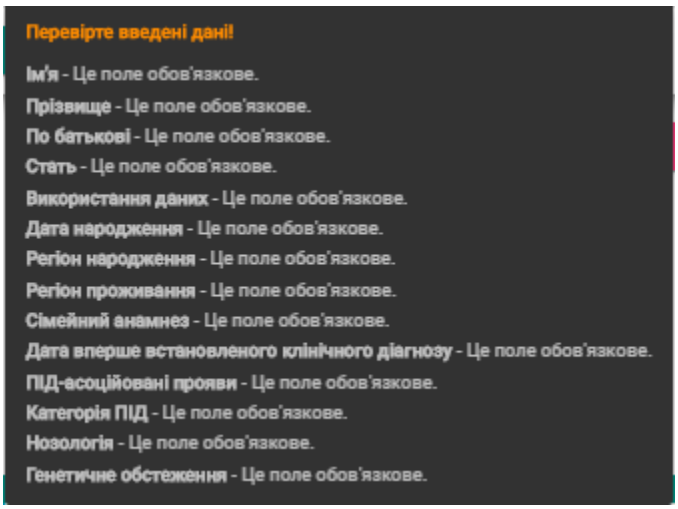


Рисунок 6.34 Вікно про допущені помилки при внесенні даних про хворого

IV. Перегляд та редагування даних про хворого:

1. Після збереження даних про хворого інформація з'являється у загальному списку непідтверджених пацієнтів (доступитися через пункт меню головної форми реєстру);

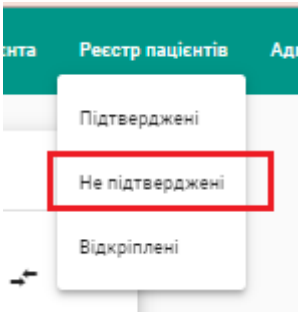


Рисунок 6.35 Перехід на таблицю із зведеною інформацією про пацієнта

2. Таблиця із списком пацієнтів з непідтвердженими діагнозами (рисунок 6.36);

Дата створення	Ідентифікатор	Вік	Стать	Область	Нозологія	Дата встановлення діагнозу	Остання дата введення імуноглобуліну	Спосіб ін'єкції	Доза введення (г/кг)	Вага (кг)	Виробник	Лікуючий лікар	Дії
07.07.2020	2CUUf1Ft6THGsKbd4nwA	18	Жіноча	Житомирська обл.	Дефіцит специфічний антитіл – має бути уточнення	03.07.2020	12.10.2017	Підшкірно	0.4	78	Біофарма	Мельникова Наталя	
03.07.2020	Kc1MIUdkJAUyUK48L9MN	53	Чоловіча	Львівська обл.	Селективний дефіцит імуноглобуліну	15.10.2014	17.07.2020	Внутрішньом'язево	1120	67	Октафарма	Скопівський Степан	

Рисунок 6.36 Таблиця із списком пацієнтів з непідтвердженими діагнозами

3. Щоб переглянути інформацію про пацієнта можна скористатися кнопкою перегляду (рисунок 6.36);

4. Після натискання кнопки перегляду переходимо на форми з даними про пацієнта (рисунок 6.37);

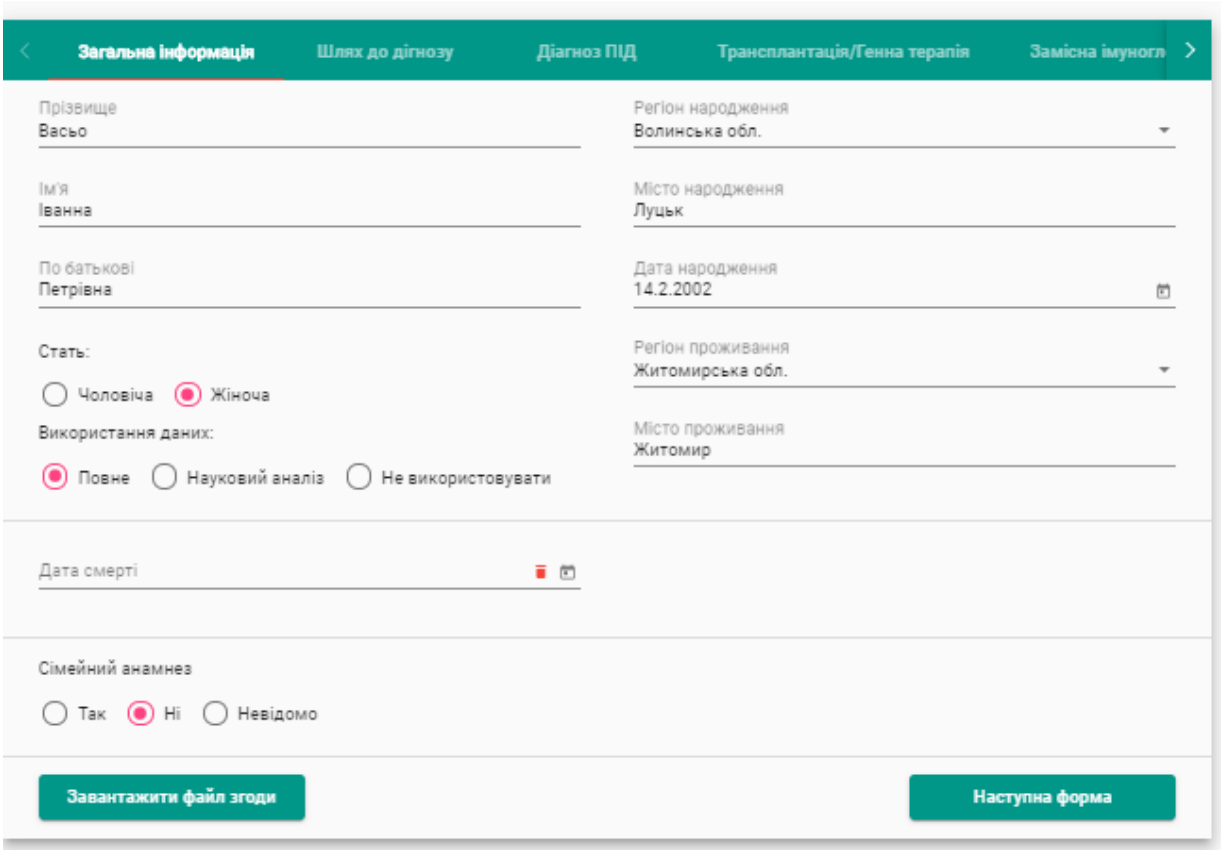
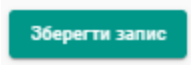
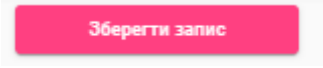


Рисунок 6.37 Форма «Загальна інформація» із попередньо внесеними даними пацієнта.

5. Якщо вносяться дані про повторну трансплантацію чи генну терапію, чи замісну імуноглобулінотерапію, потрібно зберегти поточні зміни, натиснувши на кнопку 

, або якщо необхідно відмінити внесенні дані, тоді натиснути кнопку 

(рисунок 6.21 та 6.22);
6. Якщо зберегти будь-які зміни на одній із форм, необхідно внести їх у базу даних, тоді потрібно натиснути на кнопку 

вкінці внесення на формі «Замісної імуноглобулінотерапії» (рисунок 6.32);
7. Після натискання з'явиться підтвердження внесених даних (рисунок 6.33). Якщо підтвердження з помилкою (рисунок 6.34), необхідно повернутися до даних форм та все ретельно перевірити.

V. Підтвердження діагнозу (права на цю операцію мають експерти регіону та країни):

1. Для узгодження діагнозу можна задати запитання лікуючому лікарю, натиснувши на кнопку коментар у таблиці зі списком пацієнтів. (рисунок 6.38);

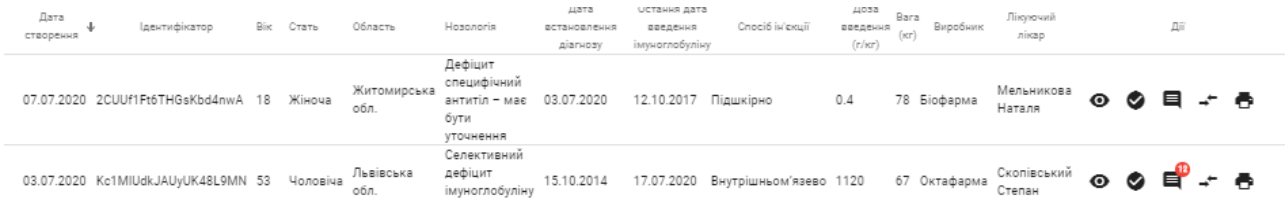


Рисунок 6.38 Таблиця із списком пацієнтів з непідтвердженими діагнозами

2. Після натискання кнопки коментар з'явиться вікно для внесення коментаря (рисунок 6.39). Для збереження коментаря потрібно натиснути кнопку Зберегти у вікні коментаря, якщо потрібно відмінити дію, тоді треба натискати на кнопку Відміна ;

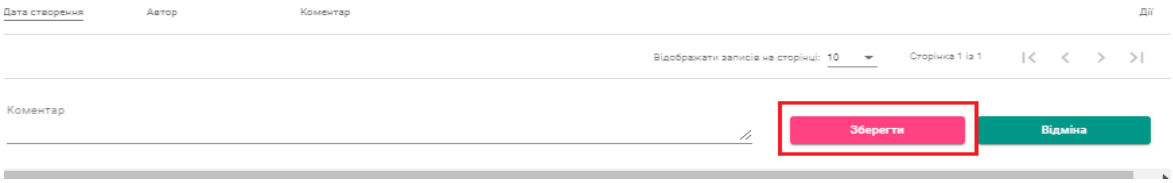


Рисунок 6.39 Вікно внесення коментарів

3. Для підтвердження діагнозу необхідно скористатися кнопкою підтвердження у таблиці зі списком пацієнтів (рисунок 6.38);
4. Після натискання з'явиться вікно схвалення процедури підтвердження діагнозу, (рисунок 6.40);

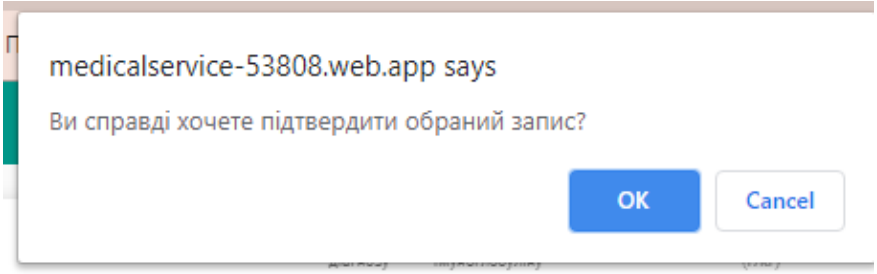


Рисунок 6.40 Вікно схвалення процедури підтвердження діагнозу

5. Після схвалення пацієнт переходить у категорію підтверджених, щоб перейти необхідно натиснути кнопку в головному меню «Реєстр пацієнтів» та позицію «Підтверджені» (рисунок 6.41, 6.42)

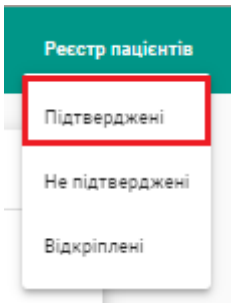


Рисунок 6.41 Перехід на таблицю із пацієнтами, що мають підтвердженні діагнози

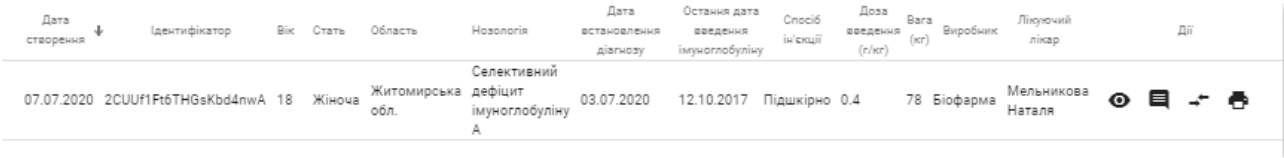


Рисунок 6.42 Таблиця зі списком пацієнтів з підтвердженими діагнозами

6. Якщо під час спостереження у пацієнта змінився діагноз, лікар може перейти у режимі перегляду та редагування п. 8 на форму «Діагноз ПІД» та обрати іншу нозологію, попередні дані будуть збережені в таблиці змін (рисунок 6.43) і пацієнт перейде у категорію із непідтвердженими діагнозами, (рисунок 6.38).

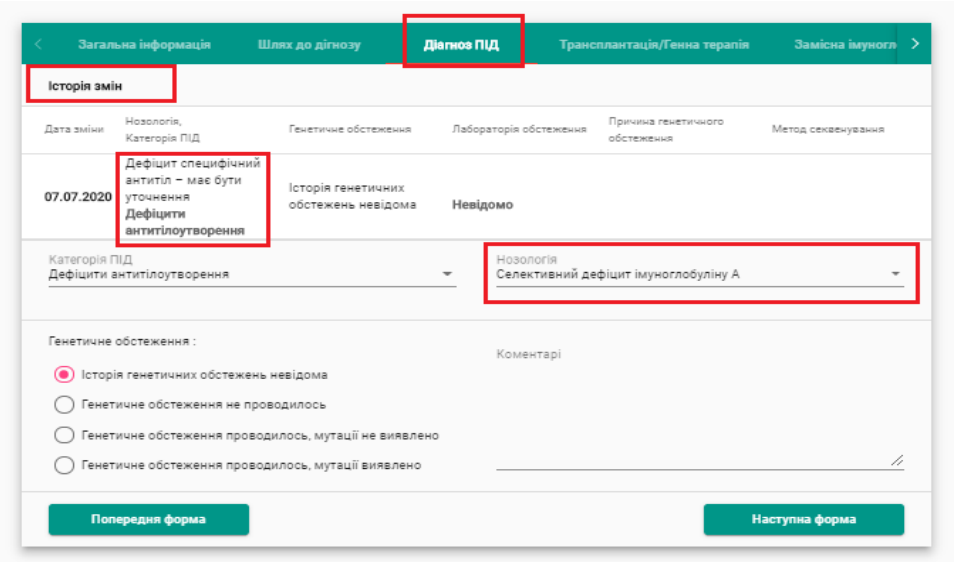


Рисунок 6.43 Форма «Діагноз ПІД» з історією змін діагнозів пацієнта

VI. Процедура відкріплення/прикріплення пацієнта або зміна лікуючого лікаря (права на цю операцію мають експерти регіону):

1. Для відкріплення пацієнта від лікуючого лікаря та прикріплення до іншого в межах одного чи різних регіонів по причині повноліття чи за потребою необхідно скористатися кнопкою відкріплення у таблиці зі списком пацієнтів. (рисунки 6.38, 6.42);
2. Після натискання на кнопку відкріплення, з'явиться діалогове вікно схвалення процедури відкріплення (рисунок 6.44);

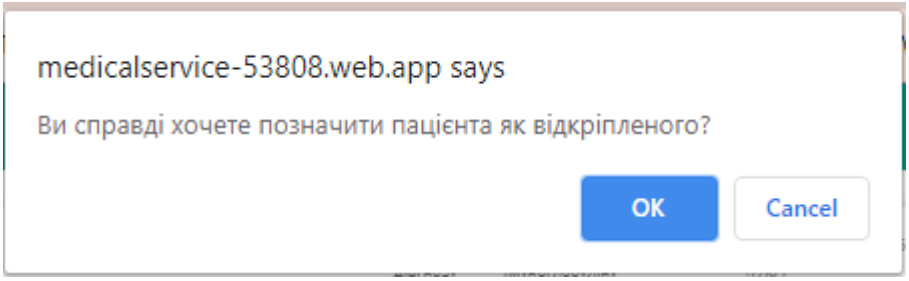


Рисунок 6.44 Діалогове вікно підтвердження відкріплення пацієнта

3. Після схвалення відкріплення пацієнт переходить в категорію відкріплених, щоб перейти необхідно натиснути кнопку в головному меню «Реєстр пацієнтів» та позицію «Відкріплені» (рисунок 6.45, 6.46);

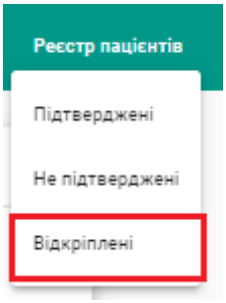


Рисунок 6.45 Перехід на таблицю із пацієнтами, що відкріплені

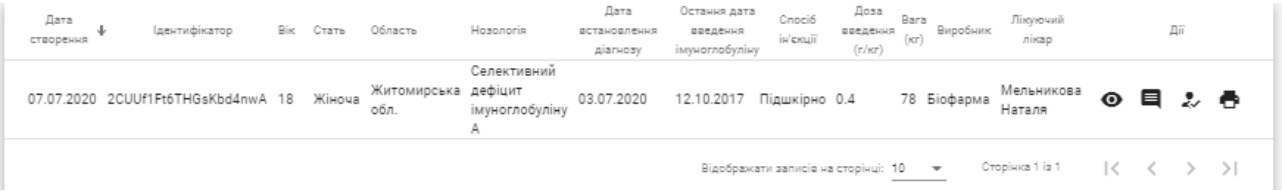


Рисунок 6.46 Таблиця із списком відкріплених пацієнтів

4. Щоб прикріпити пацієнта до іншого лікуючого лікаря, треба натиснути на кнопку навпроти конкретного пацієнта у таблиці зі списком

відкріплених (рисунок 6.46), після чого з’явиться діалогове вікно для прикріплення (рисунок 6.47):

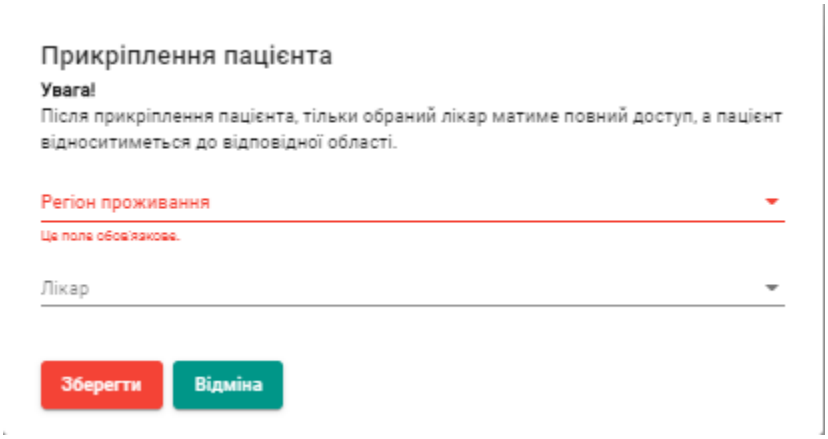


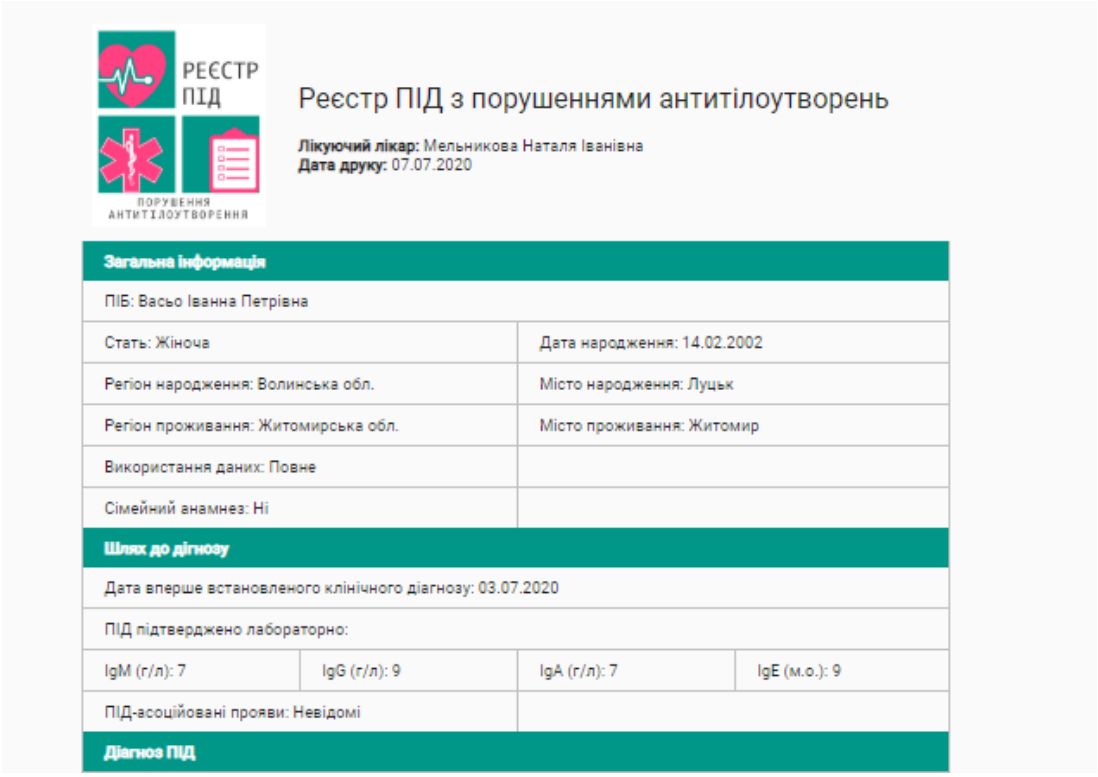
Рисунок 6.47 Діалогове вікно для зміни лікуючого лікаря

4.1. Експерт регіону обирає свій регіон та лікаря, який працює в цьому регіоні, виконує збереження, натискаючи кнопку Зберегти для зберігання інформації або копку Відміна для відхилення змін;

5. Після проведення операції зміни лікуючого лікаря. Обраний лікар матиме повний доступ до даних хворого.

VII. Друк інформації про пацієнта (доступ до друку персональних даних про хворого має лише лікар):

1. Для друку інформації про хворого необхідно у зведеній таблиці (рисунки 6.38,6.42) натиснути кнопку друк ;
2. З’явиться зведена інформація про усі дані хворого протягом спостережень, яка готова до роздруку (рисунок 6.48). Щоб роздрукувати потрібно натиснути комбінацію клавіш Ctrl+P (анг. розкладка клавіатури) та виконати операції у стандартній формі друку (рисунок 6.49).



Реєстр ПІД з порушеннями антитілоутворень

Лікуючий лікар: Мельникова Наталя Іванівна
Дата друку: 07.07.2020

Загальна інформація			
ПІБ: Васьо Іванна Петрівна			
Стать: Жіноча		Дата народження: 14.02.2002	
Регіон народження: Волинська обл.		Місто народження: Луцьк	
Регіон проживання: Житомирська обл.		Місто проживання: Житомир	
Використання даних: Повне			
Сімейний анамнез: Ні			
Шлях до дігнозу			
Дата вперше встановленого клінічного діагнозу: 03.07.2020			
ПІД підтверджено лабораторно:			
IgM (г/л): 7	IgG (г/л): 9	IgA (г/л): 7	IgE (м.о.): 9
ПІД-асоційовані прояви: Невідомі			
Діагноз ПІД			

Рисунок 6.48 Фрагмент відформатованого документу для друку

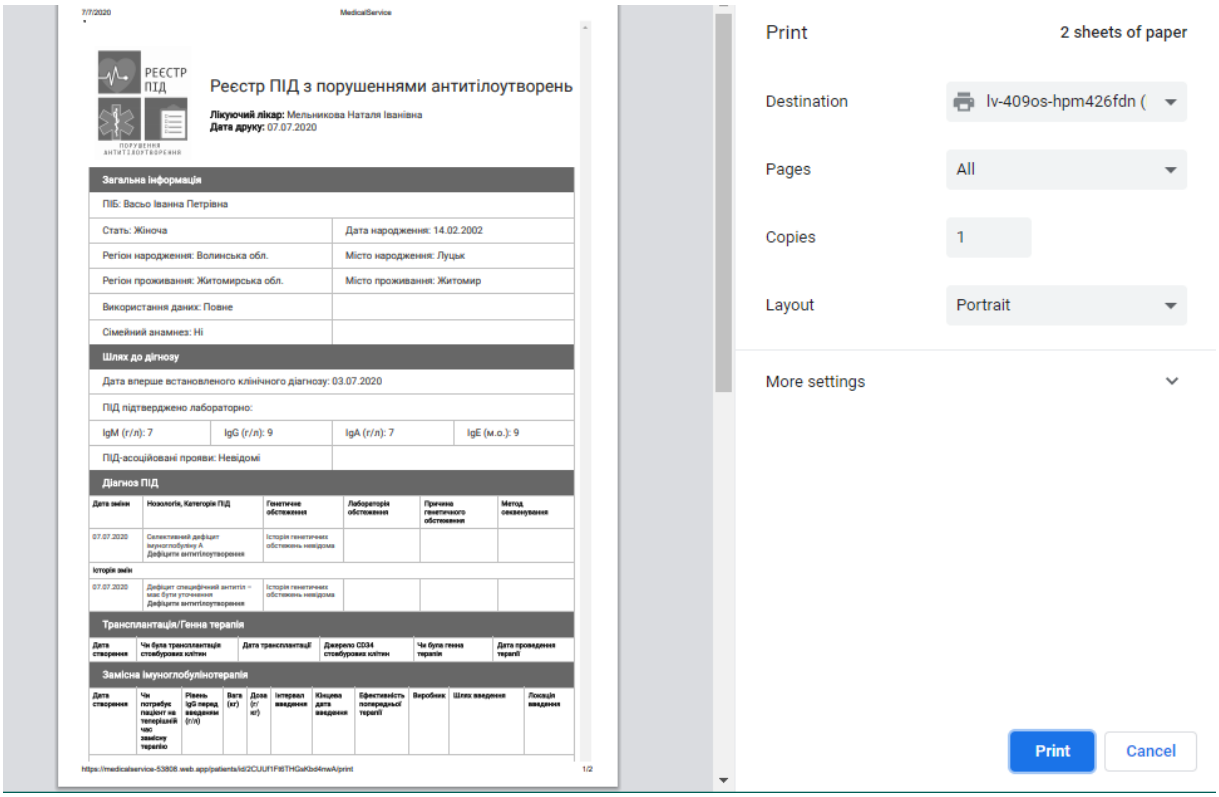


Рисунок 6.49 Стандартне вікно роздруку

VIII. Пошук пацієнтів за критеріями:

1. Для знаходження пацієнта в загальній таблиці підтверджених/непідтверджених/відкріплених потрібно внести дані у фільтр (рисунок 6.50) після чого натиснути кнопку Застосувати, або скинути дані фільтру, застосувавши Скинути:

1.1. Ідентифікатор пацієнта;

1.2. Вік;

1.3. Стать;

1.4. Область;

1.5. Нозологія;

1.6. Лікуючий лікар;

1.7. Спосіб ін’єкції;

1.8. Доза введення;

1.9. Вага;

1.10. Виробник.

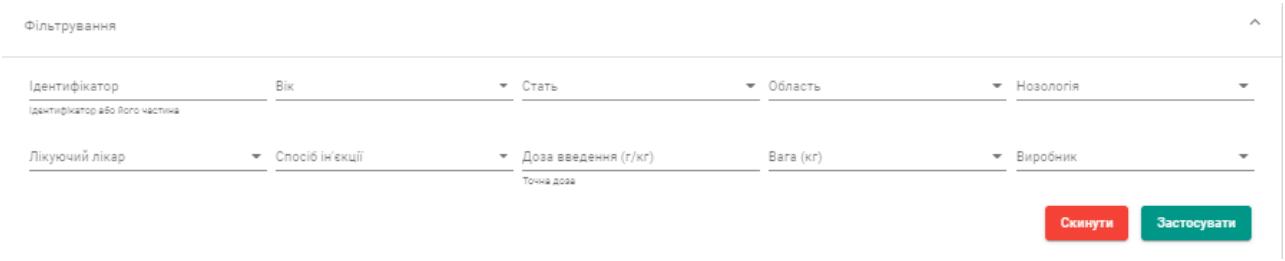


Рисунок 6.50 Фільт для пошуку пацієнта

6.2.2 Програмний модуль щодо прогнозування динаміки поширення COVID-19

Для налаштування системи обрана високорівнева об’єктно-орієнтована мова програмування Python, з застосуванням бібліотеки Pandas, що використовується як якісний, потенційний, простий та безвідмовний інструмент аналізу та обробки даних із відкритим вихідним кодом, Основне призначення якого, використання для обробки та аналізу даних. Для класифікації даних і оцінки моделі

використовувалася бібліотека scikit-learn — це простий та ефективний інструмент для прогнозного аналізу даних, доступний і придатний для повторного використання у різних контекстах. Ця бібліотека, заснована на NumPy, SciPy і matplotlib, і є відкритим кодом.

Щоб навчити модель, завжди потрібні навчальні дані. Попередньо були виконані обробка та дослідження даних, наступним етапом стала підготовка даних через усунення викидів та конвертування категоричних даних за допомогою одноразового кодувальника. Після цього дані були розділені на навчальні та тестувальні, тож у нас було 75 % навчальних даних та 25 % даних для тестування, 2325 об’єктів дослідження навчальних даних та 775 об’єктів для тестування.

Класифікацію проводили за допомогою п’яти класифікаторів, а саме: DecisionTreeClassifier, SVC, RandomForestClassifier, MultinomialNB і LogisticRegression. Усі класифікатори було імпортовано з бібліотеки sklearn.

Класифікація полягала в наступному. Спочатку були імпортовані вхідні дані, попередньо оброблені, закодовані, розділені на навчальні та тестові дані та далі передані до класифікаторів. Під час навчання обрані кращі параметри для кожного класифікатора для досягнення найвищої точності. Щоб оцінити модель, було визначено їхню точність за допомогою тестових даних і матриці помилок. У таблиці 6.1 наведені оцінки точності навчених класифікаторів.

Таблиця 6.1. Оцінки точності класифікаторів

Класифікатор	**Досягнута точність**
DecisionTreeClassifier	0.8645
RandomForestClassifier	0.8632
SVC	0.7781
LogisticRegression	0.7484
LogisticRegression	0.6194

Переглянувши таблицю 6.1, ми бачимо, що кращим виявився класифікатор DecisionTreeClassifier. У результаті роботи цього класифікатора було створено дерево, зображене на рисунку 6.51.

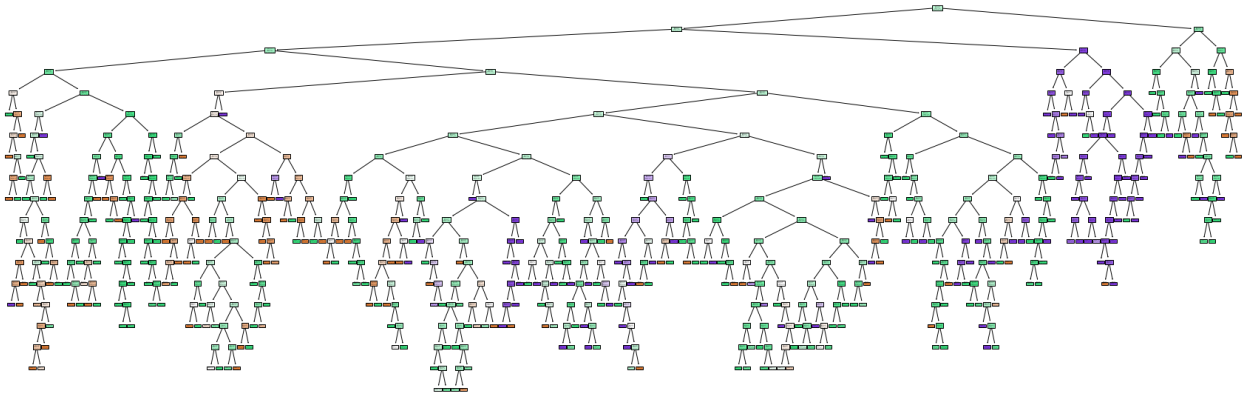


Рисунок 6.51. Результат роботи DecisionTreeClassifier

На рисунку 6.52 зображено результат застосування матриці помилок для здійснення статистичної оцінки. На рисунку 6.53 подано звіт за результатами класифікації.

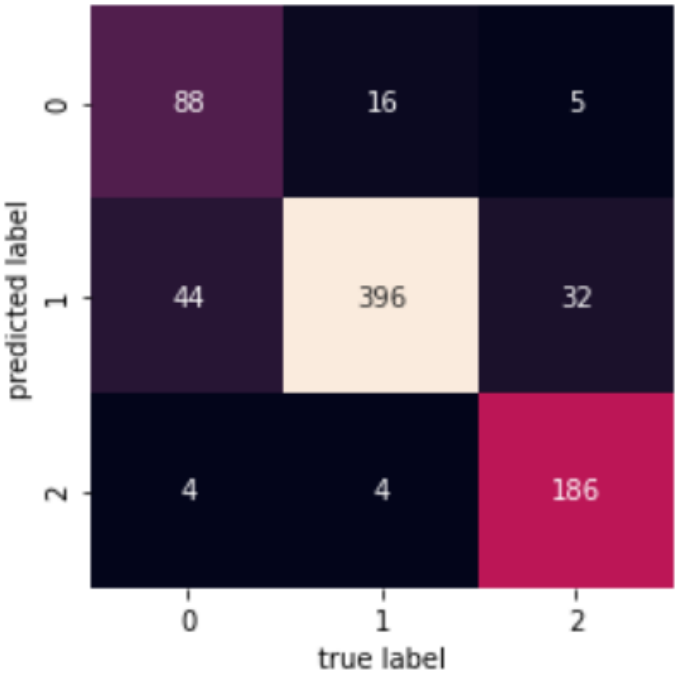


Рисунок 6.52 Матриця помилок

	precision	recall	f1-score	support
0	0.78	0.70	0.74	136
1	0.85	0.94	0.89	416
2	0.96	0.84	0.89	223
accuracy			0.87	775
macro avg	0.86	0.82	0.84	775
weighted avg	0.87	0.87	0.87	775

Рисунок 6.53 Результати класифікації

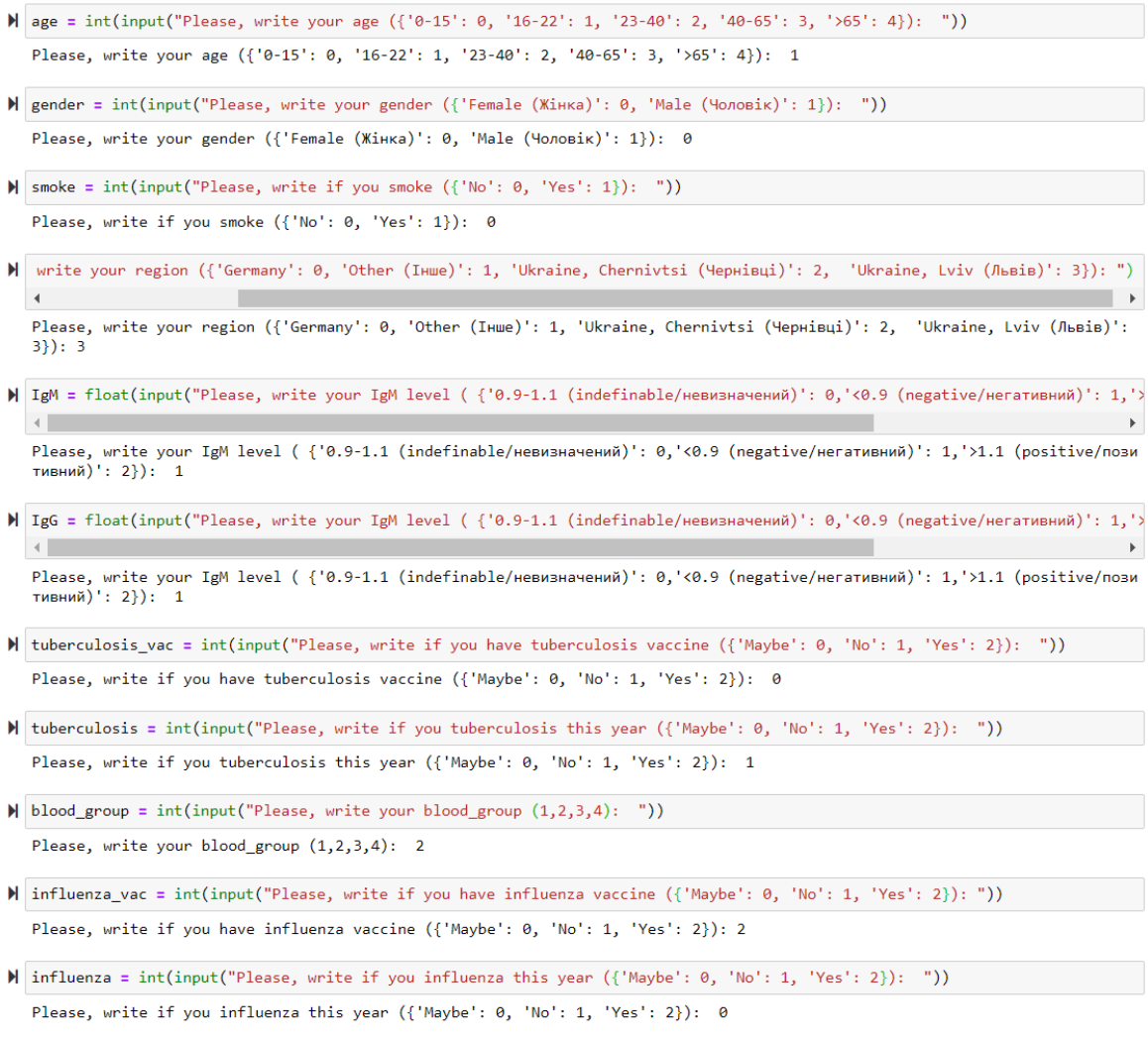
Зображення вище показують скільки прогнозів класифікатор отримав правильно, а скільки – ні. Крім того, на рисунку 6.54 подано ймовірності кожного передбачення, що належать до всіх трьох доступних класів («Можливо», «Ні», «Так»).

	Maybe	No	Yes
0	0.000000	1.000000	0.000000
1	0.100000	0.800000	0.100000
2	0.600000	0.400000	0.000000
3	0.000000	0.969697	0.030303
4	0.111111	0.888889	0.000000
5	0.000000	0.000000	1.000000
6	0.000000	1.000000	0.000000
7	0.555556	0.444444	0.000000
8	0.000000	1.000000	0.000000
9	0.000000	0.000000	1.000000

Рисунок 6.54 Ймовірності кожного передбачення

Після навчання кращий класифікатор було збережено для подальшого використання під час процесу прогнозування наявності COVID у пацієнтів. Повний код, написаний для реалізації необхідної системи, наведено в Додатку В.

Щоб перевірити збережений класифікатор на невідомих даних, людина повинна відповісти на кілька запитань. Після цього відповіді опрацьовуються і подаються на вхід класифікатора. Зображення запитань показано на рисунку 6.55.



й

Рисунок 6.55 Запитання опитника для прогнозування наявності COVID

Після відповідей на всі запитання збережений класифікатор повертає таблицю з ймовірністю прогнозування наявності COVID у кожному з трьох класів. На рисунку 6.53 відображено, що класифікатор для людини, яка відповіла на запитання, чиї відповіді показані на малюнку 6.55, дав найвищу ймовірність відповіді «Ні» – 0,65.

	Maybe	No	Yes
0	0.3125	0.65	0.0375

.

Рисунок 6.56 Результат прогнозування наявності COVID

6.2.3 Програмний модуль щодо виявлення аномалій у пацієнтів

Для реалізації системи для виявлення аномалій необхідно створити інтерфейс за допомогою якого користувач зможе:

1. Стягувати дані, що зберігаються в базі;
2. Виконати для них певне узагальнення;
3. Передача оброблених даних на вхід HTM мережі;
4. Запис результатів у базу даних;
5. Візуалізація метричних даних і результатів аналізу на графіках.

Для досягнення поставлених цілей пропонується створити бібліотеку, яка б дала користувачу можливість виконувати увесь перелічений функціонал за допомогою функцій цієї бібліотеки.

Реалізація полягає в наступному:

1. Оновити версії бібліотек, які використовувались у якості залежностей;
2. Оновити програмні рішення, відповідно до нової версії бібліотек;
3. Виправити помилки після використання підходів пошуку аномалій;
4. Додати нового необхідного функціоналу;
5. Оновити бібліотеки із загальнішою структурою даних.

6.2.3.1 Бібліотека для синхронізації з базою даних для аналізу ознак

Через клас InfluxHtmClient буде відбуватись вся взаємодія із базою даних influxDB і він буде воротами до решти функціоналу influxhtm API. Для того, щоб коректно користуватись даним інтерфейсом необхідно реалізувати:

1. Метод, який буде повертати InfluxDBClient;
2. Метод який буде повертати список з усіх послідовностей, що є у базі даних незалежно від значення measurement і tags;
3. Метод, який буде виконувати запити SELECT до бази;
4. Два методи для ініціалізації класу Sensor і HtmSensorModel;
5. Два методи для того, щоб отримувати об’єкти класу Sensor і HtmSensorModel за певними атрибутами, такими як measurement і tags, а для HtmSensorModel ще і id.

Вище наведений список дає зрозуміти ключові функції класу InfluxHtmClient. За допомогою нього два інші класи зможуть виконувати стягнення даних із бази.

6.2.3.2 Клас для підключення моделі HTM.

За допомогою класу Sensor дані будуть отримуватись, зберігатись і записуватись. Також він дає можливість підключати до цих даних HTM модель. Ключовими методами даного класу будуть:

1. Повернення значень measurement і tags, що ідентифікують послідовність яку представляє даний клас;
2. Запис даних у базу через InfluxDBClient і його метод write_point, який був ініціалізований у класі InfluxHtmClient;
3. Стягування даних із бази за відповідними значеннями measurement і tags;
4. Два методи для створення класу HtmSensorModel для відповідного сенсору та його видалення з метою використання нових значень агрегування для даних. Видалити можна як усі об’єкти цього класу, так і конкретний за значенням його id.
5. Метод для отримання всіх вхідних даних і HTM результатів в однакову структуру для легкої побудови графіків.

Висновки до 6 розділу

1. У розділі розроблено архітектуру системи підтримки прийняття медичних рішень щодо прогнозування станів пацієнта на підставі опрацювання та аналізу персоналізованих медичних даних.
2. Подано опис імплементації рішень у системі підтримки прийняття рішень для лікування хворих на ПІД з порушеннями антитілоутвореннями. Використання цієї системи дає змогу вести облік хворих, контролювати і супроводжувати їхнє лікування, досліджувати динаміку перебігу хвороби та зміни у стані пацієнта.
3. Подано опис імплементації програмного модуля щодо прогнозування динаміки поширення COVID-19 з урахуванням процесу класифікації пацієнтів відповідно до стану, проведення оцінки помилок одержаних рішень.
4. Подано результати імплементації системи для виявлення помилок, що дає змогу експериментувати із різними параметрами для моделі, знову і знову на одних і тих же самих даних. За допомогою цього користувач зможе вибрати модель, яка краще описує дані і має кращі результати: чим менший часовий інтервал між точками, тим точніша модель, яка може визначити час, коли виникла аномалія. Проте велика точність потребує більшої кількості обчислень, за допомогою даних графіків користувач зможе визначити оптимальний варіант.

ВИСНОВКИ

За результатами проведеного дослідження вирішено важливу науково-прикладну проблему розроблення та удосконалення моделей, методів і засобів машинного навчання в задачах класифікації, кластеризації, прогнозування та візуалізація результатів опрацювання персональних даних для адаптації медичних рішень до пацієнта. У результаті виконання цієї роботи одержані наступні результати:

1. Здійснено аналіз процесів опрацювання мультимодальних персоналізованих медичних даних, даних попередньої обробки та аналізу якості моделі даних, що дозволило визначити проблему та задачі дослідження.
2. Обґрунтовано актуальність розв’язання цієї проблеми на основі введення моделі відображення стану пацієнта в багатовимірному просторі умови, які за рахунок додаткових вимірів та параметрів забезпечила підвищення точності прогнозування цільових змінних та дала змогу підвищити точність прогнозування цільових показників у підмножині простору умов на 5 %, що забезпечило індивідуальний підхід до моніторингу стану пацієнта на основі тривалого спостереження та контролю лікаря.
3. Вперше розроблено метод двоетапної обробки даних на основі ієрархічного предиктора, що забезпечує підвищення точності процесу узагальнення результатів на малих наборах даних у порівнянні з результатами логістичної регресії, як кращого класифікатора для набору даних по COVID 19 та результатами поліноміального методу опорних векторів, як кращого класифікатора для набору даних по орфанних хворобах за рахунок ієрархічної класифікації та поєднання різних моделей машинного навчання.
4. Вперше розроблено метод групування моделей машинного навчання, як стекінгову модель для забезпечення точності прогнозування даних на 7-9 % та паралелізації процесу обробки даних, що на відміну від існуючих моделей убезпечує універсальність пошуку рішень щодо малих та великих наборів персоналізованих даних.

5. Вперше розроблено метод зменшення розмірності вхідних даних на основі ансамблю моделей слабких предикторів для вибору найважливіших ознак, що базується на прирості інформації з урахуванням результатів застосування декількох селекторів та агрегації кінцевих результатів на підставі урахування індекса Жакара для оцінки подібності одержаних підмножин ознак та проведення мажоритарного голосування результатів, що дає змогу уникнути кореляції результатів слабких предикторів і збільшує узагальнення моделі.
6. Удосконалено метод заповнення пропусків даних, який ґрунтується на пошуку подібності даних, що можуть впливати на доменні значення та функціональні залежності між ними і забезпечує їхнє включення у навчальні дані, що дає змогу забезпечити стійкість до помилок даних, паралелізації обчислень та аналізу різнотипних даних. Розроблений метод заповнення даних на 12 % покращує результати на відміну від моделі Випадкового Лісу (Rendom Forest) та Очікування-Максимізацією (Expectation-Maximization) для 30 % відсутніх даних.
7. Удосконалено метод пошуку зміни стану пацієнта шляхом використання простору умов та аналізу зміни приросту значень часовозалежних даних, що на відміну від методу упорядкованого пошуку надає чіткості та направленості у пошуку рішень стосовно вибору цільових схем лікування, що дає змогу зменшити ймовірність появи похибки (кількість ліжкоднів) при виборі схеми лікування за рахунок агрегованого опрацювання даних з різних просторів моделі.
8. Удосконалено метод консолідації мультимодальних даних за рахунок попереднього визначення структур даних та узгодження семантики, що на відміну від методів консолідації даних, на рівні сховища даних, забезпечило агрегування даних з різною структурою.
9. Розроблені методи використані для розроблення програмних модулів інформаційних систем підтримки прийняття рішень у лікуванні орфанних хвороб та для підтримки прийняття рішень щодо адаптації схеми лікування пацієнтів з постковідним ефектом.

10. Практичне значення результатів дисертаційного дослідження дають змогу підвищити точність прогнозування цільових показників у підмножині простору умов на 5 %, підвищити точність класифікації на 4 % порівняно з результатами логістичної регресії як кращого класифікатора у порівнянні з існуючими для набору даних по Covid-19, підвищити точність прогнозування даних, на 7-9 % та забезпечити паралелізацію процесу обробки даних як малої, так і великої розмірності, забезпечити заповнення пропусків даних за рахунок одержання додаткових значень, а також зменшити ймовірність появи похибки (кількість ліжкоднів) при пошуку шаблонів динаміки стану пацієнта шляхом використання простору умов.
11. Достовірність наукових та практичних результатів підтверджується за рахунок відповідних матеріалів про впровадження дисертаційних досліджень, а також за рахунок порівняння одержаних практичних результатів з результатами застосування існуючих класичних методів та підходів щодо опрацювання персоналізованої інформації про пацієнта.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Good, B. M., Ainscough, B. J., McMichael, J. F., Su, A. I., & Griffith, O. L. (2014). Organizing knowledge to enable personalization of medicine in cancer. Genome biology, 15(8), 1-9.;
2. ‘Personalized Medicine’. Genome.Gov, https://www.genome.gov/genetics-glossary/Personalized-Medicine.;
3. ‘Personalized Employee Experience Can Help with L&D Goals’. Gartner, https://www.gartner.com/smarterwithgartner/3-ways-personalization-can-improve-the-employee-experience.;
4. Cambridge Dictionary | English Dictionary, Translations & Thesaurus. https://dictionary.cambridge.org/.;
5. What Is Personalization and What Does It Mean to Digital Marketers? 29 Nov. 2018, https://instapage.com/blog/what-is-personalization.;
6. ‘Personalized Employee Experience Can Help with L&D Goals’. Gartner, https://www.gartner.com/smarterwithgartner/3-ways-personalization-can-improve-the-employee-experience.;
7. Sullivan, Frost &. ‘From $600 M to $6 Billion, Artificial Intelligence Systems Poised for Dramatic Market Expansion in Healthcare’. Frost and Sullivan, 5 Jan. 2016, https://www.frost.com/news/press-releases/600-m-6-billion-artificial-intelligence-systems-poised-dramatic-market-expansion-healthcare/.;
8. Ada. Survey, https://ada.com/.;
9. ‘Sensely: Conversational AI to Improve Health and Drive Member Engagement’. Sensely, https://sensely.com/.;
10. ‘SOPHiA GENETICS - Where Others See Data We See Answers’. SOPHiA GENETICS, https://www.sophiagenetics.com/.;
11. Pharmacogenomics and Personalized Medicine | NorthShore. https://www.northshore.org/pharmacogenomics/.;
12. Svitlyk, Yuri. ‘Штучний інтелект проти COVID-19’. Root Nation, 14 May 2020, https://root-nation.com/ua/articles-ua/tech-ua/ua-shtuchnij-intelekt-covid-19/.;

13. ‘China’s Facial-Recognition Giant Says It Can Crack Masked Faces during the Coronavirus’. Quartz, 18 Feb. 2020, https://qz.com/1803737/chinas-facial-recognition-tech-can-crack-masked-faces-amid-coronavirus/.;
14. Kuznietsova N. V., and Bidyuk P. I. ‘Business intelligence techniques for missing data imputation’, Наукові вісті Національного технічного університету України Київський політехнічний інститут, 2015, no. 5, pp.47-56.;
15. Y. Tang, Y. Wang, K. M., and Cooper, L. Li, ‘Towards big data Bayesian network learning-an ensemble learning based approach’, 2014 IEEE International Congress on Big Data. IEEE, 2014, pp.355-357.;
16. Lakho, Shamshad, et al. ‘Decision Support System for Hepatitis Disease Diagnosis using Bayesian Network’. Sukkur IBA Journal of Computing and Mathematical Sciences, vol. 2, no 1, 2017, pp. 11-19.;
17. 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, vol. 51, Aug. 2014, pp. 140–58. , https://doi.org/10.1016/j.compbiomed.2014.04.010.;
18. Perova, Iryna, and Yevgeniy Bodyanskiy. ‘FAST MEDICAL DIAGNOSTICS USING AUTOASSOCIATIVE NEURO-FUZZY MEMORY’. International Journal of Computing, Mar. 2017, pp. 34–40. , https://doi.org/10.47839/ijc.16.1.869.;
19. 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.;
20. 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, pp. 151-157.;
21. Silva-Ramírez, Esther-Lydia, 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, Apr. 2015, pp. 65–74. , https://doi.org/10.1016/j.asoc.2014.09.052.;

22. Chiacchio, Ferdinando, et al. ‘Coherence Region of the Priority-AND Gate: Analytical and Numerical Examples’. Quality and Reliability Engineering International, vol. 34, no. 1, Feb. 2018, pp. 107–15. , https://doi.org/10.1002/qre.2241.;
23. 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.;
24. 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.;
25. 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.
26. 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.;
27. 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.;
28. 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.;
29. 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.;
30. 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.;

31. 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.;
32. 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.;
33. 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. , https://doi.org/10.12720/jait.13.5.470-476.;
34. 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.;
35. Бідюк, П. І., et al. ‘Методи прогнозування’. Луганськ, Альма-матер, no. 1, 2008, pp. 308.;
36. 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, pp. 31-36.;
37. 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.;
38. 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.;
39. 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.;
40. Perova I, and Bodyanskiy Y. ‘Fast Medical Diagnostics Using Autoassociative Neuro-Fuzzy Memory’. International Journal of Computing, vol. 1, no. 16, 2017, pp. 34-40.;
41. 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.;

42. 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.;
43. 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.;
44. Chiacchio F., et. al. 'Coherence region of the Priority-AND gate: 5 Analytical and numerical examples'. Quality and Reliability Engineering International, vol.34, no.1, 2018, pp.107-115.;
45. 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, vol. 40, no. 23-24, 2016, pp. 10002-10019.;
46. Kadri, H., et. al. 'S. Formal approach to control design of complex and dynamical systems. Procedia Computer Science, no. 108, 2017, pp. 2512-2516.;
47. Masic, I., et. al. 'Evidence based medicine–new approaches and challenges'. Acta Informatica Medica, vol. 4, no. 16, 2008, pp. 219.;
48. 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.;
49. 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.;
50. 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.;
51. Pakhira, M. K. 'Finding number of clusters before finding clusters'. Procedia Technology, vol. 4, 2012,pp. 27-37.;
52. 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.;

53. 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.;
54. Katzir, Liran, and Stephen J. Hardiman. ‘Estimating clustering coefficients and size of social networks via random walk’. ACM Transactions on the Web (TWEB) 9.4, 2015, no. 19.;
55. Newman, M. E. ‘Mixing patterns in networks’. Physical Review E, vol. 2, no. 67, 2015, pp. 126.;
56. 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.;
57. Wirth, R., and Hipp, J. “CRISP-DM: Towards a standard process model for data mining”. In Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining, 2000, pp. 29-39.;
58. Shakhovska, N., et al. ‘Association rules mining in big data’. International Journal of Computing, vol. 1, no. 17, 2018, pp. 25-32.;
59. Hunyadi D. ‘Performance comparison of Apriori and FP-Growth algorithms in generating association rules’. European Computing Conference, 2011, pp. 376-381.;
60. Osman, A.M. “A novel big data analytics framework for smart cities”. Future Generation Computer Systems, 91, 2019, pp. 620-633.;
61. Maulik, U., and Bandyopadhyay, S. ‘Genetic algorithm-based clustering technique’. Pattern recognition, vol. 9, no. 33, 2000, pp. 1455-1465.;
62. Ramírez-Rubio, Rogelio, et al. ‘Pattern classification using smallest normalized difference associative memory’. Pattern Recognition Letters, no. 93, 2017, pp. 104-112.;
63. Baccouche M., et al. ‘Sequential Deep Learning for Human Action Recognition’. Human Behavior Understanding., HBU 2011., Lecture Notes in Computer Science, vol. 7065.;

64. Waring, Jonathan, et al. “Automated Machine Learning: Review of the State-of-the-Art and Opportunities for Healthcare.” Artificial Intelligence in Medicine, vol. 104, Apr. 2020, pp. 101822. , doi:10.1016/j.artmed.2020.101822.;
65. Vitynskyi, P., et. al. ‘Hybridization of the SGTM Neural-like Structure through Inputs Polynomial Extension’. Proceedings of the Second International Conference on Data Stream Mining Processing (DSMP), pp. 386-391.;
66. Ng, Man-Fai, et al. “Predicting the State of Charge and Health of Batteries Using Data-Driven Machine Learning.” Nature Machine Intelligence, vol. 2, no. 3, Mar. 2020, pp. 161–70. , doi:10.1038/s42256-020-0156-7.;
67. Bohdan Ye. Rytsar. ‘A New Method for Symmetry Recognition in Boolean Functions Based on the Set-Theoretical Logic Differentiation’. Control Systems and Computers, vol. 282, no. 4, Oct. 2019, pp. 3–13. , doi:10.15407/csc.2019.04.003.;
68. Syerov, Yu., et. al. ‘Method of the Data Adequacy Determination of Personal Medical Profiles. Proceedings of the International Conference of Artificial Intelligence’. Medical Engineering, Education (AIMEE2018), Advances in Artificial Systems for Medicine and Education II, Series Vol. 902. 2019.;
69. E. Molnár, R. Molnár, et. al. ‘Web Intelligence in practice’. The Society of Service Science, Journal of Service Science Research, Springer, vol. 6, no. 1, 2014, pp. 149-172.;
70. Kryvinska, N, and Gregus, M. ‘SOA and it's Business Value in Requirements, Features, Practices and Methodologies’, 2014, Comenius University in Bratislava, ISBN: 9788022337649.;
71. Tkachenko, R., and Izonin, I. ‘Model and Principles for the Implementation of Neural-Like Structures based on Geometric Data Transformations’. Advances in Computer Science for Engineering and Education. ICCSEEA2018. Advances in Intelligent Systems and Computing. Springer, Cham, vol.754, 2019, pp.578-587, https://doi.org/10.1007/978-3-319-91008-6_58
72. Tkachenko, Roman, et. al. ‘Piecewise-linear Approach for Medical Insurance Costs Prediction using SGTM Neural-Like Structure’. Informatics & Data-Driven Medicine

(IDDM 2018), Proceedings of the 1st International Workshop IDDM 2018, Lviv, Ukraine, November 28-30, 2018, pp.170-179, CEUR-WS.org.;

73. Tkachenko, Roman, et. al. ‘Development of the Non-Iterative Supervised Learning Predictor Based on the Ito Decomposition and SGTM Neural-Like Structure for Managing Medical Insurance Costs+’. Data, vol. 3, no. 4, Oct. 2018, pp. 1-14.;
74. Mulesa P., and Perova I. ‘Fuzzy Spacial Extrapolation Method Using Manhattan Metrics for Tasks of Medical Data Mining’. Computer Science and Information Technologies CSIT’2015, Lviv, Ukraine., 2015, p. 104-106.;
75. Perova, I., et. al. ‘Neo-Fuzzy Approach for Medical Diagnostics Tasks in Online-Mode’. 2016 IEEE First International Conference on Data Stream Mining & Processing (DSMP), Lviv, Ukraine, 2016, p. 34-38.;
76. Bodyanskiy, Yevgeniy, et al. ‘Adaptive Wavelet Diagnostic Neuro-Fuzzy Network for Biomedical Tasks’. 2018 14th International Conference on Advanced Trends in Radioelecrtronics, Telecommunications and Computer Engineering (TCSET), IEEE, 2018, pp. 711–715. , https://doi.org/10.1109/TCSET.2018.8336299.;
77. Izonin, I., et al. ‘The Combined Use of the Wiener Polynomial and SVM for Material Classification Task in Medical Implants Production’. International Journal of Intelligent Systems and Applications, vol. 10, no. 9, Sept. 2018, pp. 40–47. , https://doi.org/10.5815/ijisa.2018.09.05.
78. Boyko, Nataliya, et al. ‘Use of Machine Learning in the Forecast of Clinical Consequences of Cancer Diseases’. 2018 7th Mediterranean Conference on Embedded Computing (MECO), IEEE, 2018, pp. 1–6. , https://doi.org/10.1109/MECO.2018.8405985.;
79. Gregus, M. and Kryvinska. N. ‘Service Orientation of Enterprises - Aspects, Dimensions, Technologies’. 2015, Comenius University in Bratislava, ISBN: 9788022339780.;
80. Kharkiv Medical Academy of Postgraduate Education, et al. ‘About Functions of a Clinical Component of Postgraduate Education and Continuous Professional Development of Doctors’. Problems of Uninterrupted Medical Training and Science,

vol. 2018, no. 4, Nov. 2018, pp. 69–72. , https://doi.org/10.31071/promedosvity2018.04.069.;

81. Kussul, N., et al. ‘Intelligent Multi-Agent Information Security System’. International Journal of Computing, Aug. 2014, pp. 35–39. , https://doi.org/10.47839/ijc.2.2.202.;
82. Dunham, Margaret H. Data Mining Introductory and Advanced Topics. Prentice Hall/Pearson Education, 2003.;
83. Ashbacher, Charles. ‘The Art of Computer Programming, Volume 4: Generating All Trees, History of Combinatorial Generation.’ The Journal of Object Technology, vol. 6, no. 1, 2007, p. 65. , https://doi.org/10.5381/jot.2007.6.1.r1.;
84. Perova, Iryna, et al. ‘Medical Data-Stream Mining in the Area of Electromagnetic Radiation and Low Temperature Influence on Biological Objects’. 2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP), IEEE, 2018, pp. 3–6, https://doi.org/10.1109/DSMP.2018.8478577.;
85. Perova, Iryna, et al. ‘Neural Network for Online Principal Component Analysis in Medical Data Mining Tasks IEEE First International Conference on System Analysis & Intelligent Computing (SAIC) 8-12 October 2018, Kyiv, Ukraine, pp.150-154.;
86. Perova, Iryna, and Bodyanskiy, Ye. ‘Fast medical diagnostics using autoassociative neuro-fuzzy memory // International Journal of Computing, vol. 1, no. 16, 2017, pp. 34-40.;
87. https://root-nation.com/ua/articles-ua/tech-ua/ua-shtuchnij-intelekt-covid-19/
88. Jullo, E., et al. ‘A Bayesian Approach to Strong Lensing Modelling of Galaxy Clusters’. New Journal of Physics, vol. 9, no. 12, Dec. 2007, pp. 447–447. , https://doi.org/10.1088/1367-2630/9/12/447.;
89. S. Kaczor, and N. Kryvinska, ‘It is all about Services - Fundamentals, Drivers, and Business Models’. The Society of Service Science, Journal of Service Science Research, Springer, vol. 5, `no. 2, 2013, pp. 125-154.;
90. Shakhovska, Nataliya, et al. ‘ASSOCIATION RULES MINING IN BIG DATA’. International Journal of Computing, Mar. 2018, pp. 25–32. , https://doi.org/10.47839/ijc.17.1.946.;

91. Kryvinska, N. ‘Building Consistent Formal Specification for the Service Enterprise Agility Foundation’. The Society of Service Science, Journal of Service Science Research, Springer, vol. 4, no. 2, 2012, pp. 235-269.;
92. Boyko, Nataliya, et. al. ‘Modeling of the information system for processing of a large distilled data for the investigation of competitiveness of enterprises’. Kyiv, Ukraine, November 30, 2019. pp. 571-59.;
93. Demongeot, Jacques, et. al. ‘Temperature decreases spread parameters of the new Covid-19 case dynamics’. Biology, no. 9.5, 2020, pp. 94.;
94. Fukuda, Haruhisa, et al. ‘Healthcare Expenditures for the Treatment of Patients Infected with Hepatitis C Virus in Japan’. PharmacoEconomics, vol. 38, no. 3, Mar. 2020, pp. 297–306. , https://doi.org/10.1007/s40273-019-00861-x.;
95. Pham, Quoc-Viet, et al. ‘Artificial Intelligence (AI) and Big Data for Coronavirus (COVID-19) Pandemic: A Survey on the State-of-the-Arts’. IEEE Access, vol. 8, 2020, pp. 130820–39. , https://doi.org/10.1109/ACCESS.2020.3009328.;
96. Mochurad, Lesia, and Albota Solomiia. ‘Optimizing the Computational Modeling of Modern Electronic Optical Systems’. Lecture Notes in Computational Intelligence and Decision Making, edited by Volodymyr Lytvynenko et al., vol. 1020, Springer International Publishing, 2020, pp. 597–608. , https://doi.org/10.1007/978-3-030-26474-1_41.;
97. Mochurad, Lesia, et al. ‘Parallel Solving of Fredholm Integral Equations of the First Kind by Tikhonov Regularization Method Using OpenMP Technology’. Advances in Intelligent Systems and Computing IV, edited by Natalya Shakhovska and Mykola O. Medykovskyy, vol. 1080, Springer International Publishing, 2020, pp. 25–35. , https://doi.org/10.1007/978-3-030-33695-0_3.;
98. Peleshko, Dmytro, et al. ‘Design and Implementation of Visitors Queue Density Analysis and Registration Method for Retail Videosurveillance Purposes’. 2016 IEEE First International Conference on Data Stream Mining & Processing (DSMP), IEEE, 2016, pp. 159–62. , https://doi.org/10.1109/DSMP.2016.7583531.;

99. Sujath, R., et al. ‘A Machine Learning Forecasting Model for COVID-19 Pandemic in India’. Stochastic Environmental Research and Risk Assessment, vol. 34, no. 7, July 2020, pp. 959–72. , https://doi.org/10.1007/s00477-020-01827-8.;
100. Khamparia, Aditya, et al. ‘Internet of Health Things-Driven Deep Learning System for Detection and Classification of Cervical Cells Using Transfer Learning’. The Journal of Supercomputing, vol. 76, no. 11, Nov. 2020, pp. 8590–608. , https://doi.org/10.1007/s11227-020-03159-4.;
101. Waheed, Abdul, et al. ‘CovidGAN: Data Augmentation Using Auxiliary Classifier GAN for Improved Covid-19 Detection’. IEEE Access, vol. 8, 2020, pp. 91916–23. , https://doi.org/10.1109/ACCESS.2020.2994762.
102. Sakarkar, Gopal, et al. ‘Comparative Study of Ambient Air Quality Prediction System Using Machine Learning to Predict Air Quality in Smart City’. Proceedings of International Conference on IoT Inclusive Life (ICIIL 2019), NITTTR Chandigarh, India, edited by Maitreyee Dutta et al., vol. 116, Springer Singapore, 2020, pp. 175–82. , https://doi.org/10.1007/978-981-15-3020-3_16.;
103. Peters James F. ‘Computational Intelligence In Software Engineering, Advances In Fuzzy Systems.’ Applications And Theory, World Scientific, 1998, pp. 500.
104. Франклін М. ‘Принципи систем простору даних. Архів симпозіуму з принципів систем баз даних’. Матеріали двадцять п'ятого симпозіуму ACM SIGMOD–SIGACT–SIGART з принципів систем баз даних зміст, Чикаго, Нью-Йорк, Іллінойс, США, 2006. Нью-Йорк: ACM, 2006, pp. 1–9.;
105. Do Hong–Ha. ‘Comparison of Schema Matching Evaluations’. Proceedings of Web, Web–Services, and Database Systems, Erfurt, October 2002, 2002, pp. 221–237.;
106. Franklin, Michael, et al. ‘A First Tutorial on Dataspaces’. Proceedings of the VLDB Endowment, vol. 1, no. 2, Aug. 2008, pp. 1516–17. , https://doi.org/10.14778/1454159.1454217.;
107. Шаховська Н. ‘Методи опрацьовування консолідованих даних za допомогою просторів даних’. Проблеми програмування: науково - технічний журнал,

Національна академія наук України, Інститут програмних системи NAN Україна, 2011, no. 4. pp. 72-84.;

108. Шаховська Н. ‘Застосування алгоритмів класифікації для зменшення невизначеності’. Zastosowania algorytmów klasyfikacji do minimalizacji niepewności, Przeglad Elektrotechniczny, 2013, pp. 284.;

109. Harfsheno, M., et al. ‘Patients with Down Syndrome in the Coronavirus Pandemic (COVID-19)’. Armaghane Danesh, vol. 27, no. 3, Apr. 2022, pp. 407–17. , https://doi.org/10.52547/armaghanj.27.3.407.;

110. Gaudart, Jean, et. al. ‘Demographic and Spatial Factors as Causes of an Epidemic Spread, the Copule Approach: Application to the Retro-prediction of the Black Death Epidemy of 1346, 24th IEEE International Conference on Advanced Information Networking and Applications Workshops, WAINA 2010, pp. 751-758. 10.1109/WAINA.2010.79.

111. ‘Пандемія коронавірусу: скільки тестів зробили в Україні та інших країнах світу’. Слово і Діло, https://www.slovoidilo.ua/2020/09/04/infografika/suspilstvo/pandemiya-koronavirusu-skilky-testiv-zrobyly-ukrayini-ta-inshyx-krayinax-svitu. Accessed 9 Nov. 2022.;

112. Vyklyuk, Yaroslav, et al. ‘Modeling and Analysis of Different Scenarios for the Spread of COVID-19 by Using the Modified Multi-Agent Systems – Evidence from the Selected Countries’. Results in Physics, vol. 20, Jan. 2021, p. 103662. , https://doi.org/10.1016/j.rinp.2020.103662.;

113. Izonin, Ivan, et al. ‘An Approach towards Missing Data Management Using Improved GRNN-SGTM Ensemble Method’. Engineering Science and Technology, an International Journal, vol. 24, no. 3, June 2021, pp. 749–59. , https://doi.org/10.1016/j.jestch.2020.10.005.;

114. Jiang, Chao, et al. ‘Comparative Review of Respiratory Diseases Caused by Coronaviruses and Influenza A Viruses during Epidemic Season’. Microbes and Infection, vol. 22, no. 6–7, July 2020, pp. 236–44. , https://doi.org/10.1016/j.micinf.2020.05.005.;

115. Charpentier, Charlotte, et al. ‘Performance Evaluation of Two SARS-CoV-2 IgG/IgM Rapid Tests (Covid-Presto and NG-Test) and One IgG Automated Immunoassay (Abbott)’. Journal of Clinical Virology, vol. 132, Nov. 2020, p. 104618. , https://doi.org/10.1016/j.jcv.2020.104618.;

116. Muhammad, L. J., et al. ‘Predictive Data Mining Models for Novel Coronavirus (COVID-19) Infected Patients’ Recovery’. SN Computer Science, vol. 1, no. 4, July 2020, p. 206. , https://doi.org/10.1007/s42979-020-00216-w.;

117. Ivorra, B., et al. ‘Mathematical Modeling of the Spread of the Coronavirus Disease 2019 (COVID-19) Taking into Account the Undetected Infections. The Case of China’. Communications in Nonlinear Science and Numerical Simulation, vol. 88, Sept. 2020, p. 105303. , https://doi.org/10.1016/j.cnsns.2020.105303.;

118. Caruana, G., et al. ‘Diagnostic Strategies for SARS-CoV-2 Infection and Interpretation of Microbiological Results’. Clinical Microbiology and Infection, vol. 26, no. 9, Sept. 2020, pp. 1178–82. , https://doi.org/10.1016/j.cmi.2020.06.019.;

119. Ghosal, Samit, et al. ‘Linear Regression Analysis to Predict the Number of Deaths in India Due to SARS-CoV-2 at 6 Weeks from Day 0 (100 Cases - March 14th 2020)’. Diabetes & Metabolic Syndrome: Clinical Research & Reviews, vol. 14, no. 4, July 2020, pp. 311–15. , https://doi.org/10.1016/j.dsx.2020.03.017.;

120. Yang, Qiuying, et al. ‘Research on COVID-19 Based on ARIMA ModelΔ—Taking Hubei, China as an Example to See the Epidemic in Italy’. Journal of Infection and Public Health, vol. 13, no. 10, Oct. 2020, pp. 1415–18. , https://doi.org/10.1016/j.jiph.2020.06.019.;

121. Petukhova, Tatiana, et al. ‘Assessment of Autoregressive Integrated Moving Average (ARIMA), Generalized Linear Autoregressive Moving Average (GLARMA), and Random Forest (RF) Time Series Regression Models for Predicting Influenza A Virus Frequency in Swine in Ontario, Canada’. PLOS ONE, edited by Jeffrey Shaman, vol. 13, no. 6, June 2018, p. e0198313. , https://doi.org/10.1371/journal.pone.0198313.;

122. Adhikari, Ratnadip, and R. K. Agrawal. An Introductory Study on Time Series Modeling and Forecasting. 1. Aufl, LAP LAMBERT Academic Publishing, 2013.;

123. Martinez, Edson Zangiacomi, et al. ‘A SARIMA Forecasting Model to Predict the Number of Cases of Dengue in Campinas, State of São Paulo, Brazil’. Revista Da Sociedade Brasileira de Medicina Tropical, vol. 44, no. 4, Aug. 2011, pp. 436–40. , https://doi.org/10.1590/S0037-86822011000400007.;

124. Dehesh, Tania, et al. Forecasting of COVID-19 Confirmed Cases in Different Countries with ARIMA Models. preprint, Epidemiology, 18 Mar. 2020. , https://doi.org/10.1101/2020.03.13.20035345.;

125. Martinez, Edson Zangiacomi, and Elisângela Aparecida Soares da Silva. ‘Predicting the Number of Cases of Dengue Infection in Ribeirão Preto, São Paulo State, Brazil, Using a SARIMA Model’. Cadernos de Saúde Pública, vol. 27, no. 9, Sept. 2011, pp. 1809–18. , https://doi.org/10.1590/S0102-311X2011000900014.;

126. Arvind, Varun, et al. ‘Development of a Machine Learning Algorithm to Predict Intubation among Hospitalized Patients with COVID-19’. Journal of Critical Care, vol. 62, Apr. 2021, pp. 25–30. , https://doi.org/10.1016/j.jcrc.2020.10.033.;

127. Bi et al., “Prediction of severe illness due to COVID-19 based on an analysis of initial Fibrinogen to Albumin Ratio and Platelet count,” Platelets, vol. 31, no. 5, pp. 674–679, 2020, doi: 10.1080/09537104.2020.1760230.

128. Iwendi, Celestine, et al. ‘COVID-19 Patient Health Prediction Using Boosted Random Forest Algorithm’. Frontiers in Public Health, vol. 8, July 2020, p. 357. , https://doi.org/10.3389/fpubh.2020.00357.;

129. Koppu, Srinivas, et al. ‘Deep Learning Disease Prediction Model for Use with Intelligent Robots’. Computers & Electrical Engineering, vol. 87, Oct. 2020, p. 106765. , https://doi.org/10.1016/j.compeleceng.2020.106765.;

130. Li, Qiubai, et al. ‘Cancer Increases Risk of In-Hospital Death from COVID-19 in Persons <65 Years and Those Not in Complete Remission’. Leukemia, vol. 34, no. 9, Sept. 2020, pp. 2384–91. , https://doi.org/10.1038/s41375-020-0986-7.;

131. Minaee, Shervin, et al. ‘Deep-COVID: Predicting COVID-19 from Chest X-Ray Images Using Deep Transfer Learning’. Medical Image Analysis, vol. 65, Oct. 2020, p. 101794. , https://doi.org/10.1016/j.media.2020.101794.;

132. Shang, Weifeng, et al. ‘The Value of Clinical Parameters in Predicting the Severity of COVID-19’. Journal of Medical Virology, vol. 92, no. 10, Oct. 2020, pp. 2188–92. , https://doi.org/10.1002/jmv.26031.;
133. Tao, Pei-Yao, et al. ‘Determination of Risk Factors for Predicting the Onset of Symptoms in Asymptomatic COVID-19 Infected Patients’. International Journal of Medical Sciences, vol. 17, no. 14, 2020, pp. 2187–93. , https://doi.org/10.7150/ijms.47576.;
134. Wang, Yi, et al. ‘Clinical Characteristics and Laboratory Indicator Analysis of 67 COVID-19 Pneumonia Patients in Suzhou, China’. BMC Infectious Diseases, vol. 20, no. 1, Dec. 2020, p. 747. , https://doi.org/10.1186/s12879-020-05468-8.;
135. W. Wei, et al. ‘Identification of common and severe COVID-19: the value of CT texture analysis and correlation with clinical characteristics’. European Radiology, vol. 30, no. 12, pp. 6788–6796, 2020, doi: 10.1007/s00330-020-07012-3.
136. Аналіз вимог до програмного забезпечення лекціїї 1-2. визначення вимог до програмних систем. [Електронний ресурс], Режим доступу: http://baklaniv.at.ua/ANALIZ_VYMOG/lekcija_1-2.pdf
137. S. Palaniappan and R. Awang, ‘Intelligent heart disease prediction system using data mining techniques’. 2008 IEEE/ACS International Conference on Computer Systems and Applications, 2008, pp. 108-115, doi: 10.1109/AICCSA.2008.4493524.;
138. Sharma, Nikita. ‘Understanding the Mathematics behind Decision Trees’. Medium, 28 Sept. 2021, https://heartbeat.comet.ml/understanding-the-mathematics-behind-decision-trees-22d86d55906.;
139. Brownlee, Jason. ‘Tour of Data Preparation Techniques for Machine Learning’. Machine Learning Mastery, 18 June 2020, https://machinelearningmastery.com/data-preparation-techniques-for-machine-learning/.;
140. Uddin, Shahadat, et al. ‘Comparing Different Supervised Machine Learning Algorithms for Disease Prediction’. BMC Medical Informatics and Decision Making, vol. 19, no. 1, Dec. 2019, p. 281. , https://doi.org/10.1186/s12911-019-1004-8.;

141. Kalaivani Annadurai, et. al. 'Personalized Medicine: A Paradigm Shift towards Promising Health Care', Journal of Pharmacy and Bioallied Sciences, 8.1 (2016), pp. 77, https://doi.org/10.4103/0975-7406.171732.;
142. Svitlyk, Yuri. 'Штучний інтелект проти COVID-19'. Root Nation, 14 May 2020, https://root-nation.com/ua/articles-ua/tech-ua/ua-shtuchnij-intelekt-covid-19/.;
143. 'China's Facial-Recognition Giant Says It Can Crack Masked Faces during the Coronavirus'. Quartz, 18 Feb. 2020, https://qz.com/1803737/chinas-facial-recognition-tech-can-crack-masked-faces-amid-coronavirus/.;
144. Sujath, R., et al. 'A Machine Learning Forecasting Model for COVID-19 Pandemic in India'. Stochastic Environmental Research and Risk Assessment, vol. 34, no. 7, July 2020, pp. 959–72. , https://doi.org/10.1007/s00477-020-01827-8.
145. Khamparia, Aditya, et al. 'Internet of Health Things-Driven Deep Learning System for Detection and Classification of Cervical Cells Using Transfer Learning'. The Journal of Supercomputing, vol. 76, no. 11, Nov. 2020, pp. 8590–608. , https://doi.org/10.1007/s11227-020-03159-4.;
146. Sakarkar, Gopal, et al. 'Comparative Study of Ambient Air Quality Prediction System Using Machine Learning to Predict Air Quality in Smart City'. Proceedings of International Conference on IoT Inclusive Life (ICIIL 2019), NITTTR Chandigarh, India, edited by Maitreyee Dutta et al., vol. 116, Springer Singapore, 2020, pp. 175–82. , https://doi.org/10.1007/978-981-15-3020-3_16.;
147. Zheliznyak, I. J. 'Some regulations for constructing associative rules on the example of patient's physical characteristics'. Scientific Bulletin of UNFU, vol. 27, no. 9, 2017, pp. 107–10. , https://doi.org/10.15421/40270923.
148. Shakhovska, N., et al. 'The sequential associative rules analysis of patient's physical characteristics'. Proceedings of the 1st International workshop on informatics & Data-driven medicine (IDDM 2018). 2018, pp. 82–92.;
149. Fedushko S., Shakhovska N., and Syerov Yu. 'Verifying the medical specialty from user profile of online community for health-related advices'. Proceedings of the 1st International workshop on informatics & Data-driven medicine (IDDM 2018) Lviv, November 28–30, 2018, pp. 301–310.;

150. Robins, J. ‘Non-response models for the analysis of non-monotone non-ignorable missing data’. Statistics in Medicine, no. 16, pp. 21–38.;
151. Blake, C. L., and C. J. Merz. ‘UCI machine learning repository of machine learning databases’. 2007-06-25)[2008-03-20]. http:‖ www. ics. uci. edu/~ mlearn/MLSummary. html (2007).;
152. Kleissner, Charly. ‘Data mining for the enterprise’. Proceedings of the Thirty-First Hawaii International Conference on System Sciences. vol. 7. IEEE, 1998.;
153. Ionica, Oncioiu. ‘Data Mining-an Instrument Managing the Knowledge Collected for the Enterprise’. Ovidius University Annals, Series Economic Sciences 13.1 (2013).;
154. Han, Jiawei, Jian Pei, and Hanghang Tong. ‘Data mining: concepts and techniques’. Morgan kaufmann, 2022.;
155. Thuraisingham B., ‘A Primer for Understanding and Applying Data Mining’. IT Professional, 2000, pp. 28-31.;
156. Weiguo F., et. al. ‘Tapping the Power of Text Mining’. Communication of the ACM. vol 9, no. 49, 2006, pp. 77-82.;
157. Peters Wu R., and Morgan M.W. ‘The Next Generation Clinical Decision Support: Linking Evidence to Best Practice’. Journal Healthcare Information Management. vol.4, no. 16, 2002, pp. 50-55.;
158. Porkodi, R., and Shivakumar, B.L. ‘An improved association rule mining technique for xml data using Xquery and Apriori algorithm’. Advance Computing, 2009, pp. 1510-1514 (2009).
159. Woo, J. ‘Apriori-Map/Reduce algorithm’. Proceedings of the International Conference on Parallel and Distributed Processing Techniques and Applications, 2012, p. 1.;
160. Yu, Jiangsheng, et al. ‘A Bayesian Approach to Support Vector Machines for the Binary Classification’. Neurocomputing, vol. 72, no. 1–3, Dec. 2008, pp. 177–85. , https://doi.org/10.1016/j.neucom.2008.06.010.;

161. Cristianini, Nello, and John Shawe-Taylor. An Introduction to Support Vector Machines: And Other Kernel-Based Learning Methods. Cambridge University Press, 2013.;
162. Melnykova, Nataliia, et al. ‘Data-Driven Analytics for Personalized Medical Decision Making’. Mathematics, vol. 8, no. 8, July 2020, p. 1211. , https://doi.org/10.3390/math8081211.;
163. Melnykova, Nataliia, et al. ‘Personalized Data Analysis Approach for Assessing Necessary Hospital Bed-Days Built on Condition Space and Hierarchical Predictor’. Big Data and Cognitive Computing, vol. 5, no. 3, Aug. 2021, p. 37. , https://doi.org/10.3390/bdcc5030037.;
164. Melnykova, Nataliia, et al. ‘The Hierarchical Classifier for COVID-19 Resistance Evaluation’. Data, vol. 6, no. 1, Jan. 2021, p. 6. , https://doi.org/10.3390/data6010006.;
165. Melnykova, Nataliia, et al. ‘An Ensemble Methods for Medical Insurance Costs Prediction Task’. Computers, Materials & Continua, vol. 70, no. 2, 2022, pp. 3969–84. , https://doi.org/10.32604/cmc.2022.019882.;
166. Melnykova, Nataliia, et al. ‘The Ensembles of Machine Learning Methods for Survival Predicting after Kidney Transplantation’. Applied Sciences, vol. 11, no. 21, Nov. 2021, p. 10380. , https://doi.org/10.3390/app112110380.;
167. Мельникова, Н. І. ‘Особливості опрацювання медичної інформації для систем підтримки прийняття лікувальних рішень’. Вісник Національного університету Львівська політехніка. Інформаційні системи та мережі, 832, 190-204.;
168. Шаховська Н. Б., Мельникова Н. І. ‘Методи побудови моделі поведінки користувачів’. Український журнал інформаційних технологій, 2020, 1 (2), 43–51.;
169. Мельникова Н.І. ‘Розроблення інформаційної технології опрацювання персоналізованих медичних даних’. Вісник Національного університету "Львівська політехніка". Інформаційні системи та мережі: збірник наукових праць, 2015, 814, 90–99.;

170. Shakhovska, N. B., and N. I. Melnykova. ‘Нові Методи Та Рішення Щодо Побудови Моделі Поведінки Користувачів’. Scientific Bulletin of UNFU, vol. 30, no. 5, Nov. 2020, pp. 76–83. , https://doi.org/10.36930/40300513.;
171. Кривенчук Ю. П., Шаховська Н. Б., Вовк О. Б., & Мельникова Н. І. ‘Комп’ютерне моделювання функцій перетворення оптичних схем засобу вимірювання температури, побудованого на ефекті рамана та структура алгоритму їх дослідження’. Радіоелектроніка, інформатика, управління, 2018, 346, 25-33. DOI:10.15588/1607-3274-2018-3-3.;
172. Melnykova Nataliia, and Basystiuk Oleh. ‘Multimodal Speech Recognition Based on Audio and Text Data’. Вісник ХНУ. Серія: технічні науки, 2022, no 3.;
173. Мельникова Н. І., & Поберейко П. Б. ‘Дослідження методів пошуку ключових кадрів у відеопотоці з використанням нейронних мереж для систем пошуку’ Вісник Хмельницького національного університету. Серія: Технічні науки, 2022, 309(3), 55–61.;
174. Melnykova, Nataliia, et al. Specifics personalized approach in the analysis of medical information. Econtechmod : an international quarterly journal on economics in technology, new technologies and modelling processes. Lublin. Rzeszow, 2016. vol. 5, no. 2. pp. 109–116.;
175. Kryvenchuk, Y., Shakhovska, N., Melnykova, N., and Boichuk, A. ‘Organization of the network connection in the Industry 4.0’. ECONTECHMOD: An International Quarterly Journal on Economics of Technology and Modelling Processes, 8, 2019, pp. 39-45.;
176. Melnykova N. ‘Using of personalized approach for assessment of the financial condition of the company’ Econtechmod : an international quarterly journal on economics in technology, new technologies and modelling processes. Lublin, Rzeszow, 2017, vol. 6, no 2, pp. 39–44.;
177. Melnykova N. ‘Analysis of the Data Mining and Classification of Patients’ States’. Manažérska Informatika, 2020, no. 2 https://manazerskainformatika. sk/analysis-of-the-data-mining-and-classification-of-patients-states/.;

178. Melnykova N. ‘Method for Clustering and Determining the Average Distance between Clusters’. ManažérskaInformatika, 2021, no. 2, https://manazerskainformatika.sk/method-for-clustering-and-determining-the-average-distance-between-clusters/;
179. Melnykova N. ‘A New Approach to Modelling the Nature of Individual Morbidity Using Partial Functional Dependencies’. ManažérskaInformatika, 2021, no. 1, https://manazerskainformatika.sk/a-new-approach-to-modelling-the-nature-of-individual-morbidity-using-partial-functional-dependencies/;
180. Melnykova N. ‘Model of States Warehouse of a State of the State of the Object and Personalized Decisions’. ManažérskaInformatika, 2022, no. 1, https://manazerskainformatika.sk/model-of-states-warehouse-of-a-state-of-the-state-of-the-object-and-personalized-decisions/;
181. Мельникова Н. І., & Мельников В. А. ‘Персоналізований підхід до обробки та аналізу медичних даних пацієнтів’. Інформаційна безпека та інформаційні технології : монографія, Харків : Вид. Рожко С. Г., 2019., 247–259.;
182. Мельникова Н.І. ‘Методи оптимізації рішень щодо аналізу персоналізованих даних’. Кібербезпека та інформаційні технології: монографія. – Х. : ТОВ «ДІСА ПЛЮС», 2020, ISBN 978-617-7927-01-2, 210-225.;
183. Melnykova, Natalia, et al. ‘Big Data Analysis in Development of Personalized Medical System’. Procedia Computer Science, vol. 160, 2019, pp. 229–34. , https://doi.org/10.1016/j.procs.2019.09.461.;
184. Melnykova, Nataliia, et al. ‘Using Big Data for Formalization the Patient’s Personalized Data’. Procedia Computer Science, vol. 155, 2019, pp. 624–29. , https://doi.org/10.1016/j.procs.2019.08.088.;
185. Melnykova, Nataliia, et al. ‘Anomalies Detecting in Medical Metrics Using Machine Learning Tools’. Procedia Computer Science, vol. 198, 2022, pp. 718–23. , https://doi.org/10.1016/j.procs.2021.12.312.;
186. Melnykova, Nataliia. ‘A Novel Approach for the Automatic Detection of COVID in a Patient by Using a Categorization Methods’. Procedia Computer Science, vol. 198, 2022, pp. 712–17. , https://doi.org/10.1016/ j.procs.2021.12.311.;

187. Melnykova, Nataliia. ‘Model of the system of personalized analysis of financial condition of the enterprise’. Advances in Intelligent Systems and Computing, Q3, 2018, no. 689, pp. 334-345, https://www.scopus.com/inward/record.uri?eid=2-s2.0-85036467510&doi=10.1007%2f978-3-319-70581-1_24&partnerID=40&md5= 909d5e2a66315 b464c19399e7482a86e.;
188. Melnykova, Nataliia., et. al. ‘Smart Integrated Robotics System for SMEs Controlled by Internet of Things Based on Dynamic Manufacturing Processes’. Advances in Intelligent Systems and Computing, 2019, no. 871, pp. 535-549, https://doi.org/10.1007/978-3-030-01069-0_38.;
189. Melnykova, Nataliia. ‘Semantic search personalized data as special method of processing medical information’. Advances in Intelligent Systems and Computing, 2017, no. 512, pp. 315-325, https://doi.org/10.1007/978-3-319-45991-2_22.;
190. Melnykova, Nataliia., et al. ‘The special ways for processing personalized data during voting in elections’. Advances in Intelligent Systems and Computing, 1080 AISC, 2020, pp. 781-791, DOI: 10.1007/978-3-030-33695-0_52.;
191. Melnykova, Nataliia., et. al. ‘The personalized approach to the processing and analysis of patients’ medical data’. CEUR Workshop Proceedings, 2018, no. 2255, pp. 103-112, http://ceur-ws.org/Vol-2255/paper10.pdf.;
192. Shakhovska, Nataliya. and Melnykova, Nataliia. ‘Feature Engineering and Missing Data Imputation Method of Medical Data Analysis’. CEUR Workshop Proceedings, 2022, no. 3137, pp. 48–57. http://ceur-ws.org/Vol-3137/paper4.pdf.;
193. Melnykova, Nataliia., et. al. ‘The problem of analysing the relationships between individual characteristics of individuals with COVID`19’. CEUR Workshop Proceedings, 2020, no. 2753, pp. 473-482, https://pesquisa.bvsalud.org/global-literature-on-novel-coronavirus-2019-ncov/resource/pt/covidwho-984587.;
194. Melnykova, Nataliia., et. al. ‘Advisory and accounting tool for safe and economically optimal choice of online self-education services’. CEUR Workshop Proceedings, 2019, 2588, pp. 290-300, http://ceur-ws.org/Vol-2588/paper24.pdf.;
195. Melnykova, Nataliia., et. al. ‘Determination of characteristics discrete transfiguration for synthesized raster elements of non-regular structure’. CEUR

Workshop Proceedings, 2019, no. 2533, pp. 249-258, http://ceur-ws.org/Vol-2533/paper23.pdf.;

196. Melnykova, Nataliia., et. al. ‘Technologies of 3D-prototyping of Objects.’ CEUR Workshop Proceedings, 2019, no. 2533, pp. 271-281, http://ceur-ws.org/Vol-2533/paper25.pdf.;
197. Melnykova, Natalia. ‘Application of information technology for designing of treatment information systems’. 2015, Proceedings of 13th International Conference: The Experience of Designing and Application of CAD Systems in Microelectronics, CADSM 2015, pp. 156-158, https://doi.org/10.1109/CADSM.2015.7230823.;
198. Melnykova, Nataliia, and Oksana Markiv. ‘Semantic approach to personalization of medical data’. 2016 XIth International Scientific and Technical Conference Computer Sciences and Information Technologies (CSIT). IEEE, 2016., pp. 59-61, https://doi.org/10.1109/STC-CSIT.2016.7589868.;
199. Melnykova, Nataliia. ‘The basic approaches to automation of management by enterprise finances’. 2017, Proceedings of the 12th International Scientific and Technical Conference on Computer Sciences and Information Technologies, CSIT 2017, pp. 288-291, https://doi.org/10.1109/STC-CSIT.2017.8098788.;
200. Melnykova, Nataliia. et. al. ‘The personalized approach in a medical decentralized diagnostic and treatment’. 2017, 14th International Conference The Experience of Designing and Application of CAD Systems in Microelectronics, CADSM 2017, pp. 295-297, https://doi.org/10.1109/CADSM.2017.7916139.;
201. Melnykova, Nataliia. et. al. ‘The new approaches of heterogeneous data consolidation’. 2018 International Scientific and Technical Conference on Computer Sciences and Information Technologies, 2018, pp. 408-411, https://doi.org/10.1109/STC-CSIT.2018.8526677.;
202. Melnykova, Nataliia. et. al. ‘The special ways of application of neural networks for medical information processing, 2018 International Scientific and Technical Conference on Computer Sciences and Information Technologies, 2018, pp. 428-431, https://doi.org/10.1109/STC-CSIT.2018.8526708.;

203. Melnykova, Nataliia. et. al. ‘Calculation of the Exact Value of the Fractal Dimension in the Time Series for the Box-Counting Method’. 2019 9th International Conference on Advanced Computer Information Technologies, ACIT 2019, pp. 248-251, https://doi.org/10.1109/ACITT.2019.8780028.;
204. Melnykova, Nataliia. et. al. ‘The Applying Processing Intelligence Methods for Classify Persons in Identify Personalized Medication Decisions’. 2020 10th International Conference on Advanced Computer Information Technologies, ACIT 2020, pp. 422-425, https://doi.org/10.1109/ACIT49673.2020.9208822.;
205. Melnykova, Nataliia. et. al. ‘The Investigation of Artificial Intelligence Methods for Identifying States and Analyzing System Transitions between States’. 2020 International Scientific and Technical Conference on Computer Sciences and Information Technologies, 2020, vol. 1, pp. 41-75 https://doi.org/10.1109/CSIT49958.2020.9321841.;
206. Melnykova, Nataliia. et. al. ‘The Scheme Application of the Medical Expert System’. XXIV Ukrainian-Polish Conference on CAD in Machinery Design. Implementation and Educational Issues, CADMD 2016, Lviv, 21.10-22.10.2016, pp. 65-66.;
207. Melnykova, Nataliia. et. al. ‘The Intelligent System Architecture of Personalized Management’. XXIV Ukrainian-Polish Conference on CAD in Machinery Design, Implementation and Educational Issues, CADMD 2016, Lviv, 21.10-22.10.2016, pp.27-28.;
208. Мельникова Н.І. & Жилко І.В. ‘Застосування хмарних обчислень для проектування систем підтримки лікарських рішень’. Міжнародна конференція «інноваційні підходи і сучасна наука», Центр наукових публікацій, Київ, 30 квітня 2015 р., 45-46.;
209. Мельникова Н.І., & Данилів В.М. ‘Інтелектуальна інформаційна система організації інтерактивних промоакцій’. Матеріалами міжнародної науково-практичної конференції: «IV осінні наукові читання», Київ: збірник статей (рівень стандарту, академічний рівень), Центр наукових публікацій, 2015. 48–49.;

210. Мельникова Н.І., & Копач М.І. ‘Методи оптимізації аналізу персоналізованих даних підприємства Математика’. Інформаційні технології. Освіта : тези доповідей VI Міжнародної науково-практичної конференції, Східноєвропейський національний університет ім. Лесі Українки, Луцьк-Світязь, 5-7 червня 2017, 67 – 69.;
211. Мельникова Н.І., Мукалов П., & Козій Д. ‘Особливості застосування нейронних мереж щодо опрацювання даних різного походження’. Матеріали VI Всеукраїнської науково-практичної конференції «Наукові дослідження: перспективи інновацій у суспільстві і розвитку технологій», Наукове партнерство «Центр наукових технологій», Харків, 24–25лютого 2017, 104.;
212. Мельникова Н.І. & Мельников В.А. ‘Персоналізований підхід до обробки та аналізу медичних даних пацієнтів’. Матеріали Міжнародної науково-практичної конференції «Інформаційна безпека та інформаційні технології», ХНЕУ імені Семена Кузнеця, 24 – 25 квітня 2019, 56.;
213. Melnykova, Nataliia. ‘New approach to support of personalized decision making in medicine’. The 8 th International scientific and practical conference « Information, Its Impact on Social And Technical Processes, March Haifa, Israel,16-17, 2020, pp.261.;
214. Melnykova, Nataliia. et. al. ‘Semantic approach to personalization of medical data’. IX Міжнародна науково-практична конференція «Actual aspects of development in the context of globalization», Florecia, Italy, 23-24 March 2020, pp. 126-129.;
215. Мельникова Н.І. ‘Методи оптимізації рішень щодо аналізу персоналізованих даних’. Матеріали II Міжнародної науково-практичної конференції «Інформаційна безпека та інформаційні технології», Кропивницький: ЦНТУ, 2 – 3 квітня 2020, pp. 105.;
216. Melnykova, Nataliia., and Kolomyi, Anastasia. ‘Information System For Determination Of Early Symptoms Of Dementia Base On Mini-Cog Test And Mini-Mental State Examination Impact of modernity on science and practice’. XVIII International Scientific and Practical Conference, Boston, USA 2020. pp. 95-97.;

217. Melnykova, Nataliia. et. al. ‘Automatic audio to text conversion approaches along the radiology value chain’. XVII International Scientific and Practical Conference «Multidisciplinary academic notes. Theory, methodology and practice», Tokyo, Japan, 03-06 May 2022, pp. 994.;
218. Зайченко, Ю. П., & Мурга, Н. А. (2008). Застосування систем з нечіткою логікою до задачі медичної діагностики. Вісник НТУУ" КПІ": Інформатика, управління та обчислювальна техніка, 2008(49).;
219. Zaychenko, Yuriy, and Aghaei Agh Ghamish Ovi Nafas. ‘Medical Images Classification and Diagnostics Using Fuzzy Neural Networks’. American Journal of Neural Networks and Applications, vol. 5, no. 2, 2019, p. 45., https://doi.org/10.11648/j.ajnna.20190502.11.;
220. Березький, О. М., Мельник, Г. М., & Березька, К. М. (2013). Нечітка база знань інтелектуальної системи діагностування видів раку молочної залози. Вісник Хмельницького національного університету. Технічні науки, (6), 284-291.;
221. Subbotin, Sergey, et al. ‘Intelligent Data Analysis for Individual Hypertensia Patient’s State Monitoring and Prediction’. 2021 IEEE International Conference on Smart Information Systems and Technologies (SIST), IEEE, 2021, pp. 1–4. DOI.org (Crossref), https://doi.org/10.1109/SIST50301.2021.9465989.;
222. Subbotin S. A. ‘Construction of decision trees for the case of low-information features’. Radio Electronics, Computer Sci-ence, Control, 2019, №1, pp. 121-130;
223. Subbotin, Sergey. ‘A Random Forest Model Building Using A Priori Information for Diagnosis’. Computer Modeling and Intelligent Systems, vol. 2353, 2019, pp. 962–73., https://doi.org/10.32782/cmis/2353-76.;
224. Bommert, A., et. al. ‘Benchmark for filter methods for feature selection in high-dimensional classification data’. Computational Statistics & Data Analysis, 2020, no. 143.;
225. Zgurovsky, Michael Z., and Yuriy P. Zaychenko. ‘Pattern Recognition in Big Data Analysis’. Big Data: Conceptual Analysis and Applications, by Michael Z. Zgurovsky

and Yuriy P. Zaychenko, vol. 58, Springer International Publishing, 2020, pp. 97–139. DOI.org (Crossref), https://doi.org/10.1007/978-3-030-14298-8_3.;

226. Sanchez-Pinto, L. N., et. al. ‘Comparison of variable selection methods for clinical predictive modeling’. International journal of medical informatics, 2018, no. 116, pp. 10-17.;

227. Hayati Rezvan, P., Lee, K. J., and Simpson, J. A. ‘The Rise of Multiple Imputation: A Review of the Reporting and Implementation of the Method in Medical Research’. BMC Med Res Methodol, 2015, no. 15, pp. 30, https://doi.org/10.1186/s12874-015-0022-1.;

228. Ahmat Zainuri, N., Jemain, A. A., and Muda, N. A. ‘Comparison of Various Imputation Methods for Missing Values in Air Quality Data’. JSM, 2015, no. 44, pp. 449–456, https://doi.org/10.17576/jsm-2015-4403-17.;

229. Leke, C. A., and Marwala, T. ‘Introduction to Missing Data Estimation. Deep Learning and Missing Data in Engineering Systems’. 2019, pp. 1-20, https://doi.org/10.1117/12.2053057.;

230. Wang, C., Shakhovska, N., Sachenko, A., and Komar, M. ”A New Approach for Missing Data Imputation in Big Data Interface”. Information Technology and Control, 2020, vol. 4, no. 49, pp. 541-555.;

231. Azmi, M., Runger, G. C., and Berrado, A. ”Interpretable regularized class association rules algorithm for classification in a categorical data space”. Information Sciences, 2019, no. 483, pp. 313-331.;

232. Thabtah, F., Cowling, P., and Peng, Y. ‘MCAR: multi-class classification based on association rule’. In The 3rd ACS/IEEE International Conference on Computer Systems and Applications, 2005, January, IEEE, pp 33.;

233. Jagannath University Bahadurgarh, Haryana, India., and Kavita Mittal. ‘A comparative study of association rule mining techniques and predictive mining approaches for association classification’. International Journal of Advanced Research in Computer Science, vol. 8, no. 9, Sept. 2017, pp. 365–72. , https://doi.org/10.26483/ijarcs.v8i9.4984.;

234. Yan, Li, et al. ‘An Interpretable Mortality Prediction Model for COVID-19 Patients’. Nature Machine Intelligence, vol. 2, no. 5, May 2020, pp. 283–88. , https://doi.org/10.1038/s42256-020-0180-7.;

235. Trickey, Adam, et al. ‘CD4:CD8 Ratio and CD8 Count as Prognostic Markers for Mortality in Human Immunodeficiency Virus–Infected Patients on Antiretroviral Therapy: The Antiretroviral Therapy Cohort Collaboration (ART-CC)’. Clinical Infectious Diseases, vol. 65, no. 6, Sept. 2017, pp. 959–66. , https://doi.org/10.1093/cid/cix466.;

236. Hasan, Najmul. ‘A Methodological Approach for Predicting COVID-19 Epidemic Using EEMD-ANN Hybrid Model’. Internet of Things, vol. 11, Sept. 2020, p. 100228. , https://doi.org/10.1016/j.iot.2020.100228.;

237. Cristianini, Nello, and John Shawe-Taylor. An Introduction to Support Vector Machines: And Other Kernel-Based Learning Methods. Cambridge University Press, 2000.;

238. Blanchard, Gilles, et al. ‘Statistical Performance of Support Vector Machines’. The Annals of Statistics, vol. 36, no. 2, Apr. 2008. , https://doi.org/10.1214/009053607000000839.;

239. Keerthi, S. Sathiya, and Chih-Jen Lin. ‘Asymptotic Behaviors of Support Vector Machines with Gaussian Kernel’. Neural Computation, vol. 15, no. 7, July 2003, pp. 1667–89. , https://doi.org/10.1162/089976603321891855.;

240. Saiz, Fernan, and Leonardo Bernasconi. ‘Catalytic Properties of the Ferryl Ion in the Solid State: A Computational Review’. Catalysis Science & Technology, vol. 12, no. 10, 2022, pp. 3069–87. , https://doi.org/10.1039/D2CY00200K.;

241. Singh, Varshika, et al. ‘Recent Trends in Computational Tools and Data-Driven Modeling for Advanced Materials’. Materials Advances, vol. 3, no. 10, 2022, pp. 4069–87. , https://doi.org/10.1039/D2MA00067A.;

242. Jaafreh, Russlan, et al. ‘Crystal Structure Guided Machine Learning for the Discovery and Design of Intrinsically Hard Materials’. Journal of Materiomics, vol. 8, no. 3, May 2022, pp. 678–84. , https://doi.org/10.1016/j.jmat.2021.11.004.;

243. Pathrudkar, Shashank, et al. ‘Machine Learning Based Prediction of the Electronic Structure of Quasi-One-Dimensional Materials under Strain’. Physical Review B, vol. 105, no. 19, May 2022, p. 195141. , https://doi.org/10.1103/PhysRevB.105.195141.;
244. ‘Corrigendum to “Machine Learning in Energy Storage Materials”’. Interdisciplinary Materials, Oct. 2022, p. idm2.12060. , https://doi.org/10.1002/idm2.12060.;
245. Milazzo, Mario, and Flavia Libonati. ‘The Synergistic Role of Additive Manufacturing and Artificial Intelligence for the Design of New Advanced Intelligent Systems’. Advanced Intelligent Systems, vol. 4, no. 6, June 2022, p. 2100278. , https://doi.org/10.1002/aisy.202100278.;
246. Basak, Hritam, et al. ‘Cervical Cytology Classification Using PCA and GWO Enhanced Deep Features Selection’. SN Computer Science, vol. 2, no. 5, Sept. 2021, p. 369. , https://doi.org/10.1007/s42979-021-00741-2.;
247. Bhattacharyya, Trinav, et al. ‘Mayfly in Harmony: A New Hybrid Meta-Heuristic Feature Selection Algorithm’. IEEE Access, vol. 8, 2020, pp. 195929–45. , https://doi.org/10.1109/ACCESS.2020.3031718.;
248. Bommert, Andrea, et al. ‘Benchmark for Filter Methods for Feature Selection in High-Dimensional Classification Data’. Computational Statistics & Data Analysis, vol. 143, Mar. 2020, p. 106839. , https://doi.org/10.1016/j.csda.2019.106839.;
249. Gunavathi, C., et al. ‘A Survey on Feature Selection Methods in Microarray Gene Expression Data for Cancer Classification’. Research Journal of Pharmacy and Technology, vol. 10, no. 5, 2017, p. 1395. , https://doi.org/10.5958/0974-360X.2017.00249.9.;
250. El Aboudi, Naoual, and Laila Benhlima. ‘Review on Wrapper Feature Selection Approaches’. 2016 International Conference on Engineering & MIS (ICEMIS), IEEE, 2016, pp. 1–5. , https://doi.org/10.1109/ICEMIS.2016.7745366.;
251. Faramarzi, Afshin, et al. ‘Equilibrium Optimizer: A Novel Optimization Algorithm’. Knowledge-Based Systems, vol. 191, Mar. 2020, p. 105190. , https://doi.org/10.1016/j.knosys.2019.105190.;

252. Kather, Jakob Nikolas, et al. ‘Multi-Class Texture Analysis in Colorectal Cancer Histology’. Scientific Reports, vol. 6, no. 1, Sept. 2016, p. 27988, https://doi.org/10.1038/srep27988.;

253. Kou, Gang, et al. ‘Evaluation of Feature Selection Methods for Text Classification with Small Datasets Using Multiple Criteria Decision-Making Methods’. Applied Soft Computing, vol. 86, Jan. 2020, p. 105836. DOI.org (Crossref), https://doi.org/10.1016/j.asoc.2019.105836.;

254. Liang, Gaobo, and Lixin Zheng. ‘A Transfer Learning Method with Deep Residual Network for Pediatric Pneumonia Diagnosis’. Computer Methods and Programs in Biomedicine, vol. 187, Apr. 2020, p. 104964. DOI.org (Crossref), https://doi.org/10.1016/j.cmpb.2019.06.023.;

255. López, D., et al. ‘BELIEF: A Distance-Based Redundancy-Proof Feature Selection Method for Big Data’. Information Sciences, vol. 558, May 2021, pp. 124–39. DOI.org (Crossref), https://doi.org/10.1016/j.ins.2020.12.082.;

256. Mahmud, Tanvir, et al. ‘CovXNet: A Multi-Dilation Convolutional Neural Network for Automatic COVID-19 and Other Pneumonia Detection from Chest X-Ray Images with Transferable Multi-Receptive Feature Optimization’. Computers in Biology and Medicine, vol. 122, July 2020, p. 103869. DOI.org (Crossref), https://doi.org/10.1016/j.compbiomed.2020.103869.;

257. Maldonado, Sebastián, and Julio López. ‘Dealing with High-Dimensional Class-Imbalanced Datasets: Embedded Feature Selection for SVM Classification’. Applied Soft Computing, vol. 67, June 2018, pp. 94–105. DOI.org (Crossref), https://doi.org/10.1016/j.asoc.2018.02.051.;

258. Ohata, Elene Firmeza, et al. ‘A Novel Transfer Learning Approach for the Classification of Histological Images of Colorectal Cancer’. The Journal of Supercomputing, vol. 77, no. 9, Sept. 2021, pp. 9494–519. DOI.org (Crossref), https://doi.org/10.1007/s11227-020-03575-6.;

259. Paladini, Emanuela, et al. ‘Two Ensemble-CNN Approaches for Colorectal Cancer Tissue Type Classification’. Journal of Imaging, vol. 7, no. 3, Mar. 2021, p. 51. DOI.org (Crossref), https://doi.org/10.3390/jimaging7030051.;

260. Raj, R. Joshua Samuel, et al. ‘Optimal Feature Selection-Based Medical Image Classification Using Deep Learning Model in Internet of Medical Things’. IEEE Access, vol. 8, 2020, pp. 58006–17. DOI.org (Crossref), https://doi.org/10.1109/ACCESS.2020.2981337.;

261. Sharma, Harsh, et al. ‘Feature Extraction and Classification of Chest X-Ray Images Using CNN to Detect Pneumonia’. 2020 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence), IEEE, 2020, pp. 227–31. DOI.org (Crossref), https://doi.org/10.1109/Confluence47617.2020.9057809.;

262. Soheili, Majid, and Maryam Amir Haeri. ‘Scalable Global Mutual Information Based Feature Selection Framework for Large Scale Datasets’. 2021 IEEE 25th International Enterprise Distributed Object Computing Conference (EDOC), IEEE, 2021, pp. 41–50. DOI.org (Crossref), https://doi.org/10.1109/EDOC52215.2021.00015.;

263. Stephen, Okeke, et al. ‘An Efficient Deep Learning Approach to Pneumonia Classification in Healthcare’. Journal of Healthcare Engineering, vol. 2019, Mar. 2019, pp. 1–7. DOI.org (Crossref), https://doi.org/10.1155/2019/4180949.;

264. Venkatesh, B., and J. Anuradha. ‘A Review of Feature Selection and Its Methods’. Cybernetics and Information Technologies, vol. 19, no. 1, Mar. 2019, pp. 3–26. DOI.org (Crossref), https://doi.org/10.2478/cait-2019-0001.;

265. Xue, Bing, et al. ‘Particle Swarm Optimization for Feature Selection in Classification: A Multi-Objective Approach’. IEEE Transactions on Cybernetics, vol. 43, no. 6, Dec. 2013, pp. 1656–71. DOI.org (Crossref), https://doi.org/10.1109/TSMCB.2012.2227469.;

266. Xue, Bing, et al. ‘A Survey on Evolutionary Computation Approaches to Feature Selection’. IEEE Transactions on Evolutionary Computation, vol. 20, no. 4, Aug. 2016, pp. 606–26. DOI.org (Crossref), https://doi.org/10.1109/TEVC.2015.2504420.;

267. Zhang, Yong, et al. ‘Binary Differential Evolution with Self-Learning for Multi-Objective Feature Selection’. Information Sciences, vol. 507, Jan. 2020, pp. 67–85. DOI.org (Crossref), https://doi.org/10.1016/j.ins.2019.08.040.;

268. Deo, Gouri, et al. ‘Detection of COVID-19 and Prediction of Pneumonia from Chest X-Rays Using Deep Learning’. 2022 IEEE 11th International Conference on Communication Systems and Network Technologies (CSNT), IEEE, 2022, pp. 232–38. DOI.org (Crossref), https://doi.org/10.1109/CSNT54456.2022.9787593.;

269. Xiong, Tao, et al. ‘Multiple-Output Support Vector Regression with a Firefly Algorithm for Interval-Valued Stock Price Index Forecasting’. Knowledge-Based Systems, vol. 55, Jan. 2014, pp. 87–100. DOI.org (Crossref), https://doi.org/10.1016/j.knosys.2013.10.012.;

270. Wen, Zhenfu, and Yuanqing Li. ‘A Spatial-Constrained Multi-Target Regression Model for Human Brain Activity Prediction’. Applied Informatics, vol. 3, no. 1, Dec. 2016, p. 10. DOI.org (Crossref), https://doi.org/10.1186/s40535-016-0026-x.;

271. Cherman, Everton Alvares, et al. ‘Multi-Label Problem Transformation Methods: A Case Study’. CLEI Electronic Journal, vol. 14, no. 1, Apr. 2011. DOI.org (Crossref), https://doi.org/10.19153/cleiej.14.1.4.;

272. Wu, Qingyao, et al. ‘ML-FOREST: A Multi-Label Tree Ensemble Method for Multi-Label Classification’. IEEE Transactions on Knowledge and Data Engineering, vol. 28, no. 10, Oct. 2016, pp. 2665–80. DOI.org (Crossref), https://doi.org/10.1109/TKDE.2016.2581161.;

273. Tsoumakas, Grigorios, and Ioannis Katakis. ‘Multi-Label Classification: An Overview’. International Journal of Data Warehousing and Mining, vol. 3, no. 3, July 2007, pp. 1–13. DOI.org (Crossref), https://doi.org/10.4018/jdwm.2007070101.;

274. Spolaor, Newton, et al. ‘A Comparison of Multi-Label Feature Selection Methods Using the Problem Transformation Approach’. Electronic Notes in Theoretical Computer Science, vol. 292, Mar. 2013, pp. 135–51. DOI.org (Crossref), https://doi.org/10.1016/j.entcs.2013.02.010.;

275. Zhang, Min-Ling, et al. ‘Binary Relevance for Multi-Label Learning: An Overview’. Frontiers of Computer Science, vol. 12, no. 2, Apr. 2018, pp. 191–202. DOI.org (Crossref), https://doi.org/10.1007/s11704-017-7031-7.;

276. Spyromitros-Xioufis, Eleftherios, et al. ‘Multi-Target Regression via Input Space Expansion: Treating Targets as Inputs’. Machine Learning, vol. 104, no. 1, July 2016, pp. 55–98. DOI.org (Crossref), https://doi.org/10.1007/s10994-016-5546-z.;

277. Babovic, Vladan. ‘Data Mining and Knowledge Discovery in Sediment Transport’. Computer-Aided Civil and Infrastructure Engineering, vol. 15, no. 5, Sept. 2000, pp. 383–89. DOI.org (Crossref), https://doi.org/10.1111/0885-9507.00202.;

278. Syed, Farrukh Hasan, and Muhammad Atif Tahir. ‘Safe Semi Supervised Multi-Target Regression (MTR-SAFER) for New Targets Learning’. Multimedia Tools and Applications, vol. 77, no. 22, Nov. 2018, pp. 29971–87. DOI.org (Crossref), https://doi.org/10.1007/s11042-018-6367-9.;

279. Borchani, Hanen, et al. ‘A Survey on Multi-Output Regression: Multi-Output Regression Survey’. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 5, no. 5, Sept. 2015, pp. 216–33. DOI.org (Crossref), https://doi.org/10.1002/widm.1157.;

280. Gheyas, Iffat A., and Leslie S. Smith. ‘Feature Subset Selection in Large Dimensionality Domains’. Pattern Recognition, vol. 43, no. 1, Jan. 2010, pp. 5–13. DOI.org (Crossref), https://doi.org/10.1016/j.patcog.2009.06.009.;

281. ‘Correlation Based Ensemble Feature Selection Algorithm for Diagnosis of Diabetics’. International Journal of Innovative Technology and Exploring Engineering, vol. 9, no. 3S, Feb. 2020, pp. 373–373. DOI.org (Crossref), https://doi.org/10.35940/ijitee.C1080.0193S20.;

282. Zhao, Shuai, et al. ‘Ensemble Classification Based on Feature Selection for Environmental Sound Recognition’. Mathematical Problems in Engineering, vol. 2019, Feb. 2019, pp. 1–7. DOI.org (Crossref), https://doi.org/10.1155/2019/4318463.;

283. Hatzikos, Evaggelos V., et al. ‘An Empirical Study on Sea Water Quality Prediction’. Knowledge-Based Systems, vol. 21, no. 6, Aug. 2008, pp. 471–78. DOI.org (Crossref), https://doi.org/10.1016/j.knosys.2008.03.005.;

284. Bin Tahir, Hassan, et al. ‘Predicting the Permanent Deformation Behaviour of the Plant Produced Asphalt Concrete Mixtures: A First Order Regression Approach’.

Construction and Building Materials, vol. 189, Nov. 2018, pp. 629–39. DOI.org (Crossref), https://doi.org/10.1016/j.conbuildmat.2018.08.164.;

285. Tsanas, Athanasios, and Angeliki Xifara. ‘Accurate Quantitative Estimation of Energy Performance of Residential Buildings Using Statistical Machine Learning Tools’. Energy and Buildings, vol. 49, June 2012, pp. 560–67. DOI.org (Crossref), https://doi.org/10.1016/j.enbuild.2012.03.003.;

286. Tsanas, Athanasios, and Angeliki Xifara. ‘Accurate Quantitative Estimation of Energy Performance of Residential Buildings Using Statistical Machine Learning Tools’. Energy and Buildings, vol. 49, June 2012, pp. 560–67. DOI.org (Crossref), https://doi.org/10.1016/j.enbuild.2012.03.003.;

287. Nuril Izzati, Fadhila, and Catur Retnaningdyah. ‘Evaluation of River Water Quality Based on Biotic Index of Benthic Macroinvertebrate as Bioindicator (Case Study in Genjong River Wlingi Blitar East Java, Indonesia)’. Biotropika: Journal of Tropical Biology, vol. 10, no. 2, Aug. 2022, pp. 117–25. DOI.org (Crossref), https://doi.org/10.21776/ub.biotropika.2022.010.02.05.;

288. Bashir, Kamal, et al. ‘SMOTEFRIS-INFFC: Handling the Challenge of Borderline and Noisy Examples in Imbalanced Learning for Software Defect Prediction’. Journal of Intelligent & Fuzzy Systems, edited by Cengiz Kahraman, vol. 38, no. 1, Jan. 2020, pp. 917–33. DOI.org (Crossref), https://doi.org/10.3233/JIFS-179459.;

289. Soltanzadeh, Paria, and Mahdi Hashemzadeh. ‘RCSMOTE: Range-Controlled Synthetic Minority over-Sampling Technique for Handling the Class Imbalance Problem’. Information Sciences, vol. 542, Jan. 2021, pp. 92–111. DOI.org (Crossref), https://doi.org/10.1016/j.ins.2020.07.014.;

290. Sun, Yanmin, et al. ‘Cost-Sensitive Boosting for Classification of Imbalanced Data’. Pattern Recognition, vol. 40, no. 12, Dec. 2007, pp. 3358–78. DOI.org (Crossref), https://doi.org/10.1016/j.patcog.2007.04.009.;

291. Chebouba, Lokmane, et al. ‘Proteomics Versus Clinical Data and Stochastic Local Search Based Feature Selection for Acute Myeloid Leukemia Patients’ Classification’. Journal of Medical Systems, vol. 42, no. 7, July 2018, p. 129. DOI.org (Crossref), https://doi.org/10.1007/s10916-018-0972-z.;

292. Araújo, Flávio H. D., et al. ‘Using Machine Learning to Support Healthcare Professionals in Making Preauthorisation Decisions’. International Journal of Medical Informatics, vol. 94, Oct. 2016, pp. 1–7. DOI.org (Crossref), https://doi.org/10.1016/j.ijmedinf.2016.06.007.;

293. Sanchez-Pinto, L. Nelson, et al. ‘Comparison of Variable Selection Methods for Clinical Predictive Modeling’. International Journal of Medical Informatics, vol. 116, Aug. 2018, pp. 10–17. DOI.org (Crossref), https://doi.org/10.1016/j.ijmedinf.2018.05.006.;

294. Lynch, Chip M., et al. ‘Prediction of Lung Cancer Patient Survival via Supervised Machine Learning Classification Techniques’. International Journal of Medical Informatics, vol. 108, Dec. 2017, pp. 1–8. DOI.org (Crossref), https://doi.org/10.1016/j.ijmedinf.2017.09.013.;

295. Zarchi, M. S., et al. ‘SCADI: A Standard Dataset for Self-Care Problems Classification of Children with Physical and Motor Disability’. International Journal of Medical Informatics, vol. 114, June 2018, pp. 81–87. DOI.org (Crossref), https://doi.org/10.1016/j.ijmedinf.2018.03.003.;

296. Al-Shammari, Ahmed, et al. ‘An Effective Density-Based Clustering and Dynamic Maintenance Framework for Evolving Medical Data Streams’. International Journal of Medical Informatics, vol. 126, June 2019, pp. 176–86. DOI.org (Crossref), https://doi.org/10.1016/j.ijmedinf.2019.03.016.;

297. Yang, Shuo, et al. ‘An Improved Id3 Algorithm for Medical Data Classification’. Computers & Electrical Engineering, vol. 65, Jan. 2018, pp. 474–87. DOI.org (Crossref), https://doi.org/10.1016/j.compeleceng.2017.08.005.;

298. Chawla, Nitesh V., et al. ‘Editorial: Special Issue on Learning from Imbalanced Data Sets’. ACM SIGKDD Explorations Newsletter, vol. 6, no. 1, June 2004, pp. 1–6. DOI.org (Crossref), https://doi.org/10.1145/1007730.1007733.;

299. Batista, Gustavo E. A. P. A., et al. ‘A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data’. ACM SIGKDD Explorations Newsletter, vol. 6, no. 1, June 2004, pp. 20–29. DOI.org (Crossref), https://doi.org/10.1145/1007730.1007735.;

300. Dehaki, Ghoncheh Babanejad, et al. ‘A Framework for Processing Skyline Queries for a Group of Mobile Users’. Proceedings of the 20th International Conference on Information Integration and Web-Based Applications & Services, ACM, 2018, pp. 333–39. DOI.org (Crossref), https://doi.org/10.1145/3282373.3282392.;

301. Srivastava, Saurabh Kumar, et al. ‘Healthcare Text Classification System and Its Performance Evaluation: A Source of Better Intelligence by Characterizing Healthcare Text’. Journal of Medical Systems, vol. 42, no. 5, May 2018, p. 97. DOI.org (Crossref), https://doi.org/10.1007/s10916-018-0941-6.;

302. Paiva, Joana S., et al. ‘Supervised Learning Methods for Pathological Arterial Pulse Wave Differentiation: A SVM and Neural Networks Approach’. International Journal of Medical Informatics, vol. 109, Jan. 2018, pp. 30–38. DOI.org (Crossref), https://doi.org/10.1016/j.ijmedinf.2017.10.011.

ДОДАТОК А

"ЗАТВЕРДЖУЮ"
Проректор з наукової роботи
Національного університету
«Львівська політехніка»
_____ І.В.Демидов
_____ 2022 р.

АКТ
використання наукових результатів
дисертаційної роботи Мельникової Наталії Іванівні,
представленої на здобуття наукового ступеня доктора наук

Комісія в складі: голови комісії - начальника науково-дослідної частини д.т.н., с.н.с Небесного Р.В. та членів комісії - завідувача кафедри СШІ Шаховської Н.Б., професора кафедри СШІ Яковини В.С., доцента кафедри СШІ Хавалка В.М., доцента кафедри СШІ Кривенчука Ю.П. цим актом підтверджують, що результати дисертаційної роботи Мельникової Н.І., зокрема

- стекінгова модель на основі алгоритмів машинного навчання, щодо забезпечення підвищення точності прогнозування даних та паралельної обробки даних як малої, так і великої розмірності;
- гібридний ансамбль моделей машинного навчання для вибору пріорітетних ознак на великих наборах даних, щодо зниження кореляції ознак та збільшення узагальнення моделі;
- метод консолідації мультимодальних даних за рахунок попереднього визначення структур даних та узгодження семантики;

використано у науково-дослідних роботах фінансованих Міністерством освіти і науки України, що виконувались на кафедрі систем штучного інтелекту і включено до звіту:
«Розроблення інформаційної технології оцінювання та прогнозування надійності програмного забезпечення методами машинного навчання» (№ держ. реєстру 0121U109527).
Отримані автором результати використано:

- при розробленні засобів передобробки та аналізу великих наборів даних даних;
- при розробленні засобів оцінки прогнозування рішень за результатами опрацювання та аналізу персоналізованих даних;
- при розробленні засобів аналізу персоналізованих даних та стійкість до застосування в умовах викидів та шумів у нових джерелах даних.

Голова комісії:
начальник
науково-дослідної частини
д.т.н., с.н.с Р.В.Небесний

Члени комісії:
завідувача кафедри СШІ Н.Б.Шаховська

професор кафедри СШІ В.С.Яковина

доцент кафедри СШІ В.М.Хавалко

доцент каефдри СШІ Ю.П.Кривенчук

"ЗАТВЕРДЖУЮ"
Проректор з наукової роботи
Національного університету
«Львівська політехніка»
_____ І.В.Демидов
_____ 2022 р.

АКТ
використання наукових результатів
дисертаційної роботи Мельникової Наталії Іванівні,
представленої на здобуття наукового ступеня доктора наук

Комісія в складі: голови комісії - начальника науково-дослідної частини д.т.н., с.н.с Небесного Р.В. та членів комісії - завідувача кафедри СШІ Шаховської Н.Б., професора кафедри СШІ Яковини В.С., доцента кафедри СШІ Хавалка В.М., доцента кафедри СШІ Кривенчука Ю.П. цим актом підтверджують, що результати дисертаційної роботи Мельникової Н.І., зокрема

- модель відображення стану пацієнта в n-вимірному просторі умов щодо прогнозування динаміки змін цільових показників в підчастині простору з вищою точністю;
- метод заповнення пропусків даних на основі асоціативних правил щодо забезпечення аналізу мультимодальних даних при паралельній реалізації в розподілених базах даних;
- метод пошуку шаблонів зміни стану пацієнта шляхом використання простору умов та аналізу зміни приросту значень часово-залежних даних, що дало змогу зменшити імовірність появи похибки при виборі схеми лікування;

використано у науково-дослідних роботах фінансованих Міністерством освіти і науки України, що виконувались на кафедрі систем штучного інтелекту і включено до звіту:
Проєкт щодо наукових досліджень і розробок «Система підтримки прийняття рішень моделювання поширення вірусних інфекцій» № 211/01.220 ННВЦ від 06.11.2020 р. «Інформаційні технології та системи».
Отримані автором результати використано:

- при розробленні засобів підтримки прийняття рішень щодо прогнозування станів пацієнта шляхом аналізу персоналізованих даних;
- при розробленні засобів оцінки результатів препідготовки даних шляхом імпутації даних за рахунок паралелізації процесів обробки даних та навчання моделей;
- при розробленні засобів щодо пошуку та застосування шаблонів станів хворих для підвищення точності отриманих персоналізованих рішень.

Голова комісії:
начальник науково-дослідної частини
д.т.н., с.н.с Р.В.Небесний

Члени комісії:
завідувача кафедри СШІ Н.Б.Шаховська

професор кафедри СШІ В.С.Яковина

доцент кафедри СШІ В.М.Хавалко

доцент каефдри СШІ Ю.П.Кривенчук

"ЗАТВЕРДЖУЮ"
Проректор з наукової роботи
Національного університету
«Львівська політехніка»
_____ І.В.Демидов
_____ 2022 р.

АКТ
використання наукових результатів
дисертаційної роботи Мельникової Наталії Іванівні,
представленої на здобуття наукового ступеня доктора наук

Комісія в складі: голови комісії - начальника науково-дослідної частини д.т.н., с.н.с Небесного Р.В. та членів комісії - завідувача кафедри СШІ Шаховської Н.Б., професора кафедри СШІ Яковини В.С., доцента кафедри СШІ Хавалка В.М., доцента кафедри СШІ Кривенчука Ю.П. цим актом підтверджують, що результати дисертаційної роботи Мельникової Н.І., зокрема

- модель відображення стану пацієнта в n-вимірному просторі умов щодо прогнозування динаміки змін цільових показників в підчастині простору з вищою точністю;
- гібридний ансамбль моделей машинного навчання для вибору пріорітетних ознак на великих наборах даних, який дозволяє уникнути кореляції ознак та збільшує узагальнення моделі;
- розроблено програмний модуль для опрацювання та аналізу персоналізованих медичних даних старших людей для покращення їх самопочуття та догляду за ними вдома;

використано у науково-дослідних роботах фінансованих Міністерством освіти і науки України, що виконувались на кафедрі систем штучного інтелекту і включено до звіту: Міжнародного проєкту CELTIC EUROGIA PROPOSAL «Integrated care for next generation iCare4NextG».
Отримані автором результати використано:

- при розробленні засобів підтримки прийняття рішень щодо прогнозування станів пацієнта шляхом аналізу персоналізованих даних;
- при розробленні засобів щодо пошуку та застосування шаблонів станів хворих для підвищення точності отриманих персоналізованих рішень;
- при вирішенні задачі пошуку модуль щодо виявлення аномалій по значеннях параметрів пацієнтів у літніх хворих з деменцією.

Голова комісії:
начальник науково-дослідної частини
д.т.н., с.н.с Р.В.Небесний

Члени комісії:
завідувача кафедри СШІ Н.Б.Шаховська

професор кафедри СШІ В.С.Яковина

доцент кафедри СШІ В.М.Хавалко

доцент каефдри СШІ Ю.П.Кривенчук

"ЗАТВЕРДЖУЮ"
Проректор з наукової роботи
Національного університету
«Львівська політехніка»
_____ І.В.Демидов
_____ 2022 р.

АКТ
використання наукових результатів
дисертаційної роботи Мельникової Наталії Іванівні,
представленої на здобуття наукового ступеня доктора наук

Комісія в складі: голови комісії - начальника науково-дослідної частини д.т.н., с.н.с Небесного Р.В. та членів комісії - завідувача кафедри СШІ Шаховської Н.Б., професора кафедри СШІ Яковини В.С., доцента кафедри СШІ Хавалка В.М., доцента кафедри СШІ Кривенчука Ю.П. цим актом підтверджують, що результати дисертаційної роботи Мельникової Н.І., зокрема

- модель відображення стану пацієнта в n-вимірному просторі умов щодо прогнозування динаміки змін цільових показників в підчастині простору з вищою точністю;
- ієрархічний предиктор щодо забезпечення покращення стійкості моделі до нових вхідних даних;
- метод консолідації мультимодальних даних за рахунок попереднього визначення структур даних та узгодження семантики;

використано у науково-дослідних роботах фінансованих Міністерством освіти і науки України, що виконувались на кафедрі систем штучного інтелекту і включено до звіту:
«Інформаційна технологія опрацювання персоналізованої медичної інформації» (№ держ. реєстру 0119U002257).
Отримані автором результати використано:

- при розробленні засобів підтримки прийняття лікарських рішень;
- при розробленні засобів аналізу та прогнозування стану пацієнта під час лікування;
- при розробленні засобів опрацювання гетерогенних даних.

Голова комісії:

начальник
науково-дослідної частини
д.т.н., с.н.с Р.В.Небесний

Члени комісії:

завідувача кафедри СШІ Н.Б.Шаховська

професор кафедри СШІ В.С.Яковина

доцент кафедри СШІ В.М.Хавалко

доцент каефдри СШІ Ю.П.Кривенчук

"ЗАТВЕРДЖУЮ"
Проректор з наукової роботи
Національного університету
«Львівська політехніка»
І.В.Демидов
2022 р.

АКТ
використання наукових результатів
дисертаційної роботи Мельникової Наталії Іванівні,
представленої на здобуття наукового ступеня доктора наук

Комісія в складі: голови комісії - начальника науково-дослідної частини д.т.н., с.н.с Небесного Р.В. та членів комісії - завідувача кафедри СШІ Шаховської Н.Б., професора кафедри СШІ Яковини В.С., доцента кафедри СШІ Хавалка В.М., доцента кафедри СШІ Кривенчука Ю.П. цим актом підтверджують, що результати дисертаційної роботи Мельникової Н.І., зокрема

- модель відображення стану пацієнта в n-вимірному просторі умов щодо прогнозування динаміки змін цільових показників в підчастині простору з вищою точністю;
- стекінгова модель на основі алгоритмів машинного навчання, щодо забезпечення підвищення точності прогнозування даних та паралельної обробки даних як малої, так і великої розмірності;
- метод консолідації мультимодальних даних за рахунок попереднього визначення структур даних та узгодження семантики;

використано у науково-дослідних роботах фінансованих Міністерством освіти і науки України, що виконувались на кафедрі систем штучного інтелекту і включено до звіту:
«Технології та системи оброблення і зберігання персоналізованих військових медичних даних» (№ держ. реєстру 0121U107809).
Отримані автором результати використано:

- при розробленні засобів підтримки прийняття рішень шляхом аналізу персоналізованих даних;
- при розробленні засобів прогнозування стану досліджуваного об'єкта під час реабілітації;
- при розробленні засобів опрацювання мультимодальних даних для процесу препідготовки даних для аналізу.

Голова комісії:
начальник
науково-дослідної частини
д.т.н., с.н.с Р.В.Небесний

Члени комісії:
завідувача кафедри СШІ Н.Б.Шаховська

професор кафедри СШІ В.С.Яковина

доцент кафедри СШІ В.М.Хавалко

доцент каефдри СШІ Ю.П.Кривенчук

"ЗАТВЕРДЖУЮ"
Проректор з наукової роботи
Національного університету
«Львівська політехніка»
І.В.Демидов
"__" _____ 2022 р.

АКТ
використання наукових результатів
дисертаційної роботи Мельникової Наталії Іванівні,
представленої на здобуття наукового ступеня доктора наук

Комісія в складі: голови комісії - начальника науково-дослідної частини д.т.н., с.н.с Небесного Р.В. та членів комісії - завідувача кафедри СШІ Шаховської Н.Б., професора кафедри СШІ Яковини В.С., доцента кафедри СШІ Хавалка В.М., доцента кафедри СШІ Кривенчука Ю.П. цим актом підтверджують, що результати дисертаційної роботи Мельникової Н.І., зокрема

- метод заповнення пропусків даних на основі асоціативних правил щодо забезпечення аналізу мультимодальних даних при паралельній реалізації в розподілених базах даних;
- гібридний ансамбль моделей машинного навчання для вибору пріорітетних ознак на великих наборах даних, який дозволяє уникнути кореляції ознак та збільшує узагальнення моделі;
- розроблено програмний модуль для підтримки прийняття рішень у визначенні схеми лікування пацієнтів з постковідним ефектом;

використано у науково-дослідних роботах фінансованих Міністерством освіти і науки України, що виконувались на кафедрі систем штучного інтелекту і включено до звіту: Міжнародного проєкту Central European Initiatives Extraordinary call «STOP COVID 19».
Отримані автором результати використано:

- при розробленні засобів оцінки результатів препідготовки даних шляхом імпутації даних за рахунок паралелізації процесів обробки даних та навчання моделей;
- при розробленні засобів щодо пошуку та застосування шаблонів станів хворих для підвищення точності отриманих персоналізованих рішень;
- при вирішенні задачі пошуку залежностей між цільовими ознаками хворих на COVID-19, що дозволило довести закономірності поширеності COVID-19 на території різних країн та знайти рішення щодо зниження поширеності COVID-19 серед людей певних соціальних груп.

Голова комісії:
начальник науково-дослідної частини
д.т.н., с.н.с Р.В.Небесний

Члени комісії:
завідувача кафедри СШІ Н.Б.Шаховська

професор кафедри СШІ В.С.Яковина

доцент кафедри СШІ В.М.Хавалко

доцент каефдри СШІ Ю.П.Кривенчук

"ЗАТВЕРДЖУЮ"
Проректор з наукової роботи
Львівського Національного
Медичного університету
ім. Д.Галицького
д.мед.н., проф. Наконечний А.Й.
_____ 10.08 _____ 2022 р.

АКТ

про впровадження результатів дисертаційної роботи

докторанта кафедри «Системи штучного інтелекту»

Національного університету «Львівська політехніка»

Мельникової Наталії Іванівни

Ми, нижчепідписані члени комісії: завідувач кафедри сімейної медицини ФПДО д.м.н., проф. Соломенчук Т.М.,

к. мед. н. доц. Зарецька О.В.

к. мед. к. доц. Федечко М.Й.

склали даний акт про те, що результати дисертаційної роботи Мельникової Н.І. були впроваджені у навчальний процес на кафедрі сімейної медицини ФПДО, зокрема:

- метод заповнення пропусків даних на основі асоціативних правил щодо забезпечення аналізу мультимодальних даних при паралельній реалізації в розподілених базах даних;
- гібридний ансамбль моделей машинного навчання для вибору пріорітетних ознак на великих наборах даних, який дозволяє уникнути кореляції ознак та збільшує узагальнення моделі;
- програмний модуль для підтримки прийняття рішень у визначенні схеми лікування пацієнтів з постковідним ефектом;
- при розробленні засобів оцінки результатів препідготовки даних шляхом імпутації даних за рахунок паралелізації процесів обробки даних та навчання моделей;
- при розробленні засобів щодо пошуку та застосування шаблонів станів хворих для підвищення точності отриманих персоналізованих рішень;
- при вирішенні задачі пошуку залежностей між цільовими ознаками хворих на COVID-19, що дозволило довести закономірності поширеності COVID-19 на території різних країн та знайти рішення щодо зниження поширеності COVID-19 серед людей певних соціальних груп.

Голова комісії проф. Соломенчук Т.М.

Члени комісії

/Зарецька О.В./

/Федечко М.Й./

"ЗАТВЕРДЖУЮ"
Проректор з наукової роботи
Львівського Національного
Медичного університету
ім. Д.Галицького
д.мед.н., проф. Наконечний А.Й.
_____ 2022 р.

про впровадження результатів дисертаційної роботи
докторанта кафедри «Системи штучного інтелекту»
Національного університету «Львівська політехніка»
Мельникової Наталії Іванівни

Ми, нижчепідписані члени комісії: завідувач кафедри сімейної медицини ФПДО д.м.н., проф. Соломенчук Т.М.,

к. мед. н. Зарецька О.В.

к. мед. н. Федечко М.І.

склали даний акт про те, що результати дисертаційної роботи Мельникової Н.І. були впроваджені у навчальний процес на кафедрі сімейної медицини ФПДО, зокрема:

- модель відображення стану пацієнта в n-вимірному просторі умов щодо прогнозування динаміки змін цільових показників в підчастині простору з вищою точністю;
- гібридний ансамбль моделей машинного навчання для вибору пріорітетних ознак на великих наборах даних, який дозволяє уникнути кореляції ознак та збільшує узагальнення моделі;
- розроблено програмний модуль для опрацювання та аналізу персоналізованих медичних даних старших людей для покращення їх самопочуття та догляду за ними вдома;
- при розробці засобів підтримки прийняття рішень щодо прогнозування станів пацієнта шляхом аналізу персоналізованих даних;
- при розробці засобів щодо пошуку та застосування шаблонів станів хворих для підвищення точності отриманих персоналізованих рішень;
- при вирішенні задачі пошуку модуль щодо виявлення аномалій по значеннях параметрів пацієнтів у літніх хворих з деменцією.

Голова комісії проф. Соломенчук Т.М.

Члени комісії (Зарецька О.В.)

(Федечко М.І.)

"ЗАТВЕРДЖУЮ"
Проректор з наукової роботи
Львівського Національного
Медичного університету
ім. Д.Галицького
д.мед.н., проф. Наконечний А.Й.
_____ 2022 р.

АКТ

про впровадження результатів дисертаційної роботи

докторанта кафедри «Системи штучного інтелекту»

Національного університету «Львівська політехніка»

Мельникової Наталії Іванівни

Ми, нижчепідписані члени комісії: завідувач кафедри сімейної медицини ФПДО д.м.н., проф. Соломенчук Т.М.,

к. мед. н. доц. Зарецька О.В.

к. мед. н. доц. Федечко М.І.

склали даний акт про те, що результати дисертаційної роботи Мельникової Н.І. були впроваджені у навчальний процес на кафедрі сімейної медицини ФПДО, зокрема:

- модель відображення стану пацієнта в n-вимірному просторі умов щодо прогнозування динаміки змін цільових показників в підчастині простору з вищою точністю;
- метод заповнення пропусків даних на основі асоціативних правил щодо забезпечення аналізу мультимодальних даних при паралельній реалізації в розподілених базах даних;
- метод пошуку шаблонів зміни стану пацієнта шляхом використання простору умов та аналізу зміни приросту значень часово-залежних даних, що дало змогу зменшити імовірність появи похибки при виборі схеми лікування;

Отримані автором результати використані:

- при розробці засобів підтримки прийняття рішень щодо прогнозування станів пацієнта шляхом аналізу персоналізованих даних;
- при розробці засобів оцінки результатів препідготовки даних щляхом імпутації даних за рахунок паралелізації процесів обробки даних та навчання моделей;
- при розробці засобів щодо пошуку та застосування шаблонів станів хворих для підвищення точності отриманих персоналізованих рішень.

Голова комісії проф. Соломенчук Т.М.

Члени комісії /Зарецька О.В./

/Федечко М.І./

"ЗАТВЕРДЖУЮ"
Проректор з наукової роботи
Львівського Національного
Медичного університету
ім. Д.Галицького
д.мед.н., проф. Наконечний А.Й.

_____ 2022 р.

АКТ

про впровадження результатів дисертаційної роботи

докторанта кафедри «Системи штучного інтелекту»

Національного університету «Львівська політехніка»

Мельникової Наталії Іванівни

Ми, нижчепідписані члени комісії: завідувач кафедри хірургії та трансплантології ФПДО к.мед.н, доц. Щур О.О.

доц. Богар В.Т.

доц. Марічка В.Н.

склали даний акт про те, що результати дисертаційної роботи Мельникової Н.І. були впроваджені у навчальний та лікувальний процеси на кафедрі хірургії та трансплантології ФПДО, зокрема:

- метод заповнення пропусків даних на основі асоціативних правил щодо забезпечення аналізу мультимодальних даних при паралельній реалізації в розподілених базах даних;
- гібридний ансамбль моделей машинного навчання для вибору пріорітетних ознак на великих наборах даних, який дозволяє уникнути кореляції ознак та збільшує узагальнення моделі;
- розроблено програмний модуль для підтримки прийняття рішень у визначенні схеми лікування пацієнтів з постковідним ефектом;
- при розробленні засобів оцінки результатів препідготовки даних шляхом імпутації даних за рахунок паралелізації процесів обробки даних та навчання моделей;
- при розробленні засобів щодо пошуку та застосування шаблонів станів хворих для підвищення точності отриманих персоналізованих рішень;
- при вирішенні задачі пошуку залежностей між цільовими ознаками хворих на COVID-19, що дозволило довести закономірності поширеності COVID-19 на території різних країн та знайти рішення щодо зниження поширеності COVID-19 серед людей певних соціальних груп.

Голова комісії Доц. Щур О.О.

Члени комісії

"ЗАТВЕРДЖУЮ"
Проректор з наукової роботи
Львівського Національного
Медичного університету
ім. Д.Галицького
д.мед.н., проф. Наконечний А.Й.

_____ 10.08 2022 р.

АКТ

про впровадження результатів дисертаційної роботи

докторанта кафедри «Системи штучного інтелекту»

Національного університету «Львівська політехніка»

Мельникової Наталії Іванівни

Ми, нижчепідписані члени комісії: завідувач кафедри хірургії та трансплантології ФПДО к.мед.н, доц. Щур О.О.

доц. Бочар В.Т.

доц. Маріна В.Н.

склали даний акт про те, що результати дисертаційної роботи Мельникової Н.І. були впроваджені у навчальний та лікувальний процеси на кафедрі хірургії та трансплантології ФПДО, зокрема:

- модель відображення стану пацієнта в n-вимірному просторі умов щодо прогнозування динаміки змін цільових показників в підчастині простору з вищою точністю;
- гібридний ансамбль моделей машинного навчання для вибору пріорітетних ознак на великих наборах даних, який дозволяє уникнути кореляції ознак та збільшує узагальнення моделі;
- розроблено програмний модуль для опрацювання та аналізу персоналізованих медичних даних старших людей для покращення їх самопочуття та догляду за ними вдома;

Отримані автором результати використані:

- при розробленні засобів підтримки прийняття рішень щодо прогнозування станів пацієнта шляхом аналізу персоналізованих даних;
- при розробленні засобів щодо пошуку та застосування шаблонів станів хворих для підвищення точності отриманих персоналізованих рішень;
- при вирішенні задачі пошуку модуль щодо виявлення аномалій по значеннях параметрів пацієнтів у літніх хворих з деменцією.

Голова комісії Доц. Щур О.О.

Члени комісії

"ЗАТВЕРДЖУЮ"
Проректор з наукової роботи
Львівського Національного
Медичного університету
ім. Д.Галицького
д.мед.н., проф. Наконечний А.Й.

10.08. 2022 р.

АКТ

про впровадження результатів дисертаційної роботи

докторанта кафедри «Системи штучного інтелекту»

Національного університету «Львівська політехніка»

Мельникової Наталії Іванівни

Ми, нижчепідписані члени комісії: завідувач кафедри хірургії та трансплантології ФПДО к.мед.н, доц. Щур О.О.

доц. Богар В.Т

доц. Маріна В.Н.

склали даний акт про те, що результати дисертаційної роботи Мельникової Н.І. були впроваджені у навчальний та лікувальний процеси на кафедрі хірургії та трансплантології ФПДО, зокрема:

- модель відображення стану пацієнта в n-вимірному просторі умов щодо прогнозування динаміки змін цільових показників в підчастині простору з вищою точністю;
- метод заповнення пропусків даних на основі асоціативних правил щодо забезпечення аналізу мультимодальних даних при паралельній реалізації в розподілених базах даних;
- метод пошуку шаблонів зміни стану пацієнта шляхом використання простору умов та аналізу зміни приросту значень часово-залежних даних, що дало змогу зменшити імовірність появи похибки при виборі схеми лікування;

Отримані автором результати використано:

- при розробленні засобів підтримки прийняття рішень щодо прогнозування станів пацієнта шляхом аналізу персоналізованих даних;
- при розробленні засобів оцінки результатів препідготовки даних щляхом імпутації даних за рахунок паралелізації процесів обробки даних та навчання моделей;
- при розробленні засобів щодо пошуку та застосування шаблонів станів хворих для підвищення точності отриманих персоналізованих рішень.

Голова комісії Доц. Щур О.О.

Члени комісії

ЗАТВЕРДЖУЮ
Голова «Львівської асоціації алергологів,
імунологів, імунореабілітологів»
ЧОПЯК Валентина Володимирівна
«25» серпня 2022 р.

АКТ

про впровадження результатів дисертаційної роботи
докторанта кафедри «Системи штучного інтелекту»
Національного університету «Львівська політехніка»
Мельникової Наталії Іванівни

Цей акт підтверджує, що результати дисертаційної роботи Мельникової Н.І. були використані для розроблення інформаційної системи підтримки прийняття рішень для лікування хворих з орфанними хворобами «Реєстр ПІД з порушеннями антитілоутворень» в рамках виконання Госпдоговір ТОВ «БІОФАРМА ПЛАЗМА» № Т1 0625 від 11 вересня 2019 р. Львівською асоціацією алергологів, імунологів, імунореабілітологів.

Впровадження дисертаційних досліджень Н.І. Мельникової полягає у наступному:

- розроблено інформаційну модель відображення стану пацієнта в n-вимірному просторі умов, що дало змогу прогнозувати динаміку змін цільових показників в підчастині простору з вищою точністю та забезпечило індивідуальний підхід до моніторингу стану пацієнта на основі тривалого спостереження та контролю лікаря;
- розроблено ієрархічний предиктор, який включає двоетапну обробку малих наборів даних методами кластеризації об'єктів та прогнозування для кожного одержаного кластера, що забезпечує підвищити точність класифікації для набору даних по орфанних хворобах за рахунок ієрархічної класифікації та поєднання різних моделей машинного навчання;
- розроблено стекінгову модель на основі алгоритмів машинного навчання, яка на

відміну від подібних моделей, базується на деформації метаознак та повторному навчанні на розширеному наборі даних, що забезпечує підвищення точності прогнозування даних та паралельної обробки даних як малої, так і великої розмірності;

– розроблено програмні модулі інформаційної системи підтримки прийняття рішень у лікуванні орфанних хворіб «Реєстр ПІД з порушеннями антитілоутворень».

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

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

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

від 4.11.2022 № 60-0000001009
на _____ від _____

biopharma

Керівнику

Національного університету "Львівська політехніка"

Даним листом ТОВ «БІОФАРМА ПЛАЗМА» підтверджує, що не заперечувало проти того, що Національним університетом "Львівська політехніка" для виконання робіт (згідно укладених між ТОВ «БІОФАРМА ПЛАЗМА» та Національним університетом "Львівська політехніка" договорів) буде залучено (протягом 2019 -2021 рр.) громадянку Мельникову Н.І. (за її згодою) для розроблення інформаційної системи «Реєстр первинних імунодефіцитів».

В.о. директора **Плотнікова О.М.**

+380 (44) 390 08 10
info@biopharma.ua
www.biopharma.ua

Товариство з обмеженою відповідальністю
«Біофарма Плазма»
Україна, 09100, Біла Церква, ул. Київська 37-В,
IBAN UA883005280000026000455073747 в АТ "ОТП БАНК" м. Київ,
МФО 300528, код ЄДРПОУ 39000694

000976

Впровадження в межах Госпдоговору ТОВ «БІОФАРМА ПЛАЗМА» № Т1 0625 від 11 вересня 2019 р.

ДОГОВІР № Т1 /0625

м. Львів 11 вересня 2019 р.

ТОВ «БІОФАРМА ПЛАЗМА» (надалі Замовник), в особі директора Єфименко Костянтина Олексійовича, який діє на підставі Статуту, з однієї сторони, та

Національний університет «Львівська Політехніка» (надалі Виконавець), в особі проректора з наукової роботи Чухрай Наталії Іванівни, що діє на підставі наказу № 261-1-10 від 03.06.2015 р., з іншої сторони, уклали цей Договір про наступне:

1. Предмет договору

1.1. Виконавець бере на себе зобов'язання за завданням та за кошти Замовника виконати роботи з розробки інформаційної системи «Реєстр первинних імунодефіцитів» (надалі – Проект), а Замовник зобов'язується прийняти належно виконані роботи та оплатити їх в порядку та на умовах визначених даним Договором.

1.2. Вимоги до робіт, що мають бути виконані Виконавцем, визначаються в Технічних вимогам (Додаток1), що є невід'ємною складовою цього Договору.

2. Вартість робіт

2.1. Вартість робіт по Проекту за цим Договором визначена плановим кошторисом (Додаток 2), який є твердим та підлягає зміні виключно за письмовим погодженням Сторін, та узгоджена у Протоколі погодження договірної ціни (Додаток 3), що є невід'ємними частинами цього Договору і складає 362 000 грн. (триста шістдесят дві тисячі гривень), в т.ч. ПДВ – 60 333,33 грн.

2.2. Ціни встановлені у національній валюті України – гривні.

2.3. У вартість робіт по Проекту включено вартість виключних майнових прав на створенні за цим Договором об'єкти інтелектуальної власності, що будуть передані Замовнику разом з результатами робіт.

3. Порядок здійснення оплати

3.1. Оплата виконаних і прийнятих Замовником робіт здійснюється на підставі акту здавання-приймання робіт та рахунку на оплату, згідно погодженого Сторонами Календарного плану-графіку (Додаток 4).

3.2. Розрахунок здійснюється у безготівковій формі шляхом перерахування коштів на рахунок Виконавця протягом 10 (десяти) банківських днів з дати підписання Сторонами акту здавання-приймання виконаних робіт.

4. Термін виконання робіт

4.1. Термін здавання (виконання) робіт за Договором - до 25 серпня 2021 року.

4.2. Зміст і терміни виконання основних етапів Проекту визначені Календарним планом-графіком (Додаток 4), що є невід'ємною частиною цього Договору.

4.3. Виконання Проекту підтверджується підписаним Сторонами актом здавання-приймання виконаних робіт.

5. Права та обов'язки сторін

5.1. Замовник зобов'язується:

5.1.1. Надати Виконавцеві всю необхідну для виконання цього Договору інформацію.

5.1.2. Прийняти виконані роботи по Проекту та впродовж п'яти робочих днів підписати акт здавання-приймання виконаних робіт або надати мотивовану відмову від приймання виконаних робіт.

5.1.3. У разі мотивованої відмови Замовника від прийняття робіт сторони складають двосторонній Акт з переліком необхідних доопрацювань і термінів їхнього виконання.

5.1.4. Якщо Замовник у п'ятиденний термін не повернув Виконавцю підписаний акт здавання-приймання виконаних робіт і не надав мотивованої відмови від приймання робіт, то роботи вважаються виконаними, прийнятими Замовником і підлягають оплаті.

5.1.5. Своєчасно та в повному обсязі оплатити виконані роботи в порядку, визначеному у Розділі 3.

5.2. Замовник має право:

5.2.1. Здійснювати контроль за ходом і якістю виконання робіт, не втручаючись у діяльність Виконавця.

5.2.2. Повернути рахунок Виконавцю без здійснення оплати в разі неналежного оформлення документів для дооформлення документів відповідно до вимог чинного законодавства та умов цього Договору.

5.3. Виконавець зобов'язується:

5.3.1. Виконати роботи по Проекту у визначені строки згідно з Додатком 4.

5.3.2. Забезпечити високу якість продукту за Проектом.

5.3.3. Не розголошувати конфіденційну інформацію (будь-яку інформацію, яка отримана Виконавцем від Замовника для виконання цього Договору, в т. ч. сам Договір та його умови) та дотримуватися вимог по збереженню конфіденційності протягом дії цього Договору, та у разі його дострокового розірвання або припинення (закінчення строку дії договору) протягом наступних 5 (п'яти) років від такої події.

5.4. Виконавець має право:

5.4.1. Своєчасно та в повному обсязі отримувати плату за виконані роботи на умовах цього Договору.

6. Відповідальність Сторін

6.1. За невиконання або неналежне виконання зобов'язань за цим Договором Виконавець і Замовник відповідають згідно з чинним законодавством України.

6.2. У разі несвоєчасного виконання зобов'язань за цим Договором, Виконавець сплачує Замовнику неустойку у розмірі подвійної облікової ставки НБУ від суми Договору, за кожен день прострочення.

6.3. За несвоєчасне проведення розрахунків Замовник сплачує виконавцю пеню у розмірі подвійної облікової ставки НБУ від суми, з якої допущено прострочення виконання, за кожен день прострочення.

6.4. Сплата неустойки (пені) не звільняє Сторони від виконання зобов'язань за цим Договором.

6.5. Виконавець несе відповідальність за правильність підготовки кошторису.

6.6. У випадку розголошення конфіденційної інформації, за виключенням випадку, коли така інформація надається на вимогу суду чи правоохоронних органів, винна Сторона зобов'язується відшкодувати в повному обсязі іншій Стороні завданні такими діями збитки.

7. Обставини непереборної сили

7.1. Сторони звільняються від відповідальності за невиконання або неналежне виконання зобов'язань за цим Договором у разі виникнення обставин непереборної сили, які не існували під час укладання Договору та виникли поза волею Сторін (аварія, катастрофа, стихійне лихо, епідемія, епізоотія, війна тощо).

7.2. Сторона, що не може виконувати зобов'язання за цим Договором унаслідок дії обставин непереборної сили, повинна негайно з моменту їх виникнення повідомити про це іншу Сторону у письмовій формі.

7.3. Доказом виникнення обставин непереборної сили та строку їх дії є відповідні документи, які видаються уповноваженим органом.

7.4. У разі настання обставин непереборної сили кожна із Сторін в установленому порядку має право розірвати цей Договір, повідомивши іншу сторону у термін не пізніше десяти днів після настання обставин.

8. Вирішення спорів

8.1. У випадку виникнення спорів або розбіжностей Сторони зобов'язуються вирішувати їх шляхом взаємних переговорів та консультацій.

8.2. У разі не досягнення Сторонами згоди спори (розбіжності) вирішуються у судовому порядку.

9. Термін дії Договору

9.1. Договір набуває чинності з моменту його підписання і діє до 25 серпня 2021 року але в будь-якому випадку до повного виконання Сторонами своїх зобов'язань відповідно до умов Договору.

9.2. Договір укладено в двох примірниках, що мають однакову юридичну силу - один для Виконавця і один для Замовника.

10. Інші умови

10.1. Взаємовідносини Сторін, не обумовлені у Договорі, регламентуються чинним законодавством України.

10.2. Умови даного Договору можуть бути змінені за взаємною згодою Сторін з обов'язковим складанням письмового документу (додаткова угода про зміну Договору).

10.3. Виконавець має право залучати до виконання робіт за цим Договором інші організації та підприємства, залишаючись відповідальним перед Замовником за якість та своєчасність виконуваних ними робіт.

10.4. Виконавець має право на використання результатів робіт за цим Договором або їх фрагментів виключно у своїй науковій діяльності.

10.5. Сторони зобов'язуються вказувати у відповідних рекламних матеріалах, а також на продукції, створеній за цим Договором розробника (автора) Проекту.

10.6. Сторони домовилися, що всі виключні майнові права на об'єкти права інтелектуальної власності, створені на підставі цього Договору, належать Замовнику.

11. Додатки

11.1. Технічні вимоги (Додаток №1).
11.2. Плановий кошторис (Додаток №2).
11.3. Протокол погодження договірної ціни (Додаток №3).
11.4. Календарний план-графік (Додаток №4).

12. ЮРИДИЧНІ АДРЕСИ, БАНКІВСЬКІ РЕКВІЗИТИ ТА ПІДПИСИ СТОРІН:

ЗАМОВНИК:
ТОВ «БІОФАРМА ПЛАЗМА»
09100, м. Біла Церква, Київська обл.,
вул. Київська, 37
ЄДРПОУ 39000694
р/р № 26006015665702
в АТ «Альфа-Банк» м. Київ
МФО 300346
ІПН 390006910270

Директор
_______________ **К.О. Єфименко**
М.П.

ВИКОНАВЕЦЬ:
Національний університет
«Львівська політехніка»
Юридична адреса:
79013, м. Львів, вул. Ст. Бандери, 12,
р/р 31258267101057 в ДКС України,
МФО 820172,
код за ЄДРПОУ 02071010, ідентифікаційний № п/п 020710113025, свідоцтво про реєстрацію №17701600
Організація включена до Реєстру неприбуткових установ та організацій на підставі рішення контролюючого органу ДФС № 1612/0146/360 від 30.11.2016
Проректор з наукової роботи
Національного університету
«Львівська політехніка»
_______________ **Н.І. Чухрай**
М.П.
Начальник НДЧ
_______________ Л.В. Жук
Зав. відділом науково-організаційного супроводу наукових досліджень
_______________ Г.В. Лазько
Науковий керівник, керівник роботи
_______________ Н.Б. Шаховська
Начальник юридичного відділу
_______________ А.М. Мороз

Додаток 1
до Договору № Т1/06
від 11 вересня 2019 р.

Технічні вимоги

Проект призначений для формування реєстру первинних імунодефіцитів, що забезпечить облік хворих на дану патологію, внесення їхніх результатів обстеження, моніторинг стану хворого за результатами застосування імуноглобулінів, забезпечення можливості аналізу динаміки перебігу хвороби та корегування імуноглобулінотерапії у разі необхідності.

Від всіх осіб, данні про які будуть використовуватися у Проекті, відповідно до вимог чинного законодавства України Виконавцем має бути отримана згода на обробку їх персональних даних.

Об`єкти інформатизації:

1) Хворі на первинний імунодефіцит,
2) Лікарі-дослідники,
3) Нозологічні вимоги,
4) Імуноглобуліни,
5) Структура центрів всеукраїнського, регіонального та обласного рівнів.

Основними складовими Проекту є:

- Серверна частина з базою даних, яка містить інформацію про об'єкти інформатизації, роботу пресів,
- Клієнтська частина для внесення інформації про об'єкти інформатизації та формування аналітичних звітів.

ВИКОНАВЕЦЬ:

Проректор з наукової роботи
Національного університету
«Львівська політехніка»

_____ **Чмир І.І.**
(підпис) (прізвище)

М.П.

ЗАМОВНИК:

Директор
ТОВ «БІОФАРМА ПЛАЗМА»

_____ **Єфименко К.О.**
(підпис) (прізвище)

М.П.