Харківський національний університет імені В. Н. Каразіна

Міністерство освіти і науки України

Кваліфікаційна наукова праця

На правах рукопису

ДОНЕЦЬ ВОЛОДИМИР ВІТАЛІЙОВИЧ

УДК 004.67:616-71

ДИСЕРТАЦІЯ

МЕТОДИ Й МОДЕЛІ СТРАТИФІКАЦІЇ ЕЛЕМЕНТІВ КОМП'ЮТЕРНИХ СИСТЕМ МЕДИЧНОГО МОНІТОРИНГУ НА ОСНОВІ МУЛЬТИАГЕНТНОГО ПІДХОДУ

Спеціальність 122 – Комп'ютерні науки

Галузь знань 12 – Інформаційні технології

Подається на здобуття ступеня доктора філософії

Дисертація містить результати власних досліджень. Використання ідей, результатів і текстів інших авторів мають посилання на відповідне джерело.

_____ **В. В. Донець**

Науковий керівних: **Шматков Сергій Ігорович**, доктор технічних наук, професор кафедри теоретичної та прикладної системотехніки

Харків – 2024

АНОТАЦІЯ

Донець В. В. Методи й моделі стратифікації елементів комп'ютерних систем медичного моніторингу на основі мультиагентного підходу. – Кваліфікаційна наукова праця на правах рукопису.

Дисертація на здобуття ступеня доктора філософії за спеціальністю 122 – Комп'ютерні науки (Галузь знань 12 – Інформаційні технології). – Харківський національний університет імені В. Н. Каразіна, Міністерства освіти і науки України, Харків, 2024.

Дисертаційна робота присвячена розробці методів і моделей стратифікації елементів даних в комп'ютерних системах медичного моніторингу з використанням мультиагентного підходу, що є багатоетапною задачею з необхідністю узгодження протилежних цілей. Стратифікація – це багатоетапний процес визначення можливих станів пацієнтів по потоку даних, що генерує комп'ютерна система медичного моніторингу, подальша їх класифікація та виявлення впливу змінних стану. Процес стратифікації включає три етапи: кластеризації даних, класифікація стану пацієнтів та виявлення впливу змінних стану. Під мультиагентним підходом розуміється підхід елітарного відбору, що реалізований в методі кластеризації й полягає у відборі найкращих кластерів, які є агентами в просторі генерованих даних, за певною метрикою серед визначених станів.

Перший розділ присвячений огляду існуючих досліджень та наробок в сфері комп'ютерних систем медичного моніторингу. Наведено стислий огляд досліджень присвячених застосуванню методів машинного навчання для підтримки прийняття рішень в таких системах. Досліджено три типи комп'ютерних систем медичного моніторингу, такі як системи на основі нечіткої логіки, методів машинного навчання та глибинного навчання. Визначено принципи їх функціонування й застосування, а також переваги й недоліки застосування й шляхи вирішення цих проблем. Зазначено, що на сьогодні в епоху розвитку комп'ютерних систем медичного моніторингу постає проблема аналізу великого потоку різнотипних

даних. Показано, що такі дані можуть допомогти покращити якість лікування, проте обмеження в кількості спеціалістів вимагає створення систем автоматичного аналізу таких даних із можливістю виділення можливих станів пацієнтів та корекція їх лікування.

Згідно з виявленими проблемами комп'ютерних систем медичного моніторингу, існуючих методів і моделей їх вирішення було визначено мету дослідження, а саме підвищення точності діагностування стану пацієнтів за рахунок реалізації методів і моделей стратифікації елементів комп'ютерних систем медичного моніторингу. З урахуванням мети сформульовано задачу дослідження удосконалення або розробка нових математичних моделей та обчислювальних методів стратифікації елементів комп'ютерних систем медичного моніторингу, що дозволить підвищити точність діагностування стану пацієнта. Показано, що для вирішення цієї задачі необхідно вирішити завдання кластеризації даних, класифікації стану пацієнта та визначення загальної та поточної інформативності змінних стану.

З урахування зазначеної мети й завдань дослідження була запропонована процедура стратифікації, відповідно до якої була розроблена модель комп'ютерної системи медичного моніторингу з виділеною підсистемою стратифікації в ній. Пояснена роль кожного модулю в моделі комп'ютерної системи медичного моніторингу та визначено режими функціонування підсистеми стратифікації в залежності від наявності інформації про можливі стани чи їх кількості.

У другому розділі було детально розглянуто та описано компоненти підсистеми стратифікації елементів в комп'ютерних системах медичного моніторингу. Для цього було запропоновано мультиагентний метод нечіткої кластеризації, відповідно до виявлених проблем методів кластеризації. Зазначено, що розроблений мультиагентний метод нечіткої кластеризації поєднує в собі мультиагентний метод відбору еліт з базовою процедурою модифікації центрів кластерів та можливістю широкого застосування різнорідних метрик для визначення щільності й роздільності отриманих кластерів. Запропонований метод вирішує задачу кластеризації даних. Перевірку точності застосування

запропонованого методу запропоновано проводити з використанням методу класифікації на основі розробленого методу кластеризації.

Далі була розглянута архітектура можливої моделі ШНМ для вирішення проблеми класифікації даних. Показаний метод навчання цієї ШНМ для пришвидшення сходження градієнтів та підвищення точності роботи моделі. А також запропоновано використання алгоритму рою для конфігурації гіперпараметрів моделі ШНМ, що власне визначає її архітектуру.

Запропоновано метод визначення загальної інформативності змінних стану з використанням інформації про поширення градієнтів сигналу в навченій моделі ШНМ. Цей метод дозволяє обчислити інтенсивність впливу градієнтів сигналу на загальні результати функціонування навченої моделі ШНМ. Це в свою чергу дозволяє вирішити проблему визначення множини найбільш впливових змінних стану та за необхідності зменшити кількість спостережуваних параметрів. Далі представлено модифікацію методу інтегрованих градієнтів, що дозволяє визначати вплив на результати класифікації стану за конкретними змінними навченої моделі ШНМ, визначаючи поточну інформативність змінних. Визначення поточної інформативності дозволяє вирішити проблему визначення причин прийняття рішень в комп’ютерній системі медичного моніторингу.

У третьому розділі проведено аналіз програмних засобів реалізації методів і моделей стратифікації. Для цього програмне забезпечення було розглянуте як сукупність трьох компонентів, а саме компоненти мови програмування, що дозволить реалізувати розроблені методи та моделі стратифікації; компоненти доступних бібліотек обробки даних, швидких математичних обчислень та реалізованих методів машинного навчання; компоненти інтегрованого програмного забезпечення. Що в сукупності стало обґрунтуванням вибору мови програмування Python із бібліотеками NumPy, Pandas, Matplotlib, Seaborn, Tensorflow, SciKit Learn та інших; та інтегрованої середи розробки PyCharm із засобами перевірки якості коду на базі методів штучного інтелекту. Усі вище перераховані програмні інструменти дозволили швидко і точно реалізувати й перевірити методи і моделі стратифікації.

Також у розділі були наведені набори даних для проведення валідації і верифікації запропонованих методів і моделей стратифікації даних в комп'ютерних системах медичного моніторингу. Були виділені набори для перевірки точності роботи кожного з запропонованих методів і моделей по окремості та набори для загальної перевірки підсистеми стратифікації. Також розглянуті набори даних для перевірки можливості розширення сфери застосування до комп'ютерних систем економічного моніторингу.

Розділ завершено описом методу верифікації програмної реалізації розроблених методів і моделей стратифікації. Показані типові для індустрії методи й процедури перевірки точності функціонування методів кластеризації й класифікації, що допоможуть виявити якість функціонування розробленого програмного забезпечення. Також розроблено принципи перевірки розроблених методів визначення загальної та поточної інформативності.

В четвертому розділі було розглянуто результати практичного застосування та тестування розроблених методів і моделей із даними медичного моніторингу. Показано результати тестування запропонованого мультиагентного методу кластеризації; методу навчання моделі ШНМ та методів визначення загальної та поточної інформативності. Також дані медичного моніторингу були застосовані для перевірки точності визначення станів пацієнтів за допомогою розробленої підсистеми стратифікації в комп'ютерній системі медичного моніторингу.

В результаті проведення загального тестування підсистеми стратифікації показано, що розроблений мультиагентний метод кластеризації має задовільну точність формування цільових кластерів на використаному наборі даних медичного моніторингу. Це свідчить про необхідність підбору методів кластеризації для конкретних даних задля збільшення точності кластеризації. Отримано, що розроблений метод навчання та налаштування гіперпараметрів моделі ШНМ призводить до високої точності класифікації не тільки на даних розмічених методом кластеризації, а і на оригінальних даних медичного моніторингу. Далі було визначено, що розроблений метод визначення загальної інформативності здатен визначати співставно інформативність до інших існуючих методів, проте

має більш лінійну природу визначення ваг інформативності. Також шляхом кросвалідації модифікований метод інтегрованих градієнтів для визначення поточної інформативності з методом визначення загальної інформативності показав точні результати визначення впливу певних вхідних змінних на результати класифікації моделлю ШНМ. Що засвідчує можливість застосування методу визначення поточної інформативності для обґрунтування прийнятих рішень в комп'ютерній системі медичного моніторингу.

Також було перевірено можливість розширення спектру застосування запропонованих методів і моделей на даних економічного моніторингу. Було виявлено високу чутливість методу кластеризації до незбалансованих даних. Та перевірене застосування методів і моделей стратифікації до даних економічного моніторингу країн із позитивним результатом впровадження.

В кінці розділу відповідно до результатів тестування наведені практичні рекомендації щодо застосування розроблених методів і моделей стратифікації окремо і підсистеми стратифікації в комп'ютерній системі медичного моніторингу в цілому.

Сукупність отриманих у дисертації наукових результатів, підтвердження факту їх достовірності, наукової та практичної значущості дають змогу вважати, що сформульована наукова задача модифікації або розробки нових математичних моделей та обчислювальних методів для досягнення поставленої мети підвищення точності стратифікації елементів комп'ютерних системах медичного моніторингу, – розв'язаною, а поставлену мету – досягнутою.

Ключові слова: *методи машинного навчання, мультиагентний підхід, нечітка кластеризація, кластерний аналіз, штучна нейронна мережа, підбір гіперпараметрів штучних нейронних мереж, класифікація стану пацієнтів, автоматизація аналізу даних, вплив вхідних змінних, методи визначення інформативності змінних стану, аналіз даних захворювань, аналіз медичних даних, пояснення прийнятих рішень, комп'ютерна система медичного моніторингу, верифікації програмного забезпечення.*

ABSTRACT

Donets V. V. Methods and models of elements stratification of computer medical monitoring systems based on a multi-agent approach. – Qualification scholarly paper: a manuscript.

The dissertation submitted for obtaining the Doctor of Philosophy degree in Informational Technology: Speciality 122 – Computer science. V. N Karazin Kharkiv National University, Ministry of Education and Science of Ukraine, Kharkiv, 2024.

The dissertation is devoted to developing methods and models of stratification of data elements in computer medical monitoring systems using a multi-agent approach, which is a multi-stage task with the need to reconcile opposing goals. Stratification is a multi-stage process of determining the possible conditions of patients based on the flow of data generated by the computer system of medical monitoring, their further classification, and the identification of the state variables' influence. The stratification process includes three stages: data clustering, patients' state classification, and identifying the influence of state variables. The multi-agent approach refers to the elite selection approach, which is implemented in the clustering method and consists of selecting the best clusters, which are agents in the space of generated data, according to a certain metric among the defined states.

The first chapter reviews existing research and developments in the field of computer medical monitoring systems. A brief review of research devoted to the application of machine learning methods to support decision-making in such systems is given. Three types of computer-based medical monitoring systems have been researched, such as systems based on fuzzy logic, machine learning methods, and deep learning methods. The principles of their functioning and application, as well as the advantages and disadvantages of their application and ways of solving these problems, are defined. It is noted that today, in the era of the development of computer systems of medical monitoring, the problem of analyzing a large flow of various types of data arises. It has been shown that such data can help to improve the quality of treatment, however, the limitation in the number of specialists requires the creation of systems for automatic

analysis of such data with the possibility of identifying possible patients' conditions and adjusting their treatment.

According to the identified problems of computer medical monitoring systems, existing methods, and models of their solution, the research goal was determined to increase the accuracy of diagnosing the condition of patients due to the implementation of methods and models of elements stratification of computer medical monitoring systems. Considering the goal, the research task is formulated as the improvement or development of new mathematical models and computational methods of stratification of elements of computer systems of medical monitoring, which will increase the accuracy of diagnosing the patient's condition. It is shown that solving this problem is possible by solving tasks of data clustering, the patient's condition classification, and the determination of the general and current informativeness of the state variables.

Taking into account the specified goal and objectives of the study, a stratification procedure was proposed, according to which a model of a computer medical monitoring system with a separate stratification subsystem was developed. The role of each module in the computer medical monitoring system model is explained, and the modes of operation of the stratification subsystem are defined, depending on the availability of information about possible states or their number.

In the second chapter, the components of the stratification subsystem in computer medical monitoring systems were examined and described in detail. For this, a multi-agent method of fuzzy clustering was proposed, by the identified problems of clustering methods. It is noted that the developed multi-agent method of fuzzy clustering combines the multi-agent method of elite selection with the basic procedure of modification of cluster centers and the possibility of wide application of heterogeneous metrics to determine the density and resolution of the obtained clusters. The proposed method solves the problem of data clustering. It is proposed to check the accuracy of the proposed method using the classification method based on the developed clustering method.

Next, the architecture of a possible ANN model for solving the problem of data classification was considered. The method of training this ANN is shown to speed up gradient ascent and increase the accuracy of the model. Also, the swarm algorithm is

suggested to use for the configuration of the hyperparameters of the ANN model, which determines its architecture.

A method for determining the general informativeness of state variables using information about the distribution of signal gradients in the trained ANN model is offered. This method allows you to calculate the intensity of the signal gradients' influence on the general results of the functioning of the trained ANN model. This, in turn, allows solving the problem of determining the set of the most influential state variables and, if necessary, reducing the number of observed parameters. Next, a modification of the integrated gradients method is presented, which allows determining the influence on the results of state classification by specific variables by the trained ANN model, determining the current informativeness of the variables. Determining the current informativeness allows to solve the problem of determining the reasons for decision-making in the computer medical monitoring systems.

In the third section, an analysis of software tools for implementing stratification methods and models is carried out. For this, the software was considered as a set of three components: programming language, which will allow the implementation of the developed stratification methods and models; available data processing libraries, fast mathematical calculations, and implemented machine learning methods; and integrated software. These together became the rationale for choosing the Python programming language with NumPy, Pandas, Matplotlib, Seaborn, Tensorflow, SciKit Learn, and other libraries; and PyCharm integrated development environment with tools for checking the quality of code based on artificial intelligence methods. All of the software tools listed above made it possible to quickly and accurately implement and test stratification methods and models.

Datasets for validation and verification of the proposed methods and models of data stratification in computer medical monitoring systems were presented in the section. Sets for checking the accuracy of each of the proposed methods and models separately and sets for general checking of the stratification subsystem were allocated. Data sets for testing the possibility of expanding the scope of application to computer systems of economic monitoring are also considered.

The chapter concludes with a description of the verification method of the developed stratification methods and models software implementation. Industry-typical methods and procedures for checking the accuracy of the functioning of clustering and classification methods are shown, which will help to reveal the quality of the functioning of the developed software. The principles of checking the developed methods of determining general and current informativeness have also been developed.

In the fourth chapter, the results of practical application and testing of the developed methods and models on medical monitoring data were considered. The results of testing the proposed multi-agent clustering method are shown; the method of learning the ANN model and the methods of determining the general and current informativeness. Also, medical monitoring data were used to check the accuracy of determining patients' conditions using the developed stratification subsystem in the computer medical monitoring system.

As a result of general testing of the stratification subsystem, it is shown that the developed multi-agent clustering method has a satisfactory accuracy of forming target clusters on the used set of medical monitoring data. This indicates the need to select clustering methods for specific data to increase the clustering accuracy. It was further determined that the developed training and configuration hyperparameters of the ANN model methods lead to high classification accuracy not only on the data marked by the clustering method but also on the original medical monitoring data. It was determined that the developed method of determining general informativeness is capable of determining informativeness in comparison to other existing methods, but has a more linear nature of determining informativeness weights. Also, through cross-validation, the modified method of integrated gradients for determining the current informativeness with the method for determining the general informativeness showed accurate results of determining the influence of certain input variables on the results of classification by the ANN model. This proves the possibility of applying the method of determining the current informativeness to justify the decisions made in the computer system of medical monitoring.

The possibility of expanding the range of applications of the proposed methods and models on economic monitoring data was also checked. High sensitivity of the clustering method to unbalanced data was revealed. And the proven application of stratification methods and models to economic monitoring data of countries with a positive result of implementation.

Practical recommendations are given for the application of the developed stratification methods and models separately and the stratification subsystem in the computer system of medical monitoring as a whole, according to the test results.

The complex of the scientific results obtained in the dissertation, confirmation of the fact of their reliability, and scientific and practical significance allow consideration that the formulated scientific task of modification or development of new mathematical models and computational methods to achieve the set goal of increasing the accuracy of stratification of elements in computer systems of medical monitoring, – development bound, and the set goal achieved.

Keywords: *machine learning methods, multi-agent approach, fuzzy clustering, cluster analysis, artificial neural network, hyperparameters selection of artificial neural networks, patients' condition classification, automation of data analysis, input variables influence, methods of determining the informativeness of state variables, diseases data analysis, analysis of medical data, explaining the decisions made, computer medical monitoring system, software verification*.

СПИСОК ПУБЛІКАЦІЙ ЗДОБУВАЧА ЗА ТЕМОЮ ДИСЕРТАЦІЇ

Статті у наукових фахових виданнях, що входять до міжнародних наукометричних баз

1. Viktoriia Strilets, Volodymyr Donets, Mykhaylo Ugryumov, Sergii Artiukh, Roman Zelenskyi, Tamara Goncharova. Agent-oriented data clustering for medical monitoring. Radioelectronic And Computer Systems. 2022. V. 2022. Issue 1. P. 103–114. Keywords: clustering; fuzzy clustering; agent-based approach; intraclass distance; medical diagnostic.

DOI: 10.32620/reks.2022.1.08 **(Scopus)**.

URL: http://nti.khai.edu/ojs/index.php/reks/article/view/reks.2022.1.08/0

(*Особистий внесок здобувача: розробка програмної реалізації методу мультиагентної нечіткої кластеризації з впровадженням різних метрик міжелементної відстані та проведення тестування на даних Ірисів Фішера та проведення кластеризації на даних медичного діагностування. Відповідні результати наведені в теоретичній та практичній частині роботи.*

Особистий внесок Viktoriia Strilets: аналіз існуючих методів кластеризації, розробка алгоритму методу нечіткої кластеризації, а також аналіз результатів тестування розроблених методів і моделей. Відповідні результати наведені в огляді існуючих методів кластеризації, та в розробленому методі нечіткої кластеризації, обговоренні та висновках.

Особистий внесок Mykhaylo Ugryumov: розробка математичної моделі методу нечіткої кластеризації, а також аналіз результатів тестування розроблених методів і моделей. Відповідні результати наведені в розробленому методі нечіткої кластеризації, обговоренні та висновках.

Особистий внесок Sergii Artiukh: збір даних медичного моніторингу захворювання на рак простати, їх аналіз та експертне виділення цільових класів по кожному запису пацієнтів. Відповідні результати наведені в описі набору даних.

Особистий внесок Roman Zelenskyi: збір даних медичного моніторингу захворювання на рак простати, їх аналіз та експертне виділення цільових класів по кожному запису пацієнтів. Відповідні результати наведені в описі набору даних.

Особистий внесок Tamara Goncharova: переклад статі на англійську, перевірка відповідності термінів, редагування матеріалів статті.)

2. Volodymyr Donets, Viktoriia Strilets, Mykhaylo Ugryumov, Dmytro Shevchenko, Svitlana Prokopovych, Liubov Chagovets. Methodology of the countries' economic development data analysis. Data Analysis. System Research and Information Technologies. 2023. V. 2023. Issue 4. P. 21–36.

Keywords: machine learning, digital development, fuzzy clustering, radial basis neural networks, logistic regression, analysis of variables informativeness.

DOI: 10.20535/SRIT.2308-8893.2023.4.02 **(Scopus)**.

URL: http://journal.iasa.kpi.ua/article/view/297208

(*Особистий внесок здобувача: впровадження розроблених методів мультиагентної нечіткої кластеризації, класифікації на основі штучної нейромережі з модифікованим методом навчання на даних економічного розвитку країн, що дало можливість сформувати методологію стратифікації елементів в комп'ютерних системах економічного моніторингу, відповідні результаті наведені в частині практичного застосування методу та висновків.*

Особистий внесок Viktoriia Strilets: перевірка наукової достовірності отримуваних результатів, перевірка тексту роботи, редагування. Відповідні результати є матеріалами публікації.

Особистий внесок Mykhaylo Ugryumov: постановка проблеми дослідження, розробка методології стратифікації елементів, математичне обґрунтування розроблених методів і моделей, відповідні результати наведені в методологічній частині роботи.

Особистий внесок Dmytro Shevchenko: огляд методів попередньої обробки даних, що були застосовані для підготовки даних до застосування методології.

Особистий внесок Svitlana Prokopovych: збір даних економічного моніторингу цифрового розвитку країн, їх аналіз та експертне виділення цільових класів. Відповідні результати наведені в описі набору даних.

Особистий внесок Liubov Chagovets: переклад статті на англійську мову, коректування використаних термінів.)

3. Volodymyr Donets, Dmytro Shevchenko, Maksym Holikov, Viktoriia Strilets, Serhiy Shmatkov. Application of a data stratification approach in computer medical monitoring systems. Eastern-European Journal of Enterprise Technologies. 2024. 2(9 (128), 6–16.

Keywords: data stratification, anomaly detection, fuzzy clustering, neural network, sensitivity analysis.

DOI: 10.15587/1729-4061.2024.298805 **(Scopus)**.

URL: https://journals.uran.ua/eejet/article/view/298805

(*Особистий внесок здобувача: впровадження розроблених методів і моделей стратифікації даних в комп'ютерній системі медичного моніторингу, що дало можливість перевірити ефективність поєднання мультиагентного методу кластеризації, методу класифікації та методів визначення інформативності на реальних даних медичного моніторингу, відповідні результаті наведені в частині практичного застосування методу та висновків, а також переклад матеріалів статті на англійську.*

Особистий внесок Dmytro Shevchenko: розробка методів попередньої обробки даних, а саме фільтрації вхідних даних методом ізольованого лісу та автокодувальника, відповідна частина наведена в роботі.

Особистий внесок Maksym Holikov: аналіз проблемної області та робіт присвяченій цій області, відповідна частина наведена в роботі.

Особистий внесок Viktoriia Strilets: перевірка наукової достовірності отримуваних результатів, перевірка тексту роботи, редагування.

Особистий внесок Serhiy Shmatkov: перевірка наукової достовірності отримуваних результатів, перевірка тексту роботи.)

Статті у наукових фахових виданнях України

4. Донець В. В., Стрілець В. Є., Шевченко Д. О., Шматков С. І. Агентно-орієнтований метод кластеризації даних оптового дистриб'ютора. Вісник Харківського національного університету імені В. Н. Каразіна серія «Математичне моделювання. Інформаційні технології. Автоматизовані системи управління». 2022. Том 1. № 55. Стор. 6–18.

Keywords: fuzzy clustering, multi-agent approach, data processing, Box-Cox transformation, PCA method, t-SNE method, autoencoder, Kullback-Leibler divergence, Mahalanobis distance, Manhattan distance

DOI: 10.26565/2304-6201-2022-55-01.

URL: https://periodicals.karazin.ua/mia/article/view/22589

(*Особистий внесок: впровадження розробленого методу мультиагентної нечіткої кластеризації на даних оптового дистриб'ютора, що має економічне походження. Відповідні результати наведені в практичній частині роботи*

Особистий внесок Стрілець В. Є.: підготовка набору даних для тестування, перевірка наукової достовірності отримуваних результатів, перевірка тексту роботи, редагування. Відповідні результатом є матеріалами публікації.

Особистий внесок Шевченко Д. О.: попередня обробка та аналіз даних з їх візуалізацією, результати наведені у відповідній частині роботи.

Особистий внесок Шматков С. І.: перевірка наукової достовірності отримуваних результатів, перевірка тексту роботи.)

5. Володимир Донець, Сергій Шматков. Методи аналізу інформативності в медичних системах підтримки прийняття рішень. Інформаційні технології та суспільство. Рік 2023. Том 5. № 11. Стор. 6–13.

Ключові слова: аналіз чутливості, аналіз даних, штучна нейронна мережа, інтегровані градієнти, медична діагностика, прийняття рішень.

DOI: 10.32689/maup.it.2023.5.1.

URL: https://journals.maup.com.ua/index.php/it/article/view/2922

(*Особистий внесок здобувача: аналіз існуючих методів інформативності, впровадження розробленого методу визначення загальної інформативності та*

адаптація градієнтного методу визначення поточної інформативності Відповідні результати наведені в практичній частині роботи

Особистий внесок Сергій Шматков: перевірка наукової достовірності отримуваних результатів, перевірка тексту роботи.)

Наукові праці, які засвідчують апробацію матеріалів дисертації

6. Viktoriia Strilets, Nina Bakumenko, Serhii Chernysh, Mykhaylo Ugryumov, Volodymyr Donets. Application of artificial neural networks in the problems of the patient's condition diagnosis in medical monitoring systems. Advances in Intelligent Systems and Computing. AISC 1113. Харків, 2020. Pp. 173–185.
DOI: https://doi.org/10.1007/978-3-030-37618-5_16 **(Scopus).**

7. Viktoriia Strilets, Nina Bakumenko, Volodymyr Donets, Serhii Chernysh, Mykhaylo Ugryumov,Tamara Goncharova. Machine Learning Methods in Medicine Diagnostics Problem. 16th International Conference on ICT in Education, Research and Industrial Applications. Integration, Harmonization and Knowledge Transfer. Volume II: Workshops, ICTERI 2020. Харків, 2020. – Pp. 89-101.

8. Бакуменко Н. С., Донець В. В., Шевченко Д. О., Одинець О. О., Угрюмов М. Л.. Методи кластеризації даних на основі інформаційних критеріїв. Науковий збірник праці міжнародної науково-технічної конференції «Комп'ютерне моделювання у наукоємних технологіях (КМНТ -2021)». Харків, 2021. С. 20–23.

9. Donets V., Ugryumov M., Strilets V. A Measure Of Compactness For Fuzzy Clustering Based On Entropy. Науковий збірник праці міжнародної науково-технічної конференції «Комп'ютерне моделювання у наукоємних технологіях (КМНТ -2022)». Харків, 2022.

ЗМІСТ

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ

AdaBoost – Adaptive Boosting, метод класифікації

ANOVA – Набір статистичних моделей та методів для дисперсійного аналізу

AUC – Area Under the ROC Curve, одне з визначень точності класифікації

BRFSS – Behavioral Risk Factor Surveillance System

CDC – Centers for Disease Control and Prevention

DBSCAN – Density-based spatial clustering of applications with noise, метод кластеризації даних

GBSA – Gradient-based Sensitivity Analysis, що є методом визначення чутливості на основі аналізу похідних першого чи вищого порядків виходів моделей ШНМ

IG – Integrated Gradients, що є методом визначення коефіцієнтів впливу для кожної вхідної змінної, моделі ШНМ

LSTM – Long short-term memory – тип рекурентних ШНМ

PCA – Principal Component Analysis, метод стиснення даних

ROC – Receiver operating characteristic, графічна ілюстрація для оцінки точності бінарної класифікації

ROD&IDS – Machine learning methods for robust multidisciplinary optimum design problems and intellectual diagnostics of system

SHAP – SHApley Additive ExPlanations, що є методом пояснення виходів моделей машинного навчання

SVM – Support Vector Machine

t-SNE – t-distributed Stochastic Neighbor Embedding

UCI – UC, Irvine – сховище даних машинного навчання

ШНМ – Штучна Нейронна Мережа

ВСТУП

Обґрунтування вибору теми дослідження. Моделі динамічних систем допомагають в описанні будь-якого об’єкту чи процесу до певного рівня деталізації відповідно до можливостей технологій чи необхідності. Такі моделі визначають поняття стану як сукупності певних змінних стану в певний момент часу. А, отже, ці моделі дозволяють визначити закон еволюції стану з плином часу. За допомогою закону еволюції динамічної системи й початкових значень змінних можливо прогнозувати стан системи у будь-який момент часу проте із обмеженням на ступінь деталізації модельованої системи чи процесу.

Динамічні моделі широко використовуються в медицині. Для діагностування стану пацієнта чи прогнозування протікання процесу лікування й як наслідок можливості керування процесом лікування використовуються діагностичні моделі та моделі контролю стану, що є по суті інформаційними системами [1–3]. Для точного прогнозування стану пацієнтів чи прогнозування протікання процесу лікування і як наслідок надання якісних послуг лікування необхідно визначити найбільш впливові змінні стану. Знання про підмножину найбільш впливових керованих змінних стану дозволить ефективніше скеровувати процес лікування, а знання про найбільш інформативні змінні стану дозволить ліпше визначати поточний стан, а особливо якісь критичні зміни [2]. Це все в сукупності дозволить збільшити точність й своєчасність прийняття рішень щодо лікування пацієнтів. Тому дуже важливо визначати змінні стану, що є важливими в процесі прийняття рішень, проте не є очевидними для експертів. Математичне обґрунтування такому визначенню є необхідним для ліпшої інтерпретації прийнятих рішень.

Теоретичні й практичні дослідження у галузі розробки та впровадження комп’ютерних систем медичного моніторингу є важливими для покращення якості надання медичних послуг як в Україні так у всьому світі. Серед українських вчених, що приділяють увагу цим дослідженням слід зазначити: Бодянського Є. В., Угрюмова М. Л., Зеленського О. І., Бакуменко Н. С., Шматкова С. І., Яковину В. С, Ткачука А. А., Ткачука Р. А., Стрілець В. Є., Артюха С. В., Жолткевича Г. М.,

Міняйлова Є. С., Чумаченко Д. І., Бойко Д. М., Литовченко А. Д. [3–9]. Серед дослідників зарубіжжя варто відзначити: Елана Фертіг, Савлін Каур, Джиммі Сінгла, Судан Чха, Нікмал Адхікарі, Нур Заман Дханджі, Гонгфва Лі, Доєон Лі, Фаршад Фіроузі, Манодж Сінгх Адхікарі, Канагарадж Венусамі, Мохаммед Рамдані, та багато інших [1, 2, 10–21].

Роботи цих вчених є вагомим підґрунтям для розвитку комп’ютерних систем медичного моніторингу. Вони заклали основні принципи побудови й обробки даних в таких системах. Ці роботи продовжують використовуватися іншими науковцями для розвитку інформаційних технологій, обчислювальних методів для створення й модифікації комп’ютерних систем медичного моніторингу як в Україні так і поза її межами.

Передові комп'ютерні системи медичного моніторингу мають на меті визначення оптимальних стратегій лікування за допомогою аналізу великих обсягів даних пацієнтів, що зазвичай мають складний характер [10]. Ці дані, незважаючи на можливість ідентифікації стану пацієнта, потребують систематичного аналізу або експертної оцінки [11]. Використання методів машинного навчання дозволяє автоматизувати цей процес та отримати більш детальний аналіз даних, що сприяє покращенню якості надання медичних послуг [11]. Такі системи також використовуються як інструменти для підтримки прийняття рішень у випадках надзвичайних ситуацій [10, 11]. Існуючі інформаційні системи вирішують деякі аспекти стратифікації даних (як то класифікація стану чи визначення найбільш впливових змінних стану) у комп’ютерних системах медичного моніторингу, але дослідження у цій галузі дозволять розробити більш ефективні методи для класифікації станів та прийняття обґрунтованих рішень [12]. В таких комп’ютерних системах медичного моніторингу виділяються фізичні елементи (сенсори, датчики, комп’ютери, спеціалізоване обладнання) та програмні елементи (програмні модулі для отримання нових даних). Таким чином елементи комп’ютерних систем медичного моніторингу описуються даними, що продукуються ними. Під стратифікацією даних тут мається на увазі багатоетапний процес визначення

можливих станів в потоці даних, що генерує комп'ютерна система медичного моніторингу, подальша їх класифікація та виявлення впливу змінних стану.

Актуальність проблеми визначення станів пацієнтів або прогнозування протікання процесу лікування підтверджується шляхом розробки нових та модифікації математичних моделей та обчислювальних методів підтверджується рядом проектів провідних дослідницьких установ світу. Серед найбільш відомих слід зазначити:

- Програма IMPART (Inflammation & Metabolism, Physical Ability, Research Translation), що присвячена медичним технологіям і клінічним дослідженням, фінансується урядом Канади.[13].
- Програми AMED, що фінансуються урядом Японії і направлені на дослідження та впровадження медичних технологій пов'язаних з медичними пристроями моніторингу та визначення станів пацієнтів [14].
- Програми NITRD уряду США з досліджень та розвитку застосовування цифрових технологій задля покращення надання послуг лікування та медицини. Ці програми присвячені впровадженню мобільних пристроїв, побудови системи Інтернету речей, персоналізації медицини та застосування методів машинного навчання для аналізу медичних даних [15].
- Проекти міжнародних компаній Samsung, Microsoft, Google, Nvidia, Apple, що працюють на вдосконаленням та розробкою нових технологій в сфері медицини, аналізу медичних даних [16, 17].

Дослідження в цій галузі має велике значення для ефективного використання методів машинного навчання в аналізі та класифікації медичних даних. Його результати можуть автоматизувати процес визначення станів пацієнтів, підвищуючи точність та швидкість у прийнятті медичних рішень. Крім того, це сприятиме впровадженню систем оптимального планування лікування та підвищенню якості медичного обслуговування. Отже, все це визначає актуальність рішення **науково-прикладної задачі** удосконалення або розробка нових математичних моделей та обчислювальних методів стратифікації даних щодо елементів комп'ютерних систем медичного моніторингу, що має на меті

підвищення точності діагностування стану пацієнта за рахунок реалізації методів і моделей стратифікації елементів комп’ютерних систем медичного моніторингу.

Зв’язок роботи з науковими програмами, планами, темами. Тематика дисертаційної роботи пов’язана з дослідженнями:

- Участь у НДР «Моделювання інформаційних процесів у складних і розподілених системах» за 2021 – 2023 рр. (ДР № 0121U109183), у якості виконавця.

Мета і задачі дослідження. Головною ***метою*** дисертаційної роботи є підвищення точності діагностування стану пацієнтів за рахунок реалізації методів і моделей стратифікації даних щодо елементів комп’ютерних систем медичного моніторингу. Досягнення мети можливе завдяки удосконаленню або розробки нових математичних моделей та обчислювальних методів стратифікації даних щодо елементів комп’ютерних систем медичного моніторингу, що дозволить підвищити точність діагностування стану пацієнта.

Для досягнення поставленої мети та рішення поставленого наукового завдання був визначений наступний ряд ***задач***:

1. Аналіз існуючих математичних моделей, обчислювальних методів та прикладних інформаційних технологій, що використовуються для стратифікації елементів комп’ютерних систем медичного моніторингу.
2. Розробка концептуальної моделі комп’ютерної системи медичного моніторингу з підтримкою прийняття рішень та виділеною підсистемою стратифікації.
3. Розробка методів стратифікації комп’ютерних систем медичного моніторингу.

3.1. Розробка мультиагентного методу кластеризації.

3.2. Розробка методу класифікації станів пацієнтів повнозв'язною штучною нейронною мережею.

3.3. Розробка методів визначення загальної і поточної інформативності змінних стану.

4. Програмна реалізація методів стратифікації комп’ютерних систем медичного моніторингу.

5. Розробка методу верифікації програмного забезпечення.

6. Проведення тестування розроблених методів стратифікації.

7. Розробка науково-обґрунтованих практичних рекомендацій з використання розроблених методів стратифікації в діагностуванні стану пацієнта в комп’ютерних системах медичного моніторингу.

Об’єкт дослідження – це процеси стратифікації даних щодо елементів комп’ютерних систем медичного моніторингу. Характеристиками цих процесів є інформаційні змінні стану елементів комп’ютерних систем, діагностичні моделі, моделі по контролю й управлінню станом, тощо

Предмет дослідження – це математичні методи й моделі стратифікації даних щодо елементів комп’ютерної системи медичного моніторингу на основі мультиагентного підходу. Мультиагентним підходом є способом елітарного відбору найкращих станів, що є агентами в просторі генерованих даних, за певною метрикою серед визначених станів.

Методи дослідження полягають у використанні принципів та методів системного аналізу, мультиагентного підходу, а також застосування імітаційного та математичного моделювання, теорії математичної статистики, теорії множин, теорії ймовірностей, теорії графів, лінійну алгебру, методи математичної оптимізації, диференціального аналізу, теорії штучних нейронних мереж.

Наукова новизна отриманих результатів полягає в наступному.

1. **Вперше розроблено** модель комп’ютерної системи медичного моніторингу, особливістю якої є застосування та організація взаємодії методів стратифікації для вирішення проблеми кластеризації даних, класифікації станів пацієнтів та визначення інформативності змінних цього стану, що в сукупності забезпечує підвищення точності стратифікації даних в комп’ютерній системі медичного моніторингу.

2. **Удосконалено** мультиагентний метод нечіткої кластеризації, що відрізняється поєднанням нечіткої кластеризації c-means із мультиагентним відбором еліт, що дає можливість виконати модифікацію визначення щільності та роздільності отримуваних кластерів і як наслідок підвищити точність виділення станів пацієнтів в комп'ютерній системі медичного моніторингу.

3. **Удосконалено** метод класифікації станів пацієнтів повнозв'язною штучною нейронною мережею за допомогою поєднання процедур прискореного навчання та підбору гіперпараметрів моделі штучній нейронній мережі. Це дозволило ефективно оптимізувати ваги та архітектуру моделей штучній нейронній мережі для вирішення задачі класифікації станів по відповідним змінним.

4. **Удосконалено** методи визначення загальної інформативності змінних щодо стану пацієнтів комп'ютерної системи медичного моніторингу за рахунок виділення зв'язку між входами і виходами через поширення градієнтів в штучній нейронній мережі, а також поточної інформативності шляхом перетворення вагових показників методу інтегрованих градієнтів, що створює умови для виявлення найбільш впливових керованих і некерованих змінних стану й оцінки впливу виявлених змінних на конкретне прийняте рішення та дає можливість пояснити причини прийнятого медичного рішення.

5. **Дістав подальшого розвитку** метод верифікації програмного забезпечення стратифікації даних щодо елементів комп'ютерної системи медичного моніторингу, що відрізняється від існуючих виконанням комплексної перевірки як програмної реалізації, так і точності роботи розроблених методів і моделей, що дає можливість скоротити строки розробки програмного забезпечення.

Особистий внесок здобувача. Дисертаційне дослідження було виконане здобувачем самостійно, всі сформульовані в роботі положення, висновки та рекомендації були обґрунтовані особистими дослідженнями автора. Окремі положення були аргументовані з використанням робіт інших науковців, що мають відповідні посилання в тексті роботи. В індивідуальних наукових роботах використані лише авторські напрацювання та ідеї.

Автор дисертації активно брав участь у наукових дискусіях та написанні наукових статей, що були опубліковані за темою дисертації. Також автор доповідав результати досліджень на міжнародних наукових конференціях.

В роботі [1] було показано результати розробки мультиагентного методу нечіткої кластеризації, результати застосування якого були перевірені із застосуванням даних медичного моніторингу. Автору належить реалізація та тестування зазначеного методу, а також проведення аналізу й порівняння з існуючими методами нечіткого розділення даних та написання частини тексту. В роботі [2] було показано результати розробки і впровадження запропонованих методів і моделей стратифікації в сфері економічного моніторингу, перевірка відбувалась із використанням даних економічного стану країн. Автору належить поєднання розроблених методів та їх застосування на даних, а також написання відповідної частини тексту. В роботі [3] розглянуто застосування підсистеми стратифікації в комп'ютерній системі медичного моніторингу з детальним аналізом кожних з розроблених та модифікованих методів та моделей. Автору належить поєднання розроблених методів та їх застосування на даних медичного моніторингу, а також написання тексту. Робота [4] присвячена перевірці можливості розширення функціоналу мультиагентного методу нечіткої кластеризації на даних оптового дистриб'ютора. Автору належить тестування методу та написання відповідної частини тексту. Робота [5] присвячена застосуванню розроблених методів визначення загальної і поточної інформативності з використанням даних медичного моніторингу. Автору належить реалізація відповідних методів та їх порівняння з іншими методами визначення інформативності, та написання тексту роботи. В наукових працях [6–9] автор брав участь у створенні програмного забезпечення для реалізації запропонованих методів і моделей, та доповідях на конференціях. Результатом всіх зазначених наукових робіт стало написання матеріалів дисертаційної роботи.

Практичне значення отриманих результатів.

Розроблені та удосконалені методи і моделі можуть бути використані при розробці систем підтримки прийняття рішень в комп'ютерних системах медичного

моніторингу. Отримані результати зазначають, що розроблений мультиагентний метод нечіткої кластеризації в деяких випадках може показувати високу точність нечіткого розділення даних, проте було показано, що для деяких медичних даних застосування іншого методу може поліпшити точність. Отримані результати для розробленого методу навчання й підбору гіперпараметрів моделі ШНМ можуть збільшити швидкість навчання й підбору необхідної архітектури моделі ШНМ. Ці методи можуть бути застосовані не тільки для вирішення проблеми класифікації в комп'ютерній системі медичного моніторингу, а і в інших областях, де така система стає необхідною. Розроблений метод визначення загальної інформативності та модифікований метод інтегрованих градієнтів можливо використовувати для аналізу роботи моделей ШНМ, виділення найбільш впливових змінних та виділення впливу конкретних змінних.

Отримані результати дослідження також використовуються як частина навчального матеріалу курсів «Комп'ютерні інформаційні технології Data Stream Mining» та «Інтелектуальні комп'ютерно-інтегровані технології управління виробничими процесами» у Харківському національному університеті імені В. Н. Каразіна.

Апробація результатів дисертації. Теоретичні положення, результати розробки і тестування, висновки і пропозиції, що вказані в дисертації, обговорювалися та були схвалені на засіданнях кафедри теоретичної та прикладної системотехніки Харківського національного університету імені В. Н. Каразіна. Ключові положення дослідження оприлюднені у доповідях на науково-технічних конференціях всеукраїнського та міжнародного рівнів (2020–2022 роки).

– 16th International Conference on ICT in Education, Research and Industrial Applications. Integration, Harmonization and Knowledge Transfer. Volume II: Workshops, ICTERI 2020. (Харків, 06–10 жовтня 2020р.)

– Міжнародній науково-технічній конференції «Комп'ютерне моделювання в наукоємних технологіях» КМНТ–2021р (Україна, м. Харків, Харківський національний університет ім. В. Н. Каразіна, 2021р).

– Міжнародній науково-технічній конференції «Комп’ютерне моделювання в наукоємних технологіях» КМНТ–2022р (Україна, м. Харків, Харківський національний університет ім. В. Н. Каразіна, 2022р).

Публікації. Теоретичні положення, результати тестування і висновки дисертації викладені у 10 наукових працях, з яких 3 у наукових фахових виданнях, що входять до міжнародних наукометричних баз [1–3] та 2 у наукових фахових виданнях України [4–5] та 4 тези наукових доповідей [5–9].

Структура та обсяг дисертації. Дисертаційна робота складається з вступу, чотирьох розділів, висновків, списку використаних джерел і одного додатку. Загальний обсяг дисертації становить 153 сторінки: у тому числі анотації на 10 сторінках, зміст на 2 сторінках, основний текст на 129 сторінках, список використаних джерел із 112 найменувань на 15 сторінках та один додаток на 5 сторінках. Робота містить 14 таблиць, 37 рисунків, з яких 1 на окремій 1 сторінці.

РОЗДІЛ 1.
АНАЛІЗ ЗАДАЧІ СТРАТИФІКАЦІЇ ЕЛЕМЕНТІВ КОМП'ЮТЕРНИХ СИСТЕМ МЕДИЧНОГО МОНІТОРИНГУ

1.1. Аналіз методів стратифікації елементів комп'ютерних систем медичного моніторингу

Комп'ютерні системи медичного моніторингу посідають важливу роль у наданні якісних послуг з медичного контролю. Такі системи дозволяють забезпечувати постійний контроль за станом пацієнтів та надання своєчасної медичної допомоги. Системи медичного моніторингу дозволяють збирати великі набори даних станів пацієнтів, це дозволяє відслідковувати зміни в станах пацієнтів і оперативно призначати відповідне лікування [10]. Також основним напрямком розвитку таких систем останнє десятиліття було забезпечення функції віддаленого моніторингу стану пацієнта [18–20]. Реалізація такої функції дозволяє вести облік їх стану та реагувати на можливі зміни у віддаленому режимі. Спрощення реалізації цією функції було досягнуто з розвитком Інтернету речей, що дозволив розумним датчикам, сенсорам та різному спеціалізованому обладнанню обмінюватися даними через мережу Інтернет [18–20]. Де комп'ютери, сенсори та інші пристрої обмінюються даними завдяки мережі Інтернет. Таким чином комп'ютерні системи медичного моніторингу стали потужними та універсальними інструментами у сфері охорони здоров'я. Проте такі системи генерують великі об'єми даних [10]. Визначення стану пацієнта можливе в таких системах, проте із залученням деякої експертної логіки чи наявності експерта (медика), що здатен робити відповідні прогнози [11]. Застосування методів машинного навчання дозволяє аналізувати данні у таких системах без залучення людини та можливістю визначення прихованих зв'язків у даних. Завдяки чого вдається надавати якісне лікування більшому числу людей [11]. Такі комп'ютерні системи медичного моніторингу використовуються у якості систем збору даних та підтримки прийняття рішень [11].

Типова схема реалізації комплексних комп'ютерних систем медичного моніторингу показана на Рис. 1.1 [18–20]. В таких комп'ютерних системах

медичного моніторингу можливо виділити фізичні елементи (сенсори, датчики, комп'ютери, спеціалізоване обладнання) та програмні елементи (програмні модулі для отримання нових даних) [18–20]. Всі ці елементи продукують дані, які використовуються для моделювання стану пацієнтів у таких системах, та використовуються для вирішення задач діагностування стану, та підтримки прийняття рішень щодо лікування. Отже, елементи комп'ютерних систем медичного моніторингу описуються даними, що продукуються ними.

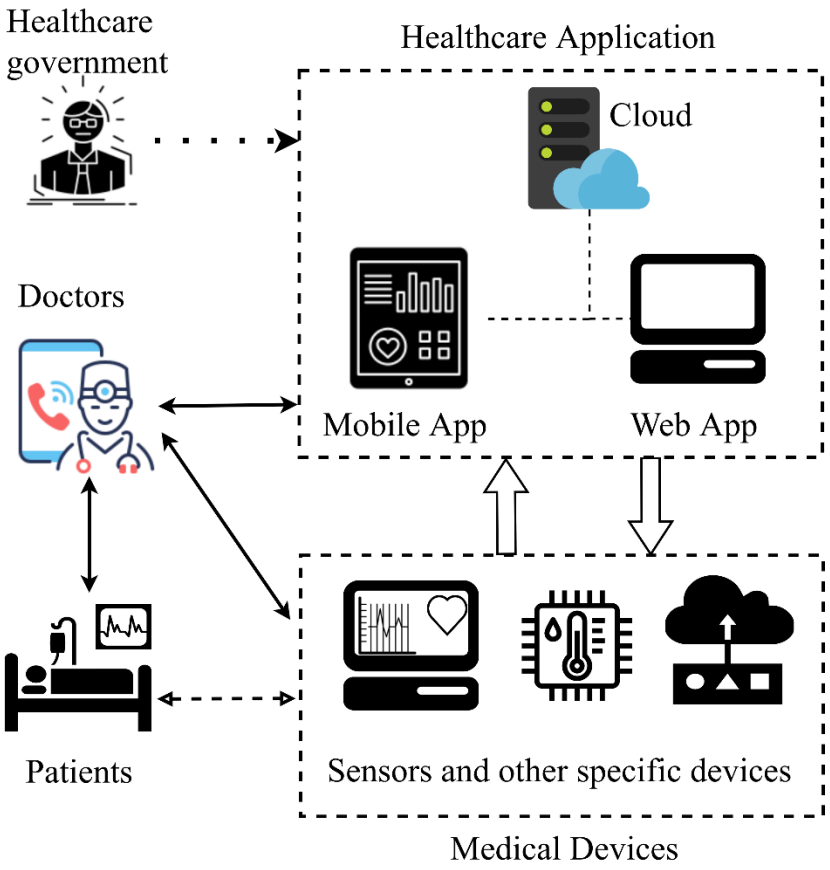


Рис. 1.1. Схема реалізації комп'ютерних систем медичного моніторингу із застосуванням Інтернету речей.

Розглянемо детальніше особливості комп'ютерних систем медичного моніторингу із можливістю підтримки прийняття рішень. В огляді комп'ютерних систем медичного моніторингу [2] визначено, що наразі існує декілька типів таких систем відповідно до методів штучного інтелекту, що в них використовуються:

1. Системи на основі нечіткої логіки. Такі системи медичного моніторингу використовують методи нечіткої логіки, що дозволяють визначати ймовірності станів. До методів, що застосовуються в таких системах входять методи кластерного аналізу, Support Vector Machines (SVM), Fuzzy Inference Systems, Self-Organizing Maps та багато інших. Типова процедура отримання результату в таких системах включає процеси перетворення вхідних даних в нечітку множину, застосування механізму прийняття рішення із застосуванням бази знань та перетворення нечітких даних назад до чіткого результату. Детально така процедура показана на Рис. 1.2 [2].

2. Системи на основі методів машинного навчання. Такі системи медичного моніторингу використовують методи машинного навчання. Головною особливістю таких систем є необхідність розробника мати уявлення про можливі входи системи та очікувані виходи, тобто розробнику необхідно мати відповідну модель. До методів, що застосовуються в таких системах, входять методи контрольованого навчання такі як Random Forest, Naïve Bayes, SVM та ШНМ. Типова процедура застосування таких систем включає етапи збору й підготовки даних до потрібного формату; конфігурація, навчання й оцінка моделі машинного навчання відповідного до даних й бажаної якості роботи; застосування підготовленої моделі для передбачення станів. Типовий підхід до застосуванням методів машинного навчання показаний на Рис. 1.3. [2].

3. Системи на основі глибинного машинного навчання. Такі системи медичного моніторингу відрізняються від систем на основі машинного навчання наявністю вбудованого екстрактора даних в модель машинного навчання. До методів, що застосовуються, входять специфічні моделі ШНМ такі як згорткові нейронні мережі, LSTM, Трансформери, а також їх поєднання. Типовий підхід до застосуванням методів машинного навчання показаний на Рис. 1.3. [2].

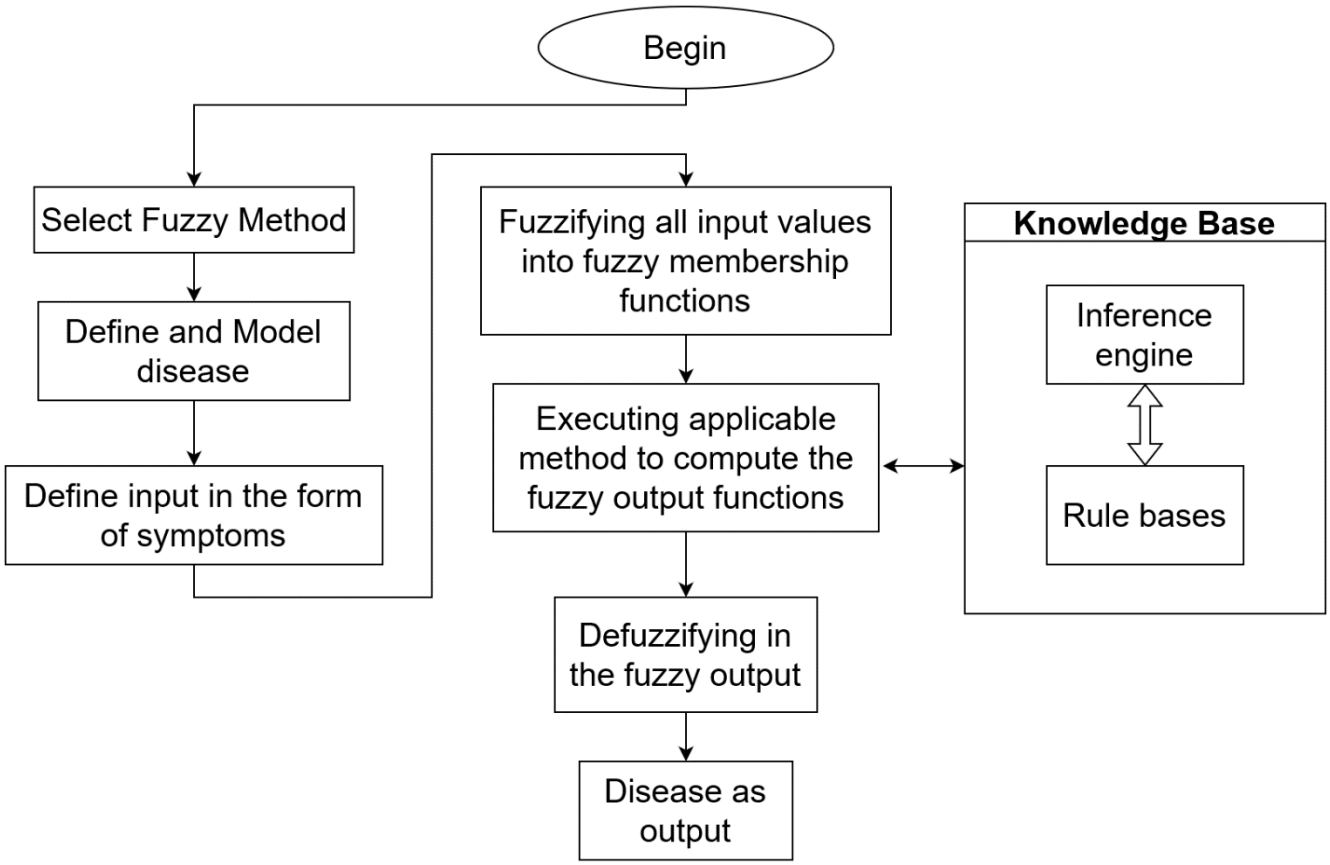


Рис. 1.2. Основні етапи процесу прийняття рішення типової медичної системи із застосуванням методів нечіткої логіки.

Розглянемо детальніше типові приклади таких систем з їх недоліками й перевагами в конкретних роботах. В роботі [21] розглядається нечітка медична діагностична система для діагностування серцевих захворювань. В основі зазначеної системи лежить дерево рішень, яке конфігурується для рішень діагностування конкретних серцевих захворювань. Перевагою такої системи є виростання нечіткої логіки в процесі побудови дерева рішень. Проте недоліком є необхідність залучення медиків для аналізу результатів роботи розробленої системи та необхідність наявності розмічених даних. Тобто така система не може бути використана в автономному режимі.

В роботі [22] розглядається нечітка експертна система з можливістю пояснення зробленого висновку для діагностування COVID-19 в автоматизованій системі. В запропонованому рішенні використана множина нечітких правил, що визначають її функціонування. Така система має суттєвий недолік, це необхідність

введення таких нечітких правил, що вимагає витрат часу й залучення експертів, тому поширення її на широке коло задач є проблематичним.

Дослідження [23] концентрується на розробці алгоритму нечіткої логіки для діагностування лихоманки. Показана реалізація методу із застосуванням правил відносно симптомів лихоманки й показана ефективність застосування такої системи. Проте залишається актуальним неможливість розширеного застосування такої системи.

В усіх розглянутих роботах [21–23] данні проходять шлях показаний на Рис. 1.2. Та як зазначено вимагають розробки правил прийняття рішень, що є достатньо трудомістким процесом із складністю перевірки отримуваних результатів.

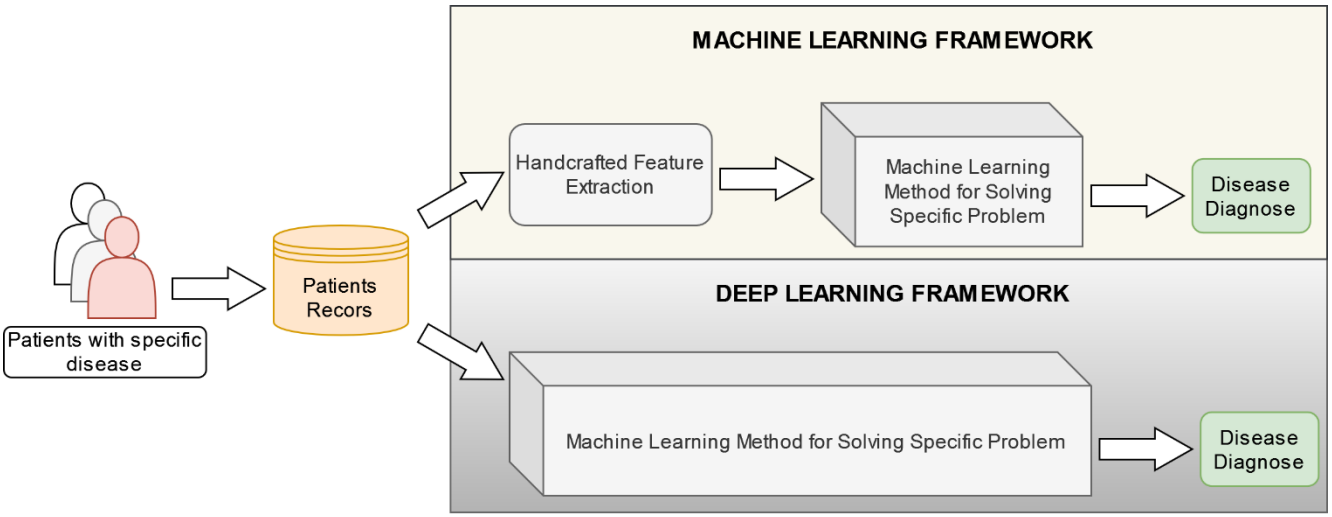


Рис. 1.3. Різниця в застосуванням методів машинного навчання і методів глибинного навчання.

Прикладом застосуванням методів машинного навчання в комп'ютерних системах медичного моніторингу може бути робота [24] в якій розглядається застосування методу класифікації Naïve Bayes на різних медичних наборах даних. Показано, що цей відносно простий метод здатний вирішувати деякі задачі з більшою точністю за метод Decision Tree. В роботі [24] не показаний етап підготовки даних у зв'язку з простотою наведених даних, проте в роботі [25] розглянуто застосування алгоритму Random Forest для діагностування захворювання Альцгеймера по зображенням магнітно-резонансної томографії.

Відповідно до особливостей застосування методів машинного навчання показаний етап підготовки даних перед застосуванням (вручну було налаштовано екстрактор специфічних ознак із зображень) моделі Random Forest. Відповідна процедура показана на Рис. 1.3. Недоліками методів машинного навчання є необхідність попереднього налаштування процедури екстракції специфічних даних необхідних для роботи обраного методу машинного навчання.

Типовим прикладом застосування глибинного навчання є робота [26]. В цій роботі розглядається модифікована згорткова нейронна мережа для виявлення й виділення злоякісних пухлин на щитовидній залозі з використанням зображень отриманих ультразвуковою діагностикою. Така робота передає суть подібних робіт [27–29] де є певні розмічені медиками отримані за допомогою різного роду сенсорів та модель глибинної нейронної мережі, що по цим даним навчається виконувати певну задачу (Рис 1.2). Перевагою такого підходу є відносна легкість реалізації, через простоту підготовки даних і валідацію результатів, а також можливість відносно простого навчання безлічі моделей глибинного навчання на вирішення поставленої проблеми. Проте моделі глибинного навчання важко застосовувати на не розмічених даних, та в процесії конфігурації їх гіперпараметрів практично неможливо позбутися взаємодії з людиною.

Після огляду методів машинного навчання розглянемо комп’ютерні системи медичного моніторингу детальніше. В працях [10–12] показані результати застосування технологій Internet of Thing в системах моніторингу стану пацієнтів. Показано, що такі системи дозволяють поліпшити надання медичних послуг за відсутності необхідної кількості лікарів для ручного нагляду. Також через високу складність розробки таких систем вони не мають вбудованих систем підтримки прийняття рішень. Тому при великій кількості пацієнтів такі системи вимагають великої кількості медиків, що будуть спостерігати за пацієнтами. Вирішити цю проблему може розробка універсальної системи прийняття рішень, що здатна в автоматичному режимі опрацьовувати отримані дані пацієнтів, визначаючи відповідні стани, скеровуючи процедуру лікування.

Прикладом успішного впровадження методів машинного навчання в системи медичного моніторингу кінцівок є робота [12]. Такі системи допомагають вирішувати проблему протезування. Показано, що ряд методів машинного навчання таких як SVM, повнозв'язна ШНМ, лінійний дискримінативний аналіз та інші спроможні вирішувати проблему класифікації рухів кінцівок на зібраних даних. Однак, в роботі розглядалися дані, що мають тільки фізичну природу. Також для аналізу використовувалися розмічені данні, тобто запропонована система не може автоматично визначати наявні стани.

Робота [30] розглядає проблеми Big Data в сфері медичного моніторингу, показані проблеми, що виникають у медиків при застосуванні даних медичного моніторингу, такі як неспроможність медиків до аналізу великого об'єму даних та виділення значущої для прийняття рішень інформації. В роботі зазначено, що дані пацієнтів можливо застосувати для покращення якості лікування, проте залишається проблемним аналіз даних з різних систем, що використовують різні стандарти їх зберігання та ведення моніторингу.

Дослідження [31, 32] присвячені проблемам схильності методів машинного й глибинного навчання до упередження в аналізі медичних даних. Прикладом такого упередження може бути вікове чи статеве упередження, коли модель машинного навчання робить заключний висновок про діагноз пацієнта використовуючи стать чи вік як головний критерій. Це дослідження показало, що передові генеративні моделі (такі як великі мовні моделі) також схильні як і моделі машинного й глибинного навчання до упереджень в аналізі медичних даних. Проте в дослідженнях не вказувались підходи до усунення чи автоматичного виявлення такої поведінки моделей. Вирішенням такої проблеми може бути аналіз інформативності змінних для виявлення їх впливу на результати роботи моделі, проте без наявності експерта для виявлення не значущих змінних тут буде не можливим.

Отже, підсумовуючи проаналізовані роботи можна зробити наступні висновки:

- Сучасні комп’ютерні системи медичного моніторингу використовують технології IoT для організації взаємозв’язку між пацієнтом та лікарем за рахунок об’єднання в суцільну систему набору різноманітних датчиків та пристроїв контролю стану із використанням мережі Інтернет та програмних засобів аналізу даних, як елементів комп’ютерних систем медичного моніторингу. Останні описуються даними, що продукуються цими елементами. Такі системи дозволяють спостерігати за великою кількістю змінних стану.
- В комп’ютерних системах медичного моніторингу впроваджуються методи для автоматизації прийняття рішення. Такі системи використовують або методи нечіткої логіки, або методи машинного та глибинного навчання. Такі системи мають принципово різний підхід до вирішення проблем. Так перші вимагають ручного налаштування правил за якими приймаються рішення. На основі методів машинного навчання вимагають налаштування тільки способу екстракції специфічних даних із наявних, а на основі методів глибинного навчання і того повністю автоматичні в своїй роботі. Проте всі ці методи об’єднує вимога в наявності розмічених даних, тобто всі методи мають мати знання про попередні прецеденти без можливості корекції під час роботи.

Враховуючи все сказане вище маємо наступні проблеми, що притаманні комп’ютерним системам медичного моніторингу, а також можливі варіанти їх вирішення:

1. Комп’ютерні системи медичного моніторингу часто оперують великими об’ємами нерозмічених даних [10], тому вимагають наявності некерованого способу розділення зібраних даних.
2. В комп’ютерних системах медичного моніторингу постійно з’являються нові прецеденти для яких постає проблема визначення (класифікації) стану. Розмічені дані можливо використовувати для виявлення стану нових прецедентів, тобто вирішити проблему класифікації стану. Також моделі класифікації можна використовувати для створення систем підтримки прийняття рішення.

3. Виявлена проблема упередженості може бути вирішена за допомогою виявлення пливу змінних на рішення про стан.

4. Також проблема обмеженості ресурсу медиків на обробку даних може бути вирішена на виявлення підмножини найбільш інформаційних змінних.

5. Проблема обґрунтування визначення стану, може бути вирішена за допомогою виявлення впливу конкретних змінних.

Виявлені проблеми визначення можливих станів в потоці даних, що генерує комп'ютерна система медичного моніторингу, подальша їх класифікація та виявлення впливу змінних стану, можна об'єднати під єдиним поняттям стратифікації даних.

1.1.1. Методи кластеризації

Вирішенням першої проблеми виявлення станів в нерозмічених даних можливе за допомогою застосування методів некерованого навчання [33], а саме методів кластеризації [33]. Розглянемо детальніше методи кластеризації, що застосовуються в комп'ютерних системах медичного моніторингу.

В дослідженні [34] показано реалізацію методу нечіткої кластеризації c-means для сегментації зображень частини легень на зображеннях комп'ютерної томографії. Результати роботи вказують на значне покращення точності сегментації, що покращує точність комп'ютерної діагностики систем, що використовують комп'ютерну томографію. Проте варто зазначити, що запропонована модифікація c-means спеціально розроблена для сегментації зображень, де присутні однотипні данні змінних. Тобто запропонований метод неможливо застосувати для розділення даних в системах з неоднотипними змінними.

В роботі [35] розглянуто порівняння точності методів кластеризації DBSCAN та k-means на двох наборах даних медичного моніторингу. Метод k-means показав більший відсоток кластеризації даних і більшу точність.

Дослідження [36] пропонує застосування методу k-means для кластеризації даних супутніх станів для спектру захворювань на аутизм в ранньому дитинстві. Показано, як кластерний аналіз допоміг виявити взаємозв'язки між змінними станів

та визначення супутнього стану. Показане точне формування кластерів незважаючи на різнотипність змінних стану.

Робота [37] присвячена застосування мережевого алгоритму спектральної кластеризації та радіально-базисної функції для підвищення точності виявлення ментального здоров'я пацієнтів. Показано, як метод кластеризації дозволяє виявляти стати стани пацієнтів, а радіально-базисна функція (що є одним з типів ШНМ) використовується для подальшої класифікації стану. Робота [37] є прикладом реалізації рішення першої і другої проблем виявлених в цьому дослідженні. Проте в роботі не розглядалося вирішення інших проблем. Також запропонована радіально-базисна функція має чітку архітектуру, що визначається експертом, це може заважати для автоматичної корекції системи при застосуванні.

В роботі [38] запропонований метод класифікації на основі зважених класів. Цей метод пропонує введення вагової функції для врівноваження розподілу елементів між кластерами та вимагає визначення кількості очікуваних кластерів. Розроблений метод також показав збільшення точності на 3–5% у визначенні кластерів у порівнянні із класичними методами кластеризації, такими як k-means, ієрархічна та нечітка кластеризація. Дослідження проводилось на різнорідних медичних даних, до яких увійшли набори діагностування діабету, виявлення раку легень, захворювань печінки та раку груди.

Підсумовуючи оглянуті роботи можна зробити висновки, що більшість робіт присвячених застосування методів кластеризації в машинному навчанні не враховують невизначеність в кількості кластерів в наборах даних, також розглянуті методи вимагають однотипних даних з однією шкалою вимірювання. Також варто зазначити відсутність універсального методу, що здатен вирішити зазначені проблеми.

1.1.2. Методи класифікації

Задача класифікації – це проблема статистичного аналізу, що полягає в ідентифікації класу в скінченній множині класів до якого належить спостереження [39]. В машинному навчанні для вирішення проблеми класифікації використовують методи контрольованого навчання [40], що вимагають наявності розмічених

навчальних даних для запам’ятовування прихованого в даних патерну. Основна ідея полягає в тому, що ці методи представляють собою деяку функцію, що під час навчання методу апроксимує наявні дані. Умовно методи контрольованого навчання можна поділити на методи з обмеженою функцією апроксимації (Linear Regression, Naïve Bayes, Linear Discriminant Analysis, Decision Trees, K-nearest neighbor algorithm і т.д.) та методи з необмеженою функцією апроксимації (ШНМ в тому числі і методи глибинного навчання). Хоча застосування певного методу визначається задачею, що він вирішує і не завжди складніший метод дає кращі результати.

Ефективність застосування методів з обмеженою функцією апроксимації показана в наступних працях. Так у роботі [41] проаналізовано точність застосування ШНМ в порівнянні з Linear Regression на декількох наборах різнотипних даних пов’язаних із захворюваннями органів травлення. Показано, що по показнику точності ШНМ в більшості випадків показує найкращі результати. Також, ШНМ показує кращу спроможність до апроксимації складних патернів даних. Проте в складних задачах таких як виділення специфічних патернів у зображеннях капсульної ендоскопії метод із зазначеного підтипу, а саме Linear Discriminant Analysis дозволив отримати результати високої точності, що показано в роботі [42]. Також у роботі [43] показане успішне застосування методу Linear Discriminant Analysis для виявлення типу травми колінного суглобу використовуючи біомеханічні змінні.

В роботі [44] розглядалося застосування методу Naïve Bayes для виявлення психологічних захворювань (таких як депресія, тривога та стресовий розлад) у студентів. Розроблена система використовувала дані тестування для визначення ймовірності психологічного захворювання та показала високу точність 86.44%. Показано, що метод Naïve Bayes може працювати із даними опитування.

Методи із необмеженою функцією апроксимації показали успішне застосування в аналізі значно складніших даних. Особливо варто виділити методи глибинного навчання такі як згорткові ШНМ, що як показано в роботах [45, 46] показали високу точність в задачах маркування зображень та здатність обробляти

не тільки 2D, а й 3D зображення. Так в роботі [45] розглянуте успішне застосування згорткових ШНМ для виявлення та класифікації пухлин головного мозку. А в дослідження [46] продемонстровано застосування згорткових ШНМ в аналізі 3D зображень ядерної медицини. Проте і звичайні багаторівневі ШНМ знаходять своє застосування, так у роботі [47] модель багаторівневої ШНМ використовувалась для контролювання рівня плазми в медичних пристроях, що використовують плазму для нагрівання підкладки.

Підсумовуючи, моделі ШНМ є ідеальні для вирішення проблем класифікації, адже здатні апроксимувати патерни даних будь-якої складності. Проте, серед розглянутих робіт варто зазначити, що всі запропоновані моделі були тонко налаштовані експертом. Тобто не розглядалось їх автоматичне налаштування. Що пов’язано із відсутністю необхідності вирішувати широке коло задач і підлаштовувати моделі ШНМ під кожну задачу окремо.

1.1.3. Методи визначення інформативності

Виявлення інформативності змінних дозволяє вирішити останні зазначені проблеми, а саме виявити упередженість в навчених моделях машинного навчання, виділити найбільш інформативні змінні та обґрунтувати визначення стану моделлю.

В роботі [48] розглядалася проблема вибору найбільш інформативних змінних в умовах нерегулярності надходження даних. Розроблено процедуру оцінки змінних на основі функцій псевдо оцінки. Дослідження проводилось із застосуванням багатоетапного клінічного дослідження направленого на лікування великого депресивного розладу. Проблема із дослідженням полягає в запропонованій моделі аналізу, що використовує підхід періодичного виміру станів пацієнтів, та отримана інформативність залежить від частоти виміру. Також в роботі не розглядались методи машинного навчання, а інформативність вимірювалась відносно процедури вимірювання станів та зміни цих станів.

Огляд [49] присвячений поточним викликам та можливостям застосування систем підтримки прийняття медичних рішень на основі машинного навчання. В роботі розглядалися техніки для обґрунтування прийняття рішень в таких системах.

Такі системи поділяються на такі, що працюють із аналізом табличних даних та на ті що присвячені аналізу текстової інформації. Відповідно методів обґрунтування прийнятих рішень для першого типу систем існує значно більше чим для другого типу. В загальному в роботі запропоновано виділення методів загального пояснення функціонування всієї системи та методів поточного пояснення. Також в роботах [49, 50] вказано, що ці методи можна поділити на методи агностики моделі, тобто вони не залежать від методу прийняття рішень в розглянутій системі і ті що залежать, тобто специфічні. Серед виділених методів слід зазначити функцію часової області (для поточного обґрунтуванням, метод агностик), SHAP (для поточного обґрунтуванням, метод агностик, Gate Recurrent Units (для поточного обґрунтуванням текстової інформації, специфічний). Серед важливих проблем варто зазначити, що серед усіх розглянутих робіт в огляді [49] визначено проблему достовірності, тобто складно визначити точність застосування методів в широкому сенсі.

Отже, враховуючи специфічність методів визначення інформативності та задач, що вони вирішують, а також проблеми достовірності притаманні існуючим методам, має сенс розглянути застосування існуючих методів для вирішення цієї проблеми.

1.2. Формування задачі дослідження

Враховуючи визначені проблеми притаманні комп’ютерним системам медичного моніторингу й огляд існуючих методів їх вирішення стає можливим визначити мету дослідження. Отже, мета дослідження полягає в підвищенні точності діагностування стану пацієнтів за рахунок реалізації методів і моделей стратифікації даних щодо елементів комп’ютерних систем медичного моніторингу. Досягнення мети можливе завдяки вирішенню задачі дослідження, а саме удосконаленню або розробки нових математичних моделей та обчислювальних методів стратифікації даних щодо елементів комп’ютерних систем медичного моніторингу, що дозволить підвищити точність діагностування стану пацієнта.

Для досягнення мети дослідження та рішення поставленого наукового завдання необхідно розв’язати наступні задачі дослідження:

1. Аналіз існуючих математичних моделей, обчислювальних методів та прикладних інформаційних технологій, що використовуються для стратифікації елементів комп’ютерних систем медичного моніторингу.

2. Розробка концептуальної моделі комп’ютерної системи медичного моніторингу з підтримкою прийняття рішень та виділеною підсистемою стратифікації.

3. Розробка методів стратифікації комп’ютерних систем медичного моніторингу.

3.1. Розробка мультиагентного методу кластеризації.

3.2. Розробка методу класифікації станів пацієнтів повнозв'язною штучною нейронною мережею.

3.3. Розробка методів визначення загальної і поточної інформативності змінних стану.

4. Програмна реалізація методів стратифікації комп’ютерних систем медичного моніторингу.

5. Розробка методу верифікації програмного забезпечення.

6. Проведення тестування розроблених методів стратифікації.

7. Розробка науково-обґрунтованих практичних рекомендацій з використання розроблених методів стратифікації в діагностуванні стану пацієнта в комп’ютерних системах медичного моніторингу.

1.3. Розробка концептуальної моделі метода стратифікації на основі мультиагентного підходу

За допомогою моделей динамічних систем (інформаційних систем) може бути описаний будь-який об’єкт або процес. Завдяки таким моделям можна визначити поняття стану як сукупність змінних стану в даних момент часу й визначити закон, який описує еволюцію стану з плином часу. Використовуючи закон еволюції складної системи та сукупність початкових значень змінних можливо прогнозувати стан системи у будь-який момент часу (в залежності від складності моделі системи).

Для діагностування стану пацієнта або прогнозування протікання процесу лікування використовують комп’ютерні системи медичного моніторингу, що включаються діагностичні моделі та моделі контроля стану, що являють собою не що інше як інформаційні системи. Для якісного прогнозування стану пацієнта необхідно визначити які змінні стану мають найбільший вплив на стан пацієнта. Знання про змінні стану, що в більшій або меншій мірі впливають на прогнозування стану пацієнта, дозволить збільшити якість прийняття рішень по керуванню станом пацієнта.

Концептуально вирішення проблеми діагностики станів пацієнтів у комп’ютерних системах медичного моніторингу можна представити як послідовність методів обробки даних, що веде до отримання моделі розв’язання задачі діагностики. Стратифікацію пацієнтів можна проводити в два етапи [51, 52]:

1. Попередня підготовка даних, що вимагає участі лікаря.

1.1. Визначення серед множити змінних стану підмножину керованих змінних.

1.2. Створення набору даних із позначенням станів пацієнтів.

1.3. Фільтрація даних від аномальних значень змінних та подальша нормалізації даних.

1.4. Кластерний аналіз даних дозволяє оцінити точність визначення станів пацієнтів. Збіг у визначенні можливих станів методом кластеризації і даних станів свідчить про повноту наданих даних змінних. Мультиагентний підхід дозволяє застосувати правило відбору еліт для формування кластерів, тобто відбору кращих з низ за певною метрикою.

1.5. Розробка стійких метамоделей при невизначеності даних у даних моніторингу можлива за умови застосування: багатовимірної логістичної регресії для визначення ймовірностей різних станів пацієнтів з урахуваннях підмножини контрольованих змінних.

2. Моніторинг стану пацієнтів за участю лікаря й пацієнта.

2.1. Вимірювання усіх змінних стану, що відповідають стану пацієнта.

2.2. Визначення стану пацієнта по найбільшій ймовірності відповідно до багатовимірної логістичної регресії.

2.3. Оцінка інформативності змінних стану, та оцінка інформативності керованих змінних стану. Визначення підмножини найбільш інформативних змінних.

2.4. Розробка багатовимірних моделей контролю стану пацієнтів.

2.5. Застосування багатовимірних часових рядів для прогнозування контрольованих змінних стану.

2.6. Зменшення розмірності підмножини контрольованих змінних стану за допомогою визначення найбільш інформативних змінних стану.

2.7. Класифікація станів пацієнтів з урахуванням прогнозованих керованих змінних стану.

2.8. Синтез індивідуальної програми лікування пацієнтів у системі медичного моніторингу за станом, визначеним на основі прогнозу.

В більшості реальних випадків можливі стани в яких можуть бути визначені пацієнти в комп’ютерній системі медичного моніторингу можуть бути не визначені. Тому для таких випадків запропонована архітектура система підтримки прийняття рішень, що може бути інтегрована в комп’ютерну систему медичного моніторингу (Рис. 1.4) [51, 52]. Відповідно до наведеної архітектури системи підтримки рішень до неї входять модулі підтримки прийняття рішень, модулі попередньої обробки даних та модулі підсистеми стратифікації (Рис. 1.4, блоки зеленого кольору). Запропонована підсистема стратифікації має блок кластерного аналізу, що вирішує проблему неконтрольованого розділення даних. Також до модулів підсистеми належить модуль класифікації для вирішення задачі класифікації та модуль аналізу чутливості для вирішення проблеми визначення впливів змінних стану.

Дані в комп’ютерну систему медичного моніторингу потрапляють від датчиків та записів лікарів. Ці дані направляються на попередню обробку та перевіряються на наявність аномальних значень. Ці аномалії даних можуть свідчити або про помилки датчиків чи мережі передачі інформації чи алгоритмів

обчислення [53]. Далі автокодувальник може додатково перевіряти аномальні дані чи використовуватися для глибшого розуміння зібраних даних [54]. За наявності аномальні дані відкидаються. Підготовлені за необхідністю таким чином дані використовуються модулем кластерного аналізу для вирішення проблеми неконтрольованого розділення даних. Модуль кластерного аналізу дозволяє отримати кластери даних, що визначаються відібраними елементами даних. Відібрані кластери визначаються розмічення навчальних даних для модуля класифікації та являються по суті станами (класами) в яких перебувають дані наявні в комп'ютерній системі медичного моніторингу. Модуль класифікації може бути використаний для визначення станів пацієнтів по новим даним, що надходять у систему, або по даних можливого стану отриманих від модулю аналізу часових рядів. Це дозволить генерувати закон управління у відповідному модулі системи. Також модуль класифікації застосовується для аналізу чутливості у відповідному модулі, тобто визначення впливу змінних на визначення стану. Це дозволяє виділяти найбільш впливові змінні, для зменшення розмірності, або виділяти вплив конкретних змінних на виявлення стану, що покращує аргументацію прийнятих рішень.

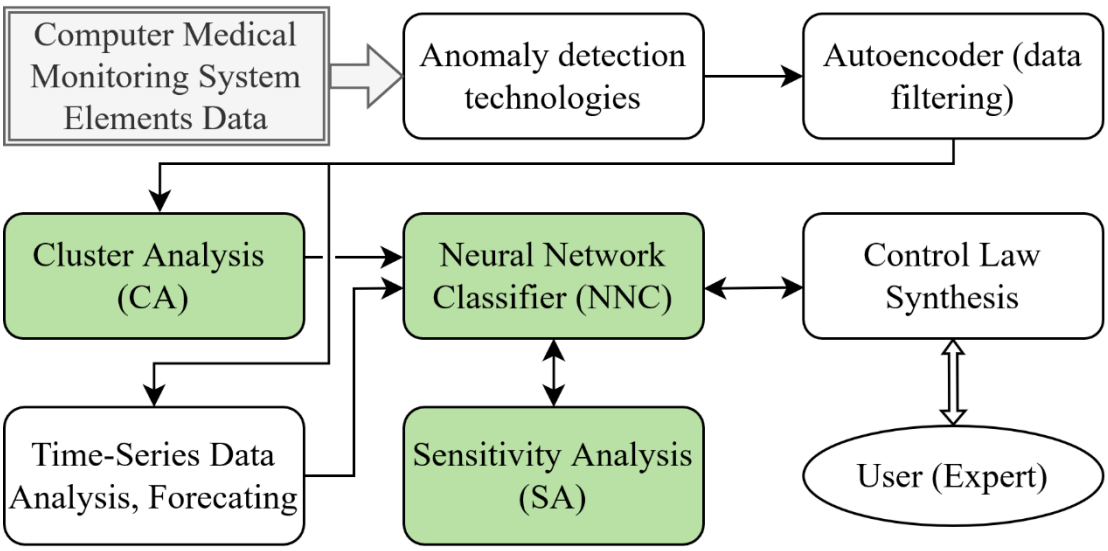


Рис. 1.4. Діаграма реалізації комп'ютерної системи медичного моніторингу з підтримкою прийняття рішень та виділеною підсистемою стратифікації.

Генерація закону управління у відповідному модулі дозволяє реалізувати підтримку прийняття медичних рішень щодо лікування пацієнта. Ці медичні рішення визначаються модулем класифікації та модулем з генерації наступного стану змінних. Використання модуля аналізу чутливості для визначення поточної інформативності дозволяє оцінювати вплив змінних стану на виявлення поточного стану, тим самим обґрунтовуючи прийняті рішення в зазначеній системі.

Підсистема стратифікації в комп'ютерній системі медичного моніторингу може бути реалізована в трьох режимах функціонування [52]:

1. ***Динамічно*** (тобто в режимі неконтрольованого навчання), визначається тим, що в підсистеми стратифікації не має конкретних навчальних даних по можливих станах системи. Іншими словами вхідні дані – це просто дані контролю стану пацієнтів без конкретного визначення експертами стану. Відповідно до такого визначення наявних даних підсистема стратифікації за допомогою модуля кластеризації визначає можливі стани і відповідно до цього визначення навчає модуль кластеризації. У такому випадку кількість можливих станів (визначених кластерів) може змінюватися в залежності від наявності нових даних, що може вести до реконфігурації модуля класифікації і змін в значеннях інформативності змінних.

2. ***Частково динамічно***, експерти визначають для комп'ютерної системи медичного моніторингу кількість можливих станів. Наприклад, для такої системи можливо визначити стани пацієнтів як «здоровий», «одужує», «ускладнення» й «хворіє». У такому випадку наявність більшої кількості даних може призводити до деяких уточненнях при визначення цільових кластерів і деякому перенавчанню модуля класифікацій й деяких змін у визначенні значень інформативності модулем аналізу чутливості.

3. ***Детерміністично*** (тобто в режимі контрольованого навчання), у такому випадку наявний набір даних, що розмічений експертами. Прикладом такому набору, можуть бути результати аналізів пацієнтів і відповідний діагноз. У такому випадку навчання модулів кластеризації та класифікації відбувається лише один раз, нові дані, що поступають в систему не впливають на результати навчання.

Цей режим є ідеальним для тестування та налаштування модулів підсистеми стратифікації адже дає змогу порівнювати отримане розділення даних з очікуваним та таким чином оцінювати точність функціонування модулів.

Висновок до розділу 1

В розділі 1 проаналізовано існуючі дослідження й розробки присвячені стратифікації елементів комп'ютерних систем медичного моніторингу. Наводиться стислий огляд робіт присвяченим комп'ютерним системам медичного моніторингу, методам машинного навчання та аналізу їх застосування. Показано, що незважаючи на розвиток методів машинного навчання в медицині досі присутні проблеми з такими системами. Тому було виведено проблеми притаманні таким системам, та проаналізовані існуючі наробки в методах вирішення розглянутих проблем, виділено переваги й недоліки існуючих методів вирішення цих проблем.

Зазначено, що на сьогодні в епоху розвитку комп'ютерних систем медичного моніторингу на основі Інтернету речей постає проблема аналізу великого потоку різнотипних даних. Показано, що такі данні можуть допомогти покращити якість лікування, проте обмеження в кількості спеціалістів вимагає створення систем автоматичного аналізу таких даних із можливістю виділення можливих станів пацієнтів та коректування їх лікування.

Відповідно до визначених проблем притаманних комп'ютерним системам медичного моніторингу та існуючим методам вирішення зазначених проблем була поставлена мета та задача дослідження. Для досягнення мети було виведено завдання дослідження.

Також відповідно до зазначеної мети дослідження була розроблена концептуальна модель методу стратифікації на основі мультиагентного підходу. Детально пояснена процедура стратифікації стану пацієнтів та визначена, відповідно до процедури, контекстна діаграма реалізації системи прийняття рішень в комп'ютерній системі медичного моніторингу, де було виділено підсистему стратифікації. Описано алгоритм функціонування та можливі режими функціонування зазначеної моделі комп'ютерної системи медичного моніторингу.

Основні положення цього розділу викладені у публікаціях автора [3, 6].

РОЗДІЛ 2
РОЗРОБКА МЕТОДІВ СТРАТИФІКАЦІЇ

2.1. Мультиагентний метод нечіткої кластеризації

2.1.1. Постановка проблеми кластеризації

Для детального аналізу запропонованого методу нечіткої кластеризації на основі мультиагентного підходу, розглянемо постановку завдання кластеризації. Означимо вхідні дані, що необхідно кластеризувати, наступним чином:

$$X = \{x_n : x_{nm} \in \dot{X}\}, m = \overline{1, M}, n = \overline{1, N}, \quad (2.1)$$

де

$\dot{X}$ – це матриця, що представляє вхідні дані,

X – це множина записів/векторів даних,

$x_i \in (x_{i1}, x_{i2}, \dots, x_{ij}, \dots, x_{iM})$ – вектор в множині вхідних даних,

N – кількість записів/векторів даних,

M – розмірність вектору даних.

Тоді визначимо, що $\mathrm{C} = \{c_k\}, k = \overline{1, K}$ – це набір можливих кластерів, де K – кількість цих кластерів. Метою є розбиття вхідних даних на різнорідні множини векторів даних, відомі як кластери. Кожен кластер повинен включати елементи, схожі за певною метрикою, при цьому загальна відмінність між кластерами також визначається цією метрикою. Таким чином, кожен запис у множині вхідних даних $x_n \epsilon X$ призначається до відповідного кластеру c_k [55].

Алгоритм кластеризації [55] описується як функція $f: X \to C$, яка ставить у відповідність кожному вектору вхідних даних $x_n \in X$ певний кластер $c_k \in C$. Зазвичай кількість кластерів K в множині C визначають заздалегідь (наприклад, експерт може визначити цю множину, розглядаючи візуалізацію даних або якщо дані описують проблему з попередньо визначеною кількістю станів). Однак на практиці часто виникають випадки, коли оптимальне число кластерів невідоме, що

породжує проблему визначення цього числа за критерієм якості кластеризації, наприклад, з використанням методу впадин [55, 56].

Розглядання проблеми кластеризації є значущим з наступних причин [55].

- Немає однозначного визначення оптимального критерію для оцінки якості кластеризації. Існує безліч евристичних критеріїв, так само як і безліч алгоритмів, які, хоча не мають чіткого критерію, забезпечують досить розумну кластеризацію для подальшого аналізу. Різні критерії можуть призводити до різних результатів на різних вхідних даних і завжди існуватиме проблема узагальнення.
- Зазвичай перед початком процесу кластеризації невідомо, скільки кластерів буде сформовано, і це число часто визначається згідно з суб'єктивними критеріями, наприклад, точне визначення числа кількості кластерів K при використанні методу впадин повністю покладається на експерта, що займається кластеризацію.
- Результати кластеризації в значній мірі визначаються обраною метрикою, що також є суб'єктивним вибором, визначеним експертною думкою (наприклад, в методах DBSCAN [57] чи Agglomerative Clustering [58] метрика є одним з гіперпараметрів, який суттєво впливає на результати їх застосування).

Ми використовуємо відстань між елементами як метрику для проведення кластерного аналізу. Тоді міру близькості (подібності) елементів можна визначити як обернене значення міжелементної відстані. У кластерному аналізі розглядається велика кількість методів обчислення міжелементної відстані. Також слід зазначити, що часто замість терміну «відстань» використовується термін «метрика», що вказує на метод обчислення певної відстані. Одним із широко використовуваних методів є «Евклідова відстань» [59, 60]:

$$d_2(x_v, x_z) = \sqrt{\sum_{m=1}^{M}(x_{vm} - x_{zm})^2}. \quad (2.2)$$

Відстань Мінковського представляє собою узагальнення Евклідової відстані, використовуючи параметр чутливості p замість значення 2 відповідно. Вираз для загальної метрики Мінковського наведено нижче [61]:

$$d_p(x_i, x_j) = \left(\sum_{m=1}^{M}\left(x_{im} - x_{jm}\right)^p\right)^{1/p} \quad (2.3)$$

В розрахунках використовуються різні метрики, такі як Мангеттенська відстань (відома також як відстань кварталів) та метрика домінування. Відстань Мінковського представляють собою сімейство метрик, яке має широке застосування в різних її формах. Однак існують також методи обчислення відстані між об'єктами, які фундаментально відрізняються від метрик Мінковського. Однією з таких метрик є відома відстань Махаланобіса [62].

Існує два типи змінних: порядкові – на прикладі метрик Spearman і Kendall; номінальні – на прикладі метрик Jacquard, Russell-Rao, Bravais, Yula [59]. Важливо зауважити, що наведені лише деякі приклади метрик, існують інші, а також не слід забувати про можливе їх поєднання [59]. У випадку роботи зі змінними різних типів (порядкових та номінальних) слід використовувати або якісь спеціалізовані метрики, або трансформувати данні до одного типу [63], а потім обирати відповідну метрику.

2.2.2. Визначення методу нечіткої кластеризації c-means

Докладніше розглянемо метод нечіткої кластеризації c-means, на якому заснований розроблений метод кластеризації. Метод нечіткої кластеризації c-means призначений для нечіткого розподілу елементів вхідної множини розмірності N на задану кількість K підмножин, що відповідає кластерам чи класам [64]. Відмінність методу нечіткої кластеризації c-means від його попередника, методу k-means, полягає в використанні нечіткого визначення приналежності елементів до кластеру (зазвичай виражається певним значенням приналежності).

Основний алгоритм c-means включає в себе наступні етапи виконання [65]:

1. Випадкове визначення кількості центрів кластерів K та їх самих центрів c_k , де $k = \overline{1, K}$.

2. Як вказано вище, існують різні методи визначення відстані між центром кластеру та елементом вхідної множини даних. У базовому алгоритмі розглядається нормальний розподіл даних, за якого обчислення матриці приналежності w_{nk} відбувається наступним чином:

$$w_{nk} = \frac{\mathcal{N}(d_2(x_n, c_k)|\mu=0, \sigma_k)}{\sum_j^k \mathcal{N}(d_2(x_n, c_k)|\mu=0, \sigma_k)}, \tag{2.4}$$

де

x_n – n-й елемент множини, $n = \overline{1, P_k}$, P_k – кількість елементів в k-ому кластері,

c_k – вектор координат в просторі X, що представляє центр k-ого кластера,

$d_2(x_n, c_k)$ – Евклідова відстань між точками x_n, c_k,

$\mathcal{N}(d_2(x_n, c_n)|\mu = 0, \sigma_k)$ – щільність ймовірності нормального розподілу для значення відстані $d_2(x_n, c_k)$.

3. Перемістити центри кластерів c_k, з урахуванням обчисленої матриці приналежності w_{nk}:

$$c_k \leftarrow \frac{\sum_{i=1}^{P_k} w_{ik} x_i}{\sum_{i=1}^{P_k} w_{ik}}. \tag{2.5}$$

4. Опираючись на принцип максимальної правдоподібності обчислюється значення оцінки якості поточної кластеризації, тобто значення функції втрат. З урахуванням виразу (2.4) й припущення про нормальний розподіл даних набуває наступного виду:

$$LOSS(C, X) = \sum_{k=1}^{K} \sum_{i=1}^{P_k} d(x_i, c_k)^2 w_{ik} \to min. \tag{2.6}$$

5. Задачею алгоритму у цьому випадку є мінімізації функції втрат, тому у випадку зменшення значення $LOSS(C, X)$ приводить до повернення до другого етапу цього алгоритму. Але автори базового алгоритму також визначили припинення виконання алгоритму при задовільнені певної умови, такої як певне порогове значення функції втрат, чи врахування обмежень обчислювальних можливостей, а саме кількості ітерацій корекції.

Серед переваг базового алгоритму c-means можна відзначити його гнучкість до використання модифікацій, легкість у реалізації та високу ефективність з точки зору використання ресурсів [59, 60, 66].

Серед недоліків базового алгоритму c-means зазначимо його неконтрольовану та непередбачувану поведінку при формуванні кластерів [59, 60, 66]. Іншими словами, часто трапляється, що один кластер привертає всі елементи даних. Крім того, метод нечіткої кластеризації c-means має серйозний недолік у застосуванні: якщо цільовий кластер має складну форму, відмінну від M-розмірної сфери, або виникає перетин цільових кластерів та присутність значного шуму в даних, то метод не може точно визначити приналежність до певного кластеру. Це пов'язано з використанням простих функцій як метрик для вимірювання міжелементних відстаней, що обмежує його здатність до адекватного моделювання форм імовірних кластерів у M-мірному просторі даних [64, 65].

Більшість реальних даних характеризуються великим розкидом можливих значень змінних стану. Як вже зазначалося, метод c-means обмежений визначенням лише M-вимірної сфери потенційного кластера, що ускладнює точне призначення елементів, особливо у випадках перемішування та великої кількості шуму [65]. Зокрема, виникають труднощі при обробці великої кількості змінних в наявних даних. Таким чином, необхідно вносити модифікації у базовий метод c-means [68, 69]. Існує кілька способів модифікації, і одним з найбільш очевидних є зміна методу оцінки відстані між центром кластеру та елементами вхідної множини даних (іншими словами, модифікація оцінки приналежності цього елементу до кластеру).

У дослідженнях [69, 70] застосовано модифікацію функції визначення приналежності елементів до кластеру, таким чином модифіковано алгоритм

обчислення матриці приналежності. Вони використовували припущення про розподіл Коші та в якості міжелементної метрики використовували відстань Махаланобіса:

$$d_{MD}^2(x_n, c_k) = MD^2(x_n, c_k) = (x_n - c_k)^T \hat{\Sigma}_j^{-1}(x_n - c_k), \quad (2.7)$$

де

$\dot{\Sigma} = \Sigma + \lambda \mathrm{E}$ – коваріаційна матриця з регуляризацію параметром λ, що визначається більшим за 0 для усунення можливості виродження оберненої матриці.

Беручи до уваги вираз (2.4) та припущення авторів роботи [70] про розподіл Коші у міжелементних відстанях отримаємо вираз для обчислення матриці приналежності:

$$w_{nk} = \frac{\rho(x_{nk}, c_k)}{\sum_{i=1}^{P_k} \rho(x_{ik}, c_k)}, \quad (2.8)$$

де

$\rho(x_{nk}, c_k) = \left(\pi\eta[1 + \frac{MD^2(x_{nk}, c_k)}{\eta^2}\right)^{-1}$ – відстань визначена відповідно до закону Коші із застосуванням параметру нормалізації η.

Аналогічно, проблема кластеризації M-вимірних сфероїдів розв'язується за допомогою методу Gaussian Mixed Model, який подібній до вищезазначеного методу та успішно використовувалися для оцінки стану працездатності чи відмови пристроїв [71].

Експерименти показали деяке покращення точності кластеризації, але також виявили сильну залежність точності від характеру розподілу вхідних змінних. Іншими словами, запропонований метод не забезпечив достатньо точного відтворення залежностей вхідних та цільових змінних [70, 72]. У роботі [65] показано, що модифікована метрика з увагою до градієнтів може допомогти мінімізувати суму квадратів відстаней в межах кластерів.

Також в роботах [64, 71] була запропонована ідея вивчення розподілів змінних у вибірках з урахуванням відносної ентропії в методі c-means. Згідно з цим автори внесли зміни в алгоритм сходження кластерів. Проте, все так само використовувалась Евклідова відстань, що не може враховувати ентропію даних через свою природу [71].

Для поліпшення точності оригінального методу с-середніх та його модифікацій, таких як Mixture models, Gaussian mixed models [73], запропонована модель зі врахуванням ентропії даних [71] за допомогою застосування інформаційної відстані Кульбака-Лейблера [73].

Відстань Кульбака-Лейблера є асиметричною мірою (не виконується нерівність трикутника), що виражає нерівність інформаційної різниці між двома розподілами даних [69]. Ця метрика широко застосовується в статистиці та машинному навчанні. Суттєвою перевагою відстані Кульбака-Лейблера над розглянутими метриками є можливість оцінювати ступінь інформаційної відмінності з урахуванням ентропії даних [73].

Перед тим як зосередитися на запропонованій модифікації методу кластеризації введемо деякі позначення. $F = \{f_q\}, q = \overline{1, Q}$ – є вектором цільових функцій, де Q – кількість цільових функцій; а x_m – m-та змінна вхідних даних X. Тоді $M_\alpha[f_q]$ та $M_\alpha[x\_m]$ визначається як математичне очікування f_q та x_m відповідно. Далі позначимо дисперсію та стандартне відхилення для f_q та x_m наступними виразами: D_{f_q}, D_{x_m} та σ_{f_q}, σ_{x_m} відповідно. Тоді визначимо дисперсію та стандартне відхилення f_q при змінній x_m за виразами:

$$D_{f_q|x\_m} = var[M_\alpha\left[f_q\left[x\_m\right]\right], \forall v = \overline{1, M}, v \neq m, x\_v = const\,, \tag{2.9}$$

$$\sigma_{f_q|x\_m} = \sqrt{D_{f_q|x\_m}}, \tag{2.10}$$

Використовуючи вираз (2.9) визначимо коефіцієнт інформативності змінних вхідних даних за виразом:

$$\beta_{f_q} = \frac{D_{f_q|x\_m}}{E_{f_q}}, \tag{2.11}$$

де

E_{f_q} – енергія сигналу функції f_q.

Коефіцієнт впливу отримано з виразу (2.10), що є по суті виразом для визначення відношення сигналу до шуму:

$$\varphi_{qm} = SNR_{f_q|x\_m} = \frac{\sigma_{f_q|x\_m}}{\sigma_{x\_m}}. \tag{2.12}$$

Тоді відстань Кульбака-Лейблера можна визначити за виразом [73]:

$$D_{KL}(f_q, x_n) = \sum_{m=1}^{M} \rho(x_{nm}|f_q) \log_2[\frac{\rho(x_{nm}|f_q)}{\rho(x_{nm})}]. \tag{2.13}$$

З урахуванням виразів (2.11) та (2.12) отримаємо вираз для оцінки взаємної інформації [59, 60, 66]:

$$H_{qm} = \frac{1}{2}\log_2\left[SNR^2_{f_q|x\_m}\right] = \frac{1}{2}\log_2[\beta_{f_q}\frac{E_{f_q}}{D_{x\_m}}]\,. \tag{2.14}$$

Модифікуємо оригінальну функцію втрат, наведену в виразі (2.6) й отримаємо вираз для оцінки зворотної інформації, що буде однією з функцій втрат у наведеній модифікації методу нечіткої кластеризації [59, 60]:

$$LOSS(X, C) = -\frac{1}{\sum_{k=1}^{K} P_k}\sum_{k=1}^{K}[\rho(\mathrm{H}_k) \times \sum_{i=1}^{P_j} D_{KL}(x_i, \mathrm{H}_k)] \to min, \tag{2.15}$$

де

H_k – це множина елементів вхідних даних, що належать до k-ого кластеру (по суті це і є k-й кластер),

$\rho(\mathrm{H}_k)$ – щільність отриманого кластеру, що визначається відповідно до обраної метрики.

2.2.3. Визначення мультиагентного методу нечіткої кластеризації

Згідно з розробленими метриками, методами модифікацій стану кластерів та уявленнями про мультиагентний підхід, визначимо позначення для подальшого опису запропонованого методу: x_n – агенти-елементи кластерів, Z – агенти-кластери.

В мультиагентному підході елементи та кластери розглядаються як агенти, які обирають свої центри або кластери на основі певної міри. Інакше кажучи, елементи-агенти обирають кластер, у якому центр знаходиться найближче до них згідно з певною міжелементною відстанню. Кількість кластерних агентів формується відповідно до певної міжкластерної відстані, забезпечуючи в загальному випадку мінімізацію функції втрат [59, 60, 74].

Задля вирішення проблем класичного методу c-means та його розглянутих модифікацій, а саме кращого врахування складної структури можливих кластерів та зменшення похибки точного визначення цільового кластеру було запропонувано мультиагентний метод нечіткої кластеризації [59, 60, 74].

Визначимо задачу, що вирішує мультиагентний метод нечіткої кластеризації як знаходження $\{K, H_k\}$ цільової кількості кластерів та елементів вхідних даних розбитим по відповідним кластерам, або в термінах мультиагентного підходу формування агентів-кластерів Z.

Було обрано чотири метрики для визначення міжелементної відстані для в розробленій моделі кластеризації [59, 60, 74]:

$$d(x_{kn}, c_k) = \begin{cases} d_1(x_{kn}, c_k), & (I) \\ MD^2(x_{kn}, c_k), & (II) \\ w_{kn}^{-1} MD^2(x_{kn}, c_k) & (III) \\ -D_{KL}(x_{kn}, c_k) & (IV) \end{cases}, \tag{2.16}$$

де

I – Мангеттенська відстань,

II – відстань Махаланобіса,

III – відстань Махаланобіса з оберненою функцією приналежності

IV – відстань заснована на несиметричній ентропії Кульбака-Лейблера.

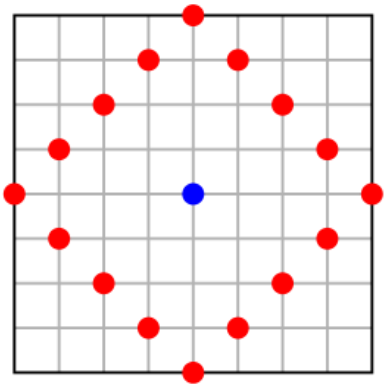


Рис. 2.1. Визначення кола в системі координат за Мангеттенською метрикою.

Розглянемо детальніше запропоновані міжелементні відстані:

I. ***Мангеттенська метрика*** (також відома як метрика прямокутного міста або метрика L1) — це вимірювання відстані між двома точками, що розраховується як сума модулів різниць їх координат [59, 60].

Перевагою цієї метрики є її низька обчислювальна складність при задовільній точності вимірювань.

Недоліками є обмеження до розглядання позицій у формі квадрата, що може обмежувати адаптацію до випадків, де агенти-кластери мають складніші форми (Рис. 2.1)

Визначається за виразом: $d_1(x_i, x_j) = |\sum_{m=1}^{M}(x_{im} - x_{jm})|$.

II. ***Відстань Махаланобіса*** представляє собою метрику у евклідовому просторі, яка узагальнює концепцію евклідової відстані та дозволяє більш гнучке врахування складної форми цільових кластерів (Рис. 2.2) [75].

Серед недоліків цього підходу слід відзначити необхідність зберігання

коваріаційної матриці, що значно збільшує обсяг використовуваної пам'яті зі зростанням розмірності даних (кількості змінних) та викликає збільшене споживання ресурсів. Крім того, важливо відзначити, що відстань Махаланобіса сильно залежить від початкового формування кластерів, і, таким чином, може часто виявлятися менш ефективною. Зберігання коваріаційної матриці може також спричиняти перенавчання на навчальних даних і обмежувати узагальнення на тестових даних або при реалізації моделі, особливо у випадку обмеженої кількості навчальних прикладів.

Визначається за виразом (2.7).

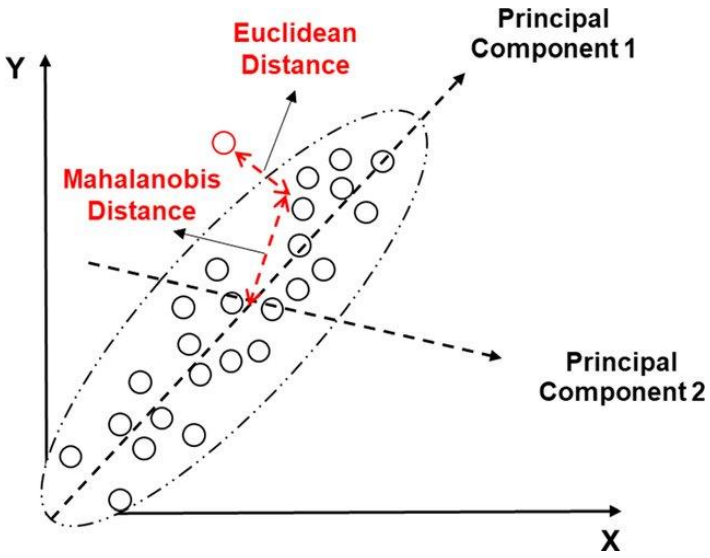


Рис. 2.2. Порівняння відстані Махаланобіса та евклідової відстані [75].

III. ***Відстань Махаланобіса з оберненою функцією приналежності*** – модифікація відстані Махаланобіса, зі згладжуванням по всій вхідній виборці через значення приналежності. Ця модифікація покращує визначення для елементів поза кластером [59, 60, 74].

Переваги: на відміну від простої відстані Махаланобіса дещо зменшує ймовірність хибного навчання та перенавчання на тестових даних шляхом більшого розмиття границь кластерів та вводить обізнаність про приналежність елементів до кластерів

IV. ***Відстань Кульбака-Лейблера*** є несиметричною мірою інформаційної розбіжності двох розподілів ймовірності, детальніше розглянута вище. Переваги

дозволяє досягти точності відстані Махаланобіса без використання значного об’єму ресурсів [59, 60, 74].

Серед недоліків варто відмітити значно більші вимоги по обчислювальним ресурсам.

Визначається виразом: (2.13).

Відповідно до того, що ми обрали декілька метрик міжелементної відстані, визначимо загальній вираз для визначення внутрішньо-кластерної відстані:

$$M(c_k, H_k) = \frac{1}{P_k} \sum_{x_n}^{H_K} d(x_n, c_k). \tag{2.17}$$

Враховуючи попередні вирази для визначення функції втрат (2.6) та (2.15), нову функцію втрат можна визначити за виразом:

$$LOSS(X, C, H) = \frac{1}{K} \sum_{k=1}^{K} M(c_k, H_k). \tag{2.18}$$

Отже в даному випадку проблему кластеризації можна виразити наступним виразом, що визначає необхідність визначення кількості кластерів і розподілу даних між кластерами так, щоб значення функції втрат було мінімальним:

$$\begin{cases} Z = \{K, H_k\} \\ \hat{Z} = \arg\min LOSS(X, C, H) \end{cases}. \tag{2.19}$$

Відповідно до класичного методу кластеризації, оптимізація центрів кластерів відбувається згідно з виразом (2.3), а корекція матриці належності обчислюється згідно з виразами (2.6) та (2.8), враховуючи припущення про розподіл Коші. На цій основі можна сформулювати алгоритм кластеризації, який визначається за мультиагентним підходом, наступним чином [59, 60, 74], візуальне представлення алгоритму методу мультиагентної нечіткої кластеризації показано на Рис. 2.3.

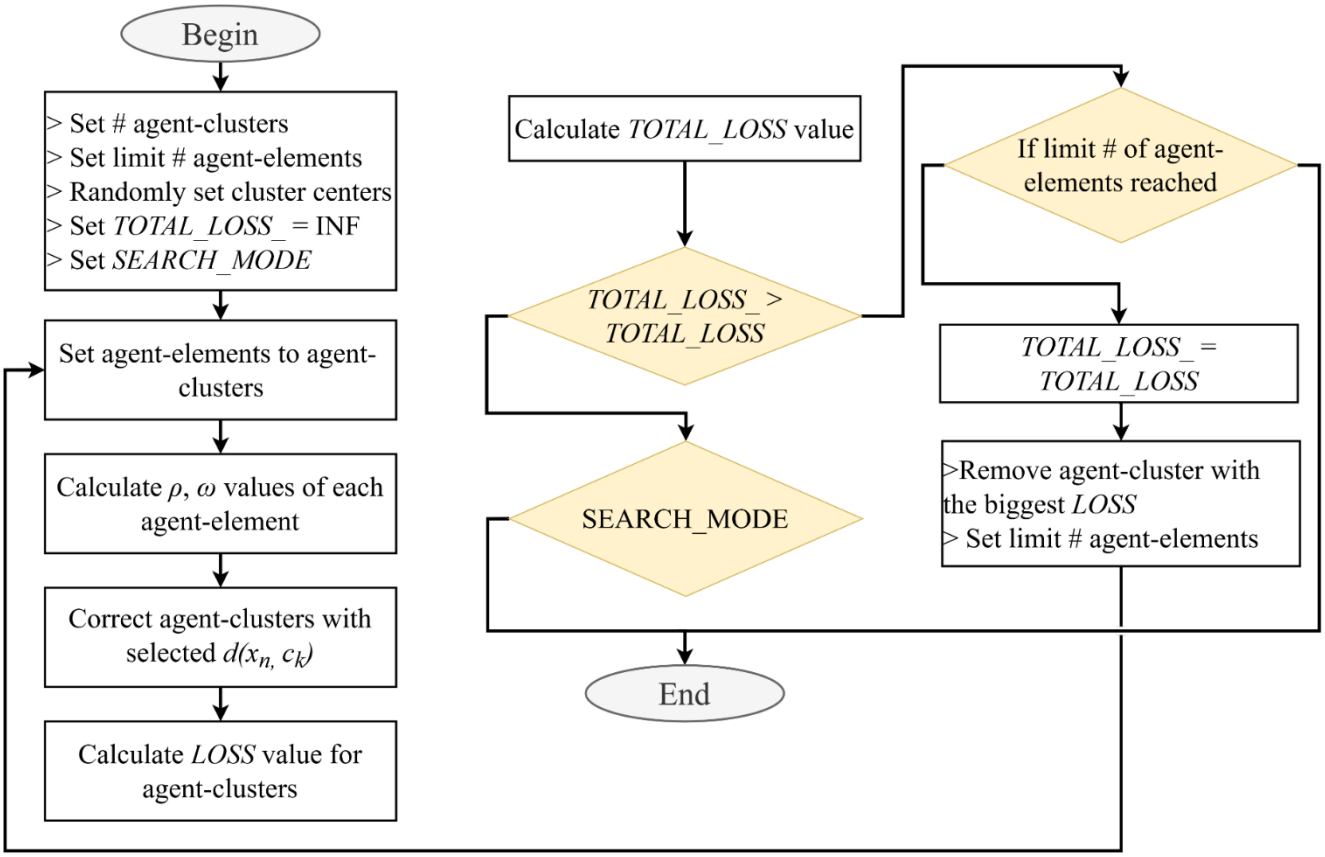


Рис. 2.3. Основні етапи мультиагентного методу нечіткої кластеризації.

1. Випадковим чином визначити початкову кількість агентів-кластерів $|Z| = K^t > K$, що більше цільової кількості кластерів, або деякої експертно визначеної межі у випадку роботи алгоритму пошуку кількості кластерів; та встановити обмеження на кількість елементів в кожному кластері $P_k^t = |H_k^t| = N/K^t$ та випадковим чином вибрати K^t центрів кластерів $\{c_j\}$.

2. Опираючись на одну з міжелементних відстаней (2.16) обрати $|P_k^t|$ найближчих елементів до кожного кластеру, тобто сформувати агенти-кластери P_k^t.

3. Для кожного кластеру обчислити значення розподілу параметрів $\rho(x_m|P_k^t)$ та значення матриці приналежності за виразами (2.8), і відповідно до виразу (2.3) відкоригувати центри кластерів.

4. До кожного агенту кластеру Z_k за вибраною мірою $d(x_n, c_k)$ вибрати $P_k^t = |H_k^t|$ нових агентів-елементів.

5. Для кожного агенту-кластеру за виразом (2.18) визначити значення функції втрат (або середню міжелементну відстань) $LOSS(X, C, H)$.

6. Оцінити поточну якість кластеризації за функцією втрат відповідно до виразу (2.18) або виразу (2.15). У випадку режиму роботи алгоритму в автопошуку оптимальної кількості кластерів, та збільшенні значення функції втрат зупинити алгоритм.

7. Провести відбір агентів-кластерів та відкинути агент-кластер з найбільшим значенням $LOSS(X, C, H)$.

8. Визначити нову кількість кластерів $K^{t+1} = K^t - 1$ та нову кількість елементів кластеру $P_k^{t+1} = |H_k^{t+1}| = N/K^{t+1}$.

9. Повернутися до 2 етапу, за умови $K^t > K$.

Модель класифікатору має видавати на вхідні дані ймовірність відповідного класу, тому з цією задачею може цілком справитися і розроблений метод кластеризації використовуючи значення одного рядку матриці приналежності тобто враховуючи вираз (2.8) отримаємо наступний вираз для визначення вектору ймовірностей приналежності елементу x_n до c_k кластеру:

$$P(x_n, c_k, H) = \frac{\rho(x_n, c_k)}{\sum_{x_i}^{H_k} \rho(x_i, c_k)}. \quad (2.20)$$

Також слід зазначити, що різні метрики, що використовуються в запропонованому методі класифікації вимагаються формування різних класифікаційних моделей:

I. Вимагає тільки наявність центрів кластерів c_k.

II. Вимагає наявність центрів кластерів c_k та обернених матриць кореляції для кожного з кластерів Σ_k^{-1}.

III. Вимагає наявність центрів кластерів c_k та обернених матриць кореляції для кожного з кластерів Σ_k^{-1}.

IV. Наявність центрів кластерів c_k та навчальних записів.

Запропонований метод класифікації може бути використаний в якості засобу тестування якості кластеризації у випадку наявності експертно-розмічених даних

[59, 60, 74]. Також його можна проваджувати в розроблену систему, в якості легко-обновлюваного класифікатора.

2.2. Модель класифікації на основі штучної нейронної мережі

Для вирішення завдання мультикласової класифікації у випадку просторово роздільних даних, також ми обрали повнозв’язну однонаправлену ШНМ із множинною логістичною регресією в якості вихідного шару, що забезпечує вирішення проблеми класифікації. Використання моделі ШНМ для мультикласової класифікації дозволить перевірити припущення про правильність визначення кластерів та протестувати здатність моделі до генералізації обраних кластерів, що в свою чергу також дозволить перевіряти інформативність змінних вхідних даних. По суті модель ШНМ функціонує як паралельний розподілений процесор, який зберігає знання шляхом апроксимації невідомої функції, що описує взаємозв'язок між вхідними даними та цільовими значеннями. Зберігання знань дозволяє проводити обробку даних в середині моделі ШНМ [5, 60].

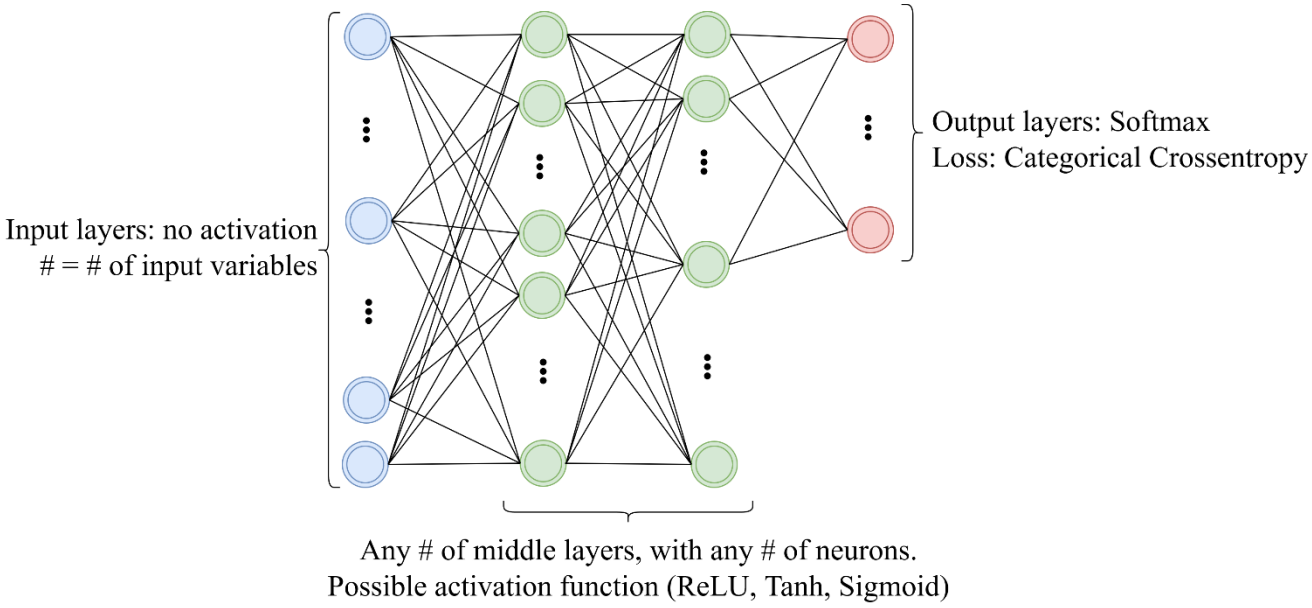


Рис. 2.4. Приклад архітектури моделі ШНМ використаної для класифікації.

Визначимо архітектуру моделі ШНМ (Рис. 2.4) наступним чином: Θ_0 – кількість нейронів вхідного шару ШНМ (кількість змінних даних), Θ_θ^{in} – кількість нейронів θ-ого проміжного шару входів моделі, отже Θ_2 – кількість нейронів

вихідного шару, при чому в контексті розробленої моделі стратифікації $\Theta_2 = K = |H|$. Визначимо вектор вхідних даних для θ-ого шару ШНМ (або вектор вихідних даних для θ-1 шару) як $\vec{Y}^{(\theta)} = \left[Y_1^{(\theta)}, \dots, Y_{\Theta_\theta}^{(\theta)}\right]^T$, звідси вектор координат центів функції активації для приховано шару визначимо як $\vec{c}_j = \left[c_{j1}, c_{j2}, \dots, c_{j\Theta_\theta}\right]^T$, де $j = 1..\Theta_{\theta+1}$, а вектор, що задає ширину вікна функції активації прихованого шару, визначимо як $\vec{\sigma}_j = \left[\sigma_{j1}, \sigma_{j2}, \dots, \sigma_{j\Theta_\theta}\right]^T$. Тоді функція активації для нейронів прихованого шару матиме вигляд [5, 60]:

$$\varphi_j = \left(\vec{Y}_p^{(\theta)}, \vec{c}_j, \vec{\sigma}_j\right) = exp\left(-\frac{1}{2}\sum_{\theta=1}^{\Theta_\theta} w_{ij} Z_{pj\theta}^2\right) \equiv \varphi_{pj}, \qquad (2.21)$$

де

$Z_{pj\theta} = \frac{Y_{ph}^{(\theta)} - c_{j\theta}}{\sigma_{j\theta}}$,

w_{ij} – вагове з'єднання між i-м нейроном вихідного шару та j-м нейроном вхідного.

Для вирішення задачі класифікації за допомогою ШНМ на виходах моделі виставляється множинна логістична регресія (або sigmoid) в якості функції активації [76], а її виходи визначаються виразом:

$$\vartheta_j = \frac{\exp(\gamma_j)}{\sum_{k=1}^{H_2}\exp(\gamma_k)}, \text{ де } \gamma_j = \sum_{\theta}^{H_2} \varphi_\theta w_{\theta j}. \qquad (2.22)$$

Метод підбору гіперпараметрів моделей ШНМ, що базується на використанні навчених ШНМ за допомогою методу прискореного навчання. Підбір проводиться з використанням адаптивного управління обчисленнями, яке здійснюється на основі принципу мінімальних збурень та використання ярового методу, методів спряженого градієнта та методу імітації руху бджолиних сімей [5, 60]. Частини моделі ШНМ, що модифікуються методом підбору гіперпараметрів показані на Рис. 2.5.

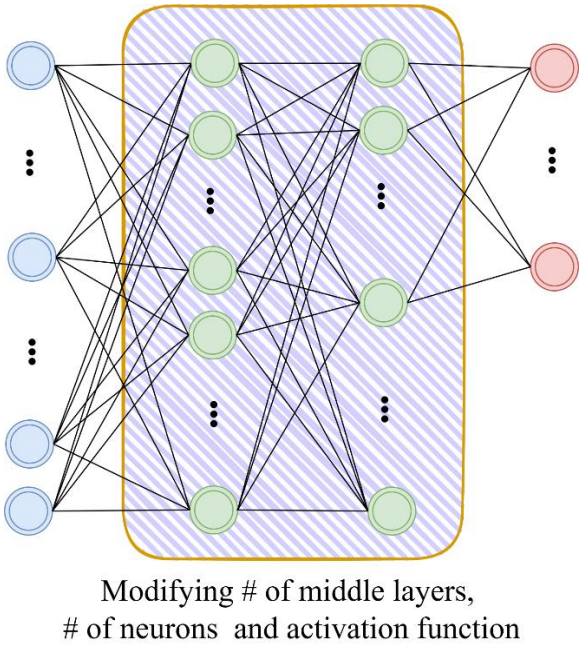


Рис. 2.5. Приклад архітектури моделі ШНМ з позначенням модифікованих частин методом підбору гіперпараметрів.

Використання запропонованих методів дозволяє уникнути виникнення неправильних заглиблень на поверхнях відгуку при великих помилках у вхідних даних. Під час порівняння моделей отриманих під час навчання оцінюється зміна дисперсії сигналу, яка визначає стійкість конкретної моделі, наведено у виразі нижче [5, 60]:

$$D_{Y_i,dB} = 10\, log_{10}\left(\frac{D_{Y_i}^{(\beta)}}{D_{Y_i}^{(0)}}\right), \beta = 1{,}2. \qquad (2.23).$$

Для тренування моделі ШНМ було використано метод прискореного навчання, що є гібридним алгоритмом, який поєднує два етапи. Повторення цих етапів зазвичай призводить до швидкого навчання мережі, особливо при належно згенерованих параметрах [77]. Частини моделі ШНМ, що модифікуються методом прискореного навчання показані на Рис. 2.6.

1. Вибір лінійних параметрів мережі (ваг) w_{ij} за допомогою методу псевдоінверсії.

2. Виправлення параметрів функцій активації (центрів та ширин вікон) із застосуванням методу оберненого поширення помилки відносно зміни значення функції витрат.

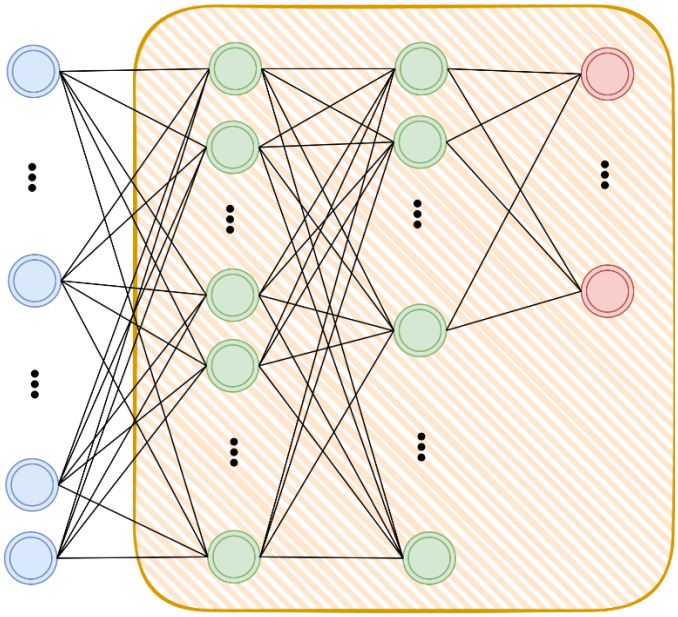


Рис. 2.6. Приклад архітектури моделі ШНМ з позначенням модифікованих частин методом прискореного навчання.

Обидва етапи тісно взаємодіють між собою. У випадку наявності P навчальних пар $\left(\vec{Y}_\theta^{(0)}, \vec{d}_\theta\right), \theta = \overline{1, \Theta_j}$ і фіксації конкретних значень центів центрів $\vec{c}_\theta$ та ширин $\vec{\sigma}_\theta$ вікон функції активації ми отримаємо систему рівнянь:

$$\Phi \vec{w}_\theta = \vec{d}_i, i = \overline{1, \Theta_j}, \tag{2.24}$$

де

$\Phi = [\varphi_{pj}], p = \overline{1, P}, j = \overline{0, \Theta_j}, \varphi_{p0} = 1,$

$\vec{w}_i = \left[w_{i0}, w_{i1}, \dots, w_{i\Theta_j}\right]^T$ та $\vec{d}_i = \left[d_{0i}, d_{1i}, \dots, d_{pi}\right]^T$.

Для визначення значень ширин вікон $\vec{\sigma}_\theta$ та зменшення часу навчання моделі ШНМ використано алгоритм формування зони покриття на основі методу k-

neighbors. Цей підхід допоміг оптимізувати процес навчання RBFN, пришвидшуючи його [5, 60]. Повторення обох етапів кілька разів призводить до повного та ефективного навчання ШНМ, зокрема у випадку, коли початкові значення параметрів функцій активації наближені до оптимальних. Цей підхід сприяє вдосконаленню швидкості та точності навчання ШНМ [5, 60].

2.3. Методи визначення інформативності змінних

Підкреслення інформативності вхідних змінних у моделі ШНМ передбачає дослідження значущості або внеску кожного вхідного параметру в прогнози моделі. Цей процес часто відомий як аналіз важливості змінних. Аналіз важливості змінних можна розділити на статистичні методи та методі на основі машинного навчання.

Наведено кілька статистичних методів, які можна використовувати для визначення інформативності в моделях ШНМ:

1. ***Дисперсійний аналіз ANOVA*** — це набір статистичних моделей та процедур для оцінки та аналізу відмінностей середніх значень змінних[78]. Цей метод використовується для статистичної перевірки відмінностей між середніми значеннями змінних. Перевірка мінливості вхідних змінних може бути пов’язана з їхнім впливом на вихідні змінні. Більша дисперсія свідчить про більший вплив, а значить, більшу інформативність [78, 79]. *Переваги*: ідентифікує ознаки з високою змінністю та невимогливий до обчислювальних ресурсів. *Недоліки*: фіксує лише лінійні зв’язки, чутливий до незбалансованих наборів даних та викидів, може неоднозначно визначати інформативність.

2. ***Аналіз коефіцієнтів кореляції*** використовує числові показники статистичних взаємозв’язків між двома змінними (у нашому випадку між вхідними змінними і цільовими значеннями) [80]. Аналіз кореляції між окремими змінними та виходами моделі ШНМ може дати розуміння їх взаємозв’язків. *Переваги*: невимогливий до обчислювальних ресурсів, легка інтерпретація. *Недолік*: фіксує лише лінійні зв’язки.

3. ***Аналіз взаємної інформації*** є потужним методом для оцінки інформативності вхідних параметрів у багатьох ситуаціях, включаючи аналіз

інформативності в нейронних мережах. Метод вимірює взаємну залежність між двома змінними, яка може бути використана для вимірювання залежності між значеннями вхідних параметрів і виходами моделі, забезпечуючи кількісну оцінку відповідності [81]. *Переваги*: можна використовувати на будь-якій моделі ШНМ, легка інтерпретація, стійкий до викидів і враховує нелінійність. *Недоліки*: вимогливий до обчислювальних ресурсів, обмеження в розмірах даних.

Наведено кілька методів на основі машинного навчання, які можна використовувати для визначення інформативності в моделях ШНМ:

1. ***Аналіз важливості змінної.*** Такі моделі, як AdaBoost, Random Forests або eXtreme Gradient Boosting, можуть оцінити важливість кожної вхідної змінної. Аналізуючи ці навчені моделі на результатах навчання моделі ШНМ можна оцінити інформативність вхідних змінних оригінальної моделі [82]. *Перевага*: забезпечує глобальний погляд на інформативність змінних. *Недоліки*: метод може не охоплювати складні зв’язки через простоту моделі для визначення інформативності порівняно з можливою складністю ШНМ.

2. ***Permutation Importance*** є популярною технікою для оцінки впливу окремих вхідних змінних на виходи моделі ШНМ [83]. Метод працює за допомогою впливу на вхідні змінні у даних і спостереженням за змінами в виходах ШНМ. Більша різниця між очікуваними й отриманими виходами моделі свідчить про більший змінної. *Переваги*: простота реалізації, працює з моделлю ШНМ, дає точну оцінку, стійкий до незбалансованих даних чи викидів. *Недоліки*: не ефективний в обчислюванні бо вимагає обчислення для кожної змінної, та метод спроможний визначити вплив лише одного параметру за один крок, не може охопити кілька змінних, неможливо використовувати на моделях ШНМ для вирішення задачі класифікації.

3. ***SHApley Additive ExPlanations (SHAP)*** — просунутий метод для пояснення виходів моделей машинного навчання, в тому числі ШНМ, працює використовуючи кооперативну теорію ігор. Метод детально аналізує вхідні змінні та їх вплив на виходи моделі та здатен виявляти впливи декількох змінних [84, 85]. *Переваги*: точність оцінки інформативності, дає пояснення для отриманих

результатів. *Недоліки*: обчислювально неефективний, обмеження по можливим моделям ШНМ.

4. ***Integrated gradients (IG)*** — це метод визначення впливу конкретних змінних на виході моделі ШНМ за допомогою визначення коефіцієнтів впливу для кожної змінної. [86, 87]. Цей метод є популярним у застосунках згроткових ШНМ для проблем комп'ютерного зору для висвітлення на зображеннях інформативних зон [88], та для пояснення роботи мовних моделей [87]. *Переваги*: обчислювально ефективний, має градієнтну інтерпретацію тому легко пояснюється. *Недоліки*: чутливий до шуму в даних, вплив базових значень змінних може визначати коефіцієнти впливу.

5. ***Gradient-based Sensitivity Analysis (GBSA)*** — це метод визначення чутливості на основі аналізу похідних першого чи вищого порядків виходів ШНМ відносно вхідних змінних для визначення впливу входів моделі на виходи [89]. *Переваги*: можливість фіксувати нелінійні зв'язки між входами та виходами, точність результатів оцінок. *Недоліки*: може бути обчислювально дорогим особливо у випадку вищих похідних, складний в реалізації особливо на нетипових моделях ШНМ, вибір порядку похідної впливає на оцінки інформативності.

2.3.1. Метод визначення загальної інформативності змінних

Оскільки для вирішення завдання класифікації пропонується застосовувати модель ШНМ та враховуючи підхід із використанням градієнтних методів, можна скористатися даною моделлю для пошуку оптимальної підмножини інформативних змінних (задача відбору змінних). Оцінка інформативності змінних при апріорній невизначеності даних передбачає синтез множини інформативних керованих змінних відповідно до стану в комп'ютерній системі медичного моніторингу. Розроблений підхід для нашої моделі ШНМ призначений для зменшення розмірності простору параметрів вхідних даних шляхом знаходження множини найбільш інформативних змінних вхідних даних S_β мінімальної розмірності, де $S_\beta \subset S$, де S – множина усіх змінних вхідних даних.

Множину змінних стану $S = \{s_m\}$, $m = \overline{1, M}$ представимо як ряди Тейлора зі збереженням лише членів нескінченно малого порядку, тоді дисперсія виходів довільної лінійної функції декількох виходів набуває виду:

$$D_{Y_k} = (grad\ Y_k)^T\ \Sigma_S\ grad\ Y_k =$$
$$= \sum_{m=1}^{M}\left(\frac{\partial \mathrm{Y_k}}{\partial \mathrm{s_m}}\right)^2 \sigma_{\mathrm{s_m}}^2 + \sum_{m=1}^{M}\sum_{l=1, l\neq m}^{M} r_{ml}\ \frac{\partial Y_k}{\partial \mathrm{s_m}}\frac{\partial Y_k}{\partial \mathrm{s_l}}\ \sigma_{\mathrm{s_m}}^2\sigma_{\mathrm{s_l}}^2, \qquad (2.25)$$

де

Σ_S – коваріаційна матриця вхідних змінних S,

$\sigma_{\mathrm{s_m}}$ – стандартне відхилення змінної s_m,

r_{ml} – коефіцієнт кореляції між змінними $\mathrm{s_m}$ та s_l,

$Y_k(S)$ – функція, що описує взаємозв'язок між вхідними та вихідними змінними.

Далі оцінюємо дисперсію та стандартне відхилення виходів навченої моделі ШНМ, та визначаємо за ними енергію сигналів [77]:

$$E_k = \sum_{\theta=1}^{\Theta_0}\left|D_{Y_k^{(2)}|\mathrm{Y}_\theta^{(0)}}\right|, \qquad (2.26)$$

де $D_{Y_k^{(2)}|\mathrm{Y}_\theta^{(0)}} = \left(\frac{\partial Y_k^{(2)}}{\partial Y_\theta^{(0)}}\right)^2 \sigma_{Y_\theta^{(0)}}^2 + \left(\sum_{n=1, n\neq\theta}^{\Theta_0} r_{hn}\ \frac{\partial Y_k^{(2)}}{\partial Y_n^{(0)}}\ \sigma_{Y_n^{(0)}}\right)\frac{\partial Y_k^{(2)}}{\partial Y_\theta^{(0)}}\ \sigma_{Y_\theta^{(0)}}$.

Далі матриця коефіцієнтів впливу $Y_\theta^{(0)}$ в $Y_k^{(2)}$ визначається виразом:

$$\beta_{k\theta} = \frac{\left|D_{Y_k^{(2)}|Y_\theta^{(0)}}\right|}{E_k}. \qquad (2.27)$$

Враховуючи матрицю впливів (2.27) можемо оцінити значення інформативності за наступним виразом:

$$GBI_\theta = \left(\sum_k^K \beta_{\theta n}\right)/\left(\sum_\theta^N \sum_k^K \beta_{\theta n}\right). \qquad (2.28)$$

Отримані коефіцієнти інформативності змінних використовуються для визначення рейтингу найбільш впливових змінних та в залежності від характеристик підсистеми прийняття рішень можуть допомогти виділити підмножину S_β найбільш інформативних змінних заданої розмірності по відсотку загальної інформативності.

2.3.2. Метод визначення поточної інформативності змінних

Розроблений метод визначення загальної інформативності по суті є представником градієнтних методів, тому він не дозволяє оцінювати інформативність вхідних змінних для конкретних даних входу й виходу моделі ШНМ, а слугує в якості загального індикатора інформативності змінних. В силу нелінійної природи поведінки ШНМ, ця інформативність може змінюватися. В контексті запропонованої моделі стратифікації елементів комп'ютерної системи медичного моніторингу нам також варто мати розуміння, чому модель ШНМ прийняла те чи інше рішення для цього добре підходить метод IG, що по суті вказує на те чому модель прийняла саме таке рішення.

Метод IG дозволяє оцінити, як кожна вхідна змінна випливає на виходи моделі. Це особливо важливо для інтерпретації рішень ШНМ, особливо в областях, де зрозумілість прийнятих рішень є критичною. Основна ідея полягає в тому, щоб інтегрувати градієнти функції витрат відносно вхідних змінних вздовж лінії від базового стану (зазвичай нульового) до конкретного вхідного стану [87–89]. Інтегрування градієнтів відбувається за виразом наведеним нижче [87–89]:

$$IG_{S_m}(\vec{x}) = (x_m - x'_m) \int_{\alpha=0}^{1} \frac{\partial F(x' + \alpha(x - x'))}{\partial x_m} d\alpha, \tag{2.29}$$

де

$\vec{x}$ – вектор вхідних даних,

F – функція, що описує ШНМ,

x_m – m-та змінна у векторі вхідних даних,

x'_m - значення m-ї змінної базового стану.

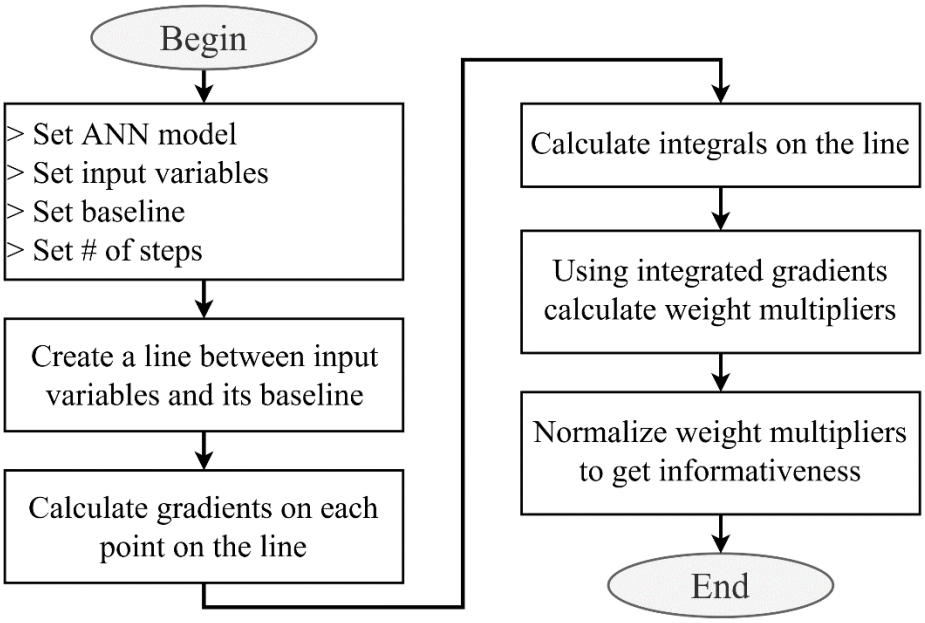


Рис. 2.7. Основні етапи методу інтегрованих градієнтів.

Враховуючи вираз (2.27), визначення алгоритму IG [87, 89] та необхідність визначення інформативності алгоритм для удосконаленого методу IG може бути визначений наступним чином [52, 90] (Рис. 2.7).

1. Визначення базового стану вхідних змінних, часто це буває нульовий стан або деякі стандартні значення.

2. Створення лінії між базовим і вхідним станом у просторі змінних.

3. Розрахунок градієнтів вздовж цієї лінії в кожній точці між базовим і вхідним станом змінних.

4. Інтеграція градієнтів вздовж зазначеної вище лінії. Це може бути зроблено, наприклад, за допомогою методів чисельного інтегрування, таких як метод трапецій або метод Сімпсона.

5. Обчислення множників ваг змінних.

6. Нормалізація множників ваг для отримання значень інформативності за наступним виразом:

$$IG_{inform} = |IG(\bar{x}) / \sum_N |IG(\bar{x})|. \quad (2.30)$$

Розроблені метод визначення загальної інформативності та модифікований метод IG дозволять вирішити проблеми визначення впливу змінних стану та опису чутливості моделі ШНМ до зміни значень вхідних змінних. Вихідні значення інформативності в обох методах приведені до одного числового рівня, що спростить їх інтерпретацію, та інтеграцію в готові рішення.

Висновок до розділу 2

У другом розділі дисертації було детально розглянуто та описано методи та моделі стратифікації елементів комп'ютерних систем медичного моніторингу. Для цього був представлений мультиагентний метод нечіткої кластеризації як вдосконалення традиційного методу c-means, відповідно до виявленої проблеми методів кластеризації в першому розділі. Що відповідно до визначення методу c-means та застосування мультиагентного підходу призвело до математичного виведення методу мультиагентної нечіткої кластеризації.

Відповідно до виявлених проблем методів класифікації була розглянута узагальнена модель ШНМ з акцентом на застосування удосконаленого методу навчання цієї моделі, що має прискорити та покращити точність навчання. Також відповідно до вимог які має задовільнять така ШНМ був розглянутий удосконалений метод підбору гіперпараметрів, що дозволить автоматично налаштовувати архітектуру моделі ШНМ відповідно до наявної складності даних задля отримання найкращих показників точності.

Далі представлено метод визначення загальної інформативності змінних з використанням інформації про поширення градієнтів в розробленій ШНМ, який дозволяє враховувати їх вплив на процес класифікації, а також представлено спосіб зменшення кількості спостережуваних параметрів. Далі в рамках вирішення задачі стратифікації, а саме визначення інформативності змінних було наведено удосконалений метод інтегрованих градієнтів для визначення поточної інформативності. Цей метод дозволяє аналізувати причини прийняття рішень навченою моделлю ШНМ, виділяючи поточні найвпливовіші змінні.

Основні положення цього розділу викладені у публікаціях автора [1–8].

РОЗДІЛ 3
РЕАЛІЗАЦІЯ МЕТОДІВ ТА МОДЕЛЕЙ СТРАТИФІКАЦІЇ НА ОСНОВІ МУЛЬТИАГЕНТНОГО ПІДХОДУ

3.1. Вибір програмного забезпечення для реалізації моделей

Для реалізації запропонованих методів і моделей стратифікації в комп'ютерній системі медичного моніторингу було проаналізовано декілька ключових критеріїв вибору програмного забезпечення:

- ***Функціональність:*** програмне забезпечення має надавати платформу для можливості реалізації запропонованих методів та моделей. Вибране програмне забезпечення має підтримувати використання сторонніх бібліотек для обробки та аналізу даних, для побудови графіків та надавати платформу для реалізації методів машинного навчання (як то наявність засобів оптимізації математичних обчислень, тощо). Також програмне забезпечення має надавати платформу для інтеграції запропонованих методів і моделей в існуючі комп'ютерні системи медичного моніторингу.
- ***Ефективність обробки даних:*** для роботи з наборами даних різних типів та різних об'ємів необхідні програмні засоби ефективні в керуванні пам'яттю, гнучкі в можливостях управління даними та такі, що надають стандартний інтерфейс для інших програмних засобів для роботи з цими даними.
- ***Простота використання:*** програмний засіб має бути легким для вивчення та використання для спеціалістів із обробки даних та машинного навчання без великого досвіду в програмуванні. Python з його чітким синтаксисом і великими бібліотеками добре підходить для цього.
- ***Спільнота та ресурси:*** велика й активна спільнота, яка підтримує обраний програмний засіб, має вирішальне значення для усунення несправностей і пошуку рішень. Також наявність великої кількості бібліотек допоможе обирати декілька варіантів реалізації в межах одного програмного засобу.

У зв’язку з форматом методів та моделей стратифікації, що розробляється вибір програмного засобу для реалізації впирається в вибір мови програмування і вибір можливих бібліотек в межах обраної мови програмування. Також не менш важливим компонентом є вибір середи розробки відповідно до обраної мови програмування. Обрана середа розробки часто має вирішальний вплив на якість реалізації методів і моделей, бо може запровадити інструменти на базі методів штучного інтелекту для аналізу якості написаного коду [94]. Тому вибір програмного забезпечення для реалізації запропонованих методів і моделей стратифікації складається з трьох компонент, це вибір мови програмування, вибір відповідних бібліотек в межах обраної мови програмування та вибір середовища розробки, що в підсумку вплине на якість реалізації.

3.1.1. Вибір мови програмування та огляд доступних бібліотек

Розглянемо спочатку вибір можливих мов програмування та наявних у них бібліотек. Також необхідно звернути увагу на відповідність обраних засобів програмного забезпечення наведеним критеріям. Очевидними кандидатами для реалізації методів і моделей є мови програмування з числа найбільш поширених мов програмування серед спеціалістів в машинному навчанні та застосуванні в комп’ютерних системах медичного моніторингу [95]: C/C++, Java, Python, R.

Придатність C/C++ та можливих бібліотек для реалізації запропонованих методів та моделей можна визначити наступними пунктами:

1. ***Продуктивність.*** Перевага C/C++ у продуктивності залишається незаперечною, особливо для обробки даних у реальному часі та критичних сценаріїв. Це може бути вирішальним для обробки великих високошвидкісних потоків медичних даних.

2. ***Контроль:*** для розробників із досвідом, детальний контроль над алгоритмами та керуванням пам’яттю може бути корисним у певних сферах, як-от низькорівнева взаємодія апаратного забезпечення або оптимізація спеціальних алгоритмів.

3. ***Наявність бібліотек для роботи з даними.*** Бібліотеки для прискорення лінійної алгебри такі як Armadillo, Eigen і Dlib надають широкі можливості у реалізації методів машинного навчання.

Але серед критичних недоліків C/C++ слід зазначити наступні:

1. ***Складність розробки:*** код C/C++ залишається більш складним для написання, налагодження та підтримки порівняно з більш високорівневими мовами програмування. Це може суттєво вплинути на час розробки та потенційно створити вразливі місця безпеки в медичному контексті.

2. ***Обмежена екосистема бібліотек для роботи з машинним навчанням:*** хоча нові бібліотеки пропонують деякі можливості, але багато з них є копіями бібліотек доступних в інших мовах програмування, що мають менший набір можливих функцій. Це призводить до збільшення об’єму виконуваної роботи, як наслідок більшої складності в налаштуванні коду.

3. ***Мала спільнота та підтримка:*** Спільнота C/C++, орієнтована на машинне навчання, менша за спільноти інших мов, що ускладнює пошук інформації та зменшує можливості з усунення проблем.

Незважаючи на потенціал продуктивності C/C++, складність розробки, обмежена екосистема Data Science, проблеми з меншими бібліотеками і менша підтримка спільноти роблять цю мову програмування небажаною для реалізації методів і моделей стратифікації в системах медичного моніторингу, особливо з огляду на критичний характер безпеки медичних програм.

Придатність Java та можливих бібліотек для реалізації запропонованих методів та моделей можна визначити наступними пунктами:

1. ***Інтеграція з існуючими системами:*** якщо система потребує інтеграції з наявною інфраструктурою лікарень або інструментами на основі Java, що спрощує впровадження підсистем в існуючі рішення.

2. ***Велика спільнота розробників:*** Java може похвалитися великою та активною спільнотою розробників, яка надає достатньо ресурсів і підтримки для усунення несправностей і пошуку експертів.

3. ***Зростаючі бібліотеки:*** такі бібліотеки, як Weka, H2O та Deeplearning4j, надають можливості машинного навчання та аналізу даних, хоча й не такі широкі, як екосистема Python.

Проте Java має найбільший недолік серед запропонованих мов програмування, а саме малу кількість бібліотек для обробки даних та реалізації методів машинного навчання, та відносно малу спільноту. Що унеможливлює реалізацію якісного програмного забезпечення на базі запропонованих методів і моделей стратифікації.

Також серед розглянутих мов слід зазначити R. Для реалізації запропонованих методів і моделей ця мова програмування має ряд переваг:

1. ***Статистичний аналіз:*** R має велику кількість бібліотек для статистичного аналізу, перевірки гіпотез і візуалізацій. Це може бути корисним для таких завдань, як дослідження даних, виявлення викидів і розробка методів у контексті систем медичного моніторингу.

2. ***Велика спільнота спеціалістів з медичних досліджень:*** R користується значною популярністю в колах спеціалістів з досліджень медичних даних, пропонуючи доступ до готових бібліотек і наборів даних, що стосуються сфери медицини.

3. ***Розширена екосистема обробки даних:*** такі бібліотеки, як dplyr, tidyr і data.table, надають потужні та ефективні інструменти для очищення від аномалій та перетворення даних.

4. ***Велика кількість бібліотек машинного навчання:*** такі бібліотеки, як caret, mlr і keras, інтегрують популярні методи машинного навчання, пропонуючи можливості для кластеризації та впровадження ШНМ.

Тим не менш, мові програмування притаманний ряд недоліків:

1. ***Низька обчислювальна ефективність:*** R є менш оптимізована за зазначені вище мови програмування, особливо для роботи з великими наборами даних або обробки даних в реальному часі. Це є суттєвим обмеженням для обробки великих потоків даних медичного моніторингу.

2. ***Низька ефективність розробки.*** Синтаксис і робочий процес R можуть бути менш інтуїтивно зрозумілими, ніж інші зазначені мови, що не підходить для швидкої та точної реалізації методів і моделей стратифікації.

3. ***Проблеми інтеграції:*** інтеграція R із зовнішніми системами чи мовами програмування може вимагати додаткових засобів.

R є підходящою мовою програмування для реалізації методів і моделей стратифікації, якщо статистичний аналіз, узгодження клінічних досліджень і досвід візуалізації є ключовими пріоритетами. Її можливості обробки даних і машинного навчання постійно зростають, пропонуючи достатню функціональність для багатьох завдань. Проте, продуктивність, швидкість розробки, і якісна інтеграція є критичними, тому R не є найкращим вибором.

Наступною мовою з її бібліотеками буде розглянута мова програмування Python. Python має найкращу відповідність по зазначеним критеріям придатності і має наступні переваги до застосування:

1. ***Універсальність:*** Python ідеально поєднує бібліотеки з аналізу даних, машинного навчання, прискорених математичних обчислень, широкі можливості з візуалізації даних та процесів функціонування методів та моделей. Також Python широко використовується в комп’ютерних система медичного моніторингу. Тому Python пропонує уніфіковане середовище для всього робочого процесу: від реалізації та тестування до впровадження запропонованих методів та моделей.

2. ***Продуктивність.*** Не дивлячись на те, що бібліотеки Python, такі як NumPy, Pandas і Dask, менш обчислювально ефективні за аналоги на C/C++, досі ефективно обробляють великі набори даних, а оптимізовані реалізації забезпечують прийнятну продуктивність для більшості медичних сценаріїв. Також Python виділяється легкою можливістю проведення обчислень на відеоадаптерах, що в сотні разів прискорює виконання однотипних операцій на великих наборах даних (як то навчання моделей ШНМ).

3. ***Простота використання:*** чіткий синтаксис Python і великі стандартні бібліотеки роблять його зручним для початківців, скорочуючи час і складність

розробки, особливо для медичних працівників, які беруть участь у наукових дослідженнях.

4. ***Екосистема Data Science.*** Такі бібліотеки, як NumPy, pandas, scikit-learn, TensorFlow і PyTorch, надають комплексні інструменти для кожного етапу проекту, від маніпулювання даними до навчання моделі та розгортання.

5. ***Спільнота та підтримка.*** Велика та активна спільнота Python із Data Science та машинного навчання пропонує миттєвий доступ до допомоги, ресурсів і найкращих практик, забезпечуючи підтримку протягом усього процесу розробки й впровадження методів і моделей машинного навчання.

Тим не менш варто враховувати відсутність інтеграції з апаратним забезпеченням. Для прямої взаємодії з апаратним забезпеченням низького рівня C/C++ може запропонувати перевагу. Однак такі бібліотеки, як Numba, можуть подолати розрив для багатьох сценаріїв інтеграції обладнання.

Враховуючи повну функціональність, простоту використання, підтримку спільноти та великі бібліотеки присвячені прискореним математичним обчисленням, обробці даних та машинного навчання, Python – є найбільш підходящою мовою для реалізації методів і моделей стратифікації комп'ютерних медичного моніторингу. Python пропонує оптимальний баланс між швидкістю розробки, продуктивністю, гнучкістю та пояснюваністю, ключовими факторами в медичному контексті. Хоча альтернативні мови можуть мати переваги в певних сферах, їх недоліки часто значною мірою перекривають переваги для цієї реалізації. Результати реалізації на обраній мові та сукупності бібліотек показані в роботах [52, 59, 60, 74, 90].

3.1.2. Вибір інтегрованого середовища розробки

Наступним компонентом програмного забезпечення для реалізації запропонованих методів і моделей є інтегроване середовище розробки. Використання розумного середовища розробки може спростити процес та покращити якість реалізації запропонованих методів і моделей. Для обраної мови програмування існують декілька видів інтегрованих середовищ розробки. Серед найбільш розповсюджених серед розробників Python та спеціалістів з машинного

навчання та Data Science, що використовують Python слід зазначити наступні середовища: Jupyter Server, PyCharm, Visual Studio Code, Spyder. Розглянемо кожну із запропонованих середовищ розробки більш детально:

- ***Jupyter Server*** – інтегроване середовище, що використовує браузер в якості робочого вікна та дозволяє поєднувати інтерактивне програмування, текст для пояснень і візуалізацію отриманих результатів із збереженням стану експериментів в HTML-подібному форматі. Підтримує декілька мов програмування в тому числі і Python, середовище є незамінним серед спеціалістів з машинного навчання та Data Science, адже дозволяє зберігати отримані результати, а наявний код, дозволяє їх відтворенню будь-яким спеціалістом з наявним відповідним середовищем.
- ***PyCharm*** – комплексне інтегроване середовище розробки, багатофункціональне, підтримує всі можливі варіанти сучасної розробки в тому числі реалізацію математичних методів і моделей. Підтримує інтелектуальну допомогу в розробці, а саме має інструменти доповнення коду, перевірки коду, та навігації по коду та документації проекту чи бібліотек. А також підтримує безліч застосунків для аналізу коду, що в цілому підвищує якість програмних продуктів. В розширеній версії має інтеграцію з Jupyter Server, що є незамінним для проведення наукових досліджень.
- ***Visual Studio Code*** – інтегроване середовище розробки, що підтримує більшість поширених мов програмування, але вимагає налаштування відповідних розширень, в тому числі доступне розширення для Python. Може реалізовувати весь функціонал PyCharm, але вимагає налаштування розширень для відповідних інструментів. Також самі розширення не являються уніфікованими тому можуть пропонувати менший функціонал, або навіть привносити помилки.
- ***Spyder*** – інтегроване середовище розробки, що було розроблено для аналізу даних та проведення наукових обчислень. Дозволяє використовувати безліч доповнень для поліпшення якості проведення експериментів. Але Spyder не є універсальним середовищем розробки й не може запропонувати комплексне керування великими проектами тому

Враховуючи все вище сказане, обрано було PyCharm, бо має підтримку розробки великих проектів та вбудований інтелектуальний аналізу коду, це в загальному покращить якість реалізації запропонованих методів і моделей стратифікації. Також вбудована підтримка Jupyter Server дозволить проводити експерименти з розробленими методами й моделями, та дозволить реалізувати їх відтворюваність.

3.2. Представлення наборів даних для проведення експериментів

У складному світі розробки систем, особливо для таких складних проектів, як реалізація запропонованих методів і моделей стратифікації в комп'ютерних системах медичного моніторингу, тестування окремих методів і моделей із різноманітними наборами даних — це абсолютна необхідність з наступних причин:

1. ***Виявлення специфічних для методу чи моделі неточностей:*** кожен запропонований метод чи модель у системі, від попередньої обробки даних до кластеризації, класифікації та визначення інформативності, можуть містити приховані помилки реалізації чи планування. Тестування кожного методу чи моделі різноманітними наборами даних, що охоплюють різноманітні демографічні дані пацієнтів, медичні стани та варіанти різноманітних змінних, дозволить виявити такі помилки. Це дозволяє вносити виправлення та вдосконалення, щоб забезпечити точне функціонування системи в різних сценаріях можливих даних.

2. ***Симуляція складності реальних даних:*** медичні дані в реальному світі не часто бувають однаковими. В даних часто наявний шум, викиди та непередбачувані взаємозв'язки. Тестування кожного методу і моделі з наборами даних, які імітують таку складність, стає можливо оцінити здатність запропонованих методів та моделей вирішувати поставлені задачі, а також порівнювати якість таких рішень з існуючими аналогами. Це підвищує загальну надійність комп'ютерних систем медичного моніторингу до виникнення непередбачуваних станів змінних під час застосування.

3. ***Перевірка спроможності до узагальнення:*** методи і моделі, навчені на обмеженому наборі даних, можуть добре працювати в цьому конкретному випадку, але давати хибні результати у випадку раніше не бачених даних. Тестування з

використанням різноманітних наборів даних допомагає виявити проблеми узагальнення, забезпечуючи адаптацію системи та її надійну роботу для широкого кола можливих застосувань комп'ютерних систем медичного моніторингу.

4. ***Сприяння безперервному вдосконаленню:*** результати тестування на різних наборах даних забезпечує зворотний зв'язок для виявленням можливостей до вдосконалення. Такі можливості спрямовують подальші зусилля щодо розробки, дозволяючи вдосконалювати алгоритми, оптимізувати параметри та підвищувати загальну якісь роботи запропонованих методів і моделей стратифікації.

3.2.1. Набори даних для тестування та налаштування

Набір даних ірисів Фішера.

Іриси Фішера, або набір даних розмірів квітки ірису – це набір даних, що був прославлений біологом Рональдом Фішером у своїй роботі «The use of multiple measurements in taxonomic problems as an example of linear discriminant analysis» [96], та зібраний Едгаром Андерсоном [96]. Набір даних включає 50 зразків ірисів та включає 3 можливі види ірисів (Setosa, Virginica та Versicolor). Для кожного зразку були виміряні 4 змінні характеристик, а саме довжина чашолистка, ширина чашолистка, довжина пелюстки, ширина пелюстки.

Цей набір даних часто використовується для перевірки працездатності запропонованих методів і моделей машинного навчання, хоча має доволі просто роздільні дані можливих класів. Але слід зазначити, що клас Setosa просторово віддільний від 2х інших, але Virginica і Versicolor неможливо відділити один від одного (Рис. 3.1). Це дає вводить додаткову складність для перевірки якості роботи запропонованих алгоритмів і свого роду вводить 2-х етапну перевірку, відділення Setosa від інших свідчить про можливості методу розділяти лінійно роздільні дані. А розділення Virginica і Versicolor є задачею з додатковою складністю і не завжди є обов'язковою для вирішення.

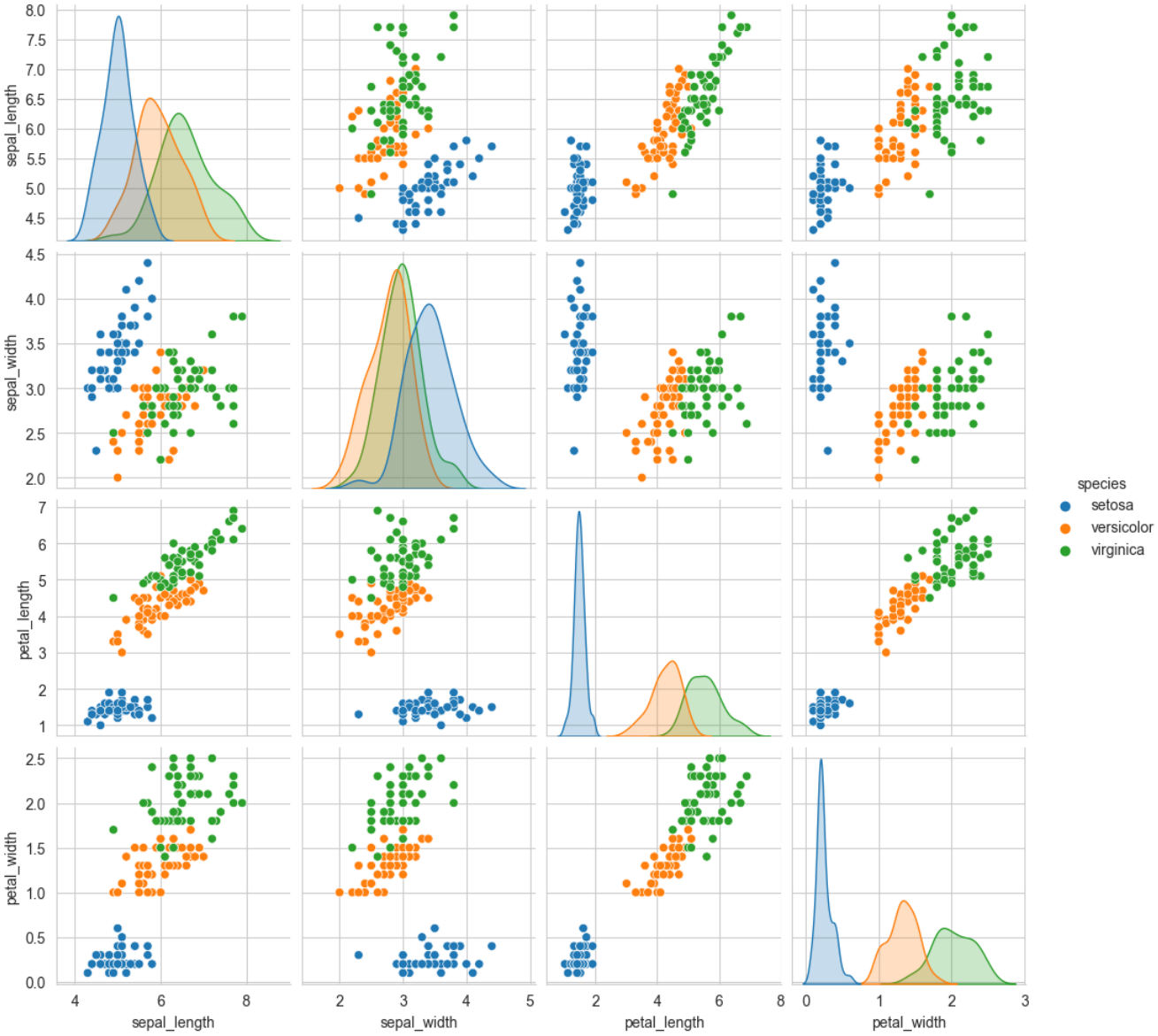


Рис. 3.1. Діаграма розсіювання ірисів Фішера.

Набір даних ідентифікації скла UCI

Набір даних ідентифікації скла – це набір даних від UCI [98], що включає 9 змінних стану та одну змінну, що визначає тип скла (7 можливих дискретних значень). Всі зміні стану є фізичними величинами, що описують вміст певних речовин у кінцевому матеріалі. Набір даних включає 214 записи. Цей набір даних є стандартним набором для тестування розроблених методів і доступний в бібліотеці UCI, що доступна в Python [98]. Змінні стану тісно пов'язані одна з одною та із цільовою змінною, що ілюструє діаграма кореляції, зображена на Рис. 3.2.

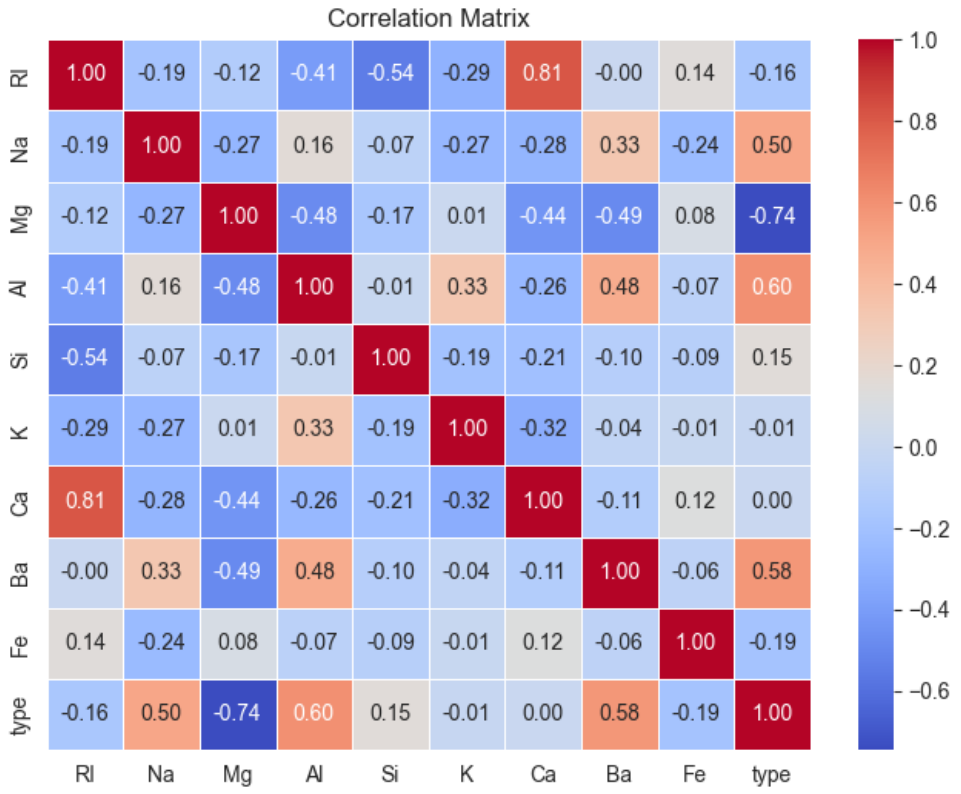


Рис. 3.2. Матриця кореляції змінних даних ідентифікації скла UCI.

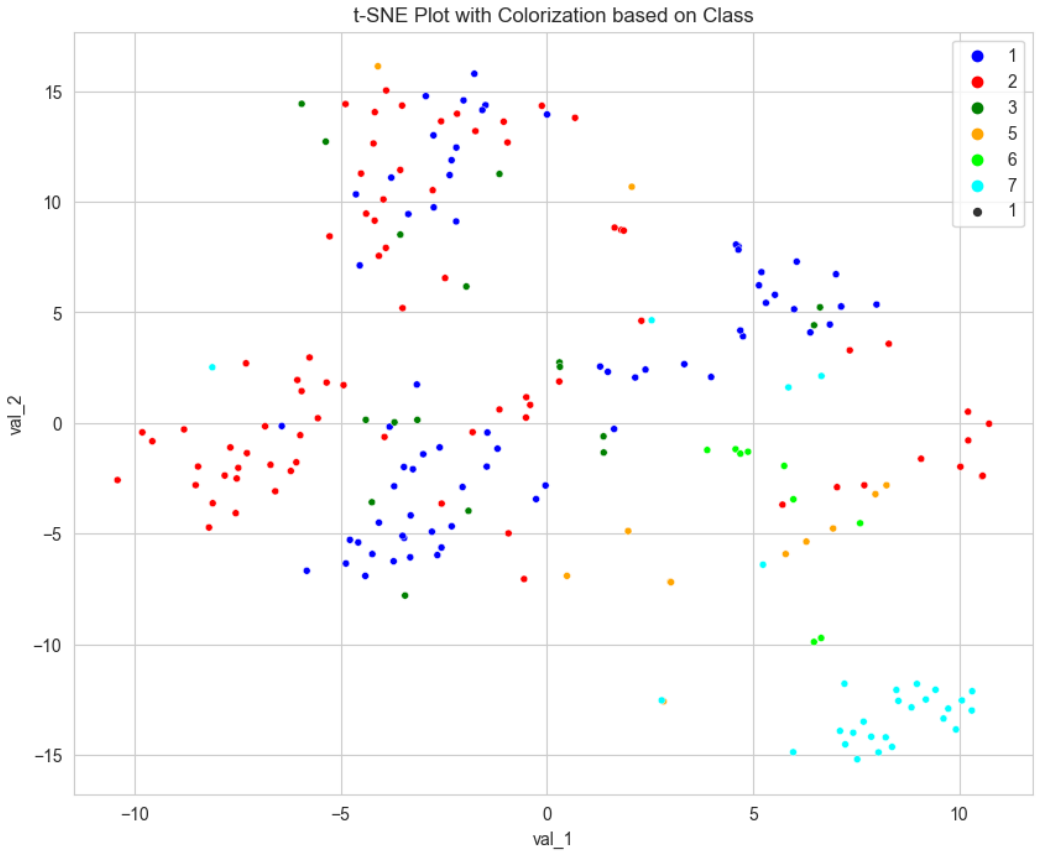


Рис. 3.3. Стиснення набору даних ідентифікації скла UCI методом t-SNE.

Представлені дані в наборі ідентифікації скла мають високий рівень спотворення, що виявляється в візуалізації їх стиснення за допомогою методу t-distributed Stochastic Neighbor Embedding (t-SNE) ((Рис. 3.3), хоча візуально деякі класи досі можливо розділити використання такого набору даних може надати перевірити спроможність розроблюваних методів і моделей працювати з реальними даними.

Набір даних типів вина UCI

Набір даних типів вина UCI – це набір результатів хімічного аналізу вин, що були виготовлені з трьох сортів винограду вирощеному в одному з регіоні Італії [99]. Дані аналізу вин включають 178 записів із 13 змінних стану та одну цільову змінну стану, що визначає один з трьох типів вин.

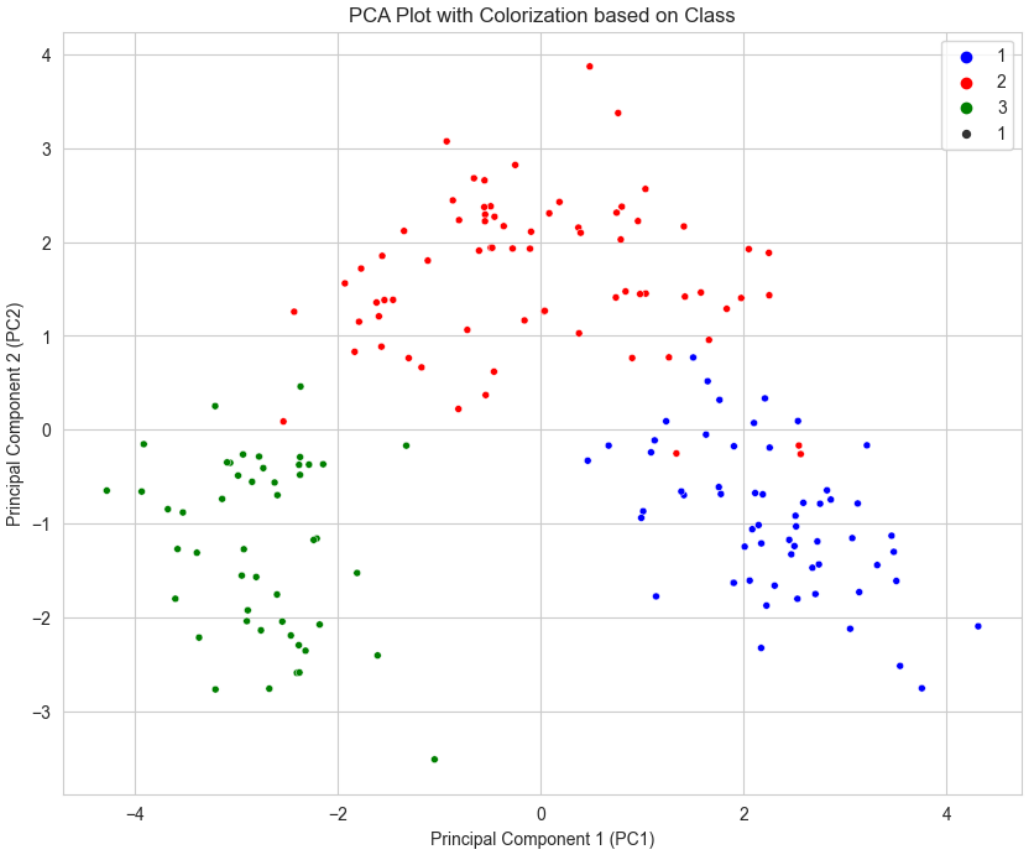


Рис. 3.4. Стиснення набору даних типу вин UCI методом PCA.

В контексті вирішення проблеми класифікації, чи кластеризації запропонований набір є ідеальним тестовим набором адже пропонує лінійно

роздільні (методом PCA, показано на Рис. 3.4) 3 класи із мінімальною кількістю шуму в даних.

Набір даних діагностики раку молочної залози Вісконсину

Набір даних діагностики раку молочної залози Вісконсину – це набір даних, що був отриманий в результаті обробки зображень клітин тканини молочної залози та описують клітинні ядра присутні на зображенні [100]. Набір даних складається з 569 записів, що включають 30 змінних стану (вимірюваних величин, що описуються дійсними числами) та одної цільової змінної, що визначає тип пухлини (злоякісна – M, та доброякісна – B).

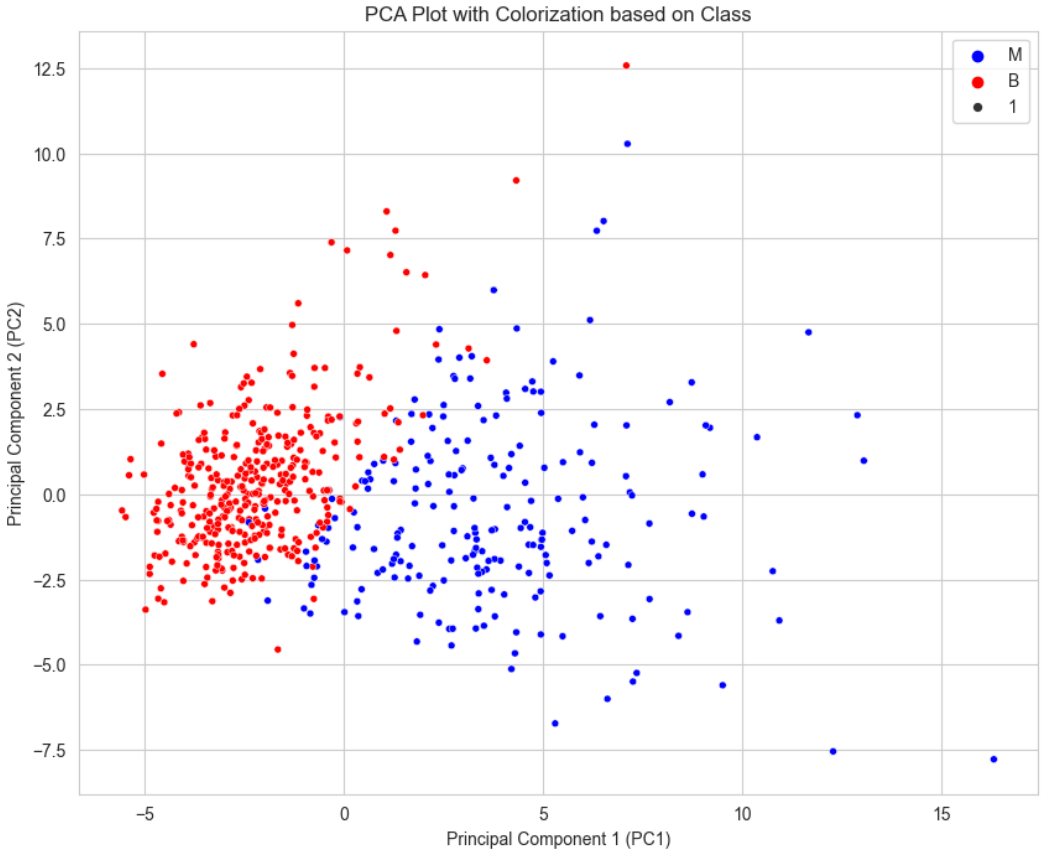


Рис. 3.5. Стиснення набору даних раку молочної залози методом PCA.

В контексті вирішення проблеми класифікації, чи кластеризації запропонований набір є ідеальним тестовим набором адже пропонує просторово нероздільні (методом PCA, показано на Рис. 3.5) 2 класи із мінімальною кількістю

шуму в даних. Нероздільність даних в такому виді, вимагає використання специфічних метрик для підвищення якості роздільності.

3.2.2. Дані для тестування методу кластеризації

Для тестування запропонованого методу мультиагентної нечіткої кластеризації було обрано набір даних про медичний моніторинг пацієнтів з захворюваннями передміхурової залози. Шляхом системного аналізу цього процесу була визначена ієрархія етапів діагностики, включаючи лабораторні дослідження, візуальну діагностику та контрольовані змінні стану пацієнтів, що відповідають кожному стану захворювання. Надалі було сформовано експериментальний набір контрольованих змінних, які характеризують стан пацієнтів, які спостерігаються у зв'язку з обраним захворюванням [59].

Набір даних складається з 180 записів пацієнтів, що містять 24 змінних стану та одну цільову змінну що приймає одне з чотирьох можливих значень. Серед 180 виділено 50 записів із доброякісною гіперплазією передміхурової залози та 130 із раком, дозволило детальніше виділити чотири можливих стани відповідно до критерію прогресування хвороби:

- здорові з доброякісним утворенням, до яких входить 50 записів,
- хворі без метастаз, до яких увійшло 45 записів,
- хворі з метастазами, до яких увійшло 52 записи,
- хворі та гормонорезистентні, до яких увійшло 33 записи.

Результати аналізу стиснення за допомогою методу t-SNE (Рис. 3.6) показали можливість нелінійного розділення даних з мінімальною кількістю шумів. Варто зазначити, що дані здорових пацієнтів є добре віддільні від даних хворих, але розділити дані хворих пацієнтів є складною задачею. Також був розглянутий кореляційний аналіз змінних стану та цільової змінної (Рис. 3.7), що показав, що деякі змінні мають високий рівень кореляції із цільової змінною, що може свідчити про можливість зменшення кількості змінних й порівняння точності роботи запропонованого мультиагентного методу нечіткої кластеризації на даних з різною наповненістю даними шуму.

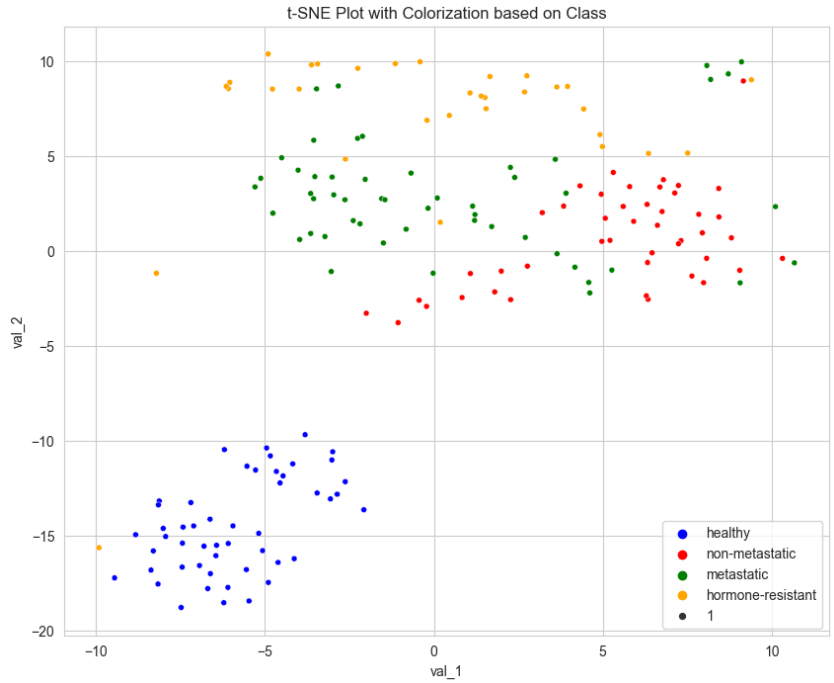


Рис. 3.6. Стиснення набору даних раку передміхурової залози методом t-SNE.

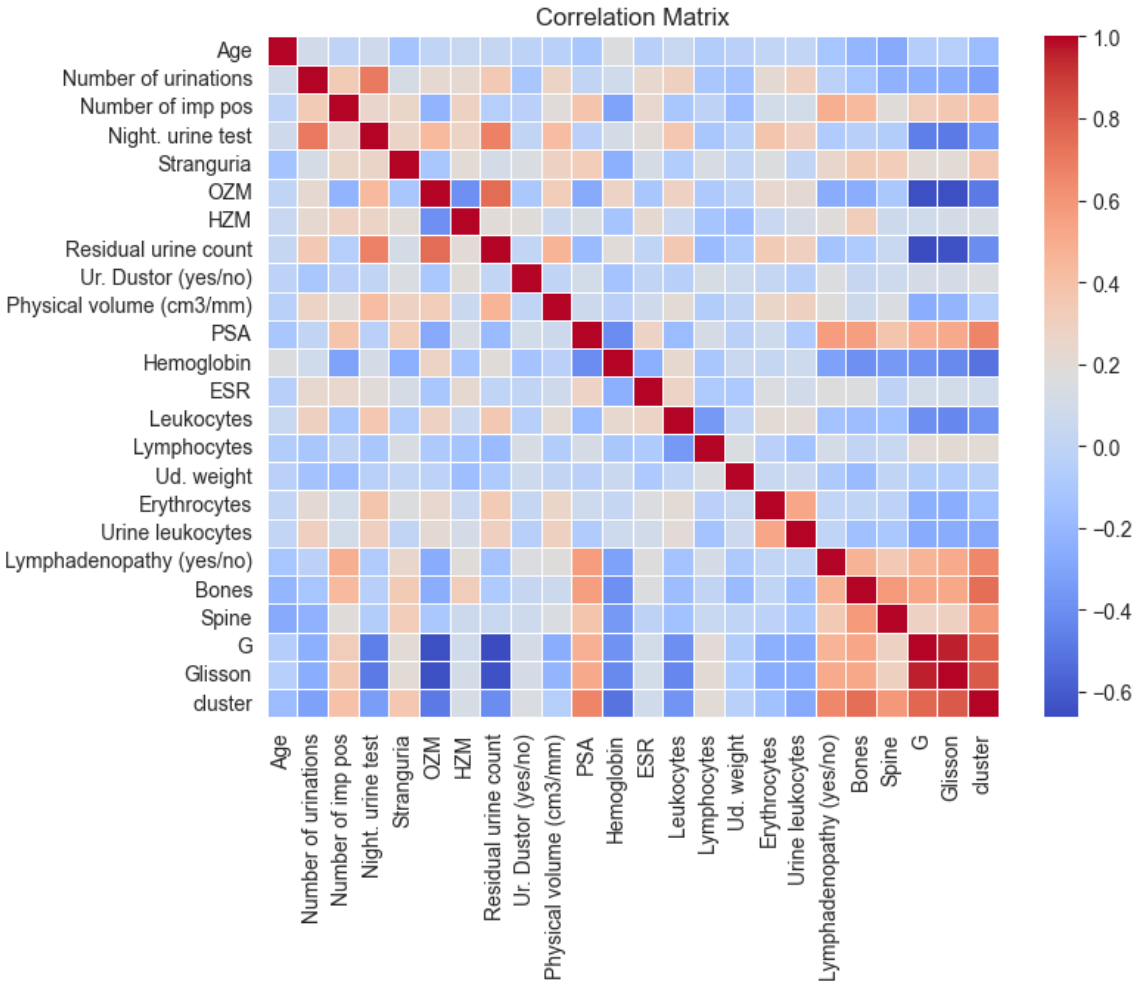


Рис. 3.7. Матриця кореляції даних раку передміхурової залози.

3.2.3. Дані для тестування моделі класифікації

Лабораторні дані урологічних досліджень

Набір лабораторних даних урологічних досліджень, що включає данні про 40 пацієнтів із 47 змінними стану, значення яких були або фізичними величинами, або перечислюваний тип. Всім записам була присвоєна одна змінна стану, що визначала один з двох можливих станів пацієнтів: здоровий чи хворий [5].

Набір даних захворювання печінки

Набір даних захворювання печінки містить дані про 590 пацієнтів, що характеризуються 10 змінними стану та однією цільовою змінною, що визначає можливий стан пацієнта як здоровий чи хворий [5].

3.2.4. Дані для тестування методів визначення інформативності

Для дослідження точності роботи запропонованих методів визначення загальної та поточної інформативності змінних стану був використаний дані серцевих захворювань UCI [101]. Цей набір даних представляє дані, що є типовими для комп'ютерних систем медичного моніторингу [10–12]. Набір даних серцевих захворювань UCI включає дані 920 пацієнтів із 76 змінними стану та цільовою змінною стану, що представляє один з 5 можливих станів пацієнтів. У зв'язку з надмірною складністю набору даних ми зосередилися на аналізі 13 змінних, що є поширеним в інших дослідженнях [102]. Слід зазначити, що записи пацієнтів були отримані з різних джерел, до яких увійшли найбільш місткі: Клівлендської бази даних – 33% та Угорщини – 32%, а інші джерела мають 35% всіх записів [101, 102]. В обраному наборі даних слід зазначити наступні змінні:

1) age – вік пацієнтів в роках;
2) origin – база даних походження;
3) gender – стать пацієнта;
4) cp – тип грудної болі: типова стенокардія, атипова стенокардія, неангінозна, безсимптомна;
5) trestbps – тиск крові в мм. рт. ст.;
6) chol – рівень холестеролу в мг/дл;
7) fbs – індикатор перевищення рівня цукру;

8) restecg – результат електрокардіографії в спокої: норма, аномалія, гіпертрофія;

9) thalach – максимально досягнута кількість серцевих скороцень;

10) exang – індикація наявності стенокардії;

11) oldpeak – наявність пригнічення серцевого м’язу спричиненого фізичним навантаженням;

12) slope – швидкість заспокоєння серця після пікового навантаження;

13) ca – кількість магістральних судин забарвлених при рентгеноскопії;

14) thal – тип дефекту серцевого м’язу: нормальний, фіксований дефект, оборотний дефект

15) class – цільова змінна: здоровий, або один з чотирьох типів захворювання серця.

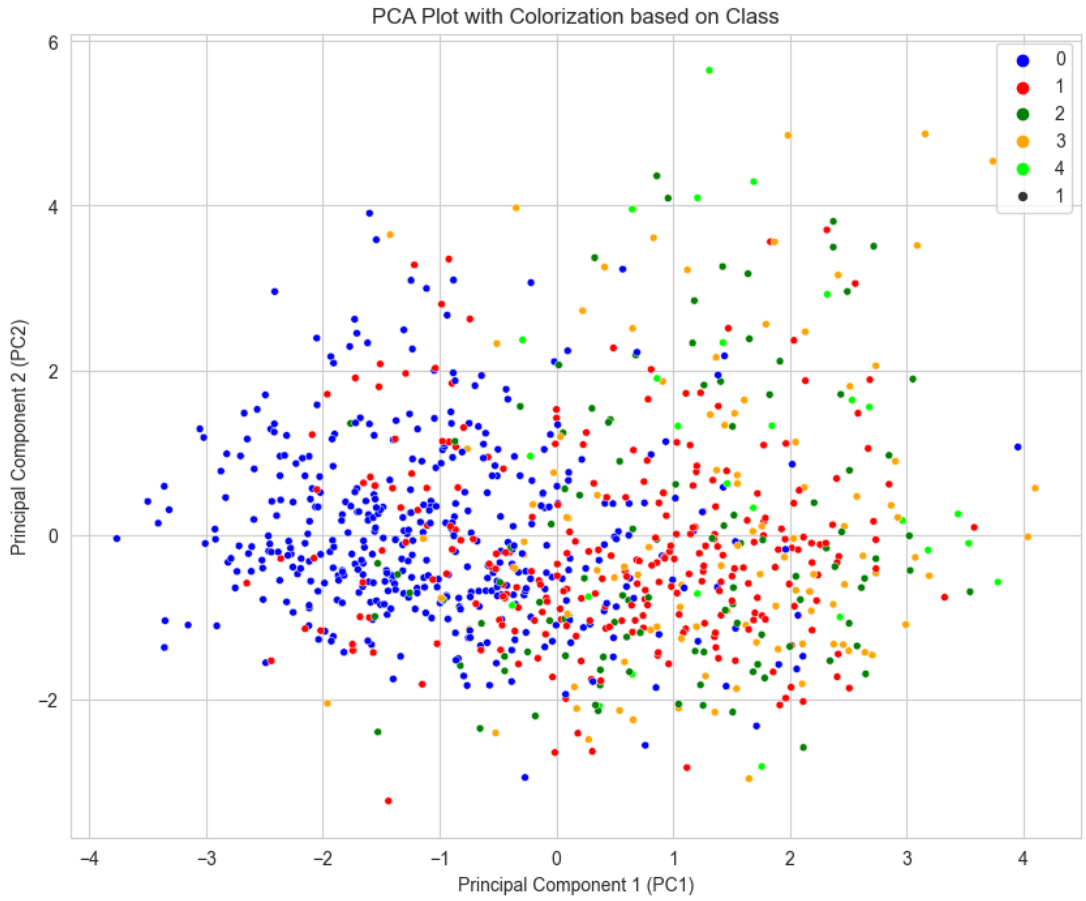


Рис. 3.8. Стиснення набору даних серцевих захворювань UCI методом PCA.

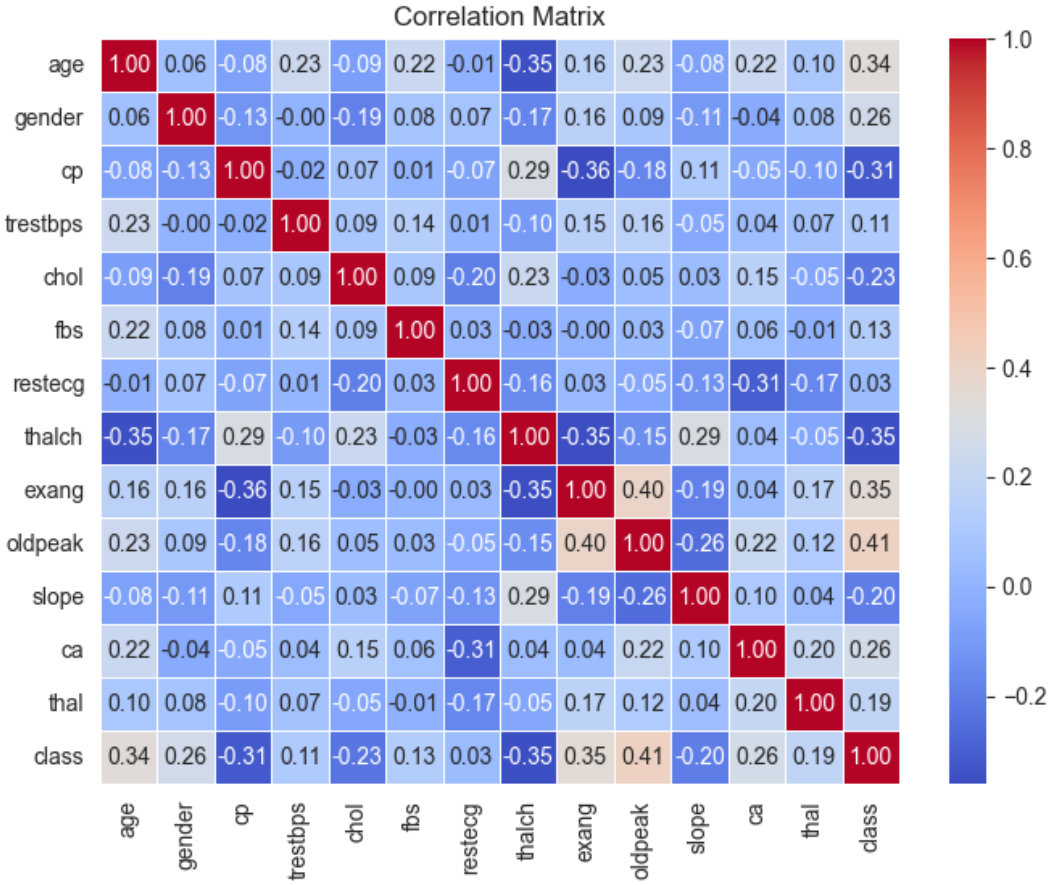


Рис. 3.9. Матриця кореляції даних серцевих захворювань UCI.

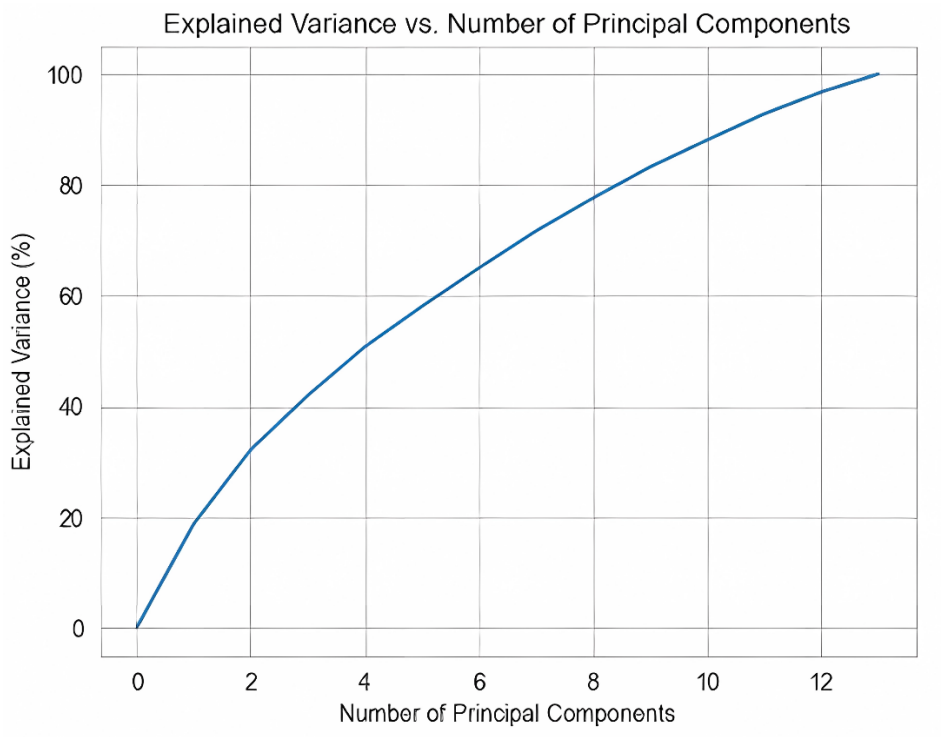


Рис. 3.10. Залежність варіативності від кількості принципових компонент, отримана за допомогою аналізу PCA.

Аналіз результатів стиснення даних серцевих захворювань UCI за допомогою методу PCA (Рис 3.8) показав лінійну нероздільність даних, та складну форму цільових кластерів, що буде випробування для запропонованих методів. Також аналіз матриці кореляції (Рис. 3.9) показав, що змінні стану між собою та з цільовою змінною мають низький рівень кореляції, що ще раз підтверджує складну природу взаємозв'язків, що лежать в середині запропонованого набору даних.

Подальший аналіз розподілів параметрів показав, що змінним age та gender притаманне зміщення щодо кількості хворих пацієнтів старшого віку та чоловічої статі відповідно. Також було проведено PCA для визначення варіативності (Рис. 3.10) змінних, та показано, що 10 принципових компонент достатньо для збереження 90% варіативності. Ще варто зазначити, що 70% даних містять відсутні значення, які були заповнені середніми величинами по відповідних змінних. Наведені результати аналізу та обробки даних вказують на складну природу запропонованих даних і те що вони є ідеальним плацдармом для перевірки запропонованих методів та моделей.

3.2.5. Дані для тестування підсистеми стратифікації

Останнім етапом роботи є перевірка точності функціонування розроблених методів і моделей стратифікації елементів комп'ютерних систем медичного моніторингу є проведення тестування загальної точності роботи. Для цього було відібрано набір даних про захворювання на діабет із загального набору медичних досліджень США із CDC BRFSS Survey 2021 [103].

Основний набір даних CDC BRFSS Survey 2021 – це дані моніторингу поведінкових факторів ризику, що є системою телефонного опитування у США. Система збирає дані про різні аспекти здоров'я та життя громадян США. Набір даних, що був використаний включає відповіді 438 тисяч осіб із 303 змінними стану, що виведені із запинань поставлених респондентам.

Захворювання на діабет можна охарактеризувати як тривалий стан організму людини, що відображається у його можливості отримувати енергію з їжі. Це захворювання можна розділити на 3 типи [103]:

- діабет I типу, характеризується негативною реакцією імунітету на клітини підшлункової залози, що виробляє інсулін, необхідний для перетворення глюкози в енергію;
- діабет II типу, найбільш поширений тип, виникає при недостатній кількості інсуліну, або ненормальної реакції організму на інсулін;
- гестаційний діабет, тимчасова форма діабету, що виникає під час вагітності й зникає після пологів.

Дані опитування про діабет були сформовані після обробки записів CDC BRFSS 2021. Отриманий набір включає 236 тисяч записів, що описуються 21 змінною стану, що експертами визначена як найбільш специфічні для діабету, та одну цільову змінну, що може приймати один з трьох можливих станів. Детальніше кожну змінну отриманого набору даних захворювання на діабет можна описати наступним чином:

1. *HighBP* – індикатор наявності високого кров'яного тиску у респондента,
2. *HighChol* – показник високого рівня холестиролу у респондента,
3. *CholCheck* – чи проводилась перевірка рівня холестиролу протягом останніх п'яти років,
4. *IMT* – індекс маси тіла,
5. *Smoker* – індикатор споживання тютюнових виробів,
6. *Stroke* – індикатор перенесення респондентом інсульту,
7. *HeartDiseaseorAttack* – індикатор перенесення серцевих нападів,
8. *PhysActivity* – індикатор наявності фізичної активності у респондента протягом останніх 30 днів,
9. *Fruits* – індикатор споживання фруктів,
10. *Veggies* – індикатор споживання овочів,
11. *HvyAlcoholConsump* – індикатор вживання алкоголю, на рівні що перевищує норму,
12. *AnyHealthcare* – індикатор наявності медичного страхування,
13. *NoDocbcCost* – індикатор пропуску огляду лікаря за останній рік,

14. *GenHlth* – оцінка рівня власного здоров’я респондента,
15. *MentHlth* – оцінка рівня власного психічного здоров’я респондента,
16. *PhysHlth* – оцінка рівня власного фізичного здоров’я респондента,
17. *DiffWalk* – індикатор складності використання сходинок,
18. *Gender* – стать респондента,
19. *Age* – вік респондента,
20. *Education* – рівень освіти респондента,
21. *Income* – рівень достатку респондента,
22. *Class* – цільова змінна, що визначає очікуваний стан пацієнта.

Подальший аналіз даних показав, що набору даних притаманна незбалансованість відносно цільового класу: 90%: записів є записами здорових респондентів. Також слід зазначити, що інші змінні не зазнали такої незбалансованості. Далі було проведено кореляційний аналіз змінних стану та цільової змінної. Результати кореляції показали (Рис. 3.12), що немає сильно пов’язаних змінних і значення кореляцій не перевищують 0.5 чи занижують -0.3. Також слід зазначити відсутність змінних з високою кореляцією до цільової змінної. Потім дані захворювань на діабет було проаналізовано з використанням методу PCA та отримано дві основні компоненти та відповідні точки позначені на Рис. 3.11. Важко розрізнити цільові класи за простими формами. На Рис. 3.13 показана залежність варіативності від кількості принципових компонент, отримана за допомогою PCA, та показано що для збереження варіативності даних на рівні 90% необхідно 15 змінних стану. Наведені результати аналізу вказують, що були представлені складні дані, які стануть справжнім викликом для запропонованої підсистеми стратифікації.

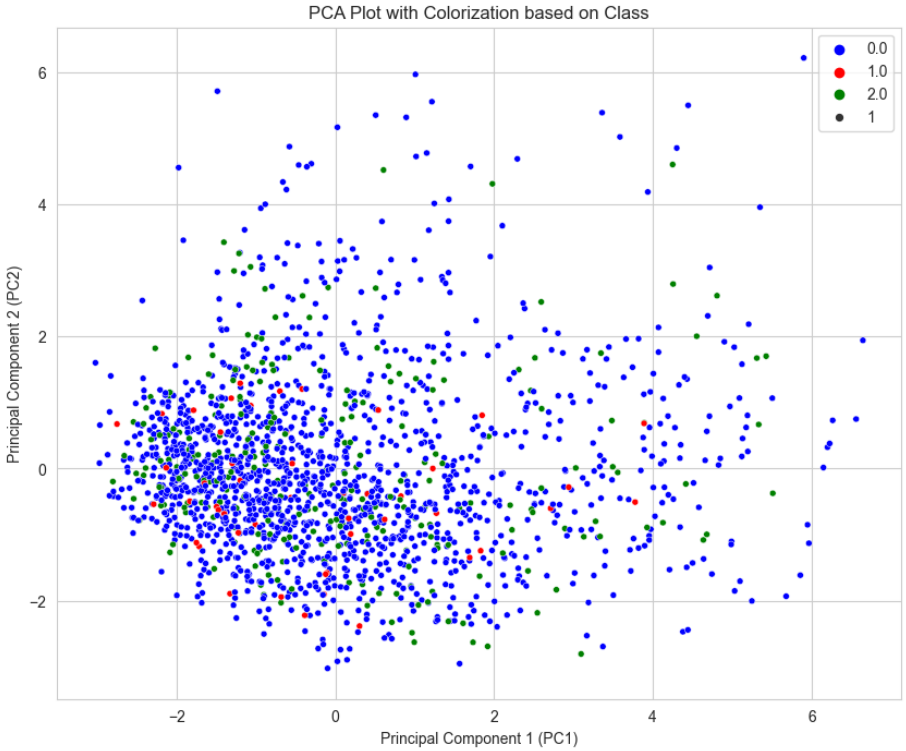


Рис. 3.11. Стиснення набору даних захворювань на діабет методом PCA.

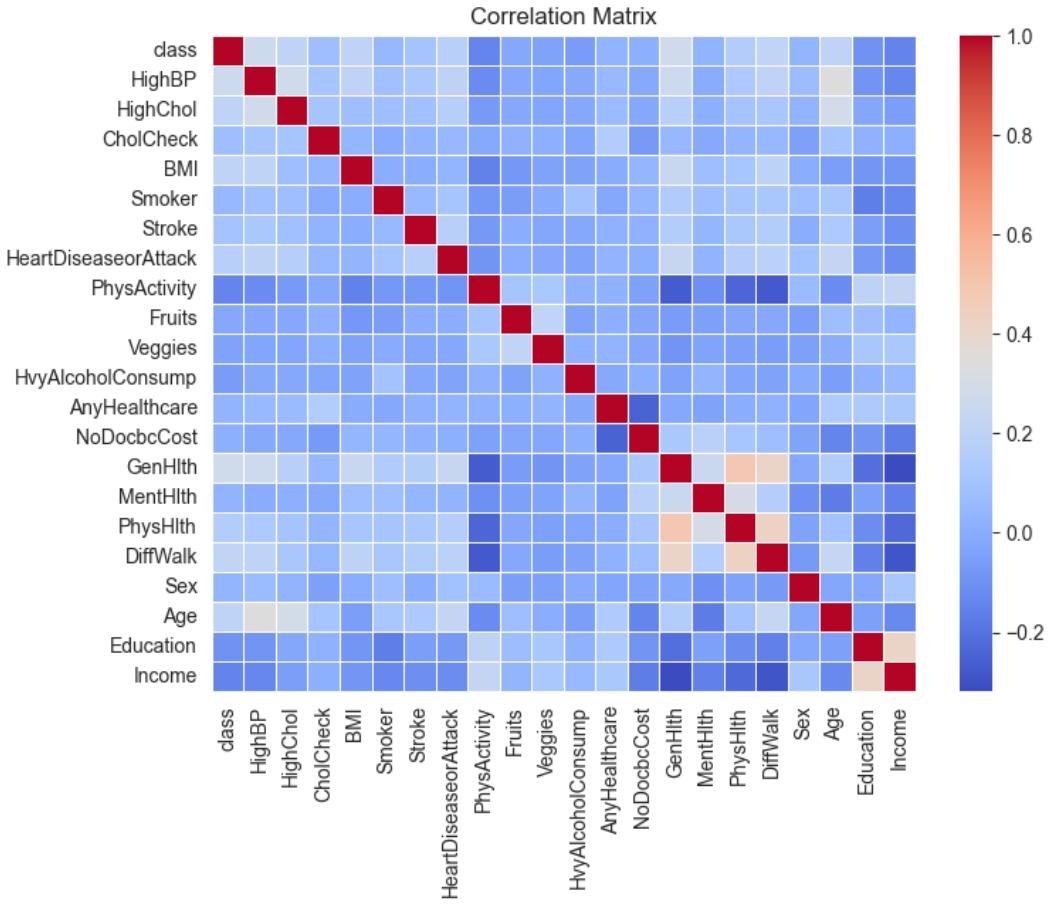


Рис. 3.12. Матриця кореляції даних захворювань на діабет.

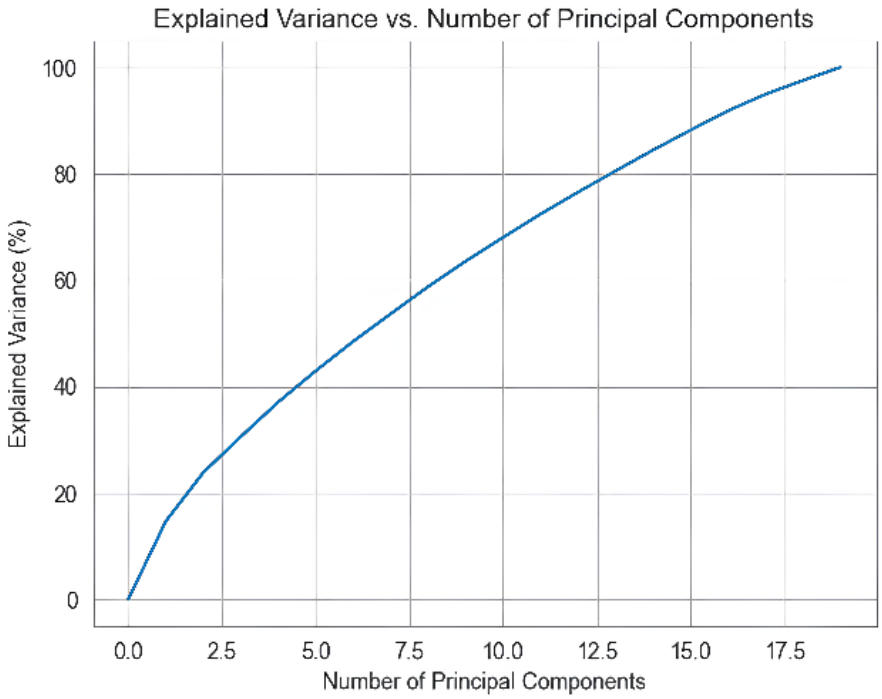


Рис. 3.13. Залежність варіативності від кількості принципових компонент, отримана за допомогою аналізу PCA.

3.2.6. Дані для розширення можливого застосування підсистеми стратифікації

Дані клієнтів оптового дистриб'ютора

Для розширення можливого варіанту застосування запропонованого методу мультиагентної нечіткої кластеризації було розглянуто набір даних клієнтів оптового дистриб'ютора [104]. Дані представленого набору даних включають 440 записів із 8 змінними стану. Дані складаються з річних витрат в умовних грошових одиницях на різні категорії продуктів. Наступні змінні стану входять до розглянутого набору даних:

- *Fresh* – річні витрати на свіжі продовольчі товари;
- *Milk* – на молочну продукцію;
- *Grocery* – на інші продовольчі товари;
- *Frozen* – на заморожені продовольчі товари;
- *Detergents and Paper* – на миючі й паперові товари;
- *Delicatessen* – на делікатесні продовольчі товари;

- *Channel* – тип поширення товару (мережева чи роздрібна торгівля);
- *Region* – регіон клієнта, що приймає один з трьох можливих значень, використаний в якості цільової змінної.

Методи PCA та t-SNE показали чітку можливість розділення даних на декілька кластерів, при стисненні даних до двох компонент (Рис. 3.14). Це є індикатором можливості розділення даних.

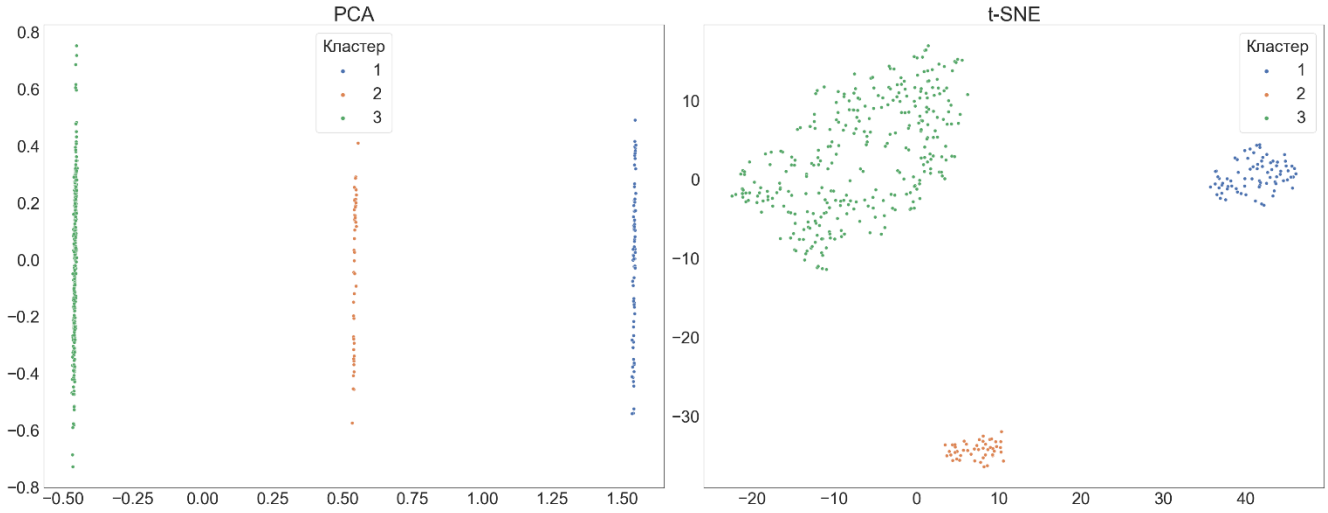


Рис. 3.14. Стиснення набору даних клієнтів оптового дистриб'ютора за допомогою методів PCA та t-SNE.

Дані економічного розвитку країн

Розроблені методи та моделі стратифікації елементів комп'ютерних систем медичного моніторингу було використано в якості методології аналізу даних економічного розвитку країн для визначення стану цифрового розвитку країн світу. Для ідентифікації рівня цифрового розвитку було перевірено гіпотезу про існування схожих країн, що можна об'єднати в групи відповідно до спеціалізованих індексів. Для цього були відібрані індекси, що показують рівень цифрового, соціального та економічного розвитку країни, а саме [60]:

- *EGlit* – глобальний індекс розвитку електронного урядування;
- *NRlit* – індекс мережевої готовності;
- *ICTit* – індекс розвитку інформаційно-комунікаційних технологій;

- *SPI* – індекс соціального прогресу;
- *Class* – обраний цільовий клас, що визначає тип економічного розвитку країни й включає наступні типи країн: «High income» – 45 записів; «Upper middle income» – 11 записів; «Lower middle income» – 25 записів та «Lower income» – 34 записи.

В загальному дані включають 115 запис спостережень економічних індексів країн по 32 змінних стану соціального, економічного та цифрового розвитку країн та 33 цільову змінну, що визначає тип економічного розвитку країни.

3.3. Метод верифікації програмного забезпечення стратифікації елементів

Для перевірки якості програмної реалізації та роботи підсистеми стратифікації елементів комп'ютерних систем медичного моніторингу необхідно розробити методологію верифікації, що має на меті перевірку якості функціонування розроблених методів та моделей з одного боку і підтвердити точність програмної реалізації з іншого боку. Підсистема стратифікації складається з трьох частин, відповідно до задач стратифікації:

1. Модель кластеризації на основі мультиагентного методу нечіткої кластеризації.
2. Модель класифікації на основі ШНМ.
3. Метод визначення загальної інформативності та метод визначення поточної інформативності змінних стану комп'ютерної системи медичного моніторингу.

Відповідно до розглянутих частин моделі стратифікації можна виділити наступні загальні завдання методології верифікації:

1. Визначення точності програмної реалізації методу кластеризації та коректності навчання моделі кластеризації на результатах отриманих з моделі кластеризації.
2. Визначення точності програмної реалізації методу навчання та методу підбору гіперпараметрів ШНМ, також точності навчання моделі ШНМ на результатах розмічених даних отриманих від моделі кластеризації.

3. Визначення точності програмної реалізації методу визначення загальної інформативності та методу визначення поточної інформативності.

Далі відповідно до поставлених задач розглянемо етапи верифікації системи стратифікації:

1. ***Збір даних.*** На цьому етапі необхідно зібрати дані, які будуть використовуватися для навчання та перевірки підсистеми стратифікації. Дані повинні бути репрезентативними для реальної ситуації, що буде моделювати комп'ютерну систему медичного моніторингу. Важливим зауваженням, тут є використання розмічених даних задля можливості проведення більш точного тестування.

2. ***Обробка даних.*** Це необов'язковий етап і його проведення залежить від типу даних, що будуть використовуватися при перевірці роботи розроблених методів та моделей. Зазвичай обробка даних вимагається у випадку даних різної природи, як то порядкові й числові дані, а також дані з різними розподілами мають бути приведені до компактних розподілів, що як показала практика приведе до швидшого навчання моделей кластеризації чи ШНМ.

3. ***Розділення даних.*** Дані необхідно розділити на навчальний, тестовий та перевірочний набори. Навчальний набір буде використовуватися для навчання моделі кластеризації та класифікації, перевірочний набір – для перевірки впливу окремих параметрів на точність моделей під час навчання, а тестовий набір – для остаточної перевірки точності функціонування системи після навчання. Зазвичай задля збільшення точності роботи моделей при остаточному навчанні задля впровадження моделей в практичне використання усі ці набори об'єднуються, аби моделі могли генералізувати найбільше доступних даних.

4. ***Навчання моделі кластеризації*** проводиться з використанням тестового та перевірочного набору даних, перевірочний набір може використовуватися в якості проміжного набору тестування для налаштування гіперпараметрів моделі кластеризації (вибір метрик міжелементної відстані, початкової кількості кластерів, тощо).

5. ***Перевірка результатів кластеризації***, на цьому етапі можуть використовуватися данні про кількість класів у розмічених даних задля перевірки якості кластеризації, а також застосування моделі класифікації на основі моделі класифікації для визначення точності навчання. Перевірка відбувається із застосуванням тестових даних, якщо вимагається тільки перевірити оригінальну модель чи перевірочні дані, якщо перевіряється модель стратифікації загалом.

6. ***Навчання моделі ШНМ*** для класифікації. На цьому етапі використовується навчальний та перевірочний набори даних задля досягнення найкращих можливих результатів навчання.

7. ***Перевірка точності моделі класифікації.*** На цьому етапі модель ШНМ перевіряється на тестовому наборі даних за допомогою типових методів та метрик перевірки якості класифікації (матриці конфузій, точність, ROC-крива, тощо).

8. ***Перевірка загального рівня інформативності вхідних даних.*** На цьому етапі перевіряється загальний рівень інформативності вхідних даних отриманих на моделі класифікації. Гарним способом такої перевірки є зменшення розмірності вхідних даних, відповідно до отриманих результаті навчання та подальша спроба навчити модель, успішне навчання буде ідентифікуватися по точності на перевірочній та навчальних вибірках. Також можливе визначення кількості інформативних змінних за допомогою методу Principal Component Analysis (PCA) та порівняння цієї кількості з отриманою, значні відмінності свідчать про неточність методу.

9. ***Перевірка поточного рівня інформативності.*** Проводиться на декількох випадково обраних записах у тестовому наборі даних. Поява одних й тих самих змінних в якості найбільш інформативних як в загальній так і в поточній свідчить про достовірність роботи запропонованого методу.

Далі детальніше роздивимось методи, що будуть використовуватися для перевірки результатів кластеризації:

- Використання алгоритмів для визначення оптимальної кількості кластерів (наприклад, метод ліктя [56]).

- Оцінку якості кластеризації можна проводити за допомогою простої візуалізації кластерів у випадку малої розмірності вхідних даних та за допомогою метрик, що визначають щільність отриманих кластерів [91] (наприклад методи Silhouette Score, Davies–Bouldin Index).
- Також оцінку якості кластеризації можна проводити з використанням розробленої на основі методу кластеризації моделі класифікації, в такому випадку показники якості кластеризації будуть такі ж самі як для тестування моделей класифікації.

Також роздивимось детальніше методи, що будуть використовуватися для перевірки результатів роботи моделей класифікації [92]:

- Оцінка точності моделей проводиться з використанням метрик: точності у випадку збалансованих по класам наборів даних, також precision, recall та F1-score – що зазвичай використовують для оцінки результатів класифікації на незбалансованих даних.
- Confusion matrix – є інструментом в оцінці точності моделі класифікації, який показує кількість правильних та неправильних класифікацій для кожного класу. Вона складається з чотирьох значень: True Positive, True Negative, False Positive, та False Negative, і є основою для розрахунку розглянутих вище метрик.
- ROC-крива, яка дозволяє порівнювати точність моделі класифікації при різних порогах класифікації, та AUC, яка є площею під ROC-кривою, що є типовою візуалізацією якості кластеризації.
- Також слід не забувати про розбиття вхідних даних на навчальний і тестовий набори даних, це дозволяє виявити спроможність моделі до генералізації в такому випадку використовуються метрики overfitting та underfitting. Overfitting в машинному навчанні відбувається, коли модель надто добре пристосовується до тренувальних даних і втрачає здатність узагальнювати (мати таку ж продуктивність і на даних, що не використовувалися для навчання) нові дані [93]. Underfitting виникає, коли модель недостатньо складна для відображення структури тренувальних даних, що призводить до поганої пристосованості моделі [93].

Overfitting виявляється, коли точність на тренувальних даних вища, ніж на перевірочних або тестових. Underfitting проявляється, коли точність на тренувальних, перевірочних та тестових даних може бути низькою. Модель не спроможна апроксимувати функцію, що пов’язує вхідні та вихідні дані.

А для перевірки якості визначення інформативності можна використовувати наступні методи:

- Використовувати методи інформованого навчання для визначення інформативності змінних в цих моделях. Наприклад, при вдалому навчанні на розмічених даних моделі Random Forest можна отримати нормалізовані ваги прийняття рішень, які можна інтерпретувати з рівнем інформативності змінних, та порівняти з іншою їх оцінкою. А використання більшого числа моделей дозволить більш комплексно поглянути на проблему. Також взаємоперевірка двох запропонованих методів визначення інформативності також можлива [52].
- Також можна порівняти отримані коефіцієнти інформативності з методом декомпозиції PCA, який продукує значення варіативності, що допоможе оцінити кількість найбільш інформативних змінних до певного рівня, наприклад 80% варіативності PCA та інформативності. Або проводити відбір найбільш інформативних змінних й оцінювати їх варіативність PCA, якщо варіативність зростатиме, це буде свідчити про точність визначення інформативності [52, 90].

Отже, запропонована процедура верифікації програмної реалізації методів і моделей стратифікації включає наступні етапи [52]:

1) Перевірка точності реалізації методу кластерного аналізу шляхом використання методу класифікації на основі моделі кластеризації та стандартних перевірочних наборів даних. Подальше порівняння із існуючими програмними реалізаціями методів кластеризації дозволить оцінити не тільки точність програмної реалізації, а і точність розробленого методу. Метрики accuracy та confusion matrix.

2) Перевірка точності реалізації модуля класифікатору із моделлю ШНМ, а саме методів навчання та підбору гіперпараметрів моделі ШНМ. Метрики для навчального і тестових наборів accuracy та confusion matrix.

3) Перевірка точності роботи методу кластерного аналізу шляхом використання моделі ШНМ навченого на оригінальному розміченні даних в якості універсального апроксиматору дозволить оцінити спроможність розробленого методу кластеризації до точного розділення даних. Метрика: різниця показників accuracy в розглянутих методів на навчальних і тестових наборах даних.

4) Перевірка методу визначення загальної інформативності, шляхом порівняльного аналізу з іншими методами ідентифікація тих самих змінних стану як самих інформативних іншими методами свідчить про точність реалізації на функціонування. Метрики: кількість визначених інформативних змінних, та відношення варіативності PCA та інформативності на кількість змінних.

5) Перевірка методу визначення поточної інформативності, шляхом порівняння визначень найбільш інформативних змінних іншими методами. Метрика: кількість визначених інформативних змінних.

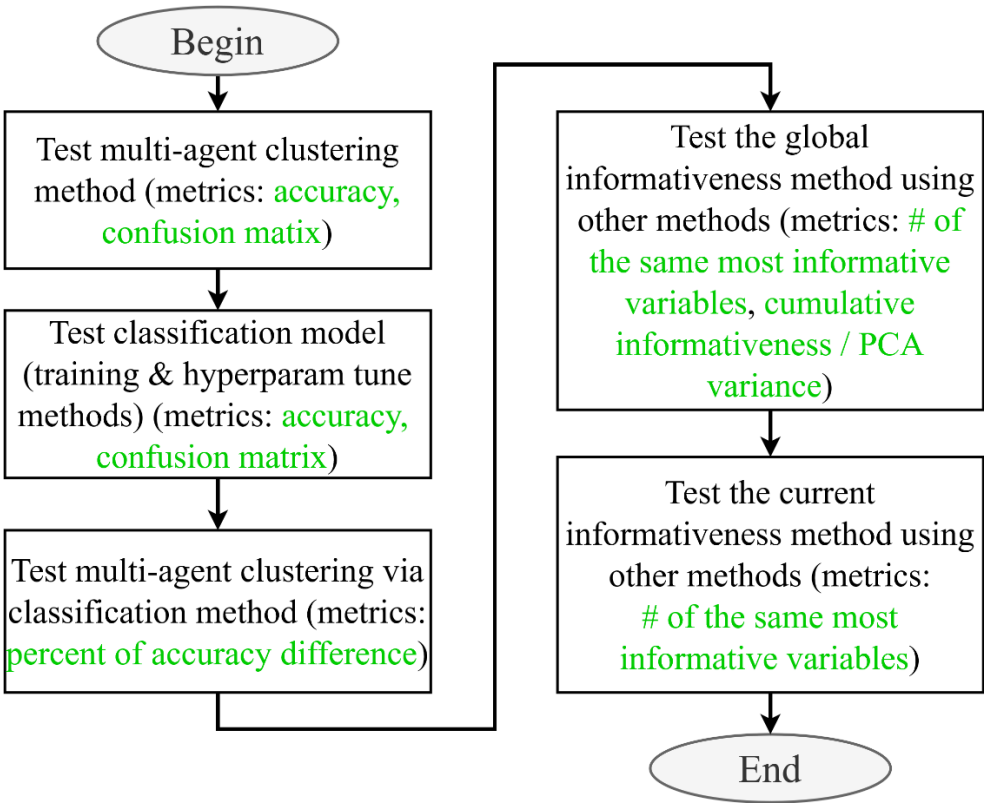


Рис. 3.15. Основні етапи методу верифікації програмної реалізації методу стратифікації елементів.

Аналіз результатів кожного етапу дозволить не тільки верифікувати програмну реалізацію, а й перевірити спроможність методів до роботи в цілому для вирішення задачі стратифікації.

Запропонована метод дозволяє перевірити якість функціонування підсистеми стратифікації елементів в комп'ютерній системі медичного моніторингу. Метод включає в себе перевірку точності методу кластеризації та класифікації, та перевірку визначення інформативності вхідних даних.

Висновок до розділу 3

В розділі 3 дисертації було детально розглянуто та описано вибір програмного забезпечення для реалізації розроблених методів та моделей стратифікації елементів комп'ютерних систем медичного моніторингу. Для цього було розглянуте програмне забезпечення для реалізації методів і моделей з точки зору трьох компонент: мови програмування в якості платформи для реалізації запропонованих методів і моделей; доступних бібліотек обробки даних, швидких математичних обчислень та методів машинного навчання; та не менш важливої компоненти інтегрованої середи розробки, що відповідно до доступного функціоналу може сильно впливати на якість реалізації запропонованих методів і моделей. Було обґрунтовано вибір мови програмування Python; обраних бібліотек NumPy, Pandas, Matplotlib, Seaborn, Tensorflow, SciKit Learn та інших; та обраної інтегрованої середи розробки PyCharm та Jupyter Notebook із засобами перевірки якості коду на базі методів штучного інтелекту.

Далі було розглянуто набори даних для перевірки точності функціонування розроблених методів і моделей. Для цього виділено набори даних для тестування базового функціоналу запропонованих методів та моделей; набори даних для перевірки кожного метода і моделі окремо і набір даних для тестування загальної точності функціонування підсистеми стратифікації. А також розглянуті набори даних для розширення функціоналу запропонованих методів і моделей до сфери застосування на економічних даних.

Завершено розділ описом методу верифікації програмної реалізації методів і моделей статифікації, зокрема використанням стандартних підходів для перевірки

точності їх застосування та стабільності на навчальних та тестових даних. Розроблений метод верифікації включає п’ять етапів верифікації, на першому етапі перевіряється реалізація мультиагентного методу нечіткої кластеризації, де метрикою є точність кластеризації та матриці конфузів, що показують точність розділення. На другому етапі перевіряється реалізація методу навчання й підбору гіперпараметрів ШНМ, де метриками також є точність класифікації та матриці конфузів. Третій етап перевіряє теоретично можливу спроможність методу кластеризації до розділення даних через модель ШНМ, де метрикою є різниця точності виявлення цільових класів. На четвертому етапі перевіряється метод загальної інформативності де метрикою є відношення сумарної інформативності й варіативності PCA, а також кількість найбільш інформативних змінних в порівнянні з іншими методами. На п’ятому етапі перевіряється метод визначення поточної інформативності, де метрикою є кількість найбільш інформативних змінних в порівнянні з методом загальної інформативності.

Основні положення цього розділу викладені у публікаціях автора [1–7, 9].

РОЗДІЛ 4.
ВИКОРИСТАННЯ РОЗРОБЛЕНИХ МЕТОДІВ ТА МОДЕЛЕЙ СТРАТИФІКАЦІЇ

4.1. Використання моделі нечіткої кластеризації на даних медичного моніторингу

Порівняння запропонованих та існуючих функцій витрат

Для порівняння різних функцій витрат, для визначення щільності кластеризації (чи переваги одного кластеру над іншим) було розглянуто ряд метрик серед яких розглянуто стандартні та запропоновані модифікації. До порівняння було віднесено наступні метрики [105]:

1. ***Partition coefficient (PC)*** –це функція витрат, що є середньою сумою приналежностей по кожному кластеру та визначає об'єм перекриття між кластерами [106, 107]. Менше значення свідчить про кращу кластеризацію.
2. ***Partition entropy coefficient (PEC)*** – це функція витрат, що подібна до PC, але із введенням логарифму приналежності [108]. Менше значення свідчить про кращу кластеризацію.
3. ***Fukuyama-Sugeno Index (FSI)*** – це функція витрат, що враховує компактність та наповненість кластерів даними (актуальну відстань між елементами в даних) [109]. Менше значення свідчить про кращу кластеризацію.
4. ***Partition coefficient and exponential separation (PCES)*** – функція витрат, що поєднує нормалізоване значення PC із експонентною щільності даних [110]. Більше значення свідчить про кращу кластеризацію.
5. ***Mahalanobis distance (Mah)*** – одна з функцій витрат, що лежить в основі мультиагентної кластеризації, що використовує метрику Махаланобіса в якості оцінки результатів кластеризації, вираз (2.17) та (2.18) із метрикою II у виразі (2.16). Менше значення свідчить про кращу кластеризацію.
6. ***Mahalanobis distance and membership values (Mah_Inv)*** – одна з функцій витрат, що лежить в основі мультиагентної кластеризації, що використовує метрику Махаланобіса із врахуванням оберненого значення

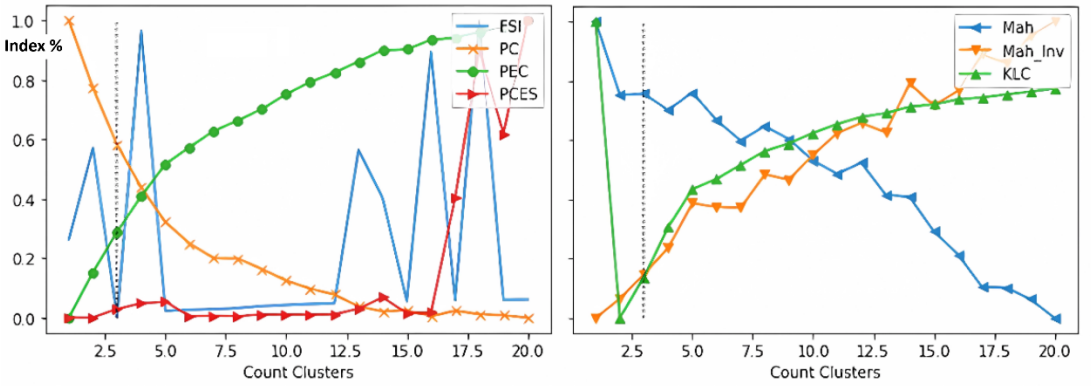
приналежності в якості оцінки результатів кластеризації, вираз (2.17) та (2.18) із метрикою III у виразі (2.16). Менше значення свідчить про кращу кластеризацію.

7. ***Kullback-Leibler entropy (KLC)*** – одна з функцій витрат, що лежить в основі мультиагентної кластеризації, що використовує Кульбака-Лейблера в якості оцінки результатів кластеризації, вираз (2.17) та (2.18) із метрикою IV у виразі (2.16). Менше значення свідчить про кращу кластеризацію.

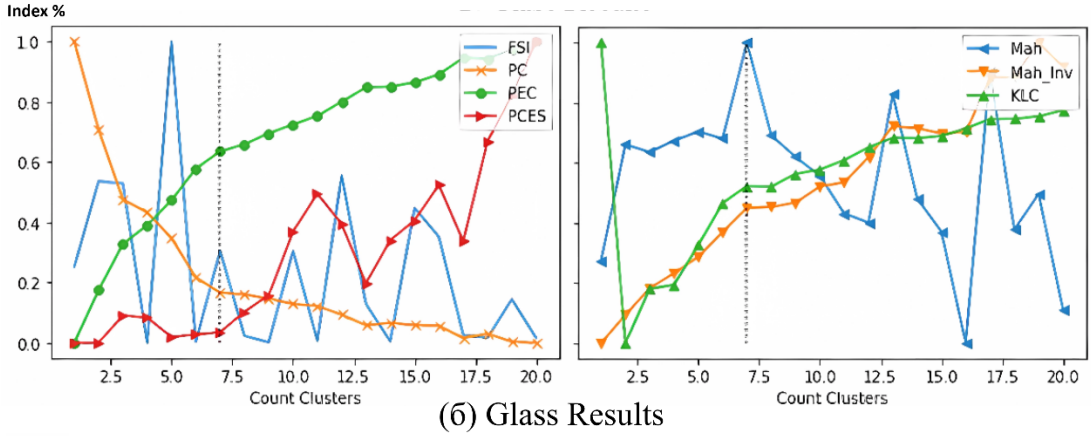
Для тестування запропонованих функцій втрат було розглянуто декілька наборів даних: ірисів Фішера, ідентифікації скла UCI, типів вина UCI та діагностики раку молочної залози Вісконсину. Усі набори даних були розглянуті в розділі 3.2.1, показано, що набори мають різну природу даних та різну складність із простою сферичною формою та складною формою, із простим просторовим розділенням кластерів та великою кількістю шумів при розділенні.

Результати оцінок значень функцій витрат для кластеризації мультиагентним методом наведені на Рис. 4.1 [105], де вказана цільова кількість кластерів вертикальною лінією. Для оцінки саме функції витрат в методі мультиагентної кластеризації було виставлено різні варіанти функції витрат при єдиній метриці для визначення міжелементної відстані. Це дало можливість порівнювати різні кластери, сформовані в обраних даних під впливом тільки функції витрат.

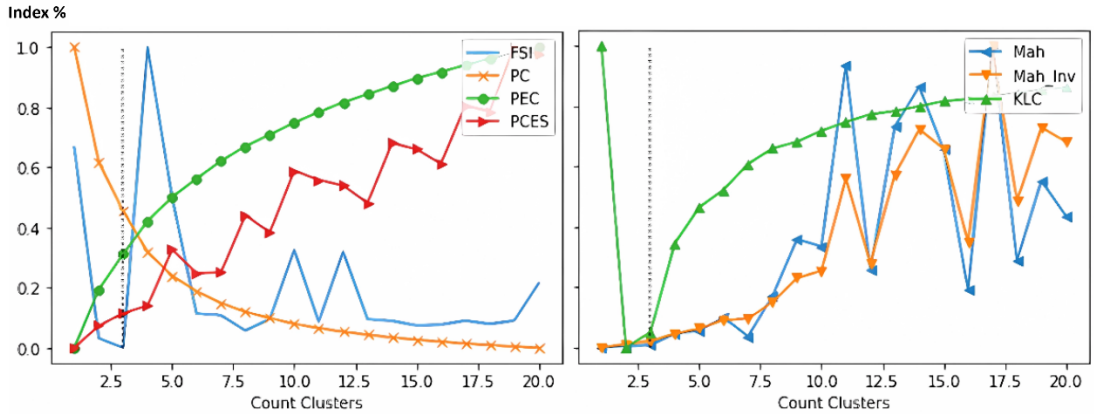
За результатами експериментів показано, що жодна з існуючих метрик не вказала чітко на необхідну кількість кластерів [105]. Функція витрат FSI дала визначити оптимальну кількість кластерів для набору даних ірисів Фішера та типів вина UCI (Рис 4.1). Але функція FSI не показала плавне зменшення до оптимальної кількості. Інші функції витрат або вказують на зменшення або збільшення своїх значень поруч з необхідною кількістю кластерів.



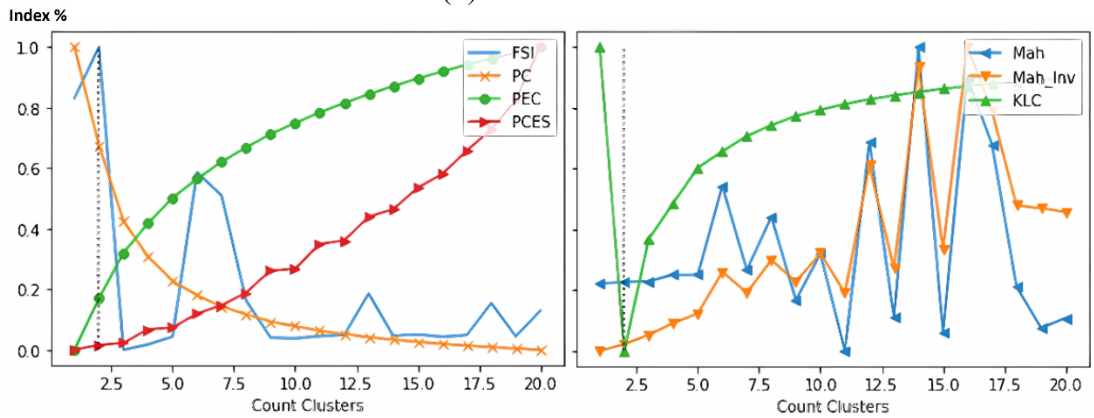
(а) Fisher Irises Results



(б) Glass Results



(в) Wine Results



(г) Breast Cancer Results

Рис. 4.1. Оцінка функцій витрат результуючих кластерів при застосуванні мультиагентного методу нечіткої кластеризації до стандартних наборів даних.

Також результати на Рис 4.1 свідчать, що використання запропонованих функцій витрат Mah_Inv і KLC дало плавне сходження до мінімумів поруч з необхідною кількістю кластерів. Функція витрат Mah мала багато коливань і не могла точно визначити кількість кластерів. Функція витрат PCES, що має досягати максимуму при оптимальній кількості кластерів, досягала максимуму лише на ірисах Фішера, що є найпростішим набором із запропонованих. На інших наборах функція PCES багато коливалась або просто зростала із зростанням кількості кластерів. Отже, в більшості випадків запропоновані функції витрат мають перевагу в плавності зменшення значення при наближенні до цільової кількості кластерів. З цієї причини було вирішено не відокремлювати метрику для визначення міжелементної відстані від цільової кількості кластерів.

Вибір найкращої метрики для подальшого тестування

Для початкового тестування точності роботи запропонованої моделі мультиагентної нечіткої кластеризації даних був використаний тестовий набір даних ірисів Фішера. Цей набір даних детальніше був розглянутий в розділі 3.2.2, як набір для базового тестування функціонування й перевірки концепції запропонованих методів чи моделей. Результати аналізу набору даних показали, що цей набір пропонує чітке виділення точок даних відповідальних за один з трьох класів у виді багатомірної сфери та поєднання двох інших класів у виді другої еліпса, із можливістю розділення на 2 класи.

Оскільки це набір даних для перевірки концепції, для перевірки не використовувалося розділення на тестовий та навчальний набори даних. Набір даних використовувався для перевірки точності роботи запропонованої моделі при різних метриках. Метрика, що отримувала найбільшу точність визначення стану використовувалася для подальшого розгляду на даних медичного моніторингу. Результати точності отримані в результаті використання моделі класифікації на основі отриманих моделей кластеризації показані в табл. 4.1 [59]. Метрики, що вказані в табл. 4.1 зазначені у виразі (2.16).

Таблиця 4.1.

Найкращі результати кластеризації ірисів Фішера із запропонованими метриками міжелементної відстані.

Метрика	Махаланобіса	Махаланобіса + обернена приналежність	Кульбака-Лейблера
Точність	80%	91.3%	98%

Як і зазначалось при аналізі набору даних ірисів Фішера, один клас був повністю віддільний, що і показано в матриці конфузів (табл. 4.2) [59], інші два були в одному еліпсі, роздільні просте із невірним визначенням деяких записів. Застосування метрики на основі ентропії Кульбака-Лейблера надало можливість розділити дані із мінімальною кількістю помилок.

Таблиця 4.2.

Матриця конфузів для результатів кластеризації ірисів Фішера з ентропією Кульбака-Лейблера.

		Результат кластеризації		
		Iris setosa	Iris virginica	Iris versicolor
Актуальний клас	Iris setosa	50	0	0
	Iris virginica	0	48	2
	Iris versicolor	0	1	49

Відповідно до специфіки запропонованого методу мультиагентної кластеризації необхідне навчання, процес навчання моделі з ентропією Кульбака-Лейблера на даних ірисів Фішера показаний на Рис. 4.2. Відповідно до процесу навчання, слід зазначити спроможність методу чітко визначати оптимальну кількість класів. Важливо зазначити, що на Рис. 4.2 показаний процес навчання в режимі пошуку оптимальної кількості кластерів [59]. Остаточний процес навчання проходив з обмеженнями по цільовій кількості кластерів, де алгоритм методу мультиагентної нечіткої кластеризації зупинився на 3-х кластерах.

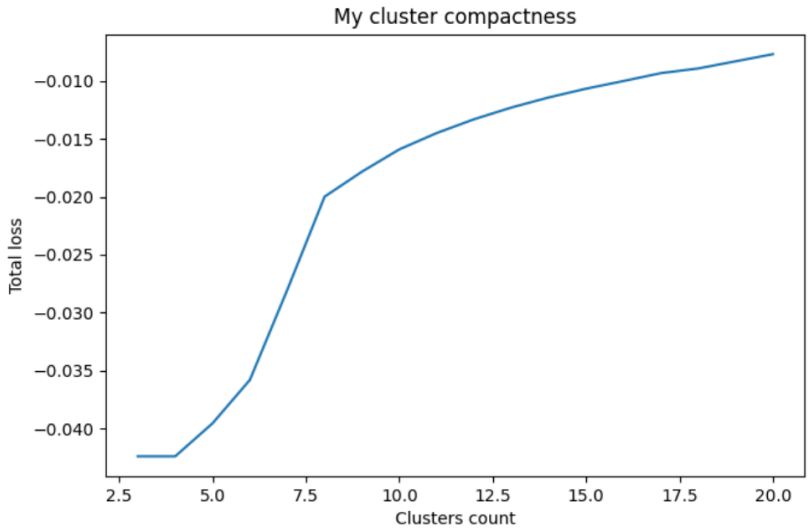


Рис. 4.2. Значення функції витрат під час навчання моделі мультиагентної нечіткої кластеризації з ентропією Кульбака-Лейблера на даних ірисів Фішера.

Тестування запропонованої моделі на даних медичного моніторингу

Для подальшого тестування точності роботи запропонованого методу мультиагентної нечіткої кластеризації даних був використаний набір даних про медичний моніторинг пацієнтів з захворюваннями передміхурової залози. Набір даних був детальніше розглянутий в розділі 3.2.2, результати аналізу даних показали, про можливість виокремлення точок даних здорових пацієнтів серед усіх наявних даних, проте показали складності у виокремленні точок даних по різним стадіям хвороби. В огляді показано, що дані набору можна розділити по стану здоровий/нездоровий, але є складності з виділення стадії хвороби. Чотири можливі класи прогресування хвороби були обрані, як цільові класи, й використані в якості перевірки точності кластеризації.

Під час навчання моделі мультиагентної нечіткої кластеризації була використана міра обрана міра Кульбака-Лейблера. Прогрес навчання показаний на Рис. 4.3, як можна помітити мінімум досягається як раз під час доходження до 4–5 кластерів, що підтверджує ефективність запропонованої функції витрат [59].

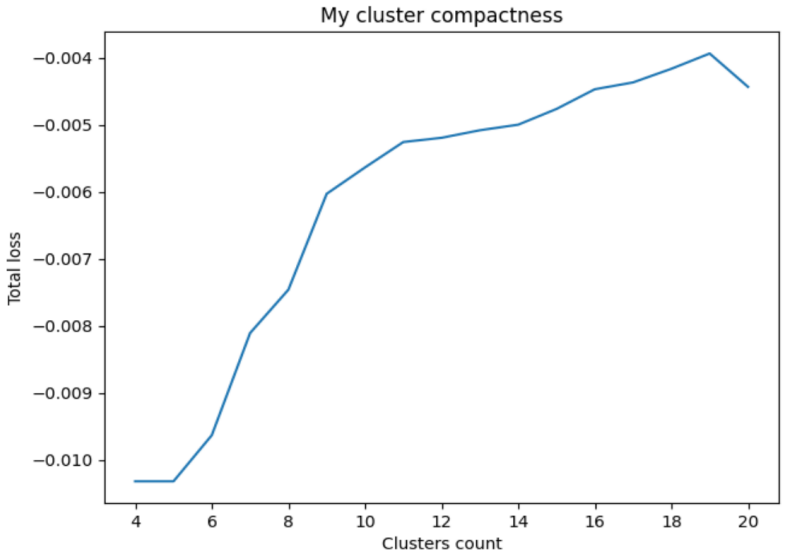


Рис. 4.3. Значення функції витрат під час навчання моделі мультиагентної нечіткої кластеризації з метрикою Кульбака-Лейблера на даних медичного моніторингу пацієнтів з захворюваннями передміхурової залози.

Таблиця 4.3.

Матриця конфузів для результатів кластеризації даних медичного моніторингу пацієнтів з захворюваннями передміхурової залози.

		Результат кластеризації			
		Здоровий	Без метастаз	З метастазами	Гормоно-резистентий
Актуальний клас	Здоровий	50	0	0	0
	Без метастаз	1	42	2	0
	З метастазами	1	0	50	1
	Гормоно-резистентий	0	1	2	30

Як показано в табл. 4.3 [59] в загальному для кожного класу хворих пацієнтів виділено 1–3 записи, що були неправильно класифіковані, але варто зазначити, що записи здорових пацієнтів були відділені повністю правильно, проте 2 записи були помилково ідентифіковані як здорові. Здатність відділяти дані здорових від не здорових є важливим аспектом систем допомоги прийняття рішень, якої задовольняє запропонований метод із рівнем помилки 0.15% визначення хворих пацієнтів. Загальна точність отримана для запропонованого методу нечіткої кластеризації складає 95.6%.

4.2. Використання моделі класифікації

Використання моделі ШНМ на лабораторних даних урологічних досліджень

Для подальшого тестування точності роботи запропонованого методу навчання на моделі ШНМ були використані лабораторні дані урологічних досліджень. Дані були розглянуті у розділі 3.2. та показано, що наявний відносно невеликий набір даних із істотною кількістю змінних стану, але не зазначено складність самого набору даних. Для навчання набір даних було розбито на навчальний та тестові набори по 30 та 10 записів відповідно [5].

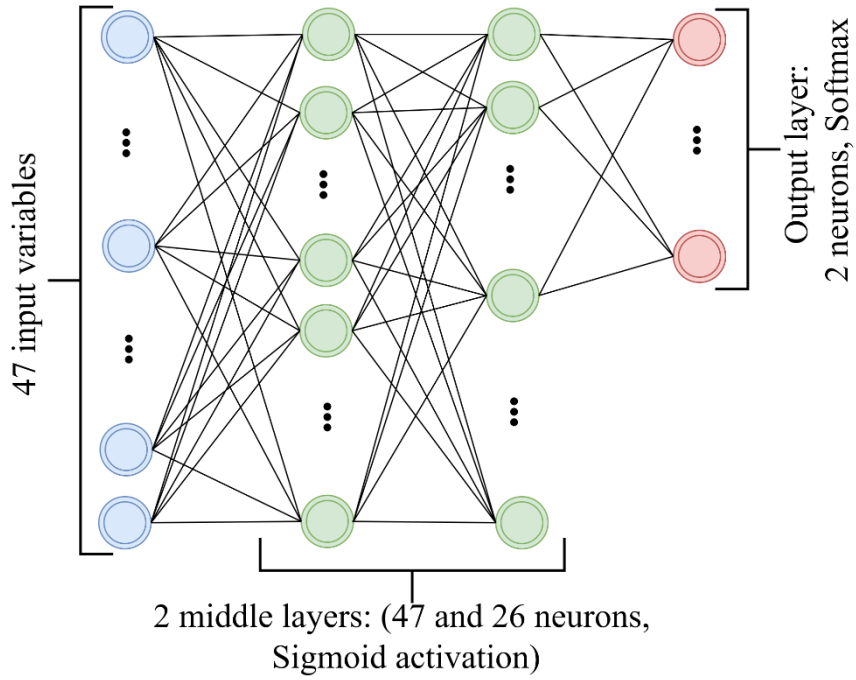


Рис. 4.4. Отримана архітектура моделі ШНМ, для класифікації даних урологічних досліджень.

Метод навчання сформував повнозв'язну нейронну мережу, що включає 3 шари: перший шар включає 47 нейронів; другий – 26 нейронів та третій – 2 нейрони. Отримана архітектура показана на Рис. 4.4. Значення функції втрат для навченої моделі ШНМ на навчальних даних складало 0.122 та стандартне відхилення 0.0024, точність розпізнавання для тестового та навчального наборів склала 100% [5].

Таблиця 4.4.

Матриця конфузів для результатів класифікації даних урологічних досліджень навченою моделлю ШНМ

		Результат класифікації	
		Здоровий	Хворий
Актуальний клас	Здоровий	11	0
	Хворий	0	29

Результати тестування моделі ШНМ наведені в табл. 4.4 [5] для всіх наявних даних, показано, що модель добре розрізняє здорових та хворих пацієнтів. Незважаючи на те, що перевірка моделі проводилась згідно методу верифікації, важко судити про здатність моделі ШНМ та методів її конфігурації узагальнювати дані, а не завчати їх, через малий розмір наявного набору даних.

Використанні моделі ШНМ на даних захворювання печінки

Також модель та метод навчання ШНМ були протестовані на захворювання печінки. Як зазначено в розділі 3.2.3 цей набір має значно більшу кількість записів проте меншу кількість змінних стану. Для навчання набір даних було розбито на навчальний та тестові набори по 550 та 40 записів відповідно [5].

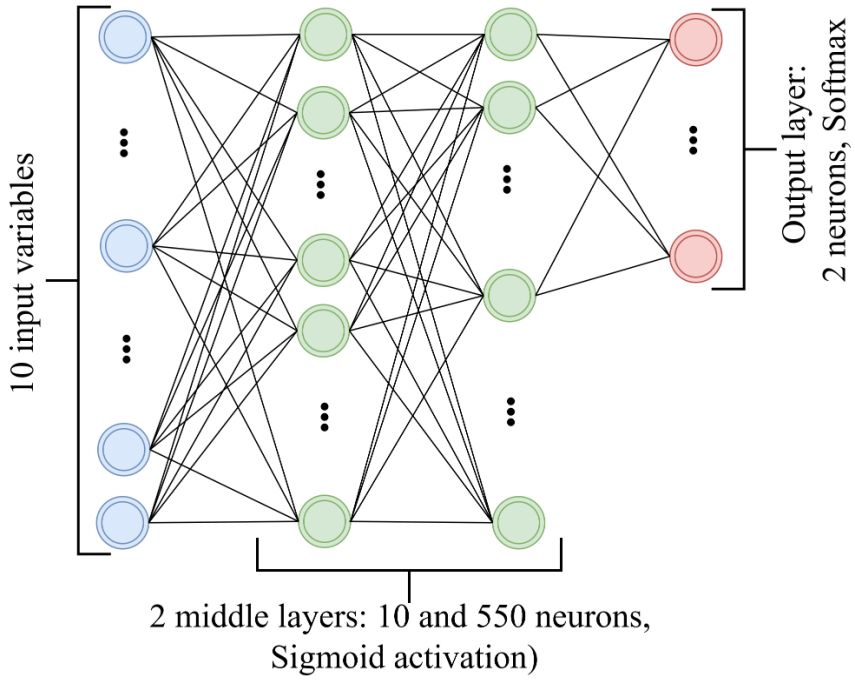


Рис. 4.5. Отримана архітектура моделі ШНМ для класифікації даних захворювань печінки.

Метод навчання сформував повнозв’язну нейронну мережу, що включає 3 шари: перший шар включає 10 нейронів; другий – 550 нейронів та третій – 2 нейрони. Отримана архітектура показана на Рис. 4.5. Значення функції втрат для навченої ШНМ на навчальному наборі складало 0.0185 та стандартне відхилення 0.0107, точність розпізнавання навчальної вибірки 100% та тестової 80% [5].

Таблиця 4.5.

Матриця конфузів для результатів класифікації тестових даних захворювань печінки навченою моделлю ШНМ

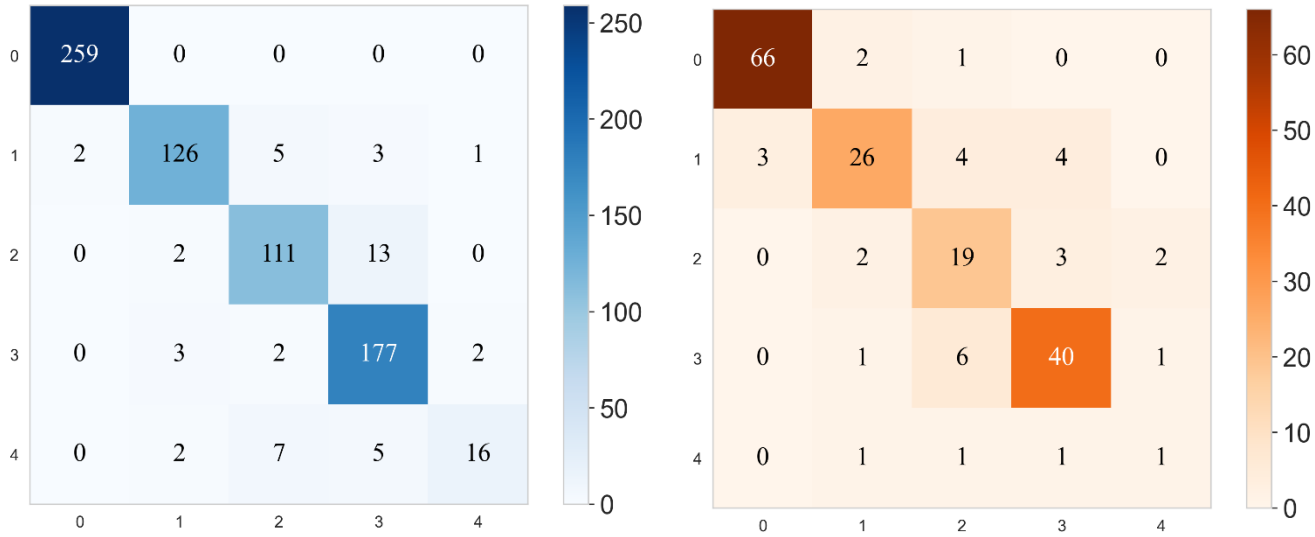
		Результат класифікації		Precision
		Здоровий	Хворий	
Актуальний клас	Здоровий	15	3	83.33%
	хворий	4	14	77.77%
Recall		78.94%	82.35%	

Результати тестування моделі ШНМ наведені в табл. 4.5 [5] для всіх наявних даних. Модель ШНМ здатна розрізнити здорових та хворих пацієнтів, проте із проблемами в 7 випадках, де 4 хворі були помилково класифіковані як здорові. Такі помилки можна пояснити складністю набору даних та помилками в змінних стану. Тому результати тестування показують, що розроблену модель та методи навчання та конфігурації ШНМ можна застосовувати в підсистемі стратифікації даних, адже вони показують високу точність на даних медичного моніторингу.

4.3. Використання моделі визначення інформативності параметрів

Головна ідея полягає в тому, що можливо використати деякий набір даних, для того аби перевірити наскільки точно розроблені методи визначення загальної та поточної інформативності. Використовуючи PCA можливо оцінити кількість змінних з найбільшою варіативністю, а отже і впливом на результати передбачень моделі ШНМ. Другим індикатором точності роботи запропонованих методів є виділення змінних, яким притаманна незбалансованість, як найбільш інформативних. А для оцінки точності визначення поточної інформативності можна використати значення отримані із загальної інформативності.

Перевірити спроможність методів до визначення інформативності неможливо без попередньо навченої моделі ШНМ на розмічених даних UCI. Для навчання ШНМ набір даних розбитий на навчальний та тестовий набори у типовому співвідношенні 80% до 20% відповідно. Результати навчання моделі ШНМ на прикладі матриць плутанини для тестового та навчального наборів даних, наведені на Рис. 4.6 [90]. Показано, що модель ШНМ майже повністю відділяє здорових пацієнтів від хворих, проте різні хвороби не може розділити між собою. Використовуючи матриці плутанини можна оцінити точність на тестовому та навчальному наборах 93.61% та 82.6% відповідно.



(а) Навчальний набір (б) Тестовий набір

Рис. 4.6. Матриці плутанини для відображення точності навчання моделі ШНМ на даних серцевих захворювань UCI для: (а) навчального, (б) тестового набору.

Результати роботи методу визначення загальної інформативності змінних наведено в табл. 4.6 [90]. Наведено 10 найбільш інформативних змінних, що визначають 83.13% інформації в змінних, що співставно з результатами визначення варіативності змінних методом PCA (Рис. 3.10). Також метод визначив змінні gender та age як найбільш інформативні, вони також визначались зміщеними при проведенні аналізу даних. Це свідчить, що метод точно визначає загальну

інформативність та висновки з оцінками інформативності є співставними з іншими методами аналізу даних.

Таблиця 4.6.

Результат роботи методу обчислення загальної інформативності, параметри відсортовані за інформативністю.

Параметр	**Навчальний**	**Тестовий**	**Загальний**	**Кумулятивний**
gender	0.0947	0.1183	0.1036	0.1036
age	0.1115	0.0950	0.1053	0.2090
chol	0.0970	0.0972	0.0971	0.3061
ca	0.0807	0.0872	0.0831	0.3892
oldpeak	0.0860	0.0801	0.0838	0.4731
exang	0.0807	0.0772	0.0793	0.5525
fbs	0.0699	0.0778	0.0729	0.6254
thalch	0.0704	0.0726	0.0712	0.6967
restecg	0.0693	0.0651	0.0677	0.7644
thal	0.0645	0.0706	0.0668	0.8313

Таблиця 4.7.

Значення інформативності вхідних параметрів конкретного запису, розрахованих модифікацією методу інтегрованих градієнтів.

Параметр	**gender**	**oldpeak**	**age**	**ca**	**thal**
Інформативність	0.4959	0.1028	0.0735	0.0733	0.0726
	fbs	**trestbps**	**chol**	**thalch**	**cp**
	0.0453	0.0386	0.0314	0.0284	0.016

Результати тестування модифікованого методу інтегрованих градієнтів для визначення поточного значення інформативності змінних наведено в табл. 4.7 [90] для одного з записів в тестовому наборі даних. Наведені змінні в табл. 4.7 мають найбільший вплив на рішення прийняте моделлю ШНМ. Слід зазначити, що змінні oldpeak та ca, найбільш впливові незміщені змінні, що також входять в топ-5 найбільш інформативних відповідно загальної інформативності. Отже, змінні, що виявилися найбільш інформативними в загальній інформативності також найбільш інформативні в поточні інформативності. Це свідчить, що запропонована модифікація методу інтегрованих градієнтів може точно визначати поточну

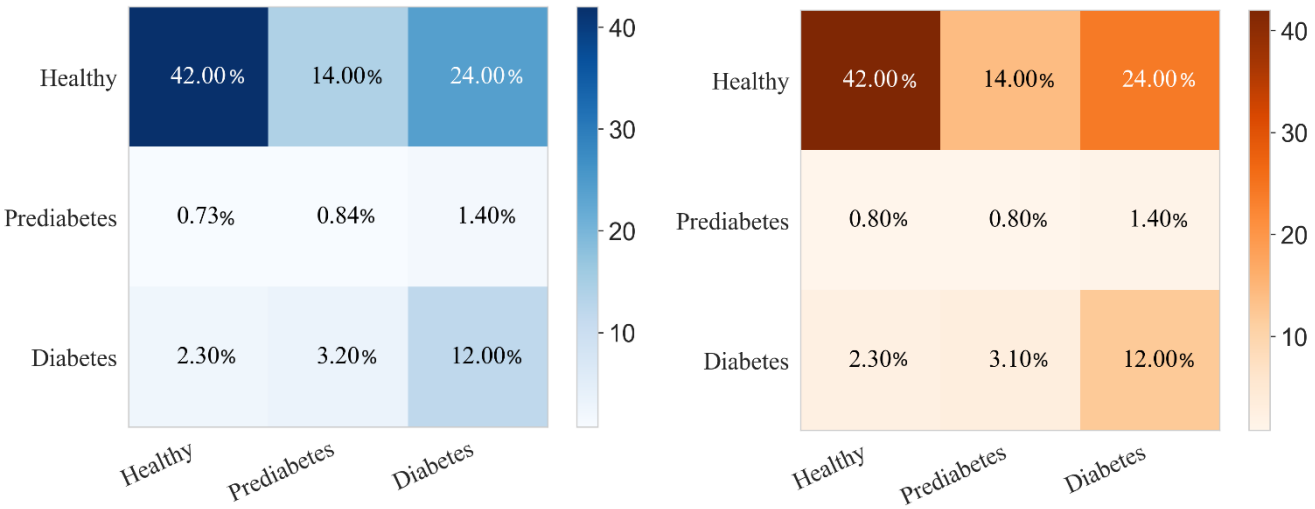
інформативність, а отже може бути використаний для обґрунтування прийняття рішень.

4.4. Аналіз результатів впровадження методів та моделей стратифікації в комп'ютерній системі медичного моніторингу

Для перевірки якості функціонування підсистеми стратифікації були використані дані респондентів з захворюванням на діабет із загального набору медичних досліджень США із CDC BRFSS Survey 2021. Що були детально проаналізовані в розділі 3.2.5. Перевірка методів та моделей стратифікації комп'ютерної системи медичного моніторингу буде проводитися на кожному виділеному модулі. Результати роботи попередніх модулів будуть використані наступними. Для навчання модулів кластеризації й класифікатора на основі ШНМ дані було розбито на тестовий та навчальні набори у співвідношенні 80% та 20%.

Тестування модуля кластерного аналізу

Тестовий набір даних було використано для навчання модулю кластеризації. Відповідно до попередніх результатів тестування методу мультиагентної нечіткої кластеризації було обрано дивергенцію Кульбака-Лейблера в якості міжелементної відстані.



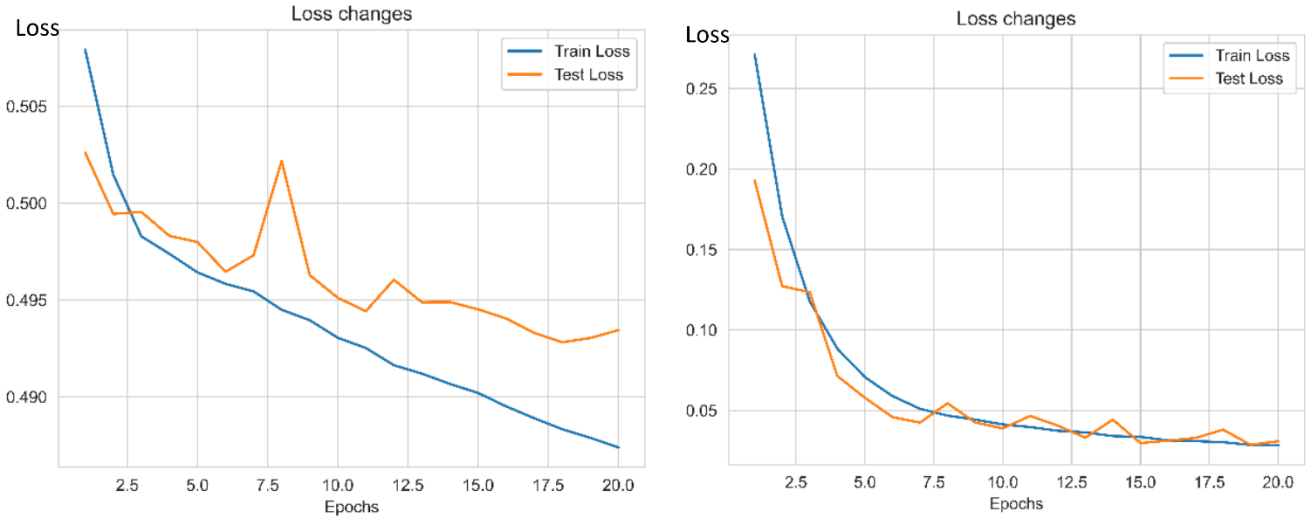
(а) Навчальний набір (б) Тестовий набір

Рис. 4.7. Матриці плутанини із нормованими значеннями, що вказують на точність кластеризації: а – на навчальних; б – на тестових наборах.

Для оцінки якості кластеризації була використана модель класифікації на основі запропонованого методу кластеризації, що дозволить оцінити точність кластеризації. Результатами навчання модуля кластеризації в такій конфігурації є матриці плутанини, що наведені на Рис. 4.7 [52]. Відповідно до результатів матриць плутанини була отримана точність кластеризації для навчального і тестових наборів даних 54.75% та 54.8% відповідно.

Тестування модуля класифікації станів

Після цього розмічені дані з модуля кластерного аналізу були направлені до модуля класифікації станів за допомогою моделі ШНМ. Модуль класифікації також був використаний для визначення теоретичної межі точності на оригінально розмічених даних. Для навчання на розмічених кластеризацію та оригінальних даних був використаний навчальний набір.



(а) Навчальний набір (б) Тестовий набір

Рис. 4.8. Зміна значень втрат при навчанні та тестуванні під час навчання модуля класифікації на даних: а – оригінально позначених; б – позначених за допомогою модуля кластерного аналізу.

Результати навчання моделей ШНМ на двох варіантах розмітки даних наведений на Рис. 4.8 [52]. Відповідно теоретично можлива точність, що може бути досягнута це 80.8% та 80.5% на навчальному і тестовому наборі оригінально

розмічених даних. А точність з використанням даних розмічених модулем кластерного аналізу це 99.8% та 99.7% для навчального і тестового набору відповідно.

Тестування модуля визначення чутливості

Модель ШНМ навчена на даних модуля кластеризації була використана для тестування двох запропонованих методів модуля визначення чутливості (для загальної і поточної інформативності). В табл. 4.8 [52] наведено 10 найбільш інформативних змінних, що згідно методу визначення загальної інформативності несуть 78.87% інформації. Згідно з методом визначення загальної інформативності найбільший вплив на рішення моделі ШНМ мають змінні HvyAlcoholConsump, CholCheck, Stroke, Age та HeartDiseaseorAttack.

Таблиця 4.8.

Результати розрахунку загальної інформативності та її кумулятивних значень, а також інформативності методами випадкового лісу та важливості перестановки

Назва змінної	**Загальна інформ.**	**Кумулятивна інформ.**	**Інформ. випадкового лісу**	**Важливість перестановки**
HvyAlcoholConsump	0.1157	0.1157	0.0711	0.0627
CholCheck	0.1068	0.2226	0.0228	0.1062
Stroke	0.0941	0.3167	0.0244	0.0914
Age	0.0930	0.4098	0.1813	0.0456
HeartDiseaseorAttack	0.0834	0.4932	0.0560	0.1621
AnyHealthcare	0.0839	0.5772	0.0219	0.0727
GenHlth	0.0665	0.6437	0.1032	0.0994
DiffWalk	0.0509	0.6947	0.0653	0.1569
BMI	0.0502	0.7450	0.0984	0.0039
HighChol	0.0437	0.7887	0.0599	0.0109

Далі метод визначення загальної інформативності було порівняно з результати визначення інформативності за допомогою методів визначення важливості змінних із застосуванням додаткової моделі Random Forest [111] та за допомогою методу Permutation Importance [83] на поточній моделі ШНМ (табл. 4.8). Результати порівняння показують, що запропонований метод та тестові мають різний характер визначення інформативності. Запропонований метод має більш

плавні зміни значення інформативності від більшого до меншого. Проте визначення більшості найбільш інформативних змінних притаманна усім методам (7/10 для інформативності випадкового лісу та 7/10 для інформативності важливості перестановки). Факт, що метод PCA показав при аналізі даних, що для збереження 80% варіативності необхідно 12 основних компонент (Рис. 3.13) також свідчить про точність запропонованого методу в оцінці інформативності.

Наступним етапом було розглянуто застосуванням модифікованого методу інтегрованих градієнтів для визначення поточної інформативності з використанням моделі ШНМ навченої на даних розмічених модулем кластерного аналізу. Типовим результатом застосування цього методу на тестових даних є наступні змінні та значення їх інформативності: AnyHealthcare – 17.6 %, PhysActivity – 9.94 %, Stroke – 9.55 %, HeartDiseaseorAttack – 8.91 % та MentHlth – 8.2 %. Видно, що більшість змінних є у найбільш інформативних по версії методу визначення загальної інформативності (табл. 4.8).

4.5. Аналіз результатів впровадження методів та моделей в комп'ютерних системах економічного моніторингу

Тестування мультиагентного методу кластеризації на даних оптового дистриб'ютора

Мета тестування полягає в перевірці можливості методу мультиагентного методу кластеризації підвищити точність кластеризації даних економічного характеру, та точність визначення цільової кількості кластерів даних. Дані для тестування детально розглянуті в розділі 3.2.6. Показано, що дані мають відносно невелику кількість записів і є просторово роздільними по визначеним кластерам. Метод мультиагентної кластеризації був перевірений із застосуванням розглянутих метрик, проте дивергенція Кульбака-Лейблера показала найбільшу точність (табл. 4.9) [74]. Такий результат може бути пояснений специфікою набору даних чи ліпшою пристосованістю методу.

Таблиця 4.9.

Результати оцінки точності кластеризації даних оптового дистриб'ютора методом мультиагентної кластеризації з різними метриками.

	Розглянута міжелемента метрика			
	Мангеттенська	Махаланобіса + обернена приналежність	Кульбака-Лейблера	Крос-ентропія
Точність	0.52	0.62	0.8	0.57

Враховуючи найбільшу точність при застосуванні дивергенції Кульбака-Лейблера отримаємо матрицю конфузів, наведену в табл. 4.10 [74]. Також показані ROC-криві для кожного з отриманих кластерів на Рис. 4.9 [74]. Згідно з отриманими результатами кожний кластер має відносно малу площу під кожною кривою, що також свідчить про низьку точність кластеризації. Відповідно до результатів наведених в матриці конфузів та рисунку ROC-кривих отримана модель кластеризації має перетягування елементів інших класів до одного, що може бути пояснене сильною незбалансованістю набору даних. Це може бути суттєвою проблемою для інших наборів даних із меншим числом прецедентів та подібною конфігурацією цільових кластерів.

Таблиця 4.10.

Матриця конфузів для оцінки точності результатів кластеризації на даних оптового дистриб'ютора за використання дивергенції Кульбака-Лейблера.

		Передбачений клас		
		Лісабон	**Опорто**	**Інші**
Актуальний клас	**Лісабон**	16	0	61
	Опорто	0	19	28
	Інші	0	1	316

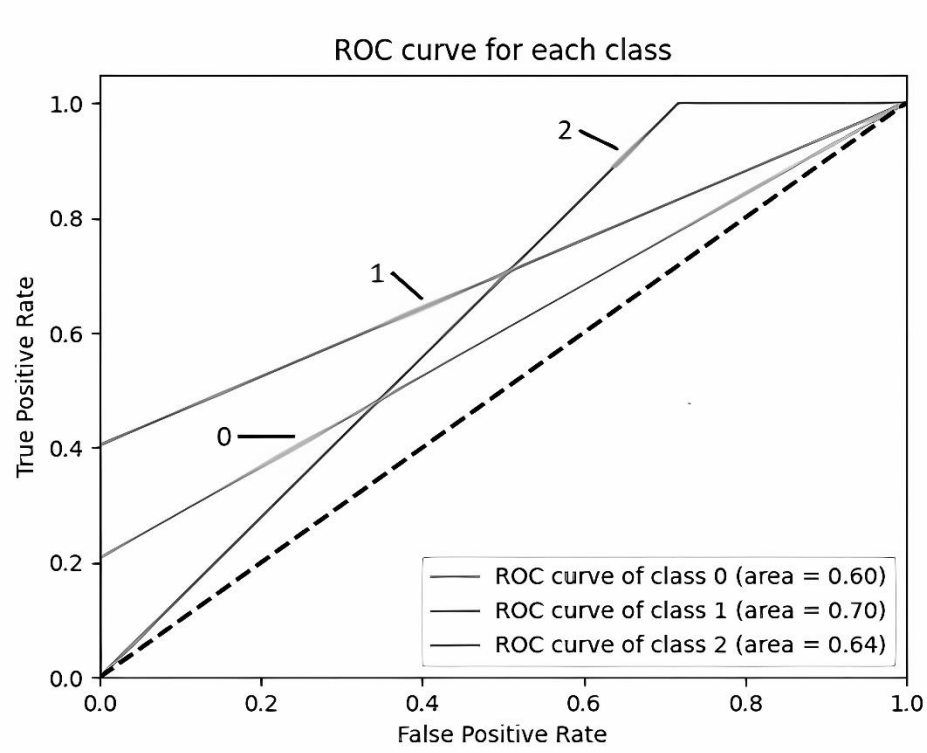


Рис. 4.9. ROC-криві для кожного з отриманих кластерів даних оптового дистриб’ютора отриманих методом мультиагентної кластеризації для дивергенції Кульбака-Лейблера.

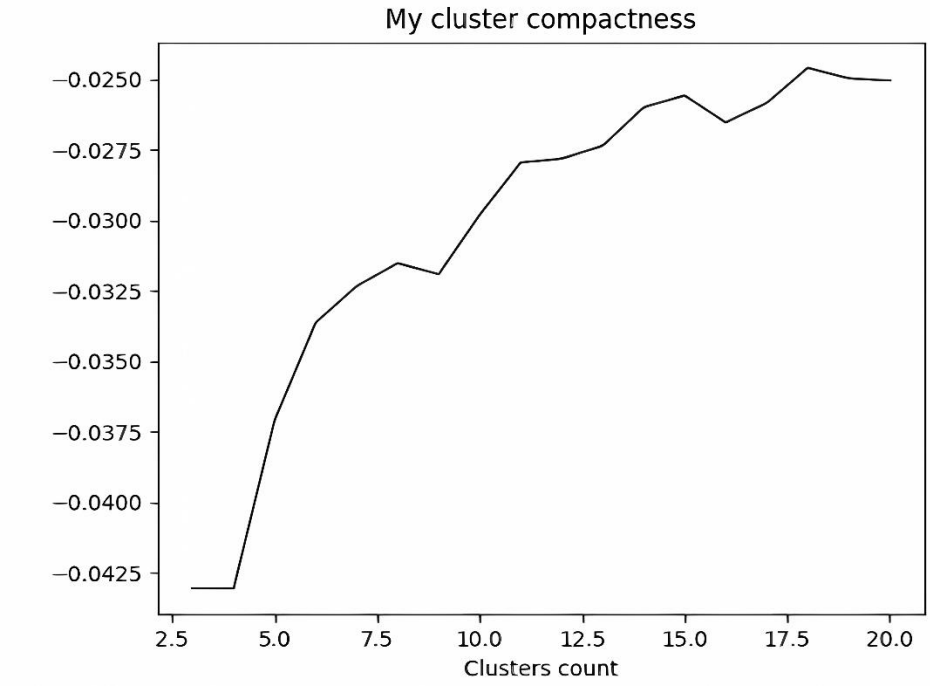


Рис. 4.10. Відношення кількості кластерів до значення функції втрат при використанні мультиагентного методу нечіткої кластеризації в режими автоматичного пошуку необхідної кількості кластерів.

Також мультиагентний метод нечіткої кластеризації із дивергенцією Кульбака-Лейблера був застосований до даних оптового дистриб'ютора в режимі автоматичного пошуку необхідної кількості кластерів. Згідно з алгоритмом пошук почався з деякого максимального числа (було обрано початкове число кластерів 20) і пошук продовжувався допоки значення функції втрат не перестане зменшуватися. Результати такого навчання моделі наведені на Рис. 4.10 [74], мінімальне значення функції витрат досягнуте при 3-х кластерах із значенням -0.0574.

Перевірка підсистеми стратифікації на даних економічного моніторингу

Відповідно до результатів тестування мультиагентного методу нечіткої кластеризації, що показали найбільшу точність відстані Кульбака-Лейблера в порівнянні з іншими існуючими методами, було вирішено зосередитися на цій відстані в подальшому тестуванні підсистеми стратифікації на даних економічного моніторингу цифрового розвитку країн. Дані економічного моніторингу детальніше розглядалися в розділі 3.2.6, де було показано їх походження та описані використані змінні, та зазначені цільові класи.

Відповідно до архітектури запропонованої підсистеми стратифікації було проведено тестування мультиагентного метода кластеризації, запропонованої моделі ШНМ із методом навчання та метода визначення загальної інформативності.

Оскільки наявний відносно невеликий об'єм даних увесь набір даних було направлено на тестування методу кластеризації. Проте для навчання ШНМ було розділено розмічені дані на тестовий та навчальний набори у співвідношенні 20% (22 прецеденти) та 80% (93 прецеденти) [60]. Точність мультиагентного методу кластеризації з обраною мірою Кульбака-Лейблера також перевірялась із застосуванням автокодувальника, що дозволило зменшити розмірність. Проте латентний простір, що надає автокодувальник неможливо використовувати для визначення інформативності змінних у зв'язку із втратою змісту в нових змінних.

Результати оцінки точності по класам мультиагентного методу кластеризації з мірою Кульбака-Лейблера показані в табл. 4.11 [60]. Процес навчання

проілюстрований на Рис. 4.9 із досягненням значення функції витрат в -0.0117 та точністю у 84.3%.

Таблиця 4.11.

Матриця конфузів для оцінки точності результатів кластеризації на даних економічного моніторингу країн.

Актуальний клас		**Передбачений клас**		
	High income	**Upper middle income**	**Lover middle income**	**Lover income**
High income	37	1	0	7
Upper middle income	1	8	1	1
Lover middle income	2	0	21	2
Lover income	0	1	2	31

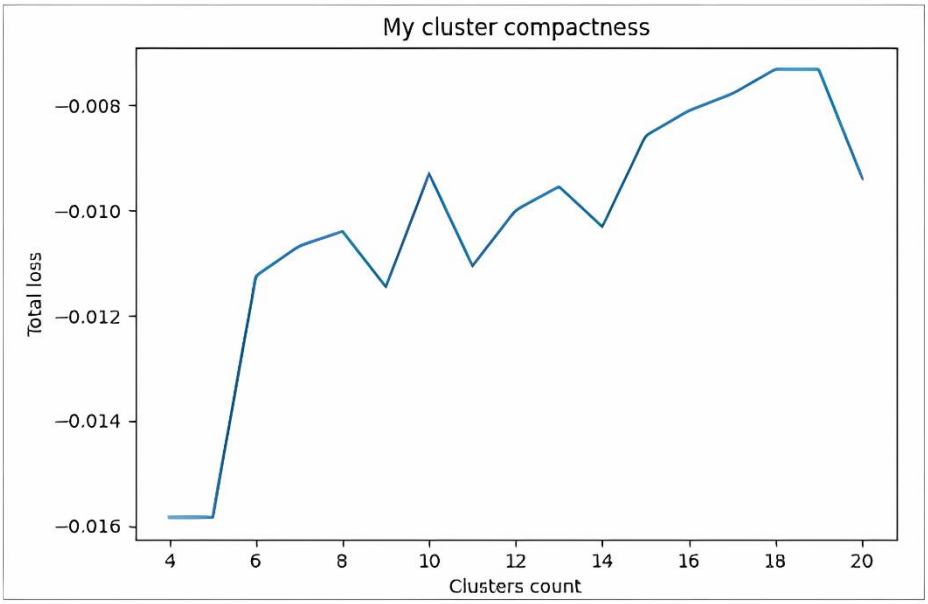


Рис. 4.11. Значення функції втрат під час навчання моделі кластеризації з відстанню Кульбака-Лейблера на даних економічного моніторингу країн.

Застосування автокодувальника дозволило зменшити розмірність даних з 32 змінних до 11 змінних латентного простору із збереженням інформативності. Застосування мультиагентного методу нечіткої кластеризації з мірою Кульбака-Лейблера дозволило дещо збільшити точність до 86.9% із досягненням значення

функції витрат -0.04827. Зазначена матриця конфузів (табл. 4.12) [60] та процес навчання обраного методу кластеризації (Рис. 4.12) [60]. Слід зазначити, що з даними, обробленими автокодувальником, запропонована модель навчається з меншою кількістю коливань значень функції витрат чим це було притаманно оригінальним даним (Рис. 4.11) [60].

Таблиця 4.12.

Матриця конфузів для оцінки точності результатів кластеризації на даних латентного простору отриманих після застосування автокодувальника до даних економічного моніторингу країн.

Актуальний клас		**Передбачений клас**		
	High income	**Upper middle income**	**Lover middle income**	**Lover income**
High income	37	1	0	7
Upper middle income	1	8	1	1
Lover middle income	2	0	21	2
Lover income	0	1	2	31

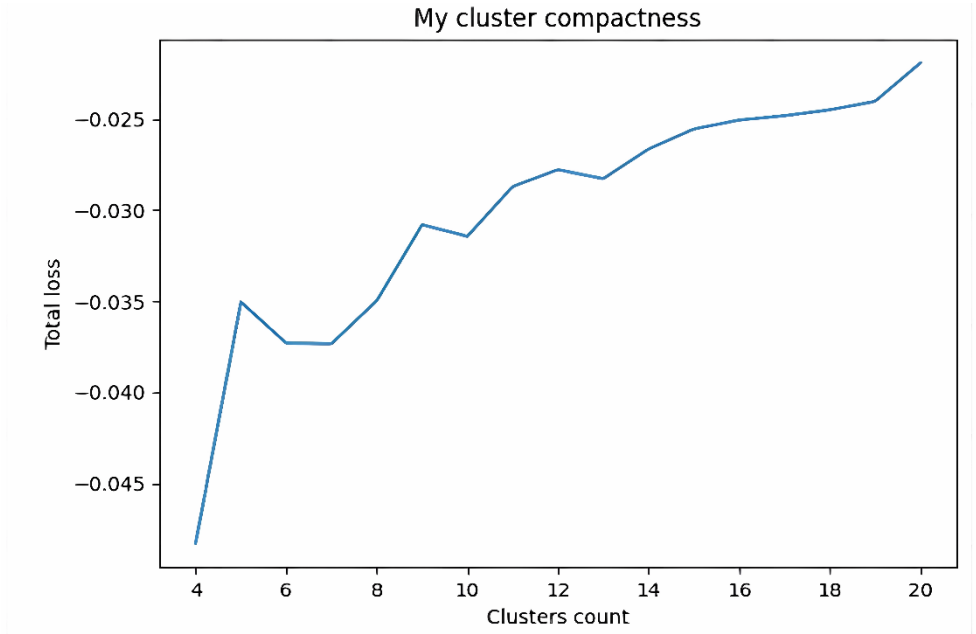


Рис. 4.12. Значення функції втрат під час навчання моделі кластеризації з відстанню Кульбака-Лейблера на даних економічного моніторингу країн

Для проведення класифікації була застосована повнозв'язна модель ШНМ, що включала 3 шари. Вхідний шар включав 7 нейронів, проміжний 90 та вихідний 4 із застосуванням Softmax функції активації для класифікації 4-х можливих станів. В результаті застосування такої моделі ШНМ отримаємо точність в 83.87% на навчальному наборі і 68.18% на тестовому. Детальніше результати тестування наведені в матриці конфузів (табл. 4.13) [60] для тестового набору даних.

Таблиця 4.13.

Матриця конфузів для оцінки точності результатів класифікації моделлю ШНМ тестових даних економічного моніторингу країн

Актуальний клас		**Передбачений клас**		
	High income	**Upper middle income**	**Lover middle income**	**Lover income**
High income	8	0	0	1
Upper middle income	0	2	0	1
Lover middle income	0	4	1	0
Lover income	2	0	0	4

Таблиця 4.14.

Матриця конфузів для оцінки точності результатів класифікації моделлю ШНМ тестових даних економічного моніторингу країн

Кластер	**Кількість записів**	**Найбільш інформативні змінні**	**Математичне очікування цільової змінної**
High income	45	TII, ICT, HCI	85.33
Upper middle income	11	TII, ICT, EGI	52.87
Lover middle income	25	EPI, HCI, OSI	63.47
Lover income	34	HCI, EPI, EGI	73.60

Також був проведений аналіз чутливості моделі ШНМ, або визначено загальну інформативність зазначених параметрів. Результати визначення загальної інформативності наведені в табл. 4.14 [60] . Відповідно до якого визначено найбільш інформативні змінні по кожному з кластерів. Розрахунки проводилися із застосуванням програми ROD&IDS, розробленої авторами [112].

4.6. Розробка практичних рекомендацій по використанню розроблених моделей та методів

В роботі були розглянуті методи і моделі стратифікації даних в комп'ютерній системі медичного моніторингу. Розроблені методи були окремо протестовані на даних медичного моніторингу. Також розроблені методи були поєднані в підсистему стратифікації, яка була перевірена із застосуванням даних медичного моніторингу. Відповідно до результатів тестування методів можна зробити наступні рекомендації щодо застосування розроблених методів та моделей:

1. Модифікований мультиагентний метод кластеризації можливо застосовувати для визначення можливих станів в нерозмічених даних медичного моніторингу. Проте цей метод також має бути використаний із іншими методами, бо як показала практика застосування, метод інколи дає неточні результати, особливо у випадку незбалансованих даних та наборів даних із специфічною роздільністю даних. Аналогом цього методу можуть бути методи DBSCAN чи Agglomerative кластеризація. Застосування методів кластеризації із різною алгоритмічною природою побудови кластерів можуть підвищити загальну стійкість системи.

2. Розроблені методи навчання й оптимізації архітектури моделей ШНМ є частиною програми ROD&IDS, розробленої авторами [112]. Тому для застосування в якості частини підсистеми стратифікації розроблена програма не підходить через специфіку концепції автономності підсистеми стратифікації. Проте застосування аналогічних ШНМ цілком можливо, а зважаючи на відносно малі об'єми таких моделей ШНМ оптимізація гіперпараметрів таких ШНМ не вимагає застосування виключно програми ROD&IDS.

3. Розроблений метод визначення загальної інформативності показав точне виявлення інформативних змінних по їх впливу на результати класифікації навченої ШНМ. Тому застосування методу цілком доцільне для аналізу чутливості змінних та визначення підмножини найбільш інформативних змінних.

4. Модифікований метод інтегрованих градієнтів для визначення поточної інформативності показав точне відтворення причин прийняття рішень

моделлю ШНМ для певних вхідних даних. Тому застосування запропонованого методу доцільне для обґрунтування прийнятих рішень в медичній комп'ютерній системі підтримки прийняття рішень.

В загальному запропонована підсистема стратифікації може працювати в трьох режимах, розглянутих раніше. Відповідно до результатів тестування підсистеми стратифікації на даних медичного моніторингу можливо зробити наступні висновки щодо можливого використання:

1. У разі наявності нерозмічених даних чи тільки припущення про кількість можливих станів в системі медичного моніторингу рекомендується використовувати систему в запропонованому виді. Проте зважаючи на недоліки запропонованого методу кластеризації необхідно розглянути також інші методи кластеризації в модулі кластерного аналізу.

2. У разі наявності розмічених даних (наприклад, даних моніторингу пацієнтів по певних захворювань, аналізів й відповідних діагнозів тощо) рекомендується виключити модуль кластерного аналізу так як зважаючи на його специфіку він буде привносити зайві неточності в функціонування системи. Також методи кластерного аналізу зазвичай спрощують форму простору змінних стану даних, що підвищує якість класифікації, проте вносить скриті помилки в функціонування підсистеми.

3. Загальна рекомендація до використання підсистеми стратифікації це проводити впровадження під наглядом спеціаліста, для перевірки якості функціонування системи. Автоматична частина не гарантує точність результатів навчання моделей в модулі кластерного аналізу та класифікації.

Висновок до розділу 4

В четвертому розділі було показано, що розроблені методи мають високу точність кластеризації та класифікації, а також високу достовірність визначення інформативності. Для перевірки використовувалися данні медичного моніторингу CDC BRFSS Survey 2021 присвячених захворюванню на діабет, що мають 21 змінну стану, 1 змінну, що визначає цільовий стан та 236 тисяч записів. Показано, що застосування мультиагентного методу нечіткої кластеризації для некерованого

розділення даних дає точність виділення станів в 54.75% та 54.8% на навчальному і тестових наборах даних. Проте теоретично можлива точність, що була досягнута з моделлю ШНМ у ролі універсального апроксиматору дала точність виявлення станів в 80.8% та 80.5% на навчальному і тестовому наборах відповідно. Метод для навчання і метод для конфігурації гіперпараметрів моделі ШНМ в результаті дали точність в 99.8% та 99.7% на тестовому і навчальному наборах. Показано, що метод визначення загальної інформативності достовірно визначає інформативність змінних, адже найбільш інформативні змінні були з невеликою різницею в рейтингу визначені так само іншими методами, а саме виявлено, що з 10 найбільш інформативних змінних 7 було виявлено методом визначення інформативності випадкового лісу та 7 було виявлено методом важливості перестановки. Також достовірність була підтверджена із використанням упереджених змінних стану та показано, що на 10 змінних стану метод визначає 78.87% загальної інформативності змінних, а метод PCA дає варіативність в 80%. Достовірність методу визначення поточної інформативності була підтверджена через порівняння з іншими методами на багатьох варіантах змінних стану. Показано, що 7 з 10 змінних також було визначено методом загальної інформативності.

Також було перевірена можливість розширення спектру застосування запропонованих методів і моделей на даних економічного моніторингу. Був перевірений метод кластеризації та виявлена висока чутливість до сильно незбалансованих даних. Та перевірене застосування методів і моделей стратифікації до даних економічного моніторингу країн із позитивним результатом впровадження.

Та в кінці розділу відповідно до результатів тестування наведені практичні рекомендації щодо застосування розроблених методів і моделей стратифікації окремо і підсистеми стратифікації в комп’ютерній системі медичного моніторингу в цілому.

Основні положення цього розділу викладені у публікаціях автора [1–6, 9].

ВИСНОВКИ

1. На етапі дослідження проведено аналіз комп’ютерних систем медичного моніторингу, в результаті якого виявлено, що таким системам притаманне оперування великим об’ємами даних. Крім того, існує обмежена кількість спеціалістів для їх аналізу. Також виявлена висока складність та вартість розробки, впровадження і підтримки таких систем. Що в загальному веде до недостатнього розуміння даних, що генеруються цими системами медичними працівниками. Показано, що через специфіку даних, що розглядаються такими системами, задача стратифікації має вирішуватися в три етапи: кластеризації даних, класифікації станів пацієнтів та аналіз чутливості, що в свою чергу поділяється на визначення загальної і поточної інформативності змінних.

2. Основними науковими результатами, що були отримані в роботі, є наступні.

✓ Вперше розроблено модель комп’ютерної системи медичного моніторингу, особливістю якої є застосування та організація взаємодії методів стратифікації для вирішення проблеми кластеризації даних, класифікації станів пацієнтів та визначення інформативності змінних цього стану, що в сукупності забезпечує підвищення точності стратифікації даних в комп’ютерній системі медичного моніторингу.

✓ Удосконалено мультиагентний метод нечіткої кластеризації, що відрізняється поєднанням нечіткої кластеризації c-means із мультиагентним відбором еліт, що дає можливість виконати модифікацію визначення щільності та роздільності отримуваних кластерів і як наслідок підвищити точність виділення станів пацієнтів в комп’ютерній системі медичного моніторингу.

✓ Удосконалено метод класифікації станів пацієнтів повнозв’язною штучною нейронною мережею за допомогою поєднання процедур прискореного навчання та підбору гіперпараметрів моделі штучній нейронній мережі. Це дозволило ефективно оптимізувати ваги та архітектуру моделей штучній нейронній мережі для вирішення задачі класифікації станів по відповідним змінним.

- ✓ Удосконалено методи визначення загальної інформативності змінних щодо стану пацієнтів комп’ютерної системи медичного моніторингу за рахунок виділення зв’язку між входами і виходами через поширення градієнтів в штучній нейронній мережі, а також поточної інформативності шляхом перетворення вагових показників методу інтегрованих градієнтів, що створює умови для виявлення найбільш впливових керованих і некерованих змінних стану й оцінки впливу виявлених змінних на конкретне прийняте рішення та дає можливість пояснити причини прийнятого медичного рішення.
- ✓ Дістав подальшого розвитку метод верифікації програмного забезпечення стратифікації даних щодо елементів комп’ютерної системи медичного моніторингу, що відрізняється від існуючих виконанням комплексної перевірки як програмної реалізації, так і точності роботи розроблених методів і моделей, що дає можливість скоротити строки розробки програмного забезпечення.

3. Розроблені методи отримали відповідну програмну реалізацію та були протестовані із застосуванням методу верифікації програмного забезпечення стратифікації елементів в комп’ютерній системі медичного моніторингу.

4. Отримані результати підтверджують можливість практичного використання запропонованої комп’ютерної системи медичного моніторингу із виділеною підсистемою стратифікації елементів. Обґрунтована можливість точного виділення можливих станів в наборі даних спостережень за допомогою розробленого мультиагентного методу нечіткої кластеризації. Показано, що застосування методу прискореного навчання і пошуку оптимального набору гіперпараметрів дозволяє автоматично налаштовувати моделі ШНМ необхідного обсягу задля забезпечення якісного й швидкого навчання для вирішення поставленої задачі. Показано, що розроблений метод визначення загальної інформативності дозволяє точно визначати впливи змінних на виходи навченої моделі ШНМ. Також показано, що модифікований метод інтегрованих градієнтів для визначення поточної інформативності дозволяє точно визначати впливи конкретних змінних на конкретні результати роботи навченої моделі ШНМ тим самим обґрунтовуючи результати її роботи.

5. Достовірність отриманих наукових положень, висновків та рекомендацій забезпечується аргументованими результатами досліджень і підтверджується співставленням з результатами експериментальних досліджень на основі методів системного аналізу, мультиагентного підходу, а також застосування імітаційного та математичного моделювання, теорії математичної статистики, теорії множин, теорії ймовірностей, теорії графів, лінійної алгебри, методів математичної оптимізації, диференціального аналізу, теорії штучних нейронних мереж.

Основні результати роботи було реалізовано в Харківському національному університеті імені В. Н. Каразіна у рамках НДР «Моделювання інформаційних процесів у складних і розподілених системах» за 2021 – 2023 рр. (ДР № 0121U109183).

В підсумку дослідження показано, що розроблені методи мають високу точність кластеризації та класифікації, а також високу достовірність визначення інформативності. Для перевірки використовувалися данні медичного моніторингу CDC BRFSS Survey 2021 присвячених захворюванню на діабет, що мають 21 змінну стану, 1 змінну, що визначає цільовий стан та 236 тисяч записів. Показано, що застосування мультиагентного методу нечіткої кластеризації для некерованого розділення даних дає точність виділення станів в 54.75% та 54.8% на навчальному і тестових наборах даних. Проте теоретично можлива точність, що була досягнута з моделлю ШНМ у ролі універсального апроксиматору дала точність виявлення станів в 80.8% та 80.5% на навчальному і тестовому наборах відповідно. Метод для навчання і метод для конфігурації гіперпараметрів моделі ШНМ в результаті дали точність в 99.8% та 99.7% на тестовому і навчальному наборах. Показано, що метод визначення загальної інформативності достовірно визначає інформативність змінних, адже найбільш інформативні змінні були з невеликою різницею в рейтингу визначені так само іншими методами, а саме виявлено, що з 10 найбільш інформативних змінних 7 було виявлено методом визначення інформативності випадкового лісу та 7 було виявлено методом важливості перестановки. Також достовірність була підтверджена із використанням упереджених змінних стану та

показано, що на 10 змінних стану метод визначає 78.87% загальної інформативності змінних, а метод PCA дає варіативність в 80% на 12 змінних. Достовірність методу визначення поточної інформативності була підтверджена через порівняння з іншими методами на багатьох варіантах змінних стану. Показано, що 7 з 10 змінних також було визначено методом загальної інформативності.

6. Результати, що були отримані в роботі можуть бути рекомендовані до застосування в науково-дослідних організаціях для розроблення комп'ютерних систем медичного моніторингу.

7. Виходячи з наведених у дисертації наукових і практичних результатів, а також підтвердження факту їх достовірності, наукової та практичної значущості, дають змогу вважати, що сформульована наукова задача удосконалення або розробки нових математичних моделей та обчислювальних методів стратифікації елементів комп'ютерних систем медичного моніторингу – розв'язаною, а поставлену мету – досягнутою.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Fertig E. J., Jaffee E. M., Macklin P., Stearns V., Wang C. Forecasting cancer: from precision to predictive medicine. Med. 2021. No. 2(9). Pp. 1004–1010. DOI: https://doi.org/10.1016/j.medj.2021.08.007

2. Simarjeet Kaur, Jimmy Singla, Lewis Nkenyereye, Sudan Jha, Deepak Prashar, Gyanendra Prasad Joshi, Shaker El-Sappagh, Md. Saiful Islam and S. M. Riazul Islam. Medical Diagnostic Systems Using Artificial Intelligence (AI) Algorithms: Principles and Perspectives. IEEE Access. 2020 No. 8. Pp. 228049–228069. DOI: https://doi.org/10.1109/ACCESS.2020.3042273

3. I. Perova, Y. Brazhnykova, N. Miroshnychenko and Y. Bodyanskiy. Information Technology for Medical Data Stream Mining. 2020 IEEE 15th International Conference on Advanced Trends in Radioelectronics, Telecommunications and Computer Engineering (TCSET). Ukraine. 2020. Pp. 93–97. DOI: https://doi.org/10.1109/TCSET49122.2020.235399

4. Veronica Goriacha, Oleksandr Sokolov, Kateryna Ugryumova, Igor Antonyan, Yuri Roshin, Alexandr Zelensky, Fedor Moshel and Taron Nalbandian. Forecasting of Patients Condition in the Monitoring Medical Systems on the Example of Prostate Diseases. Journal of Education Health and Sport. 2016. No. 6(5). DOI: https://doi.org/10.5281/zenodo.51160

5. Viktoriia Strilets, Nina Bakumenko, Serhii Chernysh, Mykhaylo Ugryumov, Volodymyr Donets. Application of artificial neural networks in the problems of the patient's condition diagnosis in medical monitoring systems. Advances in Intelligent Systems and Computing. AISC 1113. 2020. P. 173–185. DOI: https://doi.org/10.1007/978-3-030-37618-5_16F

6. Vitaliy Yakovyna, Natalya Shakhovska. Modelling and predicting the spread of COVID-19 cases depending on restriction policy based on mined recommendation rules. Mathematical Biosciences and Engineering. 2021. Vol. 18. Issue 3. Pp. 2789–2812. DOI: https://doi.org/10.3934/mbe.2021142

7. R. Tkachuk, A. Tkachuk, O. Yanenko and K. Shevchenko. Automated Implant Testing System for Intraocular Pressure Adjustment. 2020 IEEE 15th International Conference on Advanced Trends in Radioelectronics, Telecommunications and Computer Engineering (TCSET). Ukraine. 2020. Pp. 190–193, DOI: https://doi.org/10.1109/TCSET49122.2020.235420

8. D. Chumachenko, T. Chumachenko, I. Meniailov, O. Muradyan and G. Zholtkevych. Forecasting of COVID-19 Epidemic Process by Lasso Regression. 2021 IEEE International Conference on Information and Telecommunication Technologies and Radio Electronics (UkrMiCo). Ukraine. 2021. Pp. 80–83, DOI: https://doi.org/10.1109/UkrMiCo52950.2021.9716621

9. Dmytro Boyko, Dmytro Chumachenko, Tetyana Chumachenko, Sergey Lvov, Artem, Lytovchenko, Olena Muradyan and Grygoriy Zholtkevych. The Concept of Decisions Support System to Mitigate the COVID-19 Pandemic Consequences based on Social and Epidemic Processes Intelligent Analysis. Proceedings of the International Workshop of IT-professionals on Artificial Intelligence (ProfIT AI 2021) 2021. Kharkiv. 2021. Pp 55–64. URL: https://ceur-ws.org/Vol-3003/paper6.pdf

10. Logeshwaran, J., Malik, J. A., Adhikari, N., Joshi, S. S., and Bishnoi, P. IoT-TPMS: An innovation development of triangular patient monitoring system using medical internet of things. International Journal of Health Sciences. 2022. No. 6(S5). Pp. 9070–9084. DOI: https://doi.org/10.53730/ijhs.v6nS5.10765

11. Humayun, M., Jhanjhi, N.Z., Almotilag, A., & Almufareh, M.F.Agent-Based Medical Health Monitoring System. Sensors (Basel, Switzerland). 2022. No. 22. DOI: https://doi.org/10.3390/s22082820

12. Yu, M., Li, G., Jiang, D., Jiang, G., Tao, B., & Chen, D. Hand medical monitoring system based on machine learning and optimal EMG feature set. Personal and Ubiquitous Computing. 2019. Pp. 1–17. DOI: https://doi.org/10.1007/s00779-019-01285-2

13. Site of Medical Technology Clinical Research Program IMPART (Inflammation & Metabolism, Physical Ability, Research Translation). URL: https://impart.team/medical-technology-clinical-research-program

14. Site of AMED Japan Agency for Medical Research and Development. URL: https://www.amed.go.jp/en/aboutus/index.html

15. Site of NITRD: Digital Health Research and Development. URL: https://www.nitrd.gov/coordination-areas/dhrd/

16. Lee D, Kim K. Public R&D Projects-Based Investment and Collaboration Framework for an Overarching South Korean National Strategy of Personalized Medicine. Int J Environ Res Public Health. 2022. No. 19(3), 1291. DOI: https://doi.org/10.3390/ijerph19031291

17. Site of Deloitte. The next wave of innovation. Technology and value-based care are transforming medtech R&D. 2018. URL: https://www.deloitte.com/ie/en/our-thinking/insights/industry/life-sciences/medtech-research-and-development-innovation.html

18. Farshad Firouzi, Shiying Jiang, Krishnendu Chakrabarty, Bahar Farahani, Mahmoud Daneshmand, Jaeseung Song and Kunal Mankodiya. Fusion of IoT, AI, Edge–Fog–Cloud, and Blockchain: Challenges, Solutions, and a Case Study in Healthcare and Medicine. IEEE Internet of Things Journal. 2023. No. 10. Pp. 3686–3705. DOI: https://doi.org/10.1109/JIOT.2022.3191881

19. Manoj Singh Adhikari, Patan Salman Khan, Gangadhar Senapathi, Bhawna Chatterjee, Duvyu Vinodh Reddy and Praveen Kumar Malik. Design of An IoT Based Smart Medicine Box. 2023 IEEE Devices for Integrated Circuit (DevIC). 2023. Pp. 346–349. DOI: https://doi.org/10.1109/DevIC57758.2023.10134884

20. Chandramohan P., Kanagaraj Venusamy, Deepak Rishi G, Mohammed Asan Wazil A.K and Naveen Raj Um. Automated Medicine Dispenser with Personal Healthcare Monitoring Using IoT. 2023 Eighth International Conference on Science Technology Engineering and Mathematics (ICONSTEM). 2023. DOI: https://doi.org/10.1109/ICONSTEM56934.2023.10142832

21. O.Terrada, B. Cherradi, A. Raihani and O. Bouattane. A fuzzy medical diagnostic support system for cardiovascular diseases diagnosis using risk factors. 2018 International Conference on Electronics, Control, Optimization and Computer Science (ICECOCS). 2018. Pp. 1–6. DOI: https://doi.org /10.1109/ICECOCS.2018.8610649

22. O. E. Beggar, M. Ramdani and M. Kissi. Design and development of a fuzzy explainable expert system for a diagnostic robot of COVID-19. International Journal of Electrical and Computer Engineering (IJECE). 2023. DOI: https://doi.org/10.11591/ijece.v13i6.pp6940-6951

23. Bh. Nagarajasri Prof. M. Padmavathamma. A Novel Diagnostic Computer Aided Medical Tool for Breast Cancer based on Neuro Fuzzy Logic. International Journal of Scientific & Engineering Research. 2013. Volume 4. Issue 12. URL: https://www.ijser.org/paper/A-Novel-Diagnostic-Computer-Aided-Medical-Tool-for-Breast-Cancer-based-on-Neuro-Fuzzy-Logic.html

24. Nabiilah Ardini Fauziyyah, Salwani Abdullah and Siti Nurrohmah. Reviewing the consistency of the Naïve Bayes Classifier's performance in medical diagnosis and prognosis problems. Medicine, Computer Science. 2020. DOI: https://doi.org/10.1063/5.0007885

25. Dr. Chamandeep Kaur, Tuhina Panda, Subhasis Panda, Dr. Abdul Rahman, Mohammed AL Ansari, Ms. M. Nivetha and Dr. B. Kiran Bala. Utilizing the Random Forest Algorithm to Enhance Alzheimer's disease Diagnosis. 2023 Third International Conference on Artificial Intelligence and Smart Energy (ICAIS). 2023. Pp. 1662-1667. DOI: https://doi.org/10.1109/ICAIS56108.2023.10073852

26. Zhuang Jin, Yaqiong Zhu, Shijie Zhang, Fang Xie, Mingbo Zhang, Ying Zhang, Xiaoqi Tian, Jue Zhang, Yukun Luo and Junying Cao. Ultrasound Computer-Aided Diagnosis (CAD) Based on the Thyroid Imaging Reporting and Data System (TI-RADS) to Distinguish Benign from Malignant Thyroid Nodules and the Diagnostic Performance of Radiologists with Different Diagnostic Experience. Medical Science Monitor: International Medical Journal of Experimental and Clinical Research. 2020. No. 26. DOI: https://doi.org/10.12659/MSM.918452

27. Jianhui Zhao, Tianquan Chen and Bo Cai. A computer-aided diagnostic system for mammograms based on YOLOv3. Multimedia Tools and Applications. 2021. No. 81. Pp. 19257–19281. DOI: https://doi.org/10.1007/s11042-021-10505-y

28. Daniele Fresilli, Giorgio Grani, Maria Luna De Pascali, Gregorio Alagna, Eleonora Tassone, Valeria Ramundo, Valeria Ascoli, D. Bosco, Marco Biffoni, Marco

Bononi, Vito D'Andrea, Fabrizio Maria Frattaroli, Laura Giacomelli, Yana Solskaya, Giorgia Polti, Patrizia Pacini, Olga Guiban, Raffaele Gallo Curcio, Marcello Caratozzolo and Vito Cantisani. Computer-aided diagnostic system for thyroid nodule sonographic evaluation outperforms the specificity of less experienced examiners. Journal of Ultrasound. 2020. No. 23. Pp. 169-174. DOI: https://doi.org/10.1007/s40477-020-00453-y

29. Ravinder Aggarwal, Viknesh Sounderajah, Guy Martin, Daniel Shu Wei Ting, Alan Karthikesalingam, Dominic King, Hutan Ashrafian and Ara Darzi. Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis. NPJ Digital Medicine. 2021. No. 4. DOI: https://doi.org/10.1038/s41746-021-00438-z

30. V. J. Singh, P. Sharma and D. A. Mehta. Big Data Analytics in Healthcare: Opportunities and Challenges. International Journal of Advanced Research in Science, Communication and Technology. 2023. DOI: https://doi.org/10.48175/ijarsct-9414

31. J. Yang, A. A. Soltan, Y. Yang and D. A. Clifton. Algorithmic Fairness and Bias Mitigation for Clinical Machine Learning: Insights from Rapid COVID-19 Diagnosis by Adversarial Learning. medRxiv. 2022. DOI: https://doi.org/10.1101/2022.01.13.22268948

32. A. Ferrario, S. Gloeckler and N. Biller-Andorno. AI knows best? Avoiding the traps of paternalism and other pitfalls of AI-based patient preference prediction. Journal of Medical Ethics. 2023. No. 49 Pp. 185–186. DOI: https://doi.org/10.1136/jme-2023-108945

33. Site of Google Cloud: What is unsupervised learning? URL: https://cloud.google.com/discover/what-is-unsupervised-learning

34. Tharcis Paulraj, Kezi Selva Vijila Chelliah and Sundar Chinnasamy. Lung computed axial tomography image segmentation using possibilistic fuzzy C-means approach for computer aided diagnosis system. International Journal of Imaging Systems and Technology. 2019. No. 29. Pp. 374–381. DOI: https://doi.org/10.1002/ima.22340

35. N. Sh. Abu-Zeid, Rasha Kashif and Osama Badawy. Immune Based Clustering for Medical Diagnostic Systems. 2012 International Conference on Advanced

Computer Science Applications and Technologies (ACSAT). 2012. Pp. 372-375. DOI: https://doi.org/10.1109/ACSAT.2012.42

36. Troy Vargason, Richard Eugene Frye, Deborah L. McGuinness and Juergen Hahn. Clustering of co-occurring conditions in autism spectrum disorder during early childhood: A retrospective analysis of medical claims data. Autism Research. 2019. No. 12. DOI: https://doi.org/10.1002/aur.2128

37. T. Thamaraimanalan, M. Mohankumar, H. Anandakumar, M. Deepha, U. Hari Priya, G. Bhanu Priya and M. Aiswarya Devi. Machine Learning based Patient Mental Health Prediction using Spectral Clustering and RBFN Algorithms. 2022 8th International Conference on Advanced Computing and Communication Systems (ICACCS). 2022. No. 1. Pp. 1840–1843. DOI: https://doi.org/10.1109/ICACCS54159.2022.9785142

38. Sunila Godara, Rishipal Singh and Sanjeev Kumar. A Novel Weighted Class based Clustering for Medical Diagnostic Interface. Indian journal of science and technology. 2016. No. 9. DOI: https://doi.org/10.17485/IJST%2F2016%2FV9I44%2F101286

39. R. A. Fisher. The Statistical Utilization of Multiple Measurements. Annals of Eugenics. 1938. No. 8 (4). Pp. 376–386. DOI: https://doi.org/10.1111/j.1469-1809.1938.tb02189.x

40. Mehryar Mohri, Afshin Rostamizadeh, Ameet Talwalkar. Foundations of Machine Learning. The MIT Press, Second Edition. 2018. URL: https://cs.nyu.edu/~mohri/mlbook/

41. Wei Wei and Xu Yang. Comparison of Diagnosis Accuracy between a Backpropagation Artificial Neural Network Model and Linear Regression in Digestive Disease Patients: an Empirical Research. Computational and Mathematical Methods in Medicine. 2021. DOI: https://doi.org/10.1155/2021%2F6662779

42. Amit Kumar Kundu, Shaikh Anowarul Fattah and Khan Arif Wahid. Multiple Linear Discriminant Models for Extracting Salient Characteristic Patterns in Capsule Endoscopy Images for Multi-Disease Detection. IEEE Journal of Translational

Engineering in Health and Medicine. 2020. No. 8. DOI: https://doi.org/10.1109/JTEHM.2020.2964666

43. Nathan D. Schilaty, Nathaniel A. Bates, Sydney Kruisselbrink, Aaron John Krych and Timothy E. Hewett. Linear Discriminant Analysis Successfully Predicts Knee Injury Outcome From Biomechanical Variables. The American Journal of Sports Medicine. 2020. No. 48. Pp. 2447–2455. DOI: https://doi.org/10.1177/0363546520939946

44. Annisa Rahmadani, Casi Setianingsih, Fussy Mentari Dirgantara, Ayub Rosihan Ambarita, Hafid Ikhsan Arifin, Indratama Pangasian Manalu and Muhammat Lio Pratama. Depression, anxiety, and stress disorders detection in students during the Covid-19 pandemic using Naïve Bayes algorithm. BIO Web of Conferences. 2023. DOI: https://doi.org/10.1051/bioconf%2F20237501003

45. S. Saravanan, V. Vinoth Kumar, Velliangiri Sarveshwaran, Alagiri Indirajithu, D. Elangovan and Shaikh Muhammad Allayear. Computational and Mathematical Methods in Medicine Glioma Brain Tumor Detection and Classification Using Convolutional Neural Network. Computational and Mathematical Methods in Medicine. 2022. DOI: https://doi.org/10.1155/2022%2F4380901

46. P. Manimegalai , R. Suresh Kumar, Prajoona Valsalan, R. Dhanagopal, P. T. Vasanth Raj and Jerome Christhudass. 3D Convolutional Neural Network Framework with Deep Learning for Nuclear Medicine. Scanning. 2022. DOI: https://doi.org/10.1155/2022%2F9640177

47. Angelo D. Bonzanini, Joel A. Paulson, David B. Graves and Ali Mesbah. Toward Safe Dose Delivery in Plasma Medicine using Projected Neural Network-based Fast Approximate NMPC. IFAC-PapersOnLine. 2020. No. 53. Pp. 5279–5285. DOI: https://doi.org/10.1016/J.IFACOL.2020.12.1208

48. Omidali Aghababaei Jazi and Eleanor M. Pullenayegum. Variable selection in semiparametric regression models for longitudinal data with informative observation times. Statistics in Medicine. 2022. No. 41. Pp. 3281–3298. DOI: https://doi.org/10.1002/sim.9417

49. Anna Markella Antoniadi, Yuhan Du, Yasmine Guendouz, Lan Wei, Claudia Mazo, Brett A. Becker and Catherine Mooney. Current Challenges and Future Opportunities for XAI in Machine Learning-Based Clinical Decision Support Systems: A Systematic Review. Applied Sciences. 2021. No. 11. DOI: https://doi.org/10.3390/APP11115088

50. Amina Adadi and Mohammed Berrada. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). IEEE Access 6. 2018. Pp. 52138–52160. DOI: https://doi.org/10.1109/ACCESS.2018.2870052

51. Bakumenko, N., Chernysh, S., Bezlyubchenko, A., Goryachaya, V., Donets, V., Strilets, V., Meniailov, I., Ugryumova, K, Ugryumov, M. Stratification of Patients in Medical Monitoring Systems based on Machine Learning Methods. 2021. DOI: https://doi.org/10.13140/RG.2.2.11832.52482

52. Volodymyr Donets, Dmytro Shevchenko, Maksym Holikov, Viktoriia Strilets, Serhiy Shmatkov. Application of a data stratification approach in computer medical monitoring systems. Eastern-European Journal of Enterprise Technologies. 2024. Vol. 2, No. 9(128). Pp. 6–16. DOI: https://doi.org/10.15587/1729-4061.2024.298805

53. Лихач, О. Ю., Угрюмов, М. Л., Шевченко, Д. О., Шматков, С. І. етоди виявлення викидів в пробних вибірках при управлінні процесами в системах за станом. Bulletin of V.N. Karazin Kharkiv National University, series «Mathematical modeling. Information technology. Automated control systems». 2022. No. 53. DOI: https://doi.org/10.26565/2304-6201-2022-53-03

54. Shevchenko, D., Ugryumov, M., and Artiukh, S. Monitoring Data Aggregation Of Dynamic Systems Using Information Technologies. Innovative Technologies and Scientific Solutions for Industries. 2023. DOI: https://doi.org/10.30837/itssi.2023.23.123

55. M. Schlesinger, V. Hlavac. Ten lectures on statistical and structural pattern recognition. Springer. Dordrecht, 2002. P. 522. DOI: https://doi.org/10.1007/978-94-017-32

56. Edy Umargono, Jatmiko Endro Suseno and S.K Vincensius Gunawan. K-Means Clustering Optimization Using the Elbow Method and Early Centroid

Determination Based on Mean and Median Formula. Proceedings of the 2nd International Seminar on Science and Technology (ISSTEC 2019). 2020. DOI: https://doi.org/10.5220/0009908402340240

57. M. Hahsler, M. Piekenbrock and D. Doran. DBSCAN: Fast Density-Based Clustering with R. Journal of Statistical Software. 2019. DOI: https://doi.org/10.18637/jss.v091.i01

58. Monath, N., Dubey, K.A., Guruganesh, G., Zaheer, M., Ahmed, A., McCallum, A., Mergen, G., Najork, M., Terzihan, M., Tjanaka, B., Wang, Y., & Wu, Y. Scalable Hierarchical Agglomerative Clustering. Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining. 2020 DOI: https://doi.org/10.1145/3447548.3467404

59. Viktoriia Strilets, Volodymyr Donets, Mykhaylo Ugryumov, Sergii Artiukh, Roman Zelenskyi, Tamara Goncharova. Agent-oriented data clustering for medical monitoring. Radioelectronic And Computer Systems, February 2022. DOI: https://doi.org/10.32620/reks.2022.1.08

60. Volodymyr Donets, Viktoriia Strilets, Mykhaylo Ugryumov, Dmytro Shevchenko, Svitlana Prokopovych, Liubov Chagovets. Methodology of the countries' economic development data analysis. System Research and Information Technologies, December 2023. DOI: https://doi.org/10.20535/SRIT.2308-8893.2023.4.02

61. C. Aggarwal. Charu, K. Reddy Chandan (ed.). Data clustering: algorithms and applications. CRC Press, Taylor & Francis Group. 2014. P. 622.

62. W. B. Kinlaw, M. P. Kritzman, D. Turkington. A New Index of the Business Cycle. Macroeconomics: Prices. 2020. P. 30. DOI: https://doi.org/10.2139/ssrn.3521300

63. V. K. Finn, O. P. Shesternikova. The Heuristics of Detection of Empirical Regularities by JSM Reasoning. Automatic Documentation and Mathematical Linguistics. 2018. Vol. 52. Issue 5. Pp. 215–247. DOI: https://doi.org/10.3103/S0005105518050023

64. Bakumenko, N., Strilets, V., Ugryumov, M. Application of the C-Means Fuzzy Clustering Method for the Patient's State Recognition Problems in the Medicine Monitoring Systems. CEUR Workshop Proceedings of 3rd International Conference on

Computational Linguistics and Intelligent Systems, COLINS 2019. 2019. Vol. 1. Pp. 218–227. URL: https://www.researchgate.net/publication/338819685

65. Md. Abu Bakr Siddique, Rezoana Bente Arif, Mohammad Mahmudur Rahman Khan, Zahidun Ashrafi. Implementation of Fuzzy C-Means and Possibilistic C-Means Clustering Algorithms, Cluster Tendency Analysis and Cluster Validation. ArXiv e-Journal (ISSN 2331-8422). 2019. URL: https://arxiv.org/abs/1809.08417v3

66. Н. С. Бакуменко, В. В Донець, Д. О. Шевченко, О. О. Одинець, М. Л. Угрюмов. Методи кластеризації даних на основі інформаційних критеріїв. Науковий збірник праці міжнародної науково-технічної конференції «Комп'ютерне моделювання у наукоємних технологіях (КМНТ -2021)». 2021.

67. N. Amruthnath, T. Gupta. Fault Class Prediction in Unsupervised Learning using Model-Based Clustering Approach. 2018 International Conference on Information and Computer Technologies (ICICT). 2018. Pp. 5–12. DOI: https://doi.org/10.13140/RG.2.2.22085.14563

68. Askari, S. Fuzzy C-Means clustering algorithm for data with unequal cluster sizes and contaminated with noise and outliers: Review and development. Expert Systems with Applications. 2020. Vol. 165. Article no. 113856. DOI: https://doi.org/10.1016/j.eswa.2020.113856

69. F. Nielsen. On the Jensen-Shannon Symmetrization of Distances Relying on Abstract Means. Entropy. 2019. Vol. 21. Issue 5. Article no. 485. DOI: https://doi.org/10.3390/e21050485

70. Zarinbala, M., Zarandia, Fazel M. H., Turksen, I.B. Relative entropy fuzzy c-means clustering. Information Sciences. 2014. Vol. 260. Pp. 74–97. DOI: https://doi.org/10.1016/j.ins.2013.11.004

71. Menga, Yinfeng., Liangb, Jiye., Caob, Fuyuan., He, Yijun. A new distance with derivative information for functional k-means clustering algorithm. Information Sciences. 2018. Vol. 463–464. Pp. 166–185. DOI: https://doi.org/10.1016/j.ins.2018.06.035

72. Winkler, R., Klawonn, F., Kruse, R. Problems of Fuzzy c-Means Clustering and Similar Algorithms with High Dimensional Data Sets. Challenges at the Interface of

Data Analysis, Computer Science, and Optimization. 2012. Pp. 79–87. DOI: https://doi.org/10.1007/978-3-642-24466-7_9

73. Møllersen, K., Dhar, S., Godtliebsen, F. On Data-Independent Properties for Density-Based Dissimilarity Measures in Hybrid Clustering. Applied Mathematics. 2016. Vol. 7. No. 15. Pp. 1674–1706. DOI: https://doi.org/10.4236/am.2016.715143

74. Донець В. В., Стрілець В. Є., Шевченко Д. О., Шматков С. І. Агентно-орієнтований метод кластеризації даних оптового дистриб'ютора. Вісник Харківського національного університету імені В. Н. Каразіна серія «Математичне моделювання. Інформаційні технології. Автоматизовані системи управління». 2023. URL: https://periodicals.karazin.ua/mia/article/view/22589

75. Lee, W.J., Mendis, G.P., Triebe, M.J. et al. Monitoring of a machining process using kernel principal component analysis and kernel density estimation. J Intell Manuf. 2020. No. 31 Pp. 1175–1189 DOI: https://doi.org/10.1007/s10845-019-01504-w

76. Ian Goodfellow, Yoshua Bengio and Aaron Courville. Softmax Units for Multinoulli Output Distributions. Deep Learning. MIT Press. 2016. URL: https://mitpress.mit.edu/9780262035613/deep-learning/

77. Strilets V. E., Shmatkov S. I., Ugryumov M. L. et al. Methods of machine learning in the problems of system analysis and decision making: monograph. Karazin Kharkiv National University. 2020. 195 p. ISBN 978-966-285-627-9.

78. Robert Miller, Ciaran Acton, Deirdre A. Fullerton, John Maltby and Jo Campling. Analysis of Variance (ANOVA). The SAGE Encyclopedia of Research Design. 2022. DOI: https://doi.org/10.1016/B978-0-12-397025-1.00319-5

79. Mario Cozzi, Severino Romano, Mauro Viccaro, Carmelina Prete, Giovanni Persiani. Wildlife Agriculture Interactions, Spatial Analysis and Trade-Off Between Environmental Sustainability and Risk of Economic Damage. The Sustainability of Agro-Food and Natural Resource Systems in the Mediterranean Basin. 2015. Pp. 209–224. DOI: http://dx.doi.org/10.1007/978-3-319-16357-4_14

80. Patrick Schober, Christa Boer and Lothar A. Schwarte. Correlation Coefficients: Appropriate Use and Interpretation. Anesthesia & Analgesia. 2018. No. 126. Pp. 1763–1768. DOI: https://doi.org/10.1213/ANE.0000000000002864

81. R. Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Adam Trischler and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. ArXiv. 2018. URL: https://arxiv.org/abs/1808.06670

82. Heewon Chung, Hoon Ko, Wu Seong Kang, Kyung Won Kim, Hooseok Lee, Chul Park, Hyun-Ok Song, Tae-Young Choi, Jae Ho Seo and Jinseok Lee. Prediction and Feature Importance Analysis for Severity of COVID-19 in South Korea Using Artificial Intelligence: Model Development and Validation. Journal of Medical Internet Research. 2021. No. 23. DOI: https://doi.org/10.2196/27060

83. Pereira, João P. B., Erik S. G. Stroes, Aeilko H. Zwinderman and Evgeni Levin. Covered Information Disentanglement: Model Transparency via Unbiased Permutation Importance. ArXiv. 2021. DOI: https://doi.org/10.1609/aaai.v36i7.20769

84. I.U. Ekanayake, D.P.P. Meddage, Upaka Rathnayake. A novel approach to explain the black-box nature of machine learning in compressive strength predictions of concrete using Shapley additive explanations (SHAP). Case Studies in Construction Materials. 2022. Vol. 16. DOI: https://doi.org/10.1016/j.cscm.2022.e01059

85. Yanqi Wu, Yisong Zhou. Hybrid machine learning model and Shapley additive explanations for compressive strength of sustainable concrete. Construction and Building Materials. 2022. P. 127298. DOI: https://doi.org/10.1016/j.conbuildmat.2022.127298

86. Daniel Lundstrom, Tianjian Huang, Meisam Razaviyayn. A Rigorous Study of Integrated Gradients Method and Extensions to Internal Neuron Attributions. ArXiv: Machine Learning. 2022. URL: https://arxiv.org/abs/2202.11912

87. Joseph Enguehard. Sequential Integrated Gradients: a simple but effective method for explaining language models. ArXiv: Computation and Language. 2023. URL: https://arxiv.org/abs/2305.15853

88. Zhongang Qi, S. Khorram and Fuxin Li. Visualizing Deep Networks by Optimizing with Integrated Gradients. CVPR Workshops (2019). 2019. DOI: https://doi.org/10.1609/AAAI.V34I07.6863

89. Kovacs, A. Iosub, M. Ţopa, A. Buzo and G. Pelz, A Gradient-based Sensitivity Analysis Method for Complex Systems. 2019 IEEE 25th International Symposium for Design and Technology in Electronic Packaging (SIITME). Cluj-Napoca, Romania. 2019. Pp. 333–338. DOI: https://doi.org/10.1109/SIITME47687.2019.8990871

90. Володимир Донець, Сергій Шматков. Методи аналізу інформативності в медичних системах підтримки прийняття рішень. Інформаційні технології та суспільство, січень 2024. DOI: https://doi.org/10.32689/maup.it.2023.5.1

91. Site of Medium: Determining the Number of Clusters: A Comprehensive Guide. URL: https://therised.medium.com/determining-the-number-of-clusters-a-comprehensive-guide-1a2441c5a526

92. Viktoriia Strilets, Nina Bakumenko, Volodymyr Donets, Serhii Chernysh, Mykhaylo Ugryumov,Tamara Goncharova. Machine Learning Methods in Medicine Diagnostics Problem. 16th International Conference on ICT in Education, Research and Industrial Applications. Integration, Harmonization and Knowledge Transfer. Volume II: Workshops, ICTERI 2020.

93. Site of AWS Developer Guide Amazon Machine Learning. URL: https://docs.aws.amazon.com/pdfs/machine-learning/latest/dg/machinelearning-dg.pdf

94. Emery D. Berger, Celeste Hollenbeck, Petr Maj, Olga Vitek, and Jan Vitek. On the Impact of Programming Languages on Code Quality: A Reproduction Study. ACM Trans. Program. Lang. Syst. 2019. Vol. 41. Issue 4. Article 21. P. 24. https://doi.org/10.1145/3340571

95. Site of Bocasay: Top Programming Languages for the Healthcare Sector. URL: https://www.bocasay.com/top-programming-languages-healthcare-sector/

96. R. A. Fisher. The use of multiple measurements in taxonomic problems. Annals of Eugenics. No. 7(2). Pp. 179–188. DOI: https://doi.org/10.1111/j.1469-1809.1936.tb02137.x

97. Edgar Anderson. The species problem in Iris. Annals of the Missouri Botanical Garden. 1936. No. 23(3). Pp. 457–509. DOI: https://doi.org/10.2307/2394164

98. B. German. Glass Identification. UCI Machine Learning Repository. 1987. https://doi.org/10.24432/C5WW2P

99. Stefan Aeberhard and M. Forina. Wine. UCI Machine Learning Repository. 1991. DOI: https://doi.org/10.24432/C5PC7J

100. William Wolberg, Olvi Mangasarian, Nick Street and W. Street. Breast Cancer Wisconsin (Diagnostic). UCI Machine Learning Repository. 1995. DOI: https://doi.org/10.24432/C5DW2B

101. Janosi Andras, Steinbrunn William, Pfisterer Matthias, Detrano Robert. Heart Disease. UCI Machine Learning Repository. 1988. DOI: https://doi.org/10.24432/C52P4X

102. Nestor Pereira. Using Machine Learning Classification Methods to Detect the Presence of Heart Disease. Technological University, Dublin. 2019. URL: https://arrow.tudublin.ie/scschcomdis/213/

103. Daher, M., Al Rifai, M., Kherallah, R. Y., Rodriguez, F., Mahtta, D., Michos, E. D., Khan, S. U., Petersen, L. A., and Virani, S. S. Gender disparities in difficulty accessing healthcare and cost-related medication non-adherence: The CDC behavioral risk factor surveillance system (BRFSS) survey. Preventive medicine. 2021. Vol. 153. 106779. DOI: https://doi.org/10.1016/j.ypmed.2021.106779

104. Margarida Cardoso. Wholesale customers. UCI Machine Learning Repository. 2014. DOI: https://doi.org/10.24432/C5030X

105. Donets V., Ugryumov M., Strilets V. A Measure Of Compactness For Fuzzy Clustering Based On Entropy. Науковий збірник праці міжнародної науково-технічної конференції «Комп'ютерне моделювання у наукоємних технологіях (КМНТ -2022)».

106. E.Trauwaert. On the meaning of Dunn's partition coefficient for fuzzy clusters. Vrije Universiteit Brussel, Pleinlaan 2, B-1050 Brussels, Belgium. 2003. DOI: https://doi.org/10.1016/0165-0114(88)90189-3

107. Mingrui Zhang, Wei Zhang, Hugues Sicotte, and Ping Yang. A New Validity Measure for a Correlation-Based Fuzzy C-means Clustering Algorithm. Conf Proc IEEE Eng Med Biol Soc. 2010. DOI: https://doi.org/10.1109/IEMBS.2009.5332582

108. J. C. Bezdek. Cluster validity with fuzzy sets. Journal of Cybernetics. 1974. vol. 3, pp. 58–73. DOI: https://doi.org/10.1080/01969727308546047

109. Y. Fukuyama, M. Sugeno. A new method of choosing the number of clusters for the fuzzy c-means method. in Proceedings of the Fifth Fuzzy System Symposium. Kobe, Japan. 1989.

110. K.-L. Wu, M.-S. Yang. A cluster validity index for fuzzy clustering. Pattern Recognition Letters. 2005. Vol. 26. No. 9. Pp. 1275–1291. DOI: https://doi.org/10.1016/j.patrec.2004.11.022

111. Iwendi, C., Bashir, A.K., Peshkar, A., Sujatha, R., Chatterjee, J.M., Pasupuleti, S., Mishra, R., Pillai, S.K., & Jo, O. COVID-19 Patient Health Prediction Using Boosted Random Forest Algorithm. Frontiers in Public Health. 2020. No. 8. DOI: https://doi.org/10.3389/fpubh.2020.00357

112. Computer program «Nonlinear estimation methods in the multicriterion problems of system's robust optimal designing and diagnosing under parametric apriority uncertainty (methodology, methods and computer decision support and making system)» («ROD&IDS»): Copyright registration certificate № 82875 / M.L. Ugryumov, Y.S. Meniaylov, S.V. Chernysh, K.M. Ugryumova (Ukraine). –Copyright and related rights. Official bulletin. Ministry of Economic Development and Trade of Ukraine. 2018. No. 51. P. 403.

Додаток А.

СПИСОК ПУБЛІКАЦІЙ ЗДОБУВАЧА ЗА ТЕМОЮ ДИСЕРТАЦІЇ

Статті у наукових фахових виданнях, що входять до міжнародних наукометричних баз

1. Viktoriia Strilets, Volodymyr Donets, Mykhaylo Ugryumov, Sergii Artiukh, Roman Zelenskyi, Tamara Goncharova. Agent-oriented data clustering for medical monitoring. Radioelectronic And Computer Systems. 2022. V. 2022. Issue 1. P. 103–114. Keywords: clustering; fuzzy clustering; agent-based approach; intraclass distance; medical diagnostic.

DOI: 10.32620/reks.2022.1.08 **(Scopus)**.

URL: http://nti.khai.edu/ojs/index.php/reks/article/view/reks.2022.1.08/0

(*Особистий внесок здобувача: розробка програмної реалізації методу мультиагентної нечіткої кластеризації з впровадженням різних метрик міжелементної відстані та проведення тестування на даних Ірисів Фішера та проведення кластеризації на даних медичного діагностування. Відповідні результати наведені в теоретичній та практичній частині роботи.*

Особистий внесок Viktoriia Strilets: аналіз існуючих методів кластеризації, розробка алгоритму методу нечіткої кластеризації, а також аналіз результатів тестування розроблених методів і моделей. Відповідні результати наведені в огляді існуючих методів кластеризації, та в розробленому методі нечіткої кластеризації, обговоренні та висновках.

Особистий внесок Mykhaylo Ugryumov: розробка математичної моделі методу нечіткої кластеризації, а також аналіз результатів тестування розроблених методів і моделей. Відповідні результати наведені в розробленому методі нечіткої кластеризації, обговоренні та висновках.

Особистий внесок Sergii Artiukh: збір даних медичного моніторингу захворювання на рак простати, їх аналіз та експертне виділення цільових класів по кожному запису пацієнтів. Відповідні результати наведені в описі набору даних.

Особистий внесок Roman Zelenskyi: збір даних медичного моніторингу захворювання на рак простати, їх аналіз та експертне виділення цільових класів по кожному запису пацієнтів. Відповідні результати наведені в описі набору даних.

Особистий внесок Tamara Goncharova: переклад статі на англійську, перевірка відповідності термінів, редагування матеріалів статті.)

2. Volodymyr Donets, Viktoriia Strilets, Mykhaylo Ugryumov, Dmytro Shevchenko, Svitlana Prokopovych, Liubov Chagovets. Methodology of the countries' economic development data analysis. Data Analysis. System Research and Information Technologies. 2023. V. 2023. Issue 4. P. 21–36.

Keywords: machine learning, digital development, fuzzy clustering, radial basis neural networks, logistic regression, analysis of variables informativeness.

DOI: 10.20535/SRIT.2308-8893.2023.4.02 **(Scopus)**.

URL: http://journal.iasa.kpi.ua/article/view/297208

(*Особистий внесок здобувача: впровадження розроблених методів мультиагентної нечіткої кластеризації, класифікації на основі штучної нейромережі з модифікованим методом навчання на даних економічного розвитку країн, що дало можливість сформувати методологію стратифікації елементів в комп'ютерних системах економічного моніторингу, відповідні результаті наведені в частині практичного застосування методу та висновків.*

Особистий внесок Viktoriia Strilets: перевірка наукової достовірності отримуваних результатів, перевірка тексту роботи, редагування. Відповідні результати є матеріалами публікації.

Особистий внесок Mykhaylo Ugryumov: постановка проблеми дослідження, розробка методології стратифікації елементів, математичне обґрунтування розроблених методів і моделей, відповідні результати наведені в методологічній частині роботи.

Особистий внесок Dmytro Shevchenko: огляд методів попередньої обробки даних, що були застосовані для підготовки даних до застосування методології.

Особистий внесок Svitlana Prokopovych: збір даних економічного моніторингу цифрового розвитку країн, їх аналіз та експертне виділення цільових класів. Відповідні результати наведені в описі набору даних.

Особистий внесок Liubov Chagovets: переклад статті на англійську мову, коректування використаних термінів.)

3. Volodymyr Donets, Dmytro Shevchenko, Maksym Holikov, Viktoriia Strilets, Serhiy Shmatkov. Application of a data stratification approach in computer medical monitoring systems. Eastern-European Journal of Enterprise Technologies. 2024. 2(9 (128), 6–16.

Keywords: data stratification, anomaly detection, fuzzy clustering, neural network, sensitivity analysis.

DOI: 10.15587/1729-4061.2024.298805 **(Scopus)**.

URL: https://journals.uran.ua/eejet/article/view/298805

(*Особистий внесок здобувача: впровадження розроблених методів і моделей стратифікації даних в комп'ютерній системі медичного моніторингу, що дало можливість перевірити ефективність поєднання мультиагентного методу кластеризації, методу класифікації та методів визначення інформативності на реальних даних медичного моніторингу, відповідні результаті наведені в частині практичного застосування методу та висновків, а також переклад матеріалів статті на англійську.*

Особистий внесок Dmytro Shevchenko: розробка методів попередньої обробки даних, а саме фільтрації вхідних даних методом ізольованого лісу та автокодувальника, відповідна частина наведена в роботі.

Особистий внесок Maksym Holikov: аналіз проблемної області та робіт присвяченій цій області, відповідна частина наведена в роботі.

Особистий внесок Viktoriia Strilets: перевірка наукової достовірності отримуваних результатів, перевірка тексту роботи, редагування.

Особистий внесок Serhiy Shmatkov: перевірка наукової достовірності отримуваних результатів, перевірка тексту роботи.)

Статті у наукових фахових виданнях України

4. Донець В. В., Стрілець В. Є., Шевченко Д. О., Шматков С. І. Агентно-орієнтований метод кластеризації даних оптового дистриб'ютора. Вісник Харківського національного університету імені В. Н. Каразіна серія «Математичне моделювання. Інформаційні технології. Автоматизовані системи управління». 2022. Том 1. № 55. Стор. 6–18.

Keywords: fuzzy clustering, multi-agent approach, data processing, Box-Cox transformation, PCA method, t-SNE method, autoencoder, Kullback-Leibler divergence, Mahalanobis distance, Manhattan distance

DOI: 10.26565/2304-6201-2022-55-01.

URL: https://periodicals.karazin.ua/mia/article/view/22589

(*Особистий внесок: впровадження розробленого методу мультиагентної нечіткої кластеризації на даних оптового дистриб'ютора, що має економічне походження. Відповідні результати наведені в практичній частині роботи*

Особистий внесок Стрілець В. Є.: підготовка набору даних для тестування, перевірка наукової достовірності отримуваних результатів, перевірка тексту роботи, редагування. Відповідні результатом є матеріалами публікації.

Особистий внесок Шевченко Д. О.: попередня обробка та аналіз даних з їх візуалізацією, результати наведені у відповідній частині роботи.

Особистий внесок Шматков С. І.: перевірка наукової достовірності отримуваних результатів, перевірка тексту роботи.)

5. Володимир Донець, Сергій Шматков. Методи аналізу інформативності в медичних системах підтримки прийняття рішень. Інформаційні технології та суспільство. Рік 2023. Том 5. № 11. Стор. 6–13.

Ключові слова: аналіз чутливості, аналіз даних, штучна нейронна мережа, інтегровані градієнти, медична діагностика, прийняття рішень.

DOI: 10.32689/maup.it.2023.5.1.

URL: https://journals.maup.com.ua/index.php/it/article/view/2922

(*Особистий внесок здобувача: аналіз існуючих методів інформативності, впровадження розробленого методу визначення загальної інформативності та*

адаптація градієнтного методу визначення поточної інформативності Відповідні результати наведені в практичній частині роботи

Особистий внесок Сергій Шматков: перевірка наукової достовірності отримуваних результатів, перевірка тексту роботи.)

Наукові праці, які засвідчують апробацію матеріалів дисертації

6. Viktoriia Strilets, Nina Bakumenko, Serhii Chernysh, Mykhaylo Ugryumov, Volodymyr Donets. Application of artificial neural networks in the problems of the patient's condition diagnosis in medical monitoring systems. Advances in Intelligent Systems and Computing. AISC 1113. Харків, 2020. Pp. 173–185.
DOI: https://doi.org/10.1007/978-3-030-37618-5_16 **(Scopus).**

7. Viktoriia Strilets, Nina Bakumenko, Volodymyr Donets, Serhii Chernysh, Mykhaylo Ugryumov,Tamara Goncharova. Machine Learning Methods in Medicine Diagnostics Problem. 16th International Conference on ICT in Education, Research and Industrial Applications. Integration, Harmonization and Knowledge Transfer. Volume II: Workshops, ICTERI 2020. Харків, 2020. – Pp. 89–101.

8. Бакуменко Н. С., Донець В. В., Шевченко Д. О., Одинець О. О., Угрюмов М. Л.. Методи кластеризації даних на основі інформаційних критеріїв. Науковий збірник праці міжнародної науково-технічної конференції «Комп'ютерне моделювання у наукоємних технологіях (КМНТ -2021)». Харків, 2021. С. 20–23.

9. Donets V., Ugryumov M., Strilets V. A Measure Of Compactness For Fuzzy Clustering Based On Entropy. Науковий збірник праці міжнародної науково-технічної конференції «Комп'ютерне моделювання у наукоємних технологіях (КМНТ -2022)». Харків, 2022.

Онлайн сервіс створення та перевірки кваліфікованого та удосконаленого електронного підпису

ПРОТОКОЛ
створення та перевірки кваліфікованого та удосконаленого електронного підпису

Дата та час: 16:54:42 17.06.2024

Назва файлу з підписом: PhD_work_Donets.pdf
Розмір файлу з підписом: 6.2 МБ

Перевірені файли:
Назва файлу без підпису: PhD_work_Donets.pdf
Розмір файлу без підпису: 6.2 МБ

Результат перевірки підпису: Підпис створено та перевірено успішно. Цілісність даних підтверджено

Підписувач: ДОНЕЦЬ ВОЛОДИМИР ВІТАЛІЙОВИЧ
П.І.Б.: ДОНЕЦЬ ВОЛОДИМИР ВІТАЛІЙОВИЧ
Країна: Україна
РНОКПП: 3559802558
Організація (установа): ФІЗИЧНА ОСОБА
Час підпису (підтверджено кваліфікованою позначкою часу для підпису від Надавача): 16:54:33 17.06.2024
Сертифікат виданий: КНЕДП АЦСК АТ КБ "ПРИВАТБАНК"
Серійний номер: 5E984D526F82F38F04000000DD3D530180192C05
Алгоритм підпису: ДСТУ 4145
Тип підпису: Удосконалений
Тип контейнера: Підписаний PDF-файл (PAdES)
Формат підпису: З повними даними для перевірки (PAdES-B-LT)
Сертифікат: Кваліфікований

Версія від: 2024.04.15 13:00