УДК 658.512.011.56:004.42
№ держреєстрації
Інв.номер

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ “ЛЬВІВСЬКА ПОЛІТЕХНІКА”
79013, м. Львів, вул. Ст. Бандери, 12
тел. (032) 258-20-25

ЗАТВЕРДЖУЮ

Проректор з наукової роботи
Національного університету
“Львівська політехніка”
_____ Демидов І.В.
«__» _____ 2021 р.

ЗВІТ
ПРО НАУКОВО-ДОСЛІДНУ РОБОТУ

за договором № 103/01/0025 від 30.04.2021 р

«Система підтримки прийняття рішень моделювання поширення вірусних інфекцій»

(заключний)

Грантова підтримка Національного фонду досліджень України
Конкурс «Наука для безпеки та суспільства»

Науковий керівник НДР
д.т.н., проф. (Н.Б.Шаховська)

2021

Результати роботи розглянуто НТР НУ « Львівська політехніка»,
протокол від 14 грудня 2021 р. № 12/2

Список авторів

Науковий керівник НДР,

Професор, доктор технічних наук Шаховська Н.Б.

(реферат, вступ, розділ 3, розділ 4)

Професор, доктор технічних наук Виклюк Я.І.

(розділ 1, розділ 5)

Доцент, кандидат технічних наук Мельникова Н.І.

(вступ, розділ 6, розділ 7)

Доцент, кандидат технічних наук Кривенчук Ю.П.

(розділ 8)

Доцент, кандидат технічних наук Ізонін І.В.

(розділ 2)

РЕФЕРАТ

Звіт про НДР «Система підтримки прийняття рішень моделювання поширення вірусних інфекцій»: 165 стор, містить 8 розділів, 49 рис., 21 табл.

Об’єкт дослідження: процеси моделювання поширення вірусних інфекцій.

Предмет дослідження: методи та засоби моделювання поширення Covid-19, приховані залежності даних.

Метою наукового проєкту є розроблення інформаційної системи на базі двох підходів до моделювання поширення інфекцій та визначення причин відхилень даних. Перший метод зосереджений на моделюванні перебігу хвороби та визначенні основних факторів впливу на конкретного індивіда на основі модифікованих асоціативних та секвенційних залежностей, другий - імітація в режимі реального часу просторового поширення інфекції на основі агентного підходу та клітинних алгоритмів.

Методи дослідження: теорія клітинних автоматів, агентні системи, методи видобування прихованих залежностей даних, ансамблі моделей машинного навчання.

У результаті виконання проєкту отримано таку наукову та науково-технічну продукцію:

1) Метод та алгоритм формування та аналізу Великих даних для побудови інформаційного портрету досліджуваного об’єкта за необхідності їх консолідації та паралельної обробки;

2) Метод та алгоритм передбачення складності перебігу хвороби на основі секвенційних асоціативних залежностей;

3) Метод пошуку залежностей у багатовимірних даних на основі ансамблю моделей (стеккінг) та з використанням регуляризації інформаційних ознак;

4) Імітаційна модель поширення захворюваності населення, спричиненої вірусом чи інфекцією, на основі мультиагентних систем

5) Архітектура інформаційної технології для моделювання поширення захворюваності із візуалізацією на карті, а також рекомендаційна лікарська система моделювання перебігу захворюваності індивідом.

Ключові слова: інформаційна система; COVID`19; імітаційне моделювання; ансамбль моделей

ANNOTATION

Research report "Decision support system for modeling the spread of viral infections": 165 pages, contains 8 sections, 49 figures, 21 tables.

Object of research: processes of modeling the spread of viral infections.

Subject of research: methods and tools for modeling the spread of Covid-19, hidden data dependencies.

The aim of the research project is to develop an information system based on two approaches to modeling the spread of infections and determining the causes of data deviations. The first method focuses on modeling the course of the disease and determining the main factors influencing a particular individual based on modified associative and sequential dependencies, the second - real-time simulation of spatial spread of infection based on agent approach and cellular algorithms.

Research methods: the theory of cellular automata, agent systems, methods of extracting hidden data dependencies, ensembles of machine learning models.

As a result of the project the following scientific and scientific-technical products were received:

1) Method and algorithm for the formation and analysis of Big Data to build an information portrait of the object under study, if necessary, their consolidation and parallel processing;

2) Method and algorithm for predicting the complexity of the disease on the basis of sequential associative dependencies;

3) The method of finding dependencies in multidimensional data based on the ensemble of models (stacking) and using the regularization of information features;

4) Simulation model of the spread of morbidity caused by a virus or infection, based on multi-agent systems

5) Architecture of information technology for modeling the spread of the disease with visualization on the map, as well as a recommendatory medical system for modeling the course of the disease by the individual.

Keywords: information system; COVID`19; simulation modeling; ensemble of models

ЗМІСТ

Реферат **3**

Annotation **4**

Зміст **5**

Вступ **7**

Розділ 1 Розроблення імітаційної моделі на основі теорії агентних систем **19**

1.1 Опис проблематики дослідження 19

1.2 Результати моделювання 31

1.3 Висновки до розділу 55

Розділ 2 Розроблення методу прогнозування поширення інфекцій на основі ансамблів моделей GRNN **58**

Розділ 3 Пошук ймовірнісних продукційних залежностей на основі моделі великих даних **72**

Розділ 4 Рекомендаційна система для запобігання поширення інфекції: превентивні заходи на рівні держави **85**

4.1 Огляд предметної області 85

4.2 Огляд аналогів 86

4.3 Методи та засоби дослідження 89

4.4 Висновки до розділу 98

Розділ 5 Метод визначення прогностичних чинників несприятливого перебігу SARS-CoV-2-асоційованої пневмонії **100**

5.1 Опис набору даних 100

5.2 Побудова класифікаційної моделі 106

5.3 Ансамбль класифікаційних моделей 110

Розділ 6 Модель передбачення складності перебігу захворювання на основі ієрархічного предиктора **113**

6.1 Побудова експерименту та опис набору даних 113

6.2 Вибір слабких предикторів 119

6.3 Розроблення ієрархічного предиктора.....120
Розділ 7 Розпаралелення пропонованих моделей.....122
7.1 Розпаралелення ієрархічного предиктора.....122
7.2 Прискорення паралельного алгоритму.....125
7.3 Ефективність паралельного алгоритму. 127
Розділ 8 Опис системи моделювання поширення захворювань, спричинених вірусами та інфекціями.....129
8.1 Архітектура системи.....129
8.2 Візуалізація поширення інфекції.....136
Висновки.....145
Література.....150
Додатки.....157

ВСТУП

В епоху активного розвитку цифрових технологій персоналізована медицина зосереджена на обробці неоднорідних медичних даних, що забезпечує цей процес керованим. Персоналізація набуває все більшого значення в науковому співтоваристві з багатьох причин, у тому числі з новими методами та алгоритмами оптимізації, які допомагають знизити медичні витрати та забезпечують якісну та ефективну медичну допомогу. Це дослідження зосереджено на вирішенні проблеми формалізації досліджуваного об’єкта та формалізації його умов під час лікування або реабілітації, що дозволить оптимізувати лікувальні процеси, проаналізувати індивідуальні особливості пацієнта та прогнозувати можливі персоналізовані рішення на основі медичної допомоги про оцінку стану здоров’я пацієнта.

Серед масштабних бід та катастроф, які супроводжують всю історію людства, на одному рівні з голодом, війнами та стихійними лихами, завжди перше місце займають епідемії, спричинені вірусами та інфекціями. Так, за даними Всесвітньої організації охорони здоров`я (ВООЗ), ГРІ становлять 60-70% від загальної захворюваності населення із тенденцією до розвитку ускладнень та хронізації процесу. У зв`язку із надзвичайною мінливістю збудника ГРІ й досі залишається некерованою інфекцією. Іншим прикладом є коронавірусне захворювання (COVID-19), інфекційної хвороби, яку спричиняє новий штам коронавірусу. Майже один мільйон людей у всьому світі захворіли на COVID 19. З них понад 700 000 хворіють, понад 200 000 були вилікувані. Від хвороби померло понад 80 000 людей. Загалом захворювання було виявлено у 203 країнах.

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

- мінливість штамів,
- характер взаємодії,

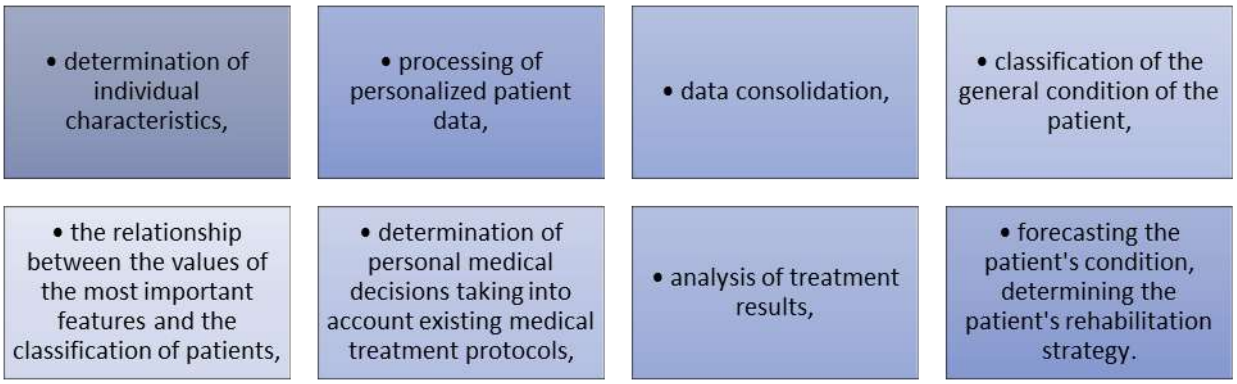
- особливості території поширення: кліматичні умови, розвиток інфраструктури та сполучення, якість медичного обслуговування, притаманні цій місцевості хронічні захворювання, політична ситуація тощо.

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

- багатокритеріальністю: тип поширення (епідемічне поширення, контрольоване поширення у легкій формі захворювання), початкові дані поширення, територія поширення..
- залежністю від часу,
- інтервалом моделювання,
- різнотипністю вхідних даних.

Водночас, важливим є персоналізація лікування.

Визначення індивідуальних характеристик, необхідних для вирішення проблеми персоналізації, залежить від ключових факторів ідентифікації об’єкта. Для формалізованого представлення досліджуваного об'єкта в медицині розглянуто основні параметри його загального стану з певними характеристиками.



Основні етапи обробки інформації пацієнтами

Орієнтовані на пацієнта дані - це дані, які пацієнт або його родичі повідомляють лікарю. Якість даних, орієнтованих на пацієнта, залежить від рівня грамотності здоров’я пацієнта, тоді як якість даних пацієнтів залежить від якості та чутливості пристроїв та технологій, що беруть участь у процесі збору [1]. Дані, орієнтовані на пацієнта, мають ряд переваг, таких як можливість управління медичними даними, підтримка подвійних перевірок медичних досліджень для зменшення частоти хибнопозитивних та хибнонегативних результатів [42].

Дані об’єкта дослідження, що надходять з різних джерел, використовуються для медичних та оздоровчих цілей. Однак дані, що підтверджують надання спеціальної медичної допомоги, переважно зосереджені і можуть бути зібрані як частина скринінгу на наявність захворювань або діагностичних процесів. Результати скринінгу або діагностики слід перевірити (принципи подвійної перевірки) медичними оглядами. Як з’ясовано в [1], зміна парадигми скринінгу на наявність захворювань має важливе значення, з одного боку, для збору достатньої кількості індивідуальних та точних даних для лікування пацієнтів, з метою попереднього виявлення, прогнозувати будь -які проблеми зі здоров'ям людини. З іншого боку, зміна парадигми допоможе зменшити медичні витрати та зменшити захворюваність та смертність. Крім того, зміна парадигми відкриє шлях до персоналізованої медицини.

Персоналізовану медицину (ПМ) можна вважати продовженням традиційних підходів до розуміння та лікування хвороб з більшою точністю. Профіль варіації генів пацієнта може керувати вибором препаратів або протоколів лікування, які мінімізують шкідливі побічні ефекти або забезпечують більш успішні результати. ПМ також може вказувати на сприйнятливість людини до певних захворювань до їх прояву, що дозволяє лікарям та пацієнтам розробити план моніторингу та профілактики.

У епоху цифрових технологій персоналізована медицина повинна підтримуватися малими та великими даними (і повинна стати процесом, керованим даними) і використовувати технологію штучного інтелекту для

підтримки запобігання ризикам, прогнозування та виявлення, а також медичного втручання. ПМ набуває все більшого значення в науковій спільноті (медицині, фельдшері та біоінженерії) з багатьох причин, особливо з огляду на більшу ефективність та ефективність медичної допомоги, зниження витрат на медичні послуги тощо.

На першому етапі виконання дослідження розроблено **і**мітаційну модель поширення інфекцій на основі теорії агентних систем та метод прогнозування поширення інфекцій на основі ансамблів моделей.

Цілями етапу є: динаміка (просторова та статистична) таких показників, як кількість інфікованих, здорових, тих, що перехворіли, летальних випадків та інші

Заплановані завдання етапу.

- Розроблення імітаційної моделі на основі теорії агентних систем
- Розроблення методу прогнозування поширення інфекцій на основі ансамблів моделей
- Пошук ймовірнісних продукційних залежностей на основі моделі великих даних

Для виконання першого завдання в роботі була вдосконалена SIR модель, що реалізується за допомогою мультиагентної системи на основі рухомих клітинних автоматів. В роботі запропоновані шляхи вдосконалення правил взаємодії та поведінки агентів. Розроблені методи співставлення параметрів SIR моделі з реальними географічними, соціальними та медичними показниками. Це дало змогу моделювати просторове поширення вірусу COVID-19 як окремо взятому реальному місті так і цілій країні що складається з окремих регіонів які взаємодіють між собою за допомогою транспортного сполучення з урахуванням таких чинників як: робота громадського транспорту, супермаркетів спортивних залів, церков, парків. Розроблена модель дозволяє також оцінити вплив карантину, обмеження транспортного сполучення між регіонами, враховувати такі чинники як інкубаційний період, масковий режим, дотримання безпечної віддалі між людьми, тощо. В роботі було проведено ряд експериментів що дали

змогу оцінити як вплив окремих заходів зупинення пандемії так і комплексне їх застосування. Запропонована методика співставлення комп'ютерного часу та динамічних параметрів моделі з реальними даними, дала змогу оцінити ефективність дії влади щодо зупинення пандемії в Чернівецькій області, Україна. Також було проведено симуляція поширення пандемії в таких країнах як Словаччина, Туреччина, Сербія. Проведені розрахунки показали високу точність моделі прогнозів з реальними даними.

Для виконання другого завдання в роботі описано результати експериментальних досліджень щодо визначення ефективності застосування трьох бустингових ансамблів для підвищення точності розв'язання задач апроксимації. Усі вони базується на використанні двох послідовно з'єднаних нейронних мереж узагальненої регресії. Основною різницею між ними є використання різних процедур формування вихідного результату роботи кожного із досліджуваних ансамблів. Подано структурно-функціональні схеми їх роботи, описано основні кроки їх застосування. Апробацію роботи досліджуваних ансамблів проведено із використанням реального набору даних якості повітряного середовища. Описано процедури та результати підбору оптимальних параметрів для ефективного використання усіх досліджуваних ансамблів. Подано ряд числових показників їх роботи щодо точності апроксимації та часових затримок в режимі застосування. Експериментальним шляхом встановлено найвищу точність роботи ансамблю з двох мереж GRNN, де формування вихідного сигналу відбувається шляхом нелінійного розширення виходів обох мереж ансамблю із використанням полінома Колмогорова-Габора із подальшим застосуванням нейроподібних структур МПГП лінійного типу. Зроблено висновки та описано перспективи подальших досліджень, зокрема щодо зменшення часу роботи усіх запропонованих ансамблевих методів.

Для виконання третього завдання в роботі розроблено метод пошуку ймовірнісних продукціних залежностей та їх використання для відновлення пропусків даних, а також пошуку прихованих залежностей даних. Особливістю

методу є обробка Великих даних. Новизна запропонованого методу обчислення відсутніх даних полягає в обробці структурованих та напівструктурованих даних на основі ієрархії об’єктів, а також набору функціональних залежностей та розробки правил асоціації. Ця проблема дуже важлива для інтерфейсів великих даних, оскільки більша частина інформації доступна в напівструктурованому вигляді. Запропонований алгоритм створює додаткові значення даних за допомогою доменних та функціональних залежностей на основі декількох методів обчислення та додає ці значення до наявних навчальних даних. Правильність обчислених значень перевіряється на класифікаторі, побудованому на вихідному наборі даних. Проведено порівняння методів аналізу та моделювання статистичних процесів за якісними критеріями.

Також проаналізовано дані існуючими методами, зокрема здійснено кластеризацію та класифікацію даних з відкритих наборів. Розглядаються рекомендаційні правила, згенеровані з ансамблю методів машинного навчання, а саме регресійне дерево та кластеризація. Для навчанні моделі використовувалися об’єднані дані, які складалися з Оксфордського урядового механізмі реагування на COVID-19 та Європейського центру з профілактики та контролю захворювань Covid-19 Набори даних використовувались із даними від 01 січня по 04 жовтня 2020 року. Висновки та результати дослідження можуть бути корисними для прийняття рішення вироблення адекватної державної політики щодо зменшення поширення нових випадків Covid-19.

Для виконання четвертого етапу вперше розроблено модель умовного багатовимірного простору для персоналізованих медичних даних. Це дозволяє прогнозувати значення цільових змінних для певних вимірів простору з вищою точністю. Розроблено новий ансамбль прогнозування тривалості стаціонарного лікування на основі ієрархічного предиктора з 5-ти кратною валідацією результатів на основі двоетапної процедури аналізу даних: кластеризації та прогнозування для кожного відокремленого кластера. Він дає в 1,47 разів кращу точність прогнозування, ніж найкращий слабкий предиктор (персептрон з 12

одиницями в одному прихованому шарі). Визначено складність розпаралеленого ансамблю та ансамблю без розпаралелення і розраховано залежність зменшення часової складності в залежності від кількості ядер. Це дає змогу застосувавати розроблені моделі для опрацювання великих наборів даних. Розроблено новий метод подвоєння входів на основі використання класичної ітеративної RBF нейронної мережі, який забезпечує підвищення точності розв'язання задачі опрацювання коротких наборів даних.

Сформовано набір даних з інформацією про особисті параметри 51 пацієнта, що дало змогу апробувати розроблені методи та алгоритми (https://doi.org/10.6084/m9.figshare.14865411.v1).

Для виконання п'ятого етапу розроблено архітектуру інформаційної системи моделювання поширення інфекцій та вірусів різного роду з використанням сховища моделей, яке складається із попередньо розроблених моделей машинного мавчання.

За результатами проекту підготовлено таку наукову продукцію:

- **Shakhovska, N., Melnykova**, N., & Chopiyak, V. (2022). An Ensemble Methods for Medical Insurance Costs Prediction Task. CMC-COMPUTERS MATERIALS & CONTINUA, 70(2), 3969-3984. **Q=1**
- **Melnykova, N.; Shakhovska**, N.; Melnykov, V.; Melnykova, K.; Lishchuk-Yakymovych, K. Personalized Data Analysis Approach for Assessing Necessary Hospital Bed-Days Built on Condition Space and Hierarchical Predictor. *Big Data Cogn. Comput.* 2021, *5*, 37. https://doi.org/10.3390/bdcc5030037 **Q=2**
- **I. Izonin**, R. Tkachenko, V. Verhun et al., An approach towards missing data management using improved GRNN-SGTM ensemble method, https://doi.org/10.1016/j.jestch.2020.10.005 **Q=1**
- Kaminskyy, R., **Shakhovska, N**., Kryvinska, N., & Younas, M. (2021). Dendrograms-based disclosure method for evaluating cluster analysis in the IoT domain. *Computers & Industrial Engineering*, *158*, 107402. **Q=1**

- **Yaroslav Vyklyuk**, Mykhailo Manylich, Miroslav Škoda, Milan M. Radovanović, Marko D. Petrović, Modeling and analysis of different scenarios for the spread of COVID-19 by using the modified multi-agent systems – Evidence from the selected countries, Results in Physics, 2020, 103662, ISSN 2211-3797, https://doi.org/10.1016/j.rinp.2020.103662 **Q=2**
- **Shakhovska, N., Izonin, I., & Melnykova, N**. (2021). The hierarchical classifier for COVID-19 resistance evaluation. Data, 6(1), 6. Data **Q=2**
- Yakovyna, V., & **Shakhovska, N**. (2021). Modelling and predicting the spread of COVID-19 cases depending on restriction policy based on mined recommendation rules. Mathematical Biosciences and Engineering: MBE, 18(3), 2789-2812. **Q=2**
- **Izonin, I.,** Tkachenko, R., **Shakhovska, N**., & Lotoshynska, N. (2021). The Additive Input-Doubling Method Based on the SVR with Nonlinear Kernels: Small Data Approach. *Symmetry*, *13*(4), 612 **Q=2**
- **Ivan Izonin,** Roman Tkachenko, Ivanna Dronyuk, Pavlo Tkachenko, Michal Gregus, Mariia Rashkevych. Predictive modeling based on small data in clinical medicine: RBF-based additive input-doubling method. Mathematical Biosciences and Engineering, 2021, 18(3): 2599-2613. doi: 10.3934/mbe.2021132 **Q=2**

- **Ivan Izonin**, Roman Tkachenko, Pavlo Vitynskyi, Khrystyna Zub, Pavlo Tkachenko, Ivanna Dronyuk Stacking-based GRNN-SGTM Ensemble Model for Prediction Tasks. Proceedings of the International Conference on Decision Aid Sciences and Applications, 8 – 9 November 2020, Kingdom of Bahrain, pp. 326-330.
- **Ivan Izonin**, Roman Tkachenko, Liliia Ryvak, Khrystyna Zub, Mariia Rashkevych, Olena Pavliuk Addressing Medical Diagnostics Issues: Essential Aspects of the PNN-based Approach Ceur, Vol-2753 , 2020, 209-218

- **Nataliia Melnykova, Natalya Shakhovska**, Volodymyr Melnykov, Mykola Logoyda, Yulia Peleshchak The Problem of Analysing the Relationships between Individual Characteristics of Individuals with COVID`19 Ceur, Vol-2753 , 2020 473-482
- Vitaliy Yakovyna, **Natalya Shakhovska**, Khrystyna Shakhovska, Jaime Campos. Recommendation Rules Mining for Reducing the Spread of COVID-19 Cases. *Ceur*, Vol 2753, p.219-229.
- **Ізонін І. В**., Ткаченко Р. О., Зуб Х. В. Застосування бустингу в ансамблях нейронних мереж узагальненої регресії для підвищення точності розв'язання задач апроксимації. На шляху до індустрії 4.0: інформаційні технології, моделювання, штучний інтелект, автоматизація : монографія / кол. авт. : В .яб. Артеменко, Л. В. Артеменко, О. В. Артеменко [та ін.]; за заг. ред. В. С. Котлика. - Одеса : Астропринт, 2021. - с. 394-409. ISBN 978-966-702-2

- **Nataliya Shakhovska**. Method for Missing Data Imputation in Big Data technologies // Інформаційні технології, економіка та право: стан і перспективи розвитку : матеріали Міжнародної науково-практичної конференції ІТЕП-2020 (Чернівці, 10 листопада 2020 р.). – 2020. – С. 67–68.
- Вергун В. Р., Вітинський П. Б., **Ізонін І. В.,** Ткаченко Р. О., Кобзар* Н. О. Дослідження ефективності використання бустингових ансамблів штучних нейронних мереж узагальненої регресії для розвязання задач прогнозування // Інформаційні технології і автоматизація – 2020 : збірник тез доповідей XIII Міжнародної науково-практичної конференції, Одеса, 22–23 жовтня 2020 р. – 2020. – С. 282–283.
- **Vyklyuk Y.,** Nevinskyi D. *Modelling And Analysis of Different Scenarios for the Spread of Covid-19 by Using bhe Modified Multi-Agent Systems - Evidence from the Selected Countries* // Міжнародна наукова конференція,

присвячена 75-річчю кафедри диференціальних рівнянь та 85-річчю від дня народження Михайла Павловича Ленюка, 28 - 30 жовтня 2021 р., Чернівці: матеріали конференції. - Чернівці, 2021. - с. 193-194

- **Ізонін І. В**., Ткаченко Р. О., Ривак Л. А. Використання імовірнісних нейронних мереж в задачах медичної діагностики // Інформаційні технології, економіка та право: стан і перспективи розвитку : матеріали Міжнародної науково-практичної конференції ІТЕП-2020 (Чернівці, 10 листопада 2020 р.). – 2020. – С. 106–107.
- **Nataliya Shakhovska**. Method for Missing Data Imputation in Big Data technologies // Інформаційні технології, економіка та право: стан і перспективи розвитку : матеріали Міжнародної науково-практичної конференції ІТЕП-2020 (Чернівці, 10 листопада 2020 р.). – 2020. – С. 67–68.
- Kunanets N., Pasichnyk V., Kunanets O., **Kryvenchuk Y., Vyklyuk Y**. The modern information technologies in the process of raising the virus pandemics. Інформаційні технології, економіка та право: стан і перспективи розвитку: матеріали Міжнародної науково-практичної конференції ІТЕП-2020 (Чернівці, 10 листопада 2020 р.). – 2020. – С. 67.
- **Ізонін І. В**., Ткаченко Р. О., Підкостельний Р. Р. Методи штучного розширення коротких вибірок даних // Інформаційні технології, економіка та право: стан і перспективи розвитку : матеріали Міжнародної науково-практичної конференції ІТЕП-2020 (Чернівці, 10 листопада 2020 р.). – 2020. – С. 104–105.
- **Ivan Izonin,** Roman Tkachenko, Michal Gregus, Khrystyna Zub, Nataliia Lotoshynska. “Input Doubling Method based on SVR with RBF kernel in Clinical Practice: Focus on Small Data”, Elsevier Journal Procedia Computer Science, Special Issue on “The 12th International Conference on Ambient Systems, Networks and Technologies (ANT) / The 4th International Conference on Emerging Data and Industry 4.0 (EDI40) / Affiliated Workshops”, vol. 184, pp. 606–613, 2021. ISSN 1877-0509, DOI: 10.1016/j.procs.2021.03.075

- **Ivan Izonin,** Roman Tkachenko, Michal Gregus ml., Khrystyna Zub, Pavlo Tkachenko. “A GRNN-based Approach towards Prediction from Small Datasets in Medical Application”, Elsevier Journal Procedia Computer Science, Special Issue on “The 12th International Conference on Ambient Systems, Networks and Technologies (ANT) / The 4th International Conference on Emerging Data and Industry 4.0 (EDI40) / Affiliated Workshops”, vol. 184, pp. 242–249, 2021. ISSN 1877-0509, DOI: 10.1016/j.procs.2021.03.075
- **Natalya Shakhovska**, Ihor Darmoriz, **Yaroslav Vyklyuk, Yurii Kryvenchuk** and Pavlo Pukach Visualization of the Epidemics Forecasting Results. IDDM 2021

Свідоцтво про авторське право:

- А. с. 103422 "Комп’ютерна програма “Гібридна система обчислювального інтелекту для підвищення точності розвязання задач класифікації” / **І. В. Ізонін,** Л. А. Ривак, Р. О. Ткаченко; заявн. І. В. Ізонін. – с202101616; заявл. 19.03.2021; опубл. 24.03.2021

  А. с. 103447 "Комп’ютерна програма “Бустинговий ансамбль з двох нейронних мереж узагальненої регресії” / **І. В. Ізонін**, Н. О. Кобзар, Р. О. Ткаченко, О. М. Павлюк; заявн. І. В. Ізонін. – с202101202; заявл. 01.03.2021; опубл. 25.03.2021

Захищено такі бакалаврські роботи:

- Яковлєва Катерина. Веб-сервіс для проведення первинної самодіагностики на COVID-19 та інформування щодо пандемії
- Славчаник Олеся. Передбачення респіраторних захворювань за допомогою алгоритмів машинного навчання
- Дарморіз Ігор. Візуалізація прогнозу поширення епідемій
- Підкостельний Роман. Дослідження методів штучного розширення коротких наборів даних.

Авторський внесок статей у журналах з Q1-Q2 вказано безпосередньо в тексті

РОЗДІЛ 1 РОЗРОБЛЕННЯ ІМІТАЦІЙНОЇ МОДЕЛІ НА ОСНОВІ ТЕОРІЇ АГЕНТНИХ СИСТЕМ

1.1 Опис проблематики дослідження

Протягом останнього пів року світом поширилась епідемія COVID-19 [1]. Цей вірус може передаватися від людини до людини повітряно крапельним шляхом, аналогічно до інших вірусів. Особливістю його є велика стійкість до зовнішніх факторів. Він осідає на пластикові та металічні поверхні та може існувати в такому стані декілька днів [2, 3].

Аналогічно до інших вірусів COVID-19 характеризується інкубаційним періодом, протягом якого інфікована людина не відчуває ознак хвороби, однак може інфікувати інших. Цей період був оцінений в 5.1 дні[4-6].

На сьогодні не існує медичних препаратів що дозволяють лікувати цей вірус. Тому основними засобами боротьби з цією хворобою є комплекс заходів що зменшує імовірність передачі вірусу від 1 людини до іншої. То таких заходів можна віднести використання масок, карантин, зупинка громадського транспорту та міжміського сполучення, блокування районів та населених пунктів де є спалахи хвороби, перекриття кордонів, обмеження роботи супермаркетів, магазинів, базарів, музеїв, ресторанів, парків, спортивних залів [7]. Однак вплив таких заходів тяжко математична оцінити, адже вони носять більшості випадків рекомендаційний характер та їх виконання залежить індивідуально від людей.

Перед початком організації чи плануванням будь-яких заходів запобігання епідемії доцільним є зімітувати та оцінити їх наслідки. Іншими словами програти в режимі реального часу можливі шляхи поширення вірусної інфекції без залучення безпосередньо живих людей. Це дозволить визначити оптимальну стратегію дій влади з мінімізацією критичних втрат серед населення та запобігання спаду економіки в цілому.

Метою роботи було, по-перших, встановлення взаємозв'язку між емпіричними параметрами існуючих моделей-симуляторів поширення епідемій та реальним параметрами соціальних систем, що дасть можливість аплікації їх до моделювання реальних територій (міст, регіонів, країн). По-других, вдосконалення моделей шляхом врахування специфіки поширення COVID-19 та особливостей функціонування різних територій, що дасть змогу змоделювати такі процеси, як наявність карантину, маскового режиму, вплив громадського транспорту, супермаркетів, шкіл, церков та інше, а також оцінити вплив різних заходів на поширення епідемії. По-третє застосувати розроблену модель для оцінки поширення COVID-19 та впливу різних факторів запобігання цій епідемії для реальних міст, регіонів та країн.

Огляд останніх досліджень

Дана ситуація може бути віднесена до класу задач моделювання поведінки з метою прийняття як групових так і централізованих рішень, а також багатоагентної оптимізації.[8] Для моделювання просторового поширення вірусної інфекції зручно скористатись моделями епідемій (SIR-model).

Існує 2 підходи до моделювання таких ситуацій:

- Аналітичний, що передбачає розв'язок систем диференціальних рівнянь. Перевагою цього методу є швидкість розв'язання (при умові створеного програмного засобу). Недоліком є те, що результат, як правило, представляє собою статистичні дані по динаміці інфікування та не дозволяє просторово зобразити поширення інфекції і є також складний до удосконалення моделі [8], [18]. Сьогодні існує велика кількість модифікацій цих моделей. Зокрема SEIR модель, що використовує чотири класи населення: моделі сприйнятливого впливу та відновлення інфекції [11]. Ці моделі найбільш підходять для опису поширення захворювань з тривалим інкубаційним періодом, що стосується спалаху COVID-19. В роботі [12] досліджувався динамічний вплив різних факторів на SEIR модель. Аналіз чутливості цих моделей був проведений

в роботах [13, 14]. В роботі [15] досліджувався вплив різних заходів на поширення COVID-19. Для оцінки різних заходів контролю в Китаї була розроблена спеціальна модель, що представлена в роботі [16]. Як відомо, багато пацієнтів не звертаються до лікаря, або відсутність тестів не дає змогу оцінити реальну статистику захворюваності. Ця ситуація була змодельована в роботі [17].

- Дискретний, що базується на теорії мультиагентних систем, може бути реалізований або клітинними автоматами або методами молекулярної динаміки. Клітинні автомати добре моделюють поширення таких просторових процесів, як пожежі, повені та інше. Є відносно швидкими. Однак ними не можливо змоделювати поведінку та рух окремого агента. Підходи молекулярної динаміки передбачають, що кожна окремо взята людина представляє агента, який переміщується в межах обмеженої області та за певними правилами взаємодіє з іншими агентами [19, 22]. Цей метод є найбільш адекватний, однак повільний, і показує найкращі результати як по візуалізації поширення інфекції, так і по статистичним результатам. Недоліком є залежність результату від генератора випадкових чисел що призводить до різних результатів симуляцій при однакових початкових умовах. Тому в якості результуючої моделі треба розуміти ансамбль моделей, що можуть бути розпаралелені на різних ядрах чи комп’ютерах. Перевагою цих моделей теж є легка адаптованість, можливість вносити зміни в ході проведення експерименту в режимі реального часу та здатність вдосконалювати систему моделювання без великого втручання в код програми[9, 10]. Існує багато різних модифікацій цих моделей зокрема: моделі соціальної сили дають більш точний опис руху пішоходів [23]. У цих моделях кожен агент прагне досягти бажаної швидкості, зберігаючи певну відстань від інших агентів. Вплив сусідніх індивідів на збудника описується за допомогою соціально-психологічної сили. Цей підхід може бути використаний для

опису складної динаміки натовпу, яка включає декілька взаємодіючих людей. Моделі соціальної сили можуть бути застосовані для вивчення динаміки поширення інфекції в натовпі та в реалістичних умовах. Наприклад, для опису поширення інфекції в літаках була використана модель соціальної сили [24]. Хоча більш складні моделі на основі агентів можуть дати результати, які добре узгоджуються з більш простими моделями ODE [25]. Для моделювання великих систем було розроблено спеціальні тип багато масштабного моделювання що має за мету об'єднати механізми різних масштабів простору та часу [26]. В роботі було розглянута [27] гібридна дискретно-безперервна модель агентів. Такі моделі використовуються для моделювання біологічних систем [28 - 32].

Сучасні комп'ютерні потужності дозволяють досить швидко змоделювати системи, що містять декілька тисяч агентів на основі молекулярної динаміки. Тому останній підхід є найбільш актуальний до нашого завдання.

Класична SIR модель

Згідно класичної мультиагентної SIR-моделі генерується множина агентів, що може вільно переміщатись в певному плоскому просторі, що імітує місто. Кожен агент може мати три стани: S – здоровий та сприйнятливий до хвороби, I – інфікований, що може заразити інших, R – видужавший, або не сприйнятливий до хвороби. Для моделювання необхідні ініціалізувати наступні дані:

1. Площа та форма території (S)
2. Кількість агентів (N)
3. Середня швидкість пересування ($\bar{v}$)
4. Початкові положення інфікованих та неінфікованих агентів ($L = l_{i=1,N}(x, y)$)
5. Дистанція менше якої можливе інфікування(d)
6. Імовірність інфікування, якщо здорова людина потрапляє на близьку відстань до інфікованої людини (p)

7. Відсоток людей, що несприйнятливі до хвороби або перехворіли (P_R).
8. Тривалість хвороби (T_{il})
9. Рівень смертності (D)

Одним з найважливіших аспектів імітації є співставлення комп'ютерного часу та лінійних розмірів з відповідними даними реального світу. При імітації реального міста такі величини як S та N – задаються емпірично. А всі інші параметри розраховуються шляхом масштабування реальних географічних масштабів та статистичних даних населеного пункту. Зокрема, якщо населення міста складає P_s, а його площа – S_s, тоді в середньому на N людей населеного пункту припадає площа:

$$S_{sim} = \frac{N}{P_s} S_s \tag{1.1}$$

Якщо в імітаційній моделі представити цю площу у вигляді квадрата зі стороною $L = 1$, то йому відповідатиме реальний розмір:

$$L_{sim} = \sqrt{S_{sim}} \tag{1.2}$$

Тоді відстань необхідна для інфікування визначатиметься як:

$$d = d_{real} \frac{L}{L_{sim}} \tag{1.3}$$

де d_{real} – реальна відстань в метрах необхідна для інфікування людей. Ця величина є відомою в медичних довідниках та залежить від вірусу, який моделюється.

Аналогічно розраховується швидкість переміщення людей. Вважаючи, що в середньому швидкість людини становить $\bar{v}_{real} = 4$ км/год в імітаційній моделі їй відповідатиме швидкість:

$$\bar{v} = \bar{v}_{real} \frac{L}{L_{sim}} \tag{1.4}$$

Наступним кроком визначається тривалість однієї ітерації. Вона має бути такою, щоб за одну ітерацію людина не могла подолати віддаль більшу за d, адже

у такому випадку агенти, які рухаються назустріч проскочуть один одного без взаємодії. Занадто малий час – суттєво сповільнює розрахунки.

$$dt < d/\bar{v} \tag{1.5}$$

Дана величина дає змогу привести до комп'ютерного часу (кількості ітерацій) реальну тривалість хвороби:

$$T_{il} = T_{real}/dt \tag{1.6}$$

де T_{real} – реальна середня тривалість хвороби.

Величини p та P_R можуть бути отримані з аналізу статистичних медичних даних. І вони для більшості хвороб є відомими.

Для аналізу поширення інфекції на основі мульти-агентних систем створюється або віртуальна квадратна площа, або за формою ідентична реальній. Далі за допомогою генератора випадкових чисел з нормальним розподілом генеруються положення агентів (L) в межах цієї площі, їх початкові швидкості та статуси:

$$L = l_{i=1,N}(x, y, v_x, v_y, S_{SIR}, r) \tag{1.7}$$

$$\sqrt{{v_x}^2 + {v_y}^2} \leq \bar{v}\ . \tag{1.8}$$

де x, y – координати агента, v_x, v_y – швидкості агента, S_{SIR} – статус агента, r – кількість людей, які інфіковав даний агент ($r = 0$ на етапі ініціалізації).

Як правило всі агенти на першому кроці ітерації мають статус S. Одному чи декільком агентам, що розташовані у конкретно визначених положеннях задається статус I (залежить від початкових умов реального запалення). Залежно від типу хвороби, деякі агенти можуть на першій ітерації мати статус R, що інтерпретуються як несприйнятливі для хвороби (мають імунітет). Наступним кроком є імітація руху людей. Для цього згідно розрахованого часу dt на кожній ітерації агент переміщується на віддаль $s = (x + v_x dt, y + v_y dt)$).

Розрахувавши положення всіх агентів на наступний крок ітерації перераховуються властивості агентів згідно наступних правил:

1. Агент знаходиться біля або перетнув межу моделювання. В цьому випадку можливо або зімітувати дзеркальне відбивання (змінити знак однієї з компонент швидкості на протилежний), або направити агента у зворотньому напрямку (змінити знаки обох компонент швидкостей на протилежні)
2. Неінфікований агент потрапив в зону зараження інфікованого агента. В такому випадку є 2 можливі ситуації:
    a. Якщо агент має статус R тоді ніяких дій з агентом не робиться
    b. Якщо агент має статус S, то згідно генератора випадкових чисел визначається число $z = [0,1]$. Якщо $z < p$ людина вважатиметься інфікованою та їй присвоюється статус I, а величина r інфікованого агента збільшується на 1. Розраховується майбутній час хвороби конкретного агента: $T_{il}^{i} = T_{il} + 2t_{il} \cdot (z_i - 0.5)$, де z_i – випадкове число в межах $[0,1]$, t_{il} – можливий діапазон тривалості хвороби, що є відомою величиною для різних хвороб.
3. Люди зіткнулися. В такому випадку найбільш адекватно ігнорувати цю ситуацію і дозволити пройти «наскрізь». Адже в реальному житті людина скоріше ухилиться від зіткнення та продовжить свій рух у попередньому напрямку. Однак можна зімітувати «пружнє зіткнення» коли компоненти швидкостей змінюються на протилежні знаки. Така ситуація можлива при великому натовпі людей, скажімо біля сцени, магазині, транспорті чи церкві.
4. Якщо статус агента I та час хвороби $t_{il}^{i} \geq T_{il}^{i}$ визначається випадкове число: $P_d = [0,1]$. Якщо $P_d > D$ агент виліковується, тобто змінює стан на R, в іншому випадку помирає та вилучається з розрахунку.

Імітація триває до тих пір, поки не залишиться агентів із статусом I.

1.1. Модифікація SIR моделі

Ізоляція.

Ізоляція передбачає створення умов при яких хворий агент не може інфікувати інших. Це можна реалізувати або додатковим статусом та перевіркою на кожному кроці ітерації, або створити іншу область моделювання (ізолятор, домашній карантин) в яке перемістити всіх хворих агентів, та дати змогу їм рухатись та взаємодіяти за класичними правилами. В момент виздоровлення агент переміщується в початкову область моделювання. В реальному житті не можливо всіх хворих агентів виявити та ізолювати за рахунок наявності інкубаційного періоду та можливості безсимптомного перебігу хвороби.

Для врахування цих факторів необхідно ввести додатковий параметр T_{inc} – інкубаційний період протягом якого інфікована людина не помічає симптомів хвороби, не ізолюється та має змогу заразити інших. Величина цього параметру визначається із врахуванням реального та відомого для різних хвороб часу T_{real}^{inc}:

$$T_{inc} = T_{real}^{inc}/dt \tag{1.9}$$

Для врахування безсимптомного перебігу хвороби достатньо на момент ініціалізаії кожному агентові присвоїти додатковий бінарний статус I_s, що визначатиме як буде переносити цей агент хворобу у разі запалення (True – безсимтомно, False – із симптомами). Цей статус визначатиметься в залежності від імовірності безсимптомного пребігу хвороби p_{sym}, що є відомою медичною величиною для різних хвороб.

$$I_s = \begin{cases} True & z < p_{sym} \\ False & z \geq p_{sym} \end{cases} \tag{1.10}$$

де $z = [0,1]$ випадкова величина.

В разі якщо $I_s = True$ агент не ізолюється в разі інфікування.

Також важливою величиною є число заражених агентів при якому починається ізоляція I_{max}. Якщо загальна кількість інфікованих є меншою за I_{max} процес ізоляції не починається.

Карантин

Карантин передбачає, що певний відсоток людей P_c залишатиметься дома. Однак агент матиме змогу контактувати з людьми, які пішли за продуктами та повернулись, або з гостями. Для моделювання цього достатньо на момент ініціалізації агента (або в ході експерименту) присвоїти його компонентам швидкості v_x, v_y=0 нульові значення.

Громадські місця з тісним контактом (Супермаркети, транспорт, церкви, спортзали, будівництво, підприємства тощо)

Спільною ознакою вищеперерахованих обє'ктів є те, що це місця, в яких на короткий час спостерігається велике скупчення людей, відстань між якими є меншою за критичну для зараження. Змоделювати це можна наступним чином. Вводиться емпірична імовірність відвідування громадських місць для агента p_{sup}. Якщо агент не на карантині на кожному кроці ітерації з імовірністю p_{sup} агент переміщується у центр модельованої комірки, що імітує громадське місце (не залежно від типу). Його попередні координати запам'ятовуються. Протягом часу t_{sup}, що залежить від типу громадського місця, йому дається змога вільно пересуватись в околі центра. Після завершення цього часу, агент повертається на свої координати.

Громадські місця скупчення людей (парки, сквери, базари, центри міст)

Відмінністю вищеперерахованих місць від попередніх є те, що в цьому випадку люди просто сповільнюють свою швидкість пересування, що призводить до більшої концентрації людей в певному місці. Моделювати це можна шляхом зменшення компонент швидкості у випадку потрапляння агента у певну область модельованої комірки.

Дотримання безпечної віддалі

Одним із найбільш ефективних заходів запобігання поширення інфекції є свідоме дотримання людьми безпечної віддалі. Для імітування цього достатньо ввести «силу відштовхування» аналогічну закону Кулона.

$$F = k\frac{q_1 q_2}{r^2} \qquad (1.11)$$

де під зарядами слід розуміти ступінь небажання індивіда стояти близько до іншого, k – коефіцієнт відштовхування (визначається емпірично), r – відстань між людьми. Для врахування цієї сили необхідно розрахувати сумарну силу, яка діє на агента з боку всіх інших агентів: $\overrightarrow{F_i}(F_x, F_y) = \sum_{j=1, j\neq i}^{N} \vec{F}_{ij}$. Далі розраховується прискорення

$$\vec{a}(a_x, a_y) = \frac{\vec{F}}{m} \tag{1.12}$$

де під масою розуміють емпіричний коефіцієнт інертності особи. Тоді зміщення агента на кожному кроці ітерації визначається як:

$$s = (x + v_x dt - \frac{1}{2} a_x dt, y + v_y dt - \frac{1}{2} a_x dt) \tag{1.13}$$

Регіони

Для моделювання великих територіальних об’єднань таких як області чи країни зручно доповнити існуючу мульти-агентну, що імітуватиме столицю додатковими автономними мультиагентними системами із власними агентами та правилами взаємодії. Ці автономні системи імітуватимуть регіони та можуть обмінюватись агентами. Як-правило агенти на короткий час переміщуються до столиці, а потім повертаються назад. Агенти, що розташовані в столиці можуть відвідувати будь-який регіон. З невеликою імовірністю агенти можуть переміщатись безпосередньо з регіону в регіон. Ці дані трансферу можна отримати у відповідних державних органах. Для адекватного моделювання необхідно правильно визначити масштаб мульти-агентних систем та їх популяцію:

$$N_{reg} = N \frac{P_{reg}}{P_S} \tag{1.14}$$

$$L_{reg} = L \sqrt{\frac{S_{reg}}{S_S}} \tag{1.15}$$

де N_{reg} – кількість агентів в системі, що моделює регіон, P_{reg} – населення регіону, P_s – населення столиці. L – довжина ребра мультиагентної системи столиці (як правило $L = 1$), S_{reg} – площа регіону, S_s – площа столиці.

Для моделювання транспортного потоку між регіонами визначається імовірність P_{trans} трансферу окремо взятого агента на кожному кроці ітерації. А також час перебування у столиці чи іншому регіоні. Як-правило цей час має бути в межах дня реального часу. Вищезазначені імовірність може бути визначена із аналізу статистичних даних транспортного потоку у відповідних державних органах. Як відомо велика кількість людей подорожує автобусами, поїздами та іншим громадським транспортом. В разі подорожу автомобілем метою теж є відвідування громадського місця. Тому при імітації трансферу агент переміщається в центр іншої мультиагентної системи, що імітує громадське місце із близьким контактом.

Міжнародний рух

Виїзд та приїзд людей із-за кордону імітується шляхом вилучення агентів, або створення нових агентів безпосередньо в центрі мульти-агентної системи.

В такому вигляді модель буде працювати до тих пір, поки не залишиться інфікованих агентів. В такому вигляді вона дозволяє імітувати різні ситуації протягом однієї симуляції. Як було зазначено вище, для отримання стабільного результату необхідно використовувати ансамблі моделей. В ході аналізу різних сценаріїв інколи необхідно запускати імітаційну модель не спочатку, а з певного фіксованого моменту часу. Тому генерація мульти-агентних систем засобами ООП дає змогу зберігати всі об’єкти на жорсткому диску в будь-який момент симуляції і використовувати їх як початкові значення при подальшому аналізі різних ситуацій методами ансамблей моделей.

Вихідними даними моделі є динаміка таких показників як сумарна кількість інфікованих, здорових, перехворівших та мертвих агентів. Така статистика може проводитись як по регіонам окремо, так і загально по всій моделі. В такому вигляді модель дозволяє легко розрахувати ці показники простим сумуванням та

перевіркою статусів агентів. Для співставлення результатів достатньо пропорційно збільшити динамічні показники до реальних даних популяції модельованої системи.

Важливим показником динаміки епідемії є базове R_0 та ефективне репродуктивне число R. Основне репродуктивне число R_0 використовується для вимірювання потенціалу передачі захворювання. Це середня кількість вторинних інфекцій, спричинених типовим випадком інфекції в популяції, де всі сприйнятливі. Загалом, для виникнення епідемії у сприйнятливій популяції R0 має бути> 1, тому кількість випадків зростає. За багатьох обставин не всі контакти будуть сприйнятливі до зараження. Це вимірюється ефективною швидкістю відтворення R. R - середня кількість вторинних випадків на один інфекційний випадок у популяції, що складається як з чутливих, так і з нечутливих господарів. Якщо R> 1, кількість випадків зросте, наприклад, на початку епідемії. Де R = 1, хвороба є ендемічною, а де R <1, кількість випадків зменшиться.

Ефективне число розмноження можна оцінити за добутком основного репродуктивного числа та часткою сприйнятливої популяції господаря (x).

В загальному функціональна схема алгоритму моделювання та прийняття рішення виглядає наступним чином:

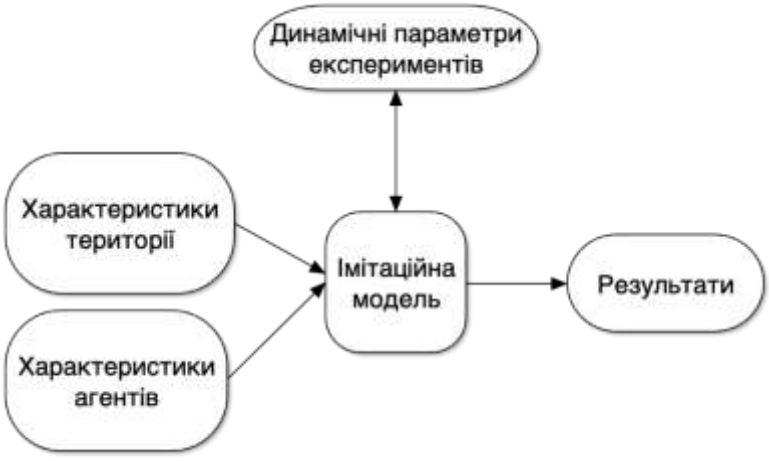


Рис. 1.1. Функціональна схема імітаційної моделі поширення вірусу

Як видно з рис. 1.1, вхідні змінні можна поділити на 2 типи: характеристики території (площа, форма, кількість агентів) та характеристики агентів – множина агентів із унікальними характеристиками. В ході проведення імітаційного експерименту можна динамічно впливати на характеристики агентів, вводити нових чи вилучати агентів в режимі реального часу. Зберігати та відновлювати дані. Це дозволить імітувати та аналізувати в режимі реального часу ефективність тих чи інших дій. Вихідними даними служать статистичні дані по кількості інфікованих агентів, коефіцієнт поширення інфекції.

1.2 Результати моделювання

Аналіз чутливості моделі.

Дослід 1. Аналіз *що-якщо*

Для аналізу чутливості захворюваності від різних факторів були проведені наступні експерименти.

Експеримент 1 (базовий). Люди вільно пересуваються серед міста. Враховано інкубаційний період близько 10 днів та час за який людина повністю одужає в середньому 1 місяць від початку захворювання. Вважається, що 20% людей переносять всю хворобу безсимптомно, тобто вони вільно пересуваються та можуть інфіковувати людей.

Експеримент 2. Збільшення дистанції між людьми на якій можлива інфікування. Фактично це означає більш активний вірус, що може вражати на вдвічі більшу віддаль повітряно-крапельним шляхом. Тобто одна людина може інфікувати іншу на більшій відстані

Експеримент 3. Зменшення імовірності зараження у випадку близького контакту з інфікованою. Передбачає покращення гігієни (миття рук), дотримання маскового режиму, тощо.

Експеримент 4. Карантин. Передбачається, що 80% людей залишається дома. Інші 20% можуть вільно пересуватись містом та контактувати з нерухомими агентами.

Експеримент 5. Наявність точок масових близьких контактів людей. До цього можна віднести наявність супермаркетів, церков, транспорту спортзалів, підприємств на кшталт будівництва, тощо. Експеримент 6 Ізоляція хворих. Передбачає, що інфіковані люди ізолюватимуться у лікарні, або перебувають дома та не контактують із здоровими людьми

Експеримент 7 Збільшення кількості людей, що переносять хворобу безсимптомно до 50%, тобто тих хто не звертається до лікарів та мають змогу інфікувати інших та . врахування ізоляції хворих

Експеримент 8. Наявність центру притягань у місті. Це місця, де люди як-правило зменшують свою швидкість пересування, що призводить до скупчення людей. Наприклад парки, базари, тощо.

Експеримент 9. Свідоме намагання людьми дотримуватись безпечної відстані, що дорівнює віддалі з експерименту 2.

Коефіцієнти моделей приведені в табл. 1.1.

Табл. 1.1. Коефіцієнти базової моделі та їх зміна в експериментах чутливості

Змінні	Змінна	Базовий експеримент	Експерименти
Дистанція між людьми	d_{real}	2 м	4 м – Ex2
Днів хвороби	T_{real}	24	
діапазон тривалості хвороби (дні)	t_{il}	7	
Кількість агентів	N	1000	
Імовірнісь зараження	p	0,1	0,05 – Ex3
Імовірність смерті	D	0,05	

Відсоток людей на карантині	P_c	0	0,8 – Ex4
Інкубаційний період (дні)	T_{real}^{inc}	10	
Імовірність відвідування громадського місця	p_{sup}	0	0,001 – Ex5
Імовірність прояву симптомів хвороби	p_{sym}	0,8	0,5 – Ex7
Ізоляція хворих		No	Yes – Ex6,7
Кількість хворих агентів для початку ізоляції	I_{max}	10	
Наявність громадського місця скупчення людей		No	Yes – Ex8
Дотримання безпечної віддалі		No	Yes – Ex9

Як видно з графіку для базового експерименту на початковому етапі спостерігається експоненційний ріст захворюваності, який поступово виходить на насичення. Після чого спостерігається спад за рахунок зменшення кількості людей, які не мають імунітету та можуть захворіти.

У випадку збільшення активності вірусу (експеримент 2) пік захворюваності зростає майже вдвічі. При чому сам пік захворюваності настає теж вдвічі раніше. Тобто спостерігається критична епідемія з переповненням лікарень та критичною смертністю.

Свідоме дотримання гігієни та маскового режиму є одним з ключових факторів стримування пандемії (експеримент 3). Як видно з графіку максимальна

кількість захворівших зменшується близько вдвічі, а сам процес епідемії розтягується в часі. І пік епідемії зсувається в часі в 1.5 рази. Це дає змогу підготуватись та розвантажити лікарів.

Аналогічний ефект від введення карантинних заходів (експеримент 4). Не дивлячись на 20% людей які мають змогу вільно пересуватись ефект від карантину є аналогічним свідомому дотриманню гігієни.

Наявність супермаркетів, церков, громадського транспорту та інших (експеримент 5) спричиняє ефект аналогічний експерименту 2. І чим більшим буде можливість близько контактувати з іншими людьми, тим критичніше буде пандемія. Тому усунення таких явищ є одним з основних факторів стримування хвороби.

Можливість ізоляції хворих людей є першочерговим фактором зупинки хвороби (експеримент 6). Це дозволяє суттєво зменшити кількість інфікованих та скоротити час пандемії. Однак наявність інкубаційного періоду та великої кількості людей, що не звертаються до лікарів можуть повністю нівелювати ефект від ізоляції (експеримент 7). Тому важливим фактором є інформування людей про симптоми та запобігання самолікуванню.

Наявність парків, базарів та інше створює ефект аналогічний функціонуванню супермаркетів та громадського транспорту (експеримент 8). Тому ізоляція парків та місць масових скупчень людей є теж вагомим фактором стримування пандемії.

Однак самим ефективним засобом струмування пандемії є дотримання людьми безпечної віддалі оди від одного. В такому разі кількість захворівших не перевищить 4% захворювання. І це лише зумовлено врахуванням випадкових флуктуацій для всіх показників. Якщо ж їх усунути або віддаль буде більшою за допустиму для поширення вірусу – епідемія припиняться не розпочавшись після одужання (або смерті) першого хворого.

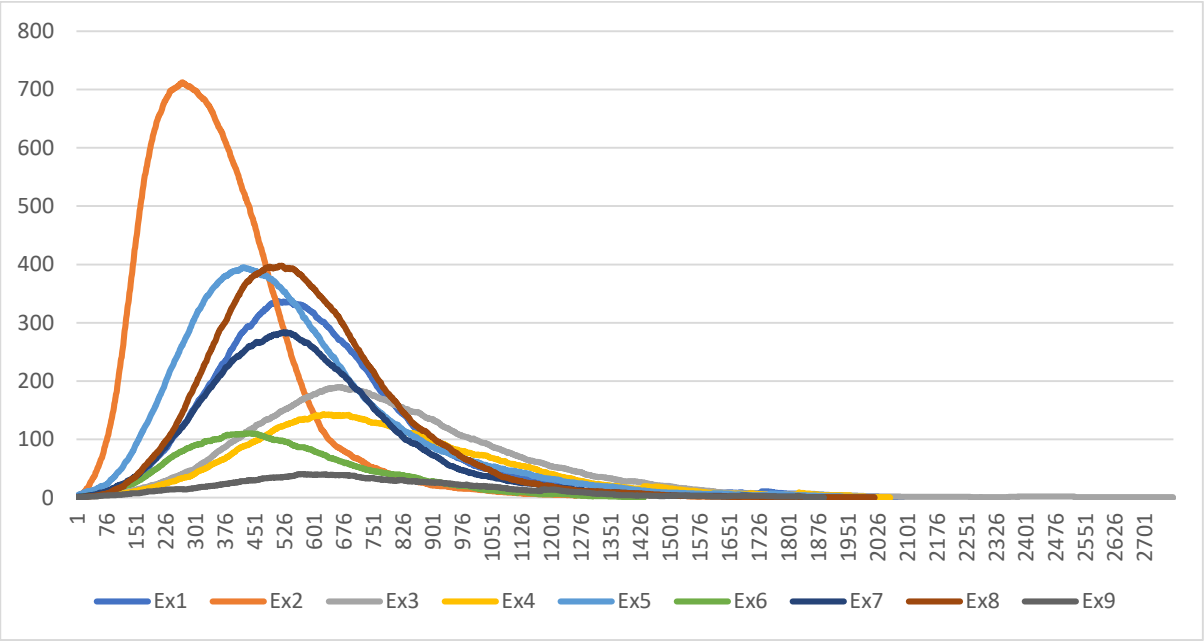


Рис. 1.2. Динаміка кількості хворих для різних експериментів

Цікавими є також статистичні дані. Так наявність супермаркетів або парків призведе до того що перехворіє близько 70% людей. Так само як при недотриманні маскового режиму. Однак пікова кількість захворівших буде меншою вдвічі. Слід зазначити, що в цих трьох експериментах спостерігається і найвища смертність. Найбільш ефективним способом запобігти інфікуванню залишається ізоляція хворих в будь-який спосіб (лікарні, дім, ізолятори тощо). При цьому спостерігається мінімум як по максимальній кількості хворих, так і по кількості не захворівших людей. Близькі дані для заходів карантину. При ньому максимальна кількість захворівших є низькою, однак перехворіє велика кількість людей.

Табл. 1.2. Статистичні дані захворювання для різних експериментів

	Здорові	**Хворі (макс)**	**Перехворіли**	**Померли**
Експеримент 1	38%	34%	59%	3%
Експеримент 2	5%	71%	89%	6%
Експеримент 3	44%	19%	53%	3%

Експеримент 4	52%	14%	46%	2%
Експеримент 5	27%	40%	70%	4%
Експеримент 6	77%	11%	21%	1%
Експеримент 7	33%	28%	64%	3%
Експеримент 8	27%	40%	69%	5%
Експеримент 9	88%	4%	11%	1%

Дослід 2. Аналіз комплексу заходів для запобігання епідемії

Наступним етапом була серія експериментів застосувань комплексу заходів щодо зупинення пандемії. Для цього в якості базового сценарію імітувалась критична пандемія при відсутності будь-яких заходів запобігання (експеримент 1). При чому враховувалась робота транспорту, лікарень, супермаркетів, церков, базарів, тощо. Далі аналізувався сукупний вплив почергового долучення різних комплексів заходів зупинення пандемії. Зокрема:

Експеримент 2. Вплив тільки маскового режиму та підсилення гігієни

Експеримент 3. Додатково зупинка громадського транспорту та обмеження на роботу супермаркетів та церков

Експеримент 4. Додаткова заборона відвідувань місць скупчення людей, на кшталт базарів, парків, скверів, спортзалів, тощо

Експеримент 5. Додавання заходів жорсткого карантину, де переважна більшість людей має залишатись дома

Експеримент 6. Дотримання людьми відстані більшої чим відстань поширення вірусу від інфікованої особи

Досліджувався вплив цих заходів для трьох випадків: 1 – якщо вони починали застосовуватись на початку пандемії, 2 – в самий розпал та 3 – безпосередньо перед максимумом захворюваності.

Коефіцієнти моделей вибирались аналогічно до табл. 1.1

Результати цих експериментів наведені на рис. 1.3, 1.4, 1.5

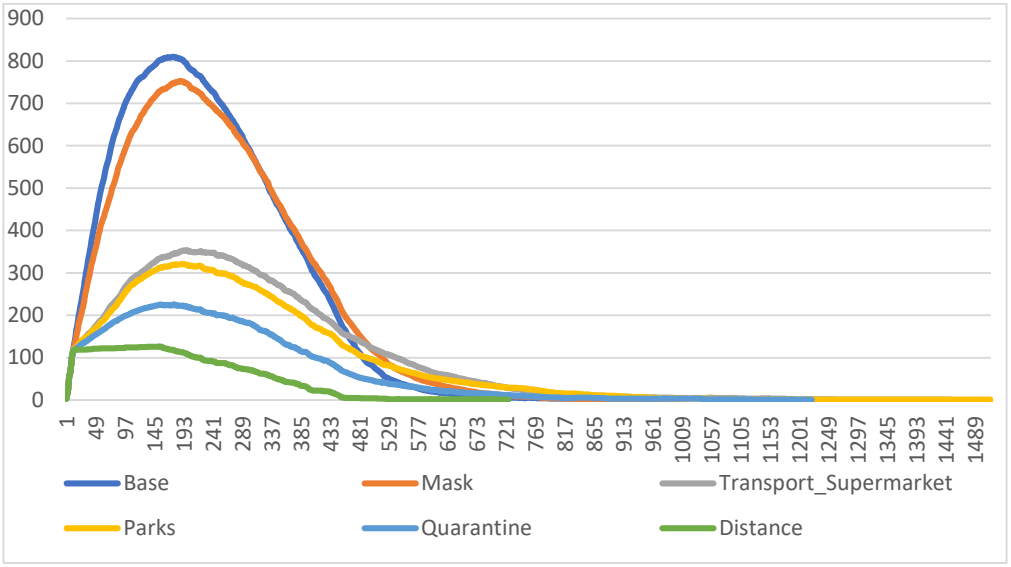


Рис. 1.3. Динаміка кількості захворювань залежно від комплексу заходів, що розпочинаються на початку пандемії

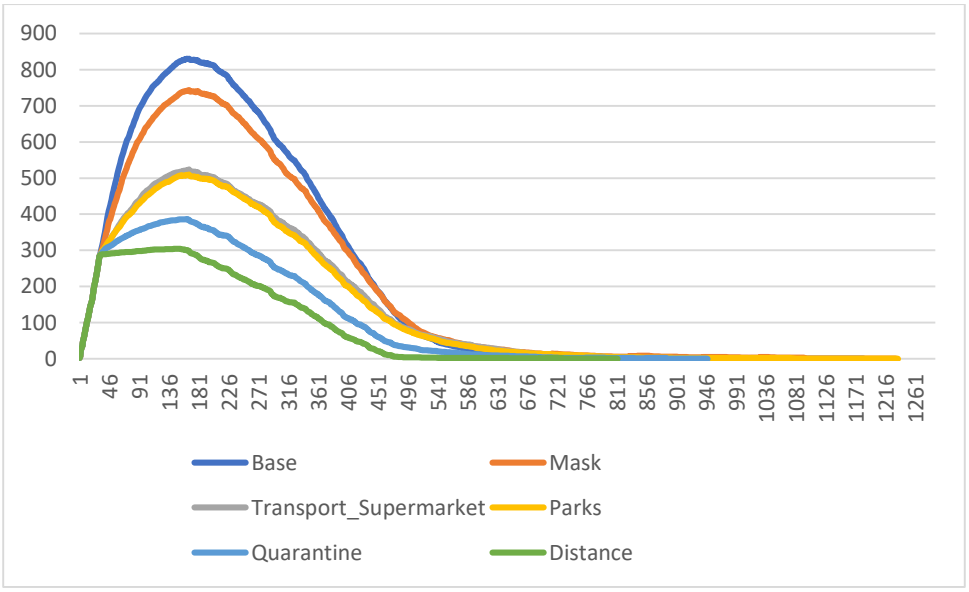


Рис. 1.4. Динаміка кількості захворювань залежно від комплексу заходів, що розпочинаються в розпал пандемії

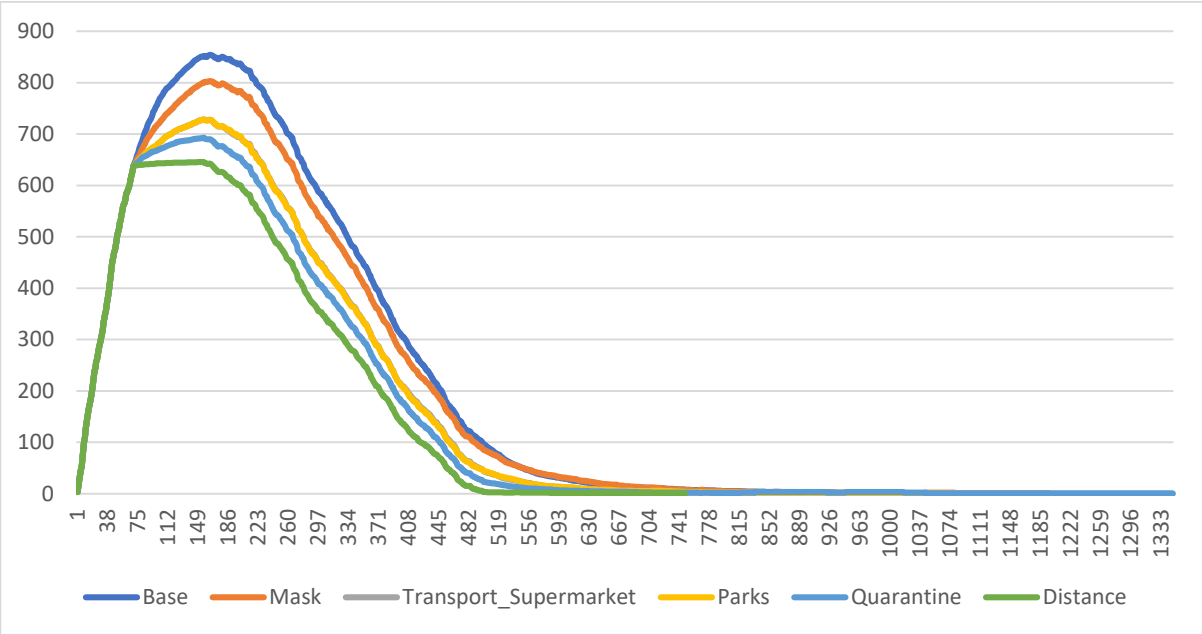


Рис. 1.5. Динаміка кількості захворювань залежно від комплексу заходів, що розпочинаються наприкінці пандемії

Як видно з рисунків, в усіх трьох ситуаціях, застосування лише маскового режиму у випадку поширення пандемії вже слабо впливає на динаміку кількості захворівших. Як показали ці та попередні експерименти, одним з найефективніших заходів запобігання виявилась зупинка транспорту та обмеження роботи супермаркетів та церков в сукупності із масковим режимом. Залучення таких заходів вдвічі сповільнює поширення вірусу. **Однак долучення до цих заходів додаткового блокування місць скупчення людей (парків, базарів, спортзалів) практично не вплинуло на пандемію. Тобто, якщо застосовувати лише цей захід, то він є досить ефективним, як було показано в попередніх експериментах. Однак, якщо розглядати сукупний ефект різних заходів зупинки пандемії, то останній виявився не ефективним в такому комплексі.** Залучення додатково заходів жорстокого карантину ще вдвічі зупиняє швидкість поширення вірусу. І такий комплекс заходів є надзвичайно ефективним, спричинюючи зменшення поширення вірусу в чотири рази. Долучення інформаційного ресурсу, результатом якого є свідомого дотримання дистанцій між людьми практично одразу зупиняє пандемію.

Якщо розглядати час початку застосування заходів, то, як видно з рисунків, застосування їх на початку пандемії дозволяє суттєво зменшити кількість захворівших, та не перевантажити лікарні. Чим раніше почати ці заходи, тим ефективнішими вони будуть і тим меншу кількість їх треба застосовувати. Якщо розпочинати боротись з пандемією в її розпал, то необхідно застосовувати максимум зусиль для її зупинки. І навпаки, якщо розпочинати боротьбу з пандемією, коли вона вийшла на насичення – то жодний з методів вже не дасть значного ефекту, адже вже більшість людей буде хвора і просто запізно щось робити. Хвороба зупиниться сама з максимальною кількістю смертельних випадків.

Цікавим висновком також є те, що не залежно від комплексу заходів та їх початку, тривалість пандемії практично залишається незмінною. Змінюється лише абсолютна кількість захворівших. Це пояснюється тривалістю хвороби та наявністю інкубаційного періоду.

Тобто:

1. Найбільш ефективним комплексом заходів є: зупинка громадського транспорту, обмеження на роботу супермаркетів та церков, жорсткий карантин та свідоме дотримання людьми безпечної віддалі
2. Неефективним є долучення до комплексу заходів закриття місць, де люди сповільнюють свій рух, але дотримуються безпечної дистанції: парки, базари та інше
3. Не залежно від комплексу заходів, загальна тривалість пандемії практично не зміниться.

Дослід 3: Вихід із карантину

Наступним важливим етапом є завершення карантинних заходів. Адже слід пам’ятати, що завжди є незафіксовані інфіковані, що можуть бути причиною другої хвилі пандемії. До таких інфікованих можна віднести людей в інкубаційному періоді та людей, що переносять хворобу безсимптомно, або у

легкій формі та не звертаються до лікаря. Для перевірки цієї гіпотези була змодельована ситуація, коли карантинні заходи залучились на початку пандемії, та були завершені, коли переважна більшість хворих була виписана з лікарні.

Дані для експерименту представлені в таблиці 1.3. Відсутні дані в таблиці є аналогічними базовій моделі таблиці 1.1. Результати представлені на рисунку.

Табл. 1.3. Коефіцієнти моделей карантину та виходу з нього

Змінні	Змінна	Карантин	Вихід з карантину
Відсоток людей на карантині	P_c	0,8	0
Імовірність відвідування громадського місця	p_{sup}	0,001	0,01
Ізоляція хворих		Yes	Yes
Наявність громадського місця скупчення людей		No	Yes
Дотримання безпечної віддалі		Yes	No

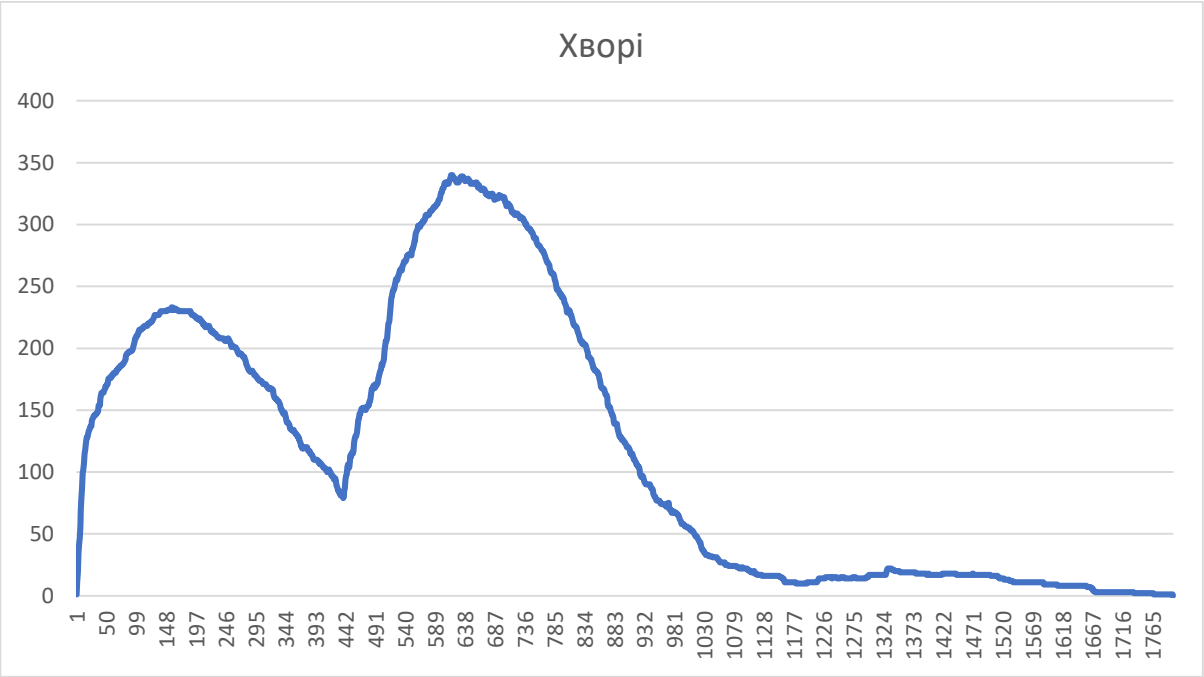


Рис. 1.6. Аналіз наявності хворих

Слід зауважити, що наявність хворих людей у лікарнях означає присутність незафіксованих носіїв хвороби і за їх межами. Різка відміна карантину миттєво призводить до другої хвилі спричиненою саме цими людьми. Ця хвиля буде в 1,5-2 рази більшою за попередню і триватиме так само як і перша хвиля. Після другої можлива і невелика третя хвиля, як видно з графіку. Однак вона вже буде не значною. Тому вихід з карантину має бути поступовий з постійною фіксацією нових захворювань.

Дослід 4. Моделювання закриття транспортного сполучення між районами Чернівецької області

Для моделювання поширення вірусу COVID-19 в Чернівецькій області було враховано:

1. Кількість населення м. Чернівці та всіх районів області
2. Густина населення обласного центру та районів
3. Соціальна дистанція між людьми
4. Тривалість хвороби
5. Імовірність захворювання при контактах людей
6. Рівень смертності

7. Наявність місць скупчення людей (супермаркети, аптеки, ринки, об’єкти будівництва)
8. Відсоток людей, що переносять хворобу безсимптомно
9. Наявність процедури ізоляції хворих людей
10. Можливість переміщення людини з району до обласного центру та назад
11. Дотримання людьми необхідної дистанції

Отже була проведена оцінка як буде розвиватись динаміка захворюваності . В якості стартового зараження виступив обласний центр, де був зафіксований перший випадок захворювання та Хотинський район, що виступив другим джерелом захворювання згідно статистичних даних.

Для боротьби з епідемією коронавірусу в Чернівецькій області були введені: строгий масковий режим, обмеження відвідування магазинів, карантин, закриття всіх масових заходів, спортзалів, кінотеатрів, кафе та ресторанів. Чернівецька область є прикордонною, тому основні джерела запалення виникли в тих селах, де робітники повернулись із Європи. Тому виникло питання оцінити ефект від закриття сіл, в яких були наявні випадки коронавірусу.

Було змодельовано два сценарії поширення інфекції:

1. Якщо між селами та районами області люди можуть вільно пересуватись
2. Обмеження в пересуванні в 10 разів, що введені на початку епідемії. (Пропускають кожну 10 машину чи людину)

Коефіцієнти моделей Чернівецької області аналогічні до табл. 1.3. Додатково були змодельовані автономні 11 мультиагентних систем, що моделюють райони Чернівецької області (табл. 1.4). Та визначені їх популяції та розмірі відповідно до вищезазначених формул. Імовірність трансферу в умовах карантину становила $P_{trans} = 0{,}001$, відповідно при відновленні руху вона була збільшена в 10 разів.

Табл. 1.4. Регіони Чернівецької області

Назви регіонів	Популяція	Площа (км^2)
Чернівці	266 533	153
Вижницький	55 381	903,4
Герцаївський	33 212	316
Глибоцький	74 104	673,2
Заставнівський	48 679	619
Кельменецький	39 666	670
Кіцманський	68 362	607
Новоселицький	77 204	738
Путильський	26 304	884
Сокирянський	42 158	661
Сторожинецький	100 895	1160
Хотинський	61 167	716

Результати динаміки активних випадків інфікування представлено на рис. На графіку на рис. 1.7 представлено відносну кількість людей, які захворіли (по вертикалі). Час (горизонтальна вісь) – комп'ютерний час симуляції.

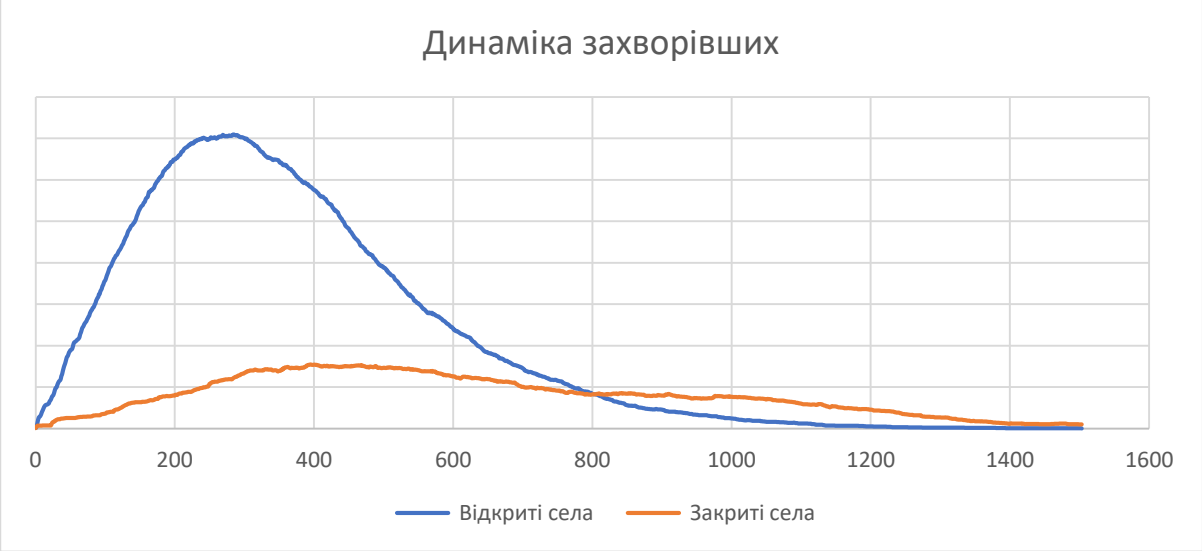


Рис. 1.7. Динаміка кількості інфікованих людей

Табл. 1.5. Відносна статистика захворюваності, отримана в результаті симуляції комп'ютерної програми поширення вірусу в Чернівецькій області

			Відносна зміна показників	
	Відкриті села	Закриті села	%	Разів
Здорових	23,58%	72,18%	-67%	0,3
Перехворіло	76,42%	27,82%	175%	2,7
Померло	4,07%	1,54%	164%	2,6
Хворих (пікове число)	49,82%	10,84%	360%	4,6

В табл. 1.5 представлені згруповані статистичні дані отримані в результаті симуляції поширення вірусу. Слід зазначити, що значення колонок «Відкриті села», «Закриті села» можуть дещо змінитись після узгодження з реальними даними статистики, адже ці розрахунки були проведені на початку епідемії. Однак це слабо вплине на загальний висновок (колонки «Відносна зміна показників»). При розрахунку відносних чисел за 100% брались абсолютні значення у випадку обмеженого пересування людей**.** Тобто **аналізувались на скільки зміниться ситуація при знятті обмеження на пересування**. (Тобто, якщо кількість людей, наприклад, з 10 зросла до 30 – в 3 рази, відносна кількість склала $\frac{30-10}{10} = 200\%$).

В ході симуляції було встановлено:

1. При відкритому русі – перехворіють приблизно 76% населення. При обмежені руху хворітимуть в основному мешканці м. Чернівці та великих громад. Більшість малих сіл або повністю уникнуть захворювання, або швидко стримають його поширення за рахунок самоізоляції та малої

густини населення. Всього перехворіє 28% населення області. Тобто при вільному русі захворіє на 175% більше людей, що в 2.7 рази більше

2. Пікове навантаження інфікованих зросте на 360% тобто в 4.6 рази. Враховуючи такий пік, треба зрозуміти, що лікарні будуть переповнені і більшість людей буде лікуватись дома. Це суттєво підвищить смертність
3. Загальна смертність зросте на 164% тобто в 2.6 рази!!!!
4. При обмеженні руху, кількість людей, які уникнуть інфекції збільшиться з 24% до 72%
5. Як видно з графіку - при обмеженні руху транспорту (червона крива) динаміка захворюваності триватиме довший час, однак без суттєвого навантаження лікарень.

Дослід 5. Відкриття будівельного бізнесу

За місяць карантину Чернівецька область понесла великі економічні збитки, тому одним із перших кроків пом’якшення карантину було відкриття будівельного бізнесу. Слід зазначити, що це відбулось в момент різкого зростання інфікування. Згідно даних УкрСтат в будівельній галузі Чернівецької області задіяно 18.9 тис.осіб., що складає 2.1% мешканців регіону. Вразі відновлення цього бізнесу вказана кількість осіб здійснюватиме робочі переміщення до обласного та районних центрів. Що в свою чергу призведе до збільшення контактів між людьми.

Для оцінки наслідків таких дій було проведено серію експериментів розвитку захворюваності як з врахуванням відновлення будівництва так і без відновлення за допомогою ансамблей моделей, що складались з 5 симуляцій кожна. Моделювання проводилось наступним чином: згідно моделі попереднього досліду моделювався сценарій закритих сіл. На момент комп’ютерного часу, що відповідав даті 6.04.2020, для 2.1% агентам дозволили щоденно переміщатись в центр мульти-агентної системи, що імітує столицю регіону. Наприкінці дня агенти повертались до своїх регіонів, або на місце попередньої локації, якщо агент мешкає в столиці. В симуляції враховувалось

дотримання санітарних норм на робочому місці, зокрема дотримання маскового режиму.

Результати моделювання представлені на рис. 1.8

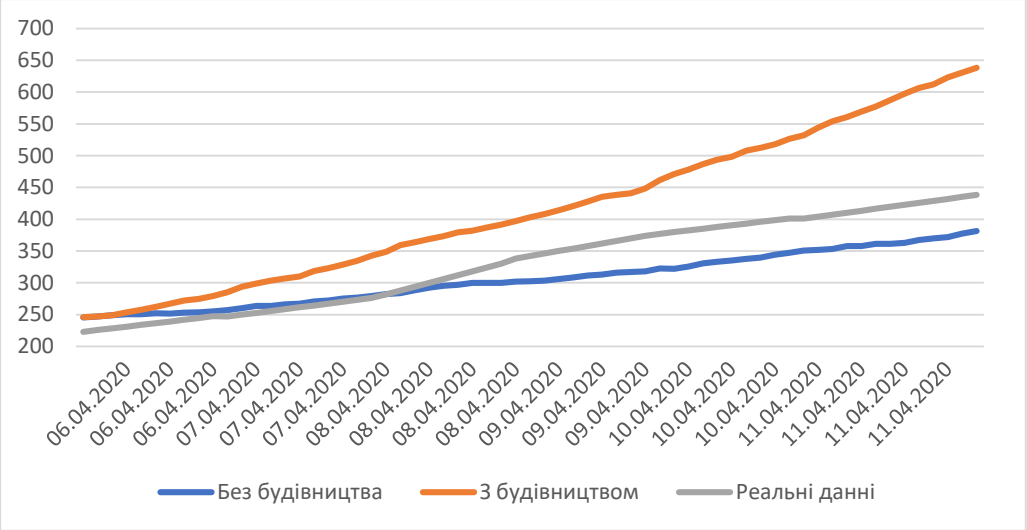


Рис. 1.8. Динаміка кількості нових випадків захворювання з відновленням будівництва та без відновлення

Як видно з рис. 1.8, за результатами моделювання без відновлення будівництва кількість нових випадків захворювання за тиждень зросте з 246 до 382 (136 осіб). Відновлення будівельного бізнесу без дотримання санітарних норм спричинить зростання з 246 до 638 (392 осіб). Тобто, у випадку відновлення будівництва загальна кількість нових випадків захворювання на кінець тижня (12.04.2020) збільшиться на 67% (з 382 до 638). Тому будівництво було відновлене з обов'язково дотриманням маскового режиму та перевірки температури працівників. Не дивлячись на це, як видно з рисунку, це призвело до зростання запалювання, хоча і не такого стрімкого як у випадку недотримання епідеміологічних норм.

Дослід 6. Закриття продовольчих магазинів та оголошення комендантської години на вихідні в м.Чернівці.

Однією із стратегій запобігання поширенню вірусу в м.Чернівці було оголошення строгого карантину на 1.5 доби. Згідно цієї стратегії у вихідні дні

люди залишаються дома та не відвідують продуктові магазини. Це призводить до усунення контактів, а отже появи нових інфікованих за ці дні. Однак на практиці у п’ятницю та суботу спостерігались великі черги за продуктами, що суттєво підвищили імовірність інфікування людей. Для оцінки цих обмежувальних заходів було промодельовано 3 варіанти подій:

Варіант 1. На вихідні дні додаткових обмежень не вводиться. Моделювання згідно базового експерименту таблиці 1. $P_c = 0.8$

Варіант 2. Вводяться додаткові обмеження. (Черги в магазинах напередодні відсутні). В момент часу, що відповідає 13:00 суботи і до 8:00 понеділка коефіцієнту P_c присвоюється значення 1 (100%).

Варіант 3. Вводяться додаткові обмеження та враховуються черги в магазинах напередодні. Для цього в п’ятницю і до суботи 13:00 збільшується коефіцієнт зараження вдвічі p=0.2.

Результати моделювання згідно даних на 9.04.2020 приведені на рис. 1.9. Нажаль в статистичних інформаційних ресурсах відсутня інформація саме про мешканців Чернівців, адже лікарні міста були переповнені мешканцями регіонів. Тому не було можливості порівняти прогнозовані дані із статистичними.

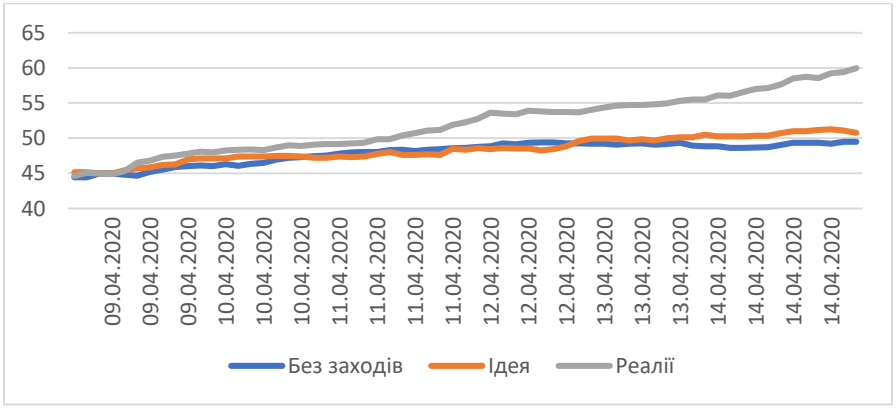


Рис. 1.9. Динаміка захворюваності мешканців м.Чернівці з врахуванням закриття продовольчих магазинів та оголошення комендантської години на вихідні дні.

Як видно зрис. 1.9, за відсутності черг напередодні вихідних та в суботу (червона крива) динаміка нових випадків захворювання практично не відрізняється від ситуації, коли ніяких додаткових обмежень не вводиться (синя крива). Відхилення знаходяться в межах допустимої похибки. Однак наявність черг напередодні вихідних призводить до зростання рівня захворюваності (сіра крива). Зокрема наприкінці тижня, згідно Варіанту 1, кількість нових випадків захворювання зросте з 44 до 49 (5 осіб за тиждень). У випадку Варіанту 3 – із 45 до 60 (15 осіб за тиждень). Тобто введення таких обмежувальних заходів спричинить зростання нових випадків захворювання на 21% за тиждень.

Отже можна зробити висновок: навіть при завчасному інформуванні мешканців про закриття супермаркетів на вихідні дні (тобто усунення ажіотажу і черг) та комендантської години – такі дії не призведуть до жодного ефекту.

Дослід 7. Моделювання поширення COVID-16 в Словаччині

У Словаччині на 15 квітня було 863 інфікованих, 6 людей загинули і 159 вже одужали. 121 людина перебуває в лікарні, 16 - у критичному стані, з них 5 - на вентиляційних апаратах. Як правило, найбільші проблеми є у двох групах:

1. Циганські громади. Багато з них повернулися з-за кордону (особливо з Великобританії), вони живуть у циганських селах, не мають гігієнічних норм. Вони не дотримуються встановлених правил (про правила далі в цьому документі). 14 циганських сіл повністю заблоковані, їх контролюють солдати та поліція. Вони можуть зайти, але не можуть вийти. Майже половина заражених людей - з циганських громад.

2. Будинки для старших. У Словаччині є два будинки для старших (обидва - у місті Пезінок, за 30 км від Братислави). Існує інфекція, яка широко поширена серед людей похилого віку з цих будинків. 4 із 6 задокументованих смертей - від цих двох будинків для старших.

Перший випадок коронавірусу в Словаччині датується 6 березня 2020 року. Це був чоловік, якому було 52 роки, який повернувся додому з катання на лижах

в Італії. Він живе в 20 км від Братислави, в місті, що називається Малацький. Наступний розвиток справ ви можете побачити на веб-сторінці

https://www.statista.com/statistics/1104627/slovakia-coronavirus-covid-19-new-cases-by-date/

З 8 березня всі спортивні заходи, публічні заходи, магазини, торгові зали були закриті. Залишились відкритими єдині магазини: магазини з продуктами харчування, аптеки, паливні баки та аптеки. Після певного тиску з боку підприємців 22 березня також були відкриті магазини із садівничим інвентарем.

З 15 березня було обов'язково носити маски на вулиці, також у магазинах. Поверхня магазинів була розділена на 25м квадратну поверхню. Наприклад, якщо в магазині є 75 метрів у квадраті, максимум людей, яких допускається до магазину, - це троє. Готівки уникають, кредитній картці надають перевагу.

Якщо виходити на вулицю, інших суворих обмежень немає. Єдине блокування було під час Великодніх вихідних. З середи 8 квітня до понеділка опівночі 13 квітня їхати за кордон округів не дозволялося. На жаль, прем'єр-міністр допустив велику помилку. Він повідомив людей за два дні до цього, тож у вівторок (за день до заявленого блокування) люди залишили всю Братиславу.

Кордони в Словаччині закриті. Єдина можливість подорожувати - це маятники, які працюють на відстані максимум 30 км від кордонів (особливо в Угорщині, Австрії та Чехії). Люди, які повернулися додому з-за кордону, повинні пройти 14-денний державний карантин. Це правило діє з 13 березня. Перебувати в карантині вдома не можна.

Є два основні епіцентри в Словаччині - Братислава (приблизно 20% інфікованих людей) та циганські громади (приблизно 50% інфікованих людей - мандрівники).Для моделювання ситуації в Словаччині, імітаціна модель була поділена на 7 регіонів та столицю: Bratislava, Trnava, Trenčín, Nitra, Žilina, Banská Bystrica, Prešov, Košice. Популяція агентів та розмірність автономних комірок були вибрані пропорційно реальним даним (таблиця 1.6). Найбільша популяція циган проживає в регіонах: Košice, Prešov, Banská Bystrica. Як було зазначено

вище, вони слабо дотримуються гігієнічних норм, тому в моделі для цих регіонів коефіцієнти імовірності зараження p були підвищені пропорційно до проценту населення циган в цих регіонах. В Братиславі – як столиці, де контроль був найвищим, коефіцієнт імовірності зараження був зменшений вдвічі. Початкове інфікування моделювалось у зазначених регіонах та столиці. Інші коефіцієнти моделі наведені в табл. 1.7.

Табл. 1.6. Регіони Словаччини

Назви регіонів	Популяція	Площа (км^2)
Bratislava	659598	2052,6
Trnava	563591	4172,2
Trenčín	585882	4501,9
Nitra	676672	6343,4
Žilina	691368	6808,4
Banská Bystrica	647875	9454,8
Prešov	825022	8974,5
Košice	800414	6751,9

Табл. 1.7. Коефіцієнти мультиагентної системи моделювання COVID-19 для Словаччини, Туреччини, Сербії

Змінні	Змінна	Словаччина	Туреччина	Сербія
Кількість агентів столиці	N	1000	500	1000
Відсоток людей на карантині	P_c	0.3	0.5	0.3
Імовірність відвідування громадського місця	p_{sup}	0.0005	0.0005	0.0005
Ізоляція хворих		Yes	Yes	Yes
Кількість хворих агентів для початку ізоляції	I_{max}	1	1	1

Наявність громадського місця скупчення людей		No	No	No
Дотримання безпечної віддалі		Yes	Yes	Yes
Імовірність трансферу	P_{trans}	0.005	0.001	0.005

Було проведено 4 експерименти по 5 симуляцій кожен. Результати прогнозування нових випадків захворювання станом на 6.05.2020 наведено на рисунку. Як видно з рис. 1.10, результати зростання нових випадків захворювання практично збігаються із реальними даними. Динаміка спаду прогнозних даних є дещо повільнішою. Це можливо пояснити тим, що в моделі враховуються всі 100% нових випадків захворювання. На практиці не можливо провести тести 100% населення, тому реальні показники активних випадків будуть нижчими.

Для порівняння був проведений додатковий експеримент дозволу вільного пересування людьми між регіонами країни. У випадку збільшення автомобільного потоку в 10 разів кількість активних випадків зросте близько в тричі. Пік захворювання припав би на 1 квітня, а сам процес видужання розтягнувся би в часі. Аналогічно можна дослідити ефективність і інших заходів.

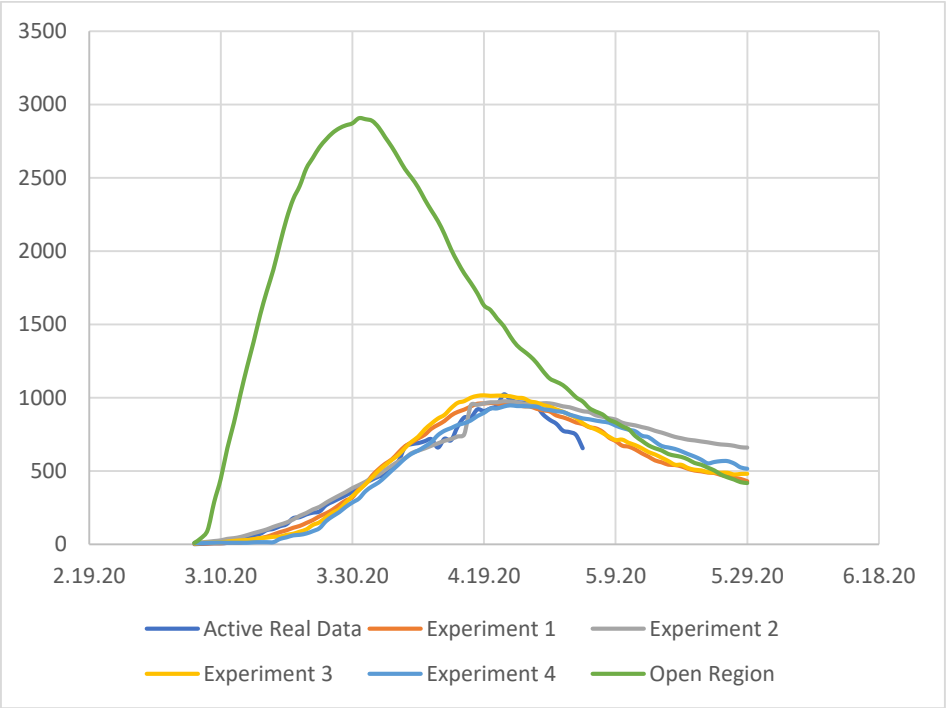


Рис. 1.10. Динаміка активних випадків у Словаччині в порівнянні з реальними даними

Дослід 8. Моделювання поширення COVID-16 в Туреччині

Методи боротьби з коронавірусаом у Туреччині були аналогічні Словацьким. Тому основною відмінністю є розмір та населення регіонів, а також початкове джерело запалювання. Як відомо Туреччина поділяється на 11 адміністративних регіонів (West Marmara, Aegean, East Marmara, West Anatolia, Mediterranean, Central Anatolia, West Black Sea, East Black Sea, Northeast Anatolia, Central East Anatolia, Southeast Anatolia) та столицю – Стамбул (Таблиця 1.9). Як відомо – аеропорт Стамбула є величезним терміналом, що міг бути джерелом потрапляння вірусу у цю країну. Тому в якості початкових умов зараження почалось у агентній системі, що імітує столицю. Всі інші системи були інфіковані за рахунок обміну агентами.

Коефіцієнти моделі приведені в табл. 1.8

Табл. 1.8. Регіони Туреччини

Назви регіонів	Популяція	Площа (км^2)

İstanbul	2513819	5,343
West Marmara	3338203	41,28
Aegean	8045602	90,456
East Marmara	5371207	49,404
West Anatolia	4743395	73,126
Mediterranean	8843959	89,983
Central Anatolia	7143628	90,868
West Black Sea	8771749	73,946
East Black Sea	5669399	37,551
Northeast Anatolia	5446157	70,903
Central East Anatolia	6267470	78,458
Southeast Anatolia	11023037	76,192

Результати моделювання представлені на рис. 1.11.

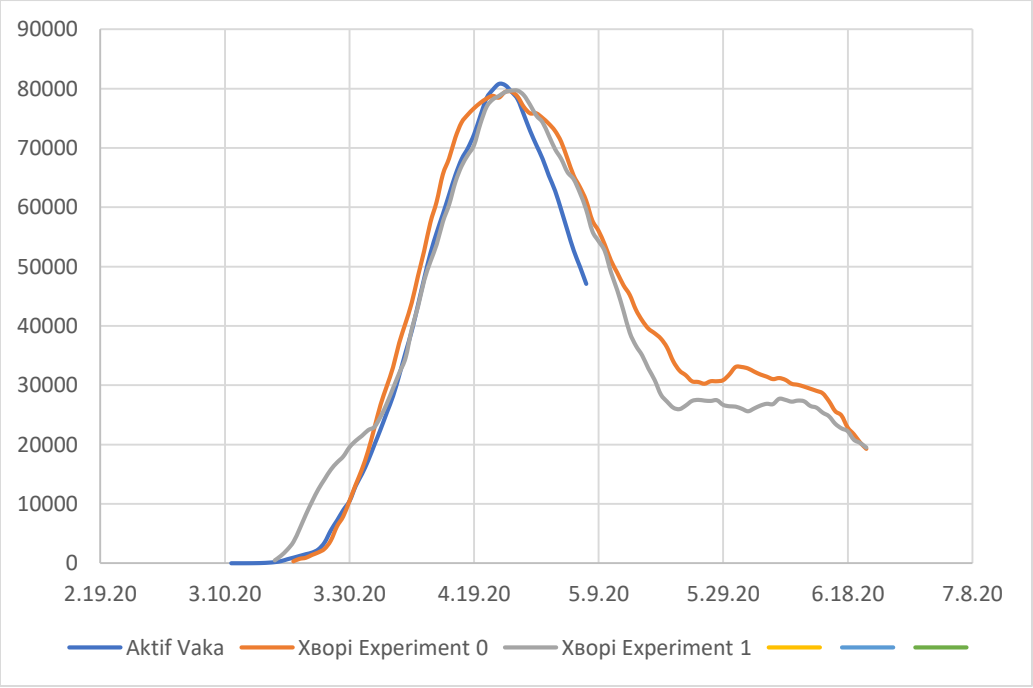


Рис. 1.11. Динаміка активних випадків інфікування для Туреччини

Як видно з рис. 1.11, результати моделювання досить добре збігаються з реальними даними. Особливістю результуючих прогнозів є наявність другої хвилі захворювання. Це пов’язано з тим, що за рахунок складного рельєфу,

сполучення між регіонами є складними і здійснюється в більшості випадків через аеропорт Стамбула. Що і було враховано в моделі. Це призводить до високої автономності регіонів, за рахунок чого деякі з них взагалі зможуть уникнути інфікування. Другий пік виникає за рахунок того, що в момент спаду епідемії та послаблення карантинних умов один з неінфікованих регіонів, де є багато нехворівших агентів зазнає інфікування. Це призводить до локального спалаху епідемії в одному з районів. Це може статись зокрема при відкритті туристичного сезону.

Дослід 9. Моделювання поширення COVID-16 в Туреччині

Сербія – маленька країна, що містить 4 адміністративні регіони: Vojvodina, Šumadija and Western Serbia, Southern and Eastern Serbia, Autonomous Province of Kosovo and Metohija та столицю – Белград (табл. 1.9). Регіони поєднані автобамами та мають добре транспортне сполучення, що призводить до швидкого поширення коронавірусу по всій країні. В якості джерела запалювання обиралась столиця країни, що є туристичним центром та крупним транспортним вузлом. Результати моделювання представлені на рис. 1.12.

Табл. 1.9. Регіони Сербії

Назви регіонів	Популяція	Площа ($км^2$)
Belgrade	1687132	3222,6
Vojvodina	1931809	21614
Šumadija and Western Serbia	2031697	26493
Southern and Eastern Serbia	1505732	26245
Autonomous Province of Kosovo and Metohija	1780021	10910

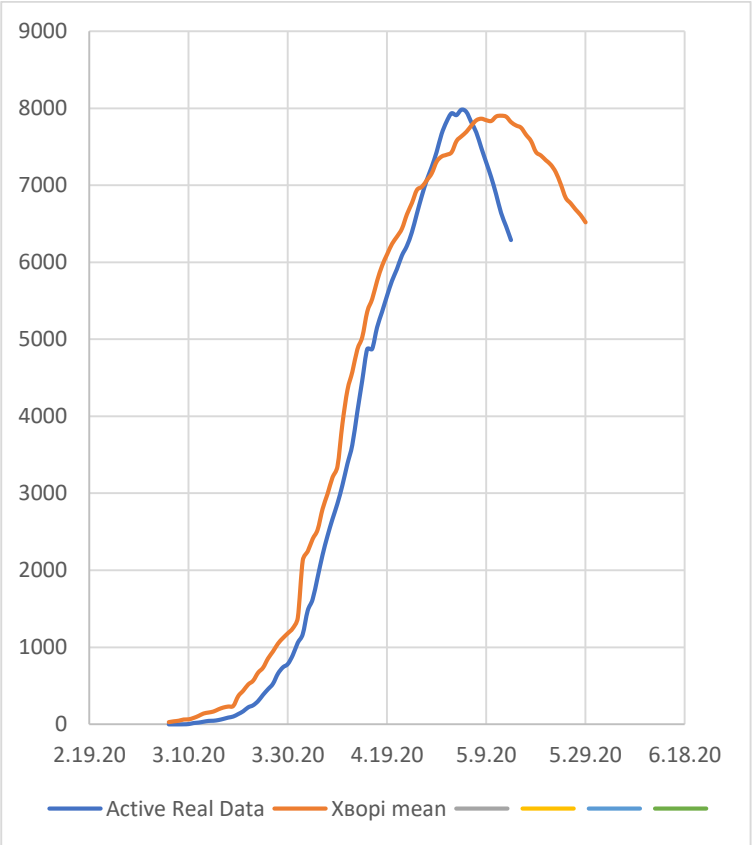


Рис. 1.12. Динаміка активних випадків інфікування для Словаччини

Як видно з рис. 1.12, початкова динаміка захворюваності повністю повторює реальну статистику. Відмінністю, як і в інших розрахунках більш повільне спадання прогнозованої пандемії, що зумовлено як затримкою в часі виявлення хворих, так і не звертанням багатьом з них до лікарів та наявністю безсимптомних хворих.

1.3 Висновки до розділу

В результаті дослідження запропоновано методи співставлення комп'ютерого час розрахунку, динамічних параметрів мультиагентної SIR моделі з реальним часом та такими показниками як кількість активних хворих, рівень смертності, ефективне репродуктивне число. Це дало змогу максимально реалістично побудавати імітаційну модель, що показала достатній рівень адекватності і точності прогнозування поширення пандемії.

Запропоновано модифікацію класичної мульти-агентної SIR моделі, що дало змогу врахувати час інкубаційного періоду, дотримання людьми безпечної віддалі під час пересування, змоделювати карантин, ізоляцію, відвідування громадських місць, таких як супермаркети, парки, церкви, спортзали, змоделювати транспорт, будівництво, спортзали, тощо.

Запропоновано метод моделювання великих регіонів, що складається з декількох адміністративних одиниць шляхом симуляції множини автономних мульти-агентних систем. Запропоновано шляхи моделювання транспортного потоку між регіонами та врахування поширення вірусу між ними.

Був проведений аналіз чутливості моделі до різних заходів запобігання інфекції, визначені як абсолютна так і комплексна ефективність цих заходів. Оцінено ефективність їх використання на різних стадіях поширення вірусної інфекції.

Вдосконалена SIR модель дала змогу змоделювати та провести прогнозні симуляції поширення COVID-19 в Чернівцях та Чернівецькій області та оцінити ефективність різних заходів обласної державної адміністрації для запобігання його поширенню, зокрема відкриттю будівництва та введення жорсткого карантину на вихідні дні. Була проведена симуляція наслідків масового відвідування церков на Пасху та Вербну неділю.

Розроблена модель дала змогу спрогнозувати поширення COVID-19 в таких країнах як Туреччина, Словаччина, Сербія. Отримані результати показали добре узгодження з експериментом та дали змогу передбачити другу хвилю поширення хвороби.

Отже в результаті дослідження вдалось створити та програмно реалізувати уніфіковану SIR-модель, що дає змогу оцінити та змоделювати процеси поширення COVID-19 на реальних соціальних системах різного розміру та структури (міста, регіони, країни). Розроблена модель дає змогу врахувати різні фактори та особливості як самого вірусу (інкубаційний період, рівень смертності, рівень зараження, тощо), так і особливості функціонування

територій (рух транспорту, робота громадських закладів, масковий режим, трансфер між територіями, тощо). Великою перевагою цього методу є можливість адаптації та здатність легко розширюватись шляхом врахування нових факторів. До основних недоліків цієї моделі можна віднести неможливість точного співставлення результатів моделювання із реальними даними. Адже модель дозволяє точно встановити кількість хворих в комп'ютерній моделі в любий момент часу. В реальних системах цього зробити неможливо, адже тести проводять, як-правило, вже людям з підозрою на хворобу, контактним особам, подорожуючим, або іншій невеликій частинці населення. Величезний відсоток людей є непротестованим. Крім того, для реального повного співставлення реальних даних з моделлю необхідно 100% людей робити тести по декілька разів на день. Не дивлячись на це, отримані результати показали як адекватність, так і точність моделі.

РОЗДІЛ 2 РОЗРОБЛЕННЯ МЕТОДУ ПРОГНОЗУВАННЯ ПОШИРЕННЯ ІНФЕКЦІЙ НА ОСНОВІ АНСАМБЛІВ МОДЕЛЕЙ GRNN

В мовах сучасного постіндустріального суспільства, при зборі даних, їх консолідації та передачі не вдається уникати чималих втрат, спотворень в каналах зв'язку та інших неочікуваних ситуацій, через які виникають пропуски у наборах даних, призначених для аналізу. Така форма невизначеності в постановці задач нівелює ефективність подальшого аналізу даних, а проблема їх достовірності, надійності та повноти набуває особливого значення.

Попереднє опрацювання даних вважається одним з найважливіших кроків в їх життєвому циклі. Цей етап відіграє величезну роль для подальшого інтелектуального аналізу, зокрема із застосуванням алгоритмів машинного навчання з вчителем [33]. Це пояснюється тим, що зашумлені дані з великою кількістю пропусків або просто надмірна кількість даних зменшує ефективність або взагалі унеможливлює інтелектуальний аналіз на їх основі. Попереднє опрацювання даних включає ряд процедур, які необхідно виконати над даними для подальшого опрацювання. Зазвичай, серед них виділяють очищення даних, їх інтеграцію, зменшення розмірності та перетворення.

Проблема очищення даних, а саме однин із її ключових етапів – усунення пропусків у даних виникає у різноманітних галузях, починаючи від психології і завершуючи системами розумного будинку, підприємства чи міста. Серед основних стратегій розв'язання цієї задачі виділяють [34]:

- видалення частини набору, у якій містяться пропущені дані;

- заповнення пропущених даних вибірковими статистиками (середнє значення, медіана, тощо);

- заповнення пропущених значень із врахування кореляції між факторними змінними, або на основі міри близькості між спостереженнями.

Загалом, кожен із підходів має своє місце застосування. Це залежить від конкретної задачі, структури даних, їх обсягу, тощо. Проте, якщо аналізувати ці стратегії загалом, можна сказати наступне. При видаленні частини набору із пропущеними значеннями існує велика ймовірність втрати важливої для подальшого аналізу інформації. Використання середнього значення, медіани, тощо для заповнення пропусків у наборах даних базується на припущенні про відсутність залежності між змінними. Проте це не завжди відповідає дійсності [35]. Остання із перерахованих стратегій враховує взаємозв'язки між змінними. Вона дає можливість застосування регресійного підходу, а також методів обчислювального інтелекту для ефективного розв'язання поставленої задачі. В епоху Великих даних і даних великих обсягів саме вона виглядає найбільш перспективною серед розглянутих.

Суттєвий ріст обсягів даних, зокрема зібраних пристроями IoT, складні взаємозв'язки в середині набору, а також неефективні існуючі техніки опрацювання таких наборів даних вимагають розроблення нових методів вірогідного заповнення пропущених даних. Це збільшить ефективність аналітики на їх основі і забезпечить можливість прийняття правильного рішення оператором. Також, це може слугувати надійним фундаментом для автоматизації окремих трудо- та часомістких процедур в різних smart системах.

Нейронну мережу узагальненої регресії (GRNN) розроблено у 1991 р. Дональд Ф. Сперхт. Топологію та деталі її роботи подано у [36]. До основних її переваг серед нейромереж інших відомих нейропарадигм слід віднести такі:

- фактична відсутність процедури навчання і налагодження, що суттєво зменшує час попередньої підготовки до використання мережі на різнотипних даних;
- необхідність налаштування лише одного параметра роботи нейромережі – коефіцієнту розмаху функції активації σ, можливий діапазон зміни якого для кожної задачі є достатньо широким;

- максимально можливу серед штучних нейронних мереж здатність до генералізації.

Топологію цього інструменту обчислювального інтелекту подано на рис. 2.1.

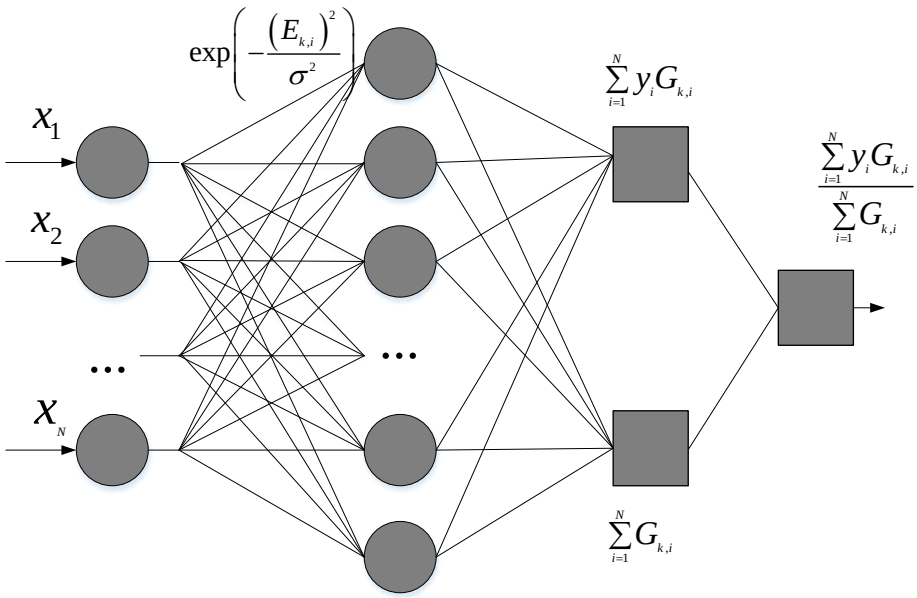


Рис. 2.1. Топологія нейронної мережі узагальненої регресії (GRNN)

До основних недоліків GRNN відносять значні похибки під час її роботи для розв’язання задач конкретного прикладного застосування.

У [37] авторами використано нейронну мережу узагальненої регресії [36] для розв’язання задачі заповнення пропусків у даних. Метод прогнозування на її основі показує вищу точність роботи в порівнянні з існуючими алгоритмами машинного навчання. Для ефективного опрацювання даних великих обсягів, автори з [38] досліджували комбінації різних методів кластеризації та GRNN для розв’язання поставленої задачі. Експериментальним шляхом встановлено найвищу точність роботи саме нейромережі цього типу серед усіх розглянутих, за умови правильного підбору алгоритму кластеризації даних.

Дисертаційну роботу [39] присвячено розробці методів машинного навчання для ефективного розв’язання задачі заповнення пропусків у даних. Автором розроблено ряд ансамблів із використанням GRNN. Серед них слід

виділити гомогенні та гетерогенні ансамблі мереж GRNN. Ідея першого з них полягає у тому, де кожна GRNN з ансамблю опрацьовує свій набір особливостей, отриманих із набору даних. Членами другого ансамлю є набір нейронних мереж різних типів, в тому числі і GRNN, які опрацьовують заданий набір даних. У [40] розроблено новий непараметричний алгоритм на основі ансамблю мереж GRNN. Ефективність його використання підтверджено шляхом порівняння із 25 існуючими методами на близько 100 різноманітних вибірках даних.

Загалом привабливість застосування нейронних мереж цього типу до розв'язання задачі прогнозування полягає у тому, що вони є універсальним засобом, не вимагають наявності великих навчальних вибірок та адаптації до особливостей даних, а також не передбачають ні налагодження, ні навчання [41, 42]. Проте перспективним виглядає саме ансамблевий підхід до апроксимації даних із використанням мереж GRNN.

Зокрема, з метою підвищення точності розвязання задач апроксимації даних, зокрема для прогнозування поширення інфекцій, у [43, 44] Ізоніним І. В., Ткаченком Р. О та Вітинським П. В. розроблено ансамбль з двох мереж GRNN, де перша нейромережа прогнозує значення y_k^{pred} для k-того вектора, а друга - похибку роботи Δ_k^{pred} першої мережі ансамблю. Слід зазначити, що k вектор – це вектор для якого відомо лише вхідні атрибути. Остаточний вихід ансамблю формується таким чином:

$$y_k \approx y_k^{pred} + \Delta_k^{pred} . \qquad (2.1)$$

В основу роботи таких бустингових ансамблів покладено існуючу можливість достатньо гладкої апроксимації методичної похибки роботи нейронної мережі узагальненої регресії, що доведено аналітично [43, 44]. Це забезпечує можливість розроблення відповідних засобів для часткового усунення цієї похибки. Зокрема, дві послідовно з'єднані нейронні мережі узагальненої регресії, що утворюють ансамбль, забезпечують більш точні результати в порівнянні із роботою однієї нейронної мережі цього типу.

Узагальнену структурно-функціональну схема роботи першого бустингового ансамблю наведено на рис. 2.2.

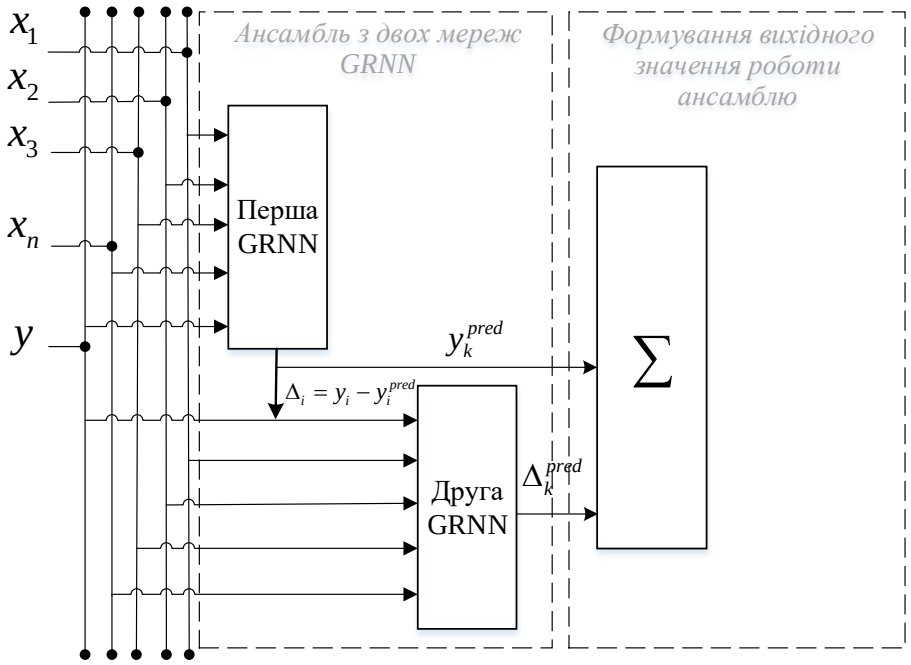


Рис. 2.2. Структурно-функціональна схема роботи ансамблю базового бустингового ансамблю, побудованого із використанням двох послідовно з’єднаних мереж GRNN

У дослідженні [43] за рахунок оптимізації роботи цього ансамблю, вдалося отримати параметри обох його компонентів, які забезпечують підвищену точність його роботи.

У [44] Ісоніном І. В. та Зуб Х. В. розроблено ансамблевий метод прогнозування на основі топології двох послідовно з’єднаних мереж GRNN за рахунок застосування нейроподібної структури Моделі Послідовних Геометричних Перетворень (МПГП). В цьому випадку, фінальний результат роботи ансамблю формується не звичайним сумуванням, а зваженим сумуванням зі зміщенням. Це відбувається шляхом застосування нейроподібних структур Моделі Послідовних Геометричних Перетворень (МПГП) з не ітеративним навчанням. Вхідними параметрами роботи останньої в цьому випадку є виходи обох GRNN ансамблю з [43]. Як показали результати

експериментальних досліджень, лінійна апроксимації поверхні відгуку шуканого вихідного параметру у відповідних координатах вихідних сигналів мереж GRNN ансамблю забезпечує позитивний ефект у порівнянні з тривіальним сумуванням відповідних сигналів. Процедуру формування результату роботи такого ансамблю можна записати так:

$$y_k = \beta_0 + \beta_1 y_k^{pred} + \beta_2 \Delta_k^{pred} \quad , \qquad (2.2)$$

де пошук коефіцієнтів $\beta_0, \beta_1, \beta_2$ з (2.2) відбувається із використанням цієї неітеративної нейроподібної структури.

Узагальнену структурно-функціональну схему роботи другого ансамблю наведено на рис. 2.3.

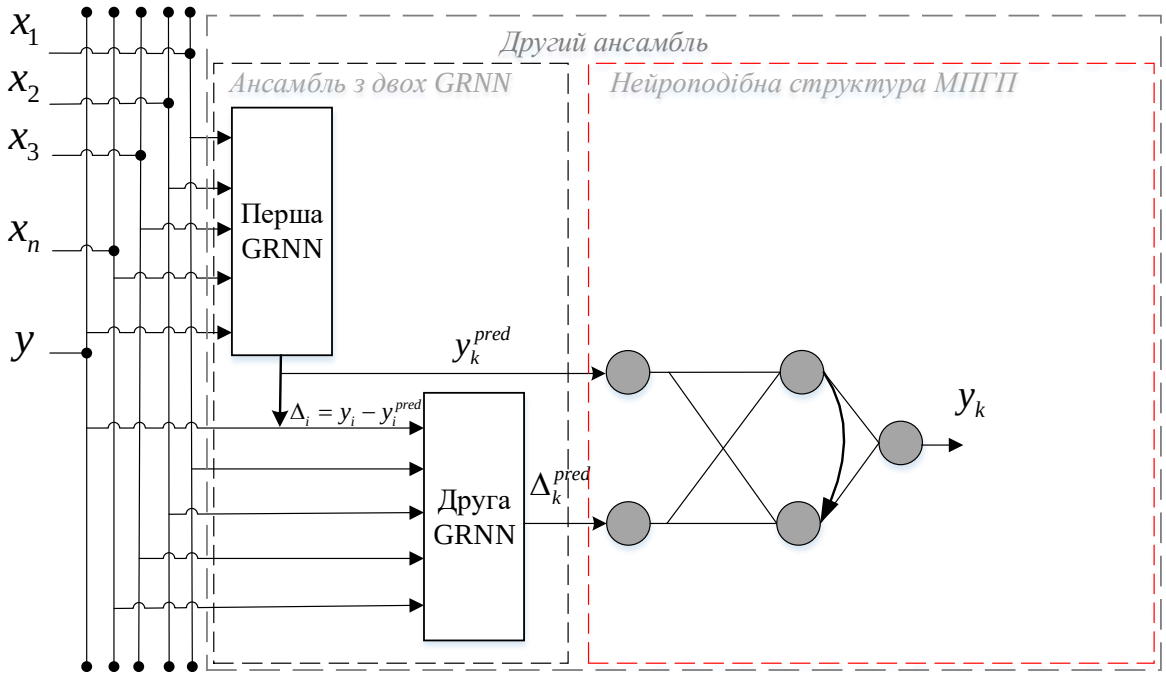


Рис. 2.3. Структурно-функціональна схема роботи ансамблю з двох послідовно з'єднаних мереж GRNN та нейроподібної структури МПГП

У [45, 48] Ізоніном І. В. розроблено ефективнішу гібридну версію нейроподібної структури МПГП. Вона полягає у нелінійному розширенні та

заміні початкових входів нейроподібної структури МПГП членами полінома Колмогорова-Габора. Якщо зробити запис полінома Колмогорова-Габора другого степеня для (2.1) то він виглядатиме так:

$$y_k(y_k^{pred}, \Delta_k^{pred}) = \beta_0 + \beta_1 y_k^{pred} + \beta_2 \Delta_k^{pred} + \beta_3 \left(y_k^{pred}\right)^2 + \beta_5 \left(y_k^{pred} * \Delta_k^{pred}\right) + \beta_5 \left(\Delta_k^{pred}\right)^2 \quad (2.3)$$

За рахунок високих апроксимаційних властивостей цього полінома, гібридна версія нейроподібної структури МПГП демонструє значно вищі результати в точності роботи в порівнянні з базовим інструментом. Деталі її топології, високошвидкісного алгоритму навчання та ефективність застосування подано у [45].

Саме тому, ідея третього бустингового ансамблю базується на її використанні [46]. Зокрема, замість виразу (1) у базовому бустинговому ансамблі (рис. 2.2) та замість виразу (2.2) удосконаленого ансамблю (рис. 2.3) використовується вираз (2.3) для формування вихідного значення роботи ансамблю. Тобто формування остаточного результату роботи ансамблю в цьому випадку відбувалося шляхом побудови нелінійної поверхні відгуку у вигляді степеневого полінома. Це забезпечує підвищення точності роботи ансамблю загалом при незначному рості часу його роботи (оскільки використовується другий степінь цього полінома). Загальну структурно-функціональну схему роботи третього ансамблю наведено на рис. 2.4.

Для моделювання роботи усіх трьох ансамблевих методів прогнозування обрано набір даних моніторингу стану повітряного середовища біля італійського містечка [48]. Цей набір містить ряд атрибутів (табл. 2.1), значення яких зібрано сенсорами датчика пристрою інтернету речей [48]. Основні характеристики обраного набору даних подано у таблиці 2.1.

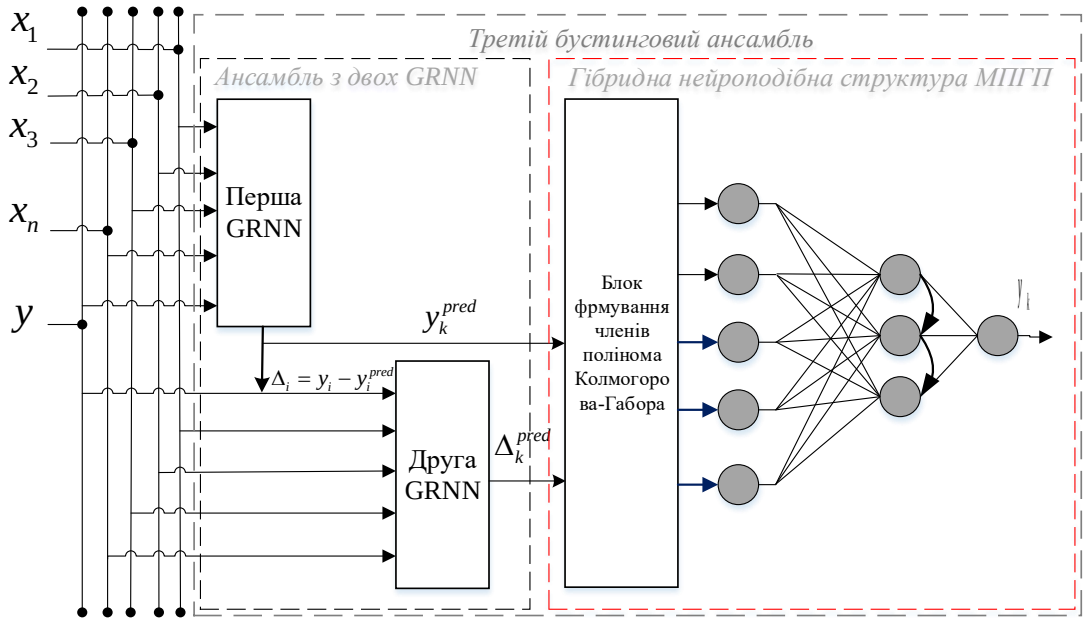


Рис. 2.4. Структурно-функціональна схема роботи третього ансамблю з двох послідовно з’єднаних мереж GRNN та нейроподібної структури МПГП з нелінійним розширенням входів

Табл. 2.1. Характеристика набору даних

Назва атрибуту	Мінімальне значення	Максимальне значення	Середнє значення
Неметанові вуглеводні (SnO2)	2040	647	1119,626
Бензол (C6H6)	63,7	0,2	10,54635
Титан (Ti)	2214	390	958,2302
Оксид вольфраму (WO)	2683	322	817,0748
Двоокис вольфраму (WO2)	2775	551	1452,494
Оксид індія (InO)	2523	221	1057,363
Температура (T)	44,6	0,1	17,75942
Відносна вологість (RH)	88,7	9,2	48,90163
Абсолютна вологість (AH)	2,2345	0,1847	0,986315
Окис азоту (NO)	1479	2	250,465
Двоокис азоту (NO2)	333	2	113,7894
Оксид вуглецю (CO)	11,9	0,1	2,19059

Атрибут *Оксид вуглецю (CO)* обрано як залежну змінну оскільки саме він містить найбільше пропущених значень. Усі інші атрибути із таблиці 2.1 обрано як факторні (незалежні). Після процедур попереднього опрацювання [47] обраний набір даних було випадковим чином розподілено на дві вибірки – навчальну та тестову. Перша вибірка містила 5560 векторів, друга - 1390.

Моделювання відбувалося на ПК з наступними характеристиками: Intel Core i5, 2.40 GHz processor, 8 GB RAMM.

Ефективність роботи досліджуваних методів оцінювалася на основі трьох показників: MAPE, RMSE і часу роботи (застосування). Тривалість процедури навчання до уваги не бралася з огляду на те, що перший ансамблевий метод не вимагає процедури навчання.

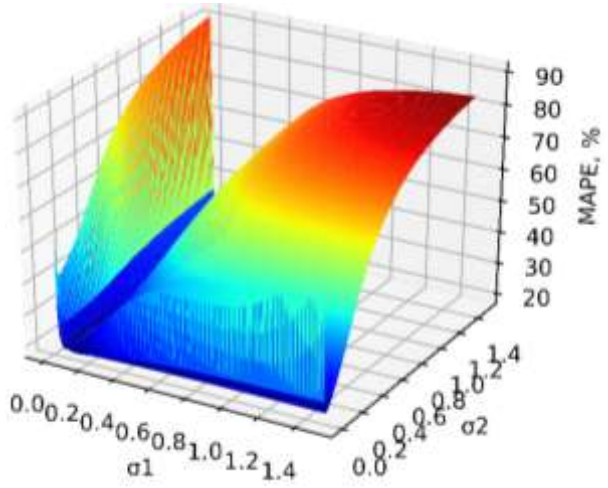
Підбір оптимальних параметрів. Топологія обох GRNN ансамблю для розв'язання поставленої задачі передбачає використання 11 нейронів у вхідному шарі та 1 у вихідному. В результаті роботи обох мереж GRNN отримуємо два значення – спрогнозоване значення та спрогнозовувану похибку, які надходять на вхід наступної компоненти ансамблю. Кожен із досліджуваних ансамблів використовує свою процедуру формування вихідного сигналу. У випадку використання першого ансамблю відбувається звичайне сумування. У випадку другого - сумування зі зміщення за рахунок використання нейроподібної структури МПГП. ЇЇ топологія – два нейрони у вхідному та прихованому шарах. Для реалізації процедури зваженого сумування зі зміщенням у випадку використання третього ансамблю, розробленого Ізоніном І. В. та Зуб Х. В. відбувається заміна початкових входів задачі на розширені [46]. У випадку нелінійного розширення входів на основі використання полінома Колмогорова-Габора другого степеня, кількість нових входів m можна визначити згідно формули [48]:

$$m = n + \frac{n(n+1)}{2} \tag{2.4}$$

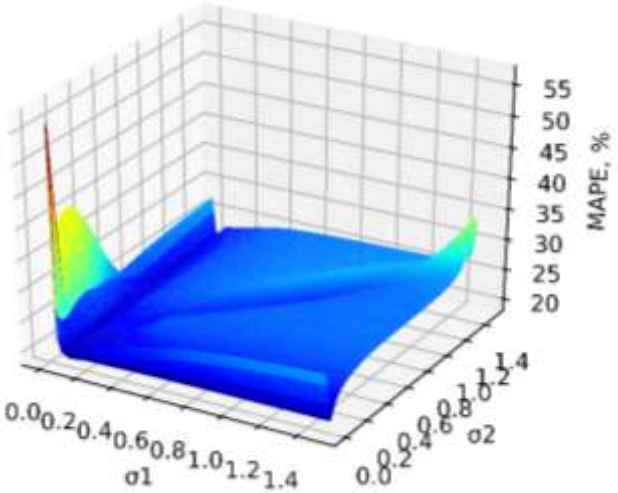
де n описує число початкових входів задачі ($n = 11$).

Таким чином, вихідна компонента третього ансамблю містить 5 нейронів у вхідному шарі, 3 у прихованому (обґрунтовано у [48] та [45]) та 1 у вихідному.

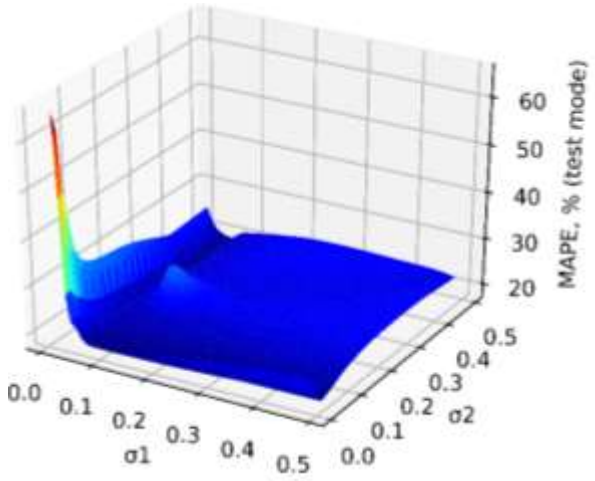
Як відомо, ефективна робота GRNN окрім іншого [36] визначається підбором правильного значення параметру розмаху гаусівської функції σ. Очевидно, що точність роботи ансамблю, в основі якого покладено дві мережі GRNN буде визначатися двома такими параметрами: σ_1 для першої GRNN та σ_2 для другої [44]. Саме тому, метою експерименту було визначення оптимальних значень σ_1 та σ_2 при яких досліджувані ансамблі забезпечуватимуть найвищу точність роботи на основі обох показників (MAPE and RMSE).



а)



б)



в)

Рис. 2.5. Поверхні похибок MAPE, % при зміні параметрів σ_1 та σ_2 обох мереж GRNN для: а) першого ансамблю; б) другого ансамблю; в) третього ансамблю.

Проаналізуємо дані експериментів щодо підбору оптимальних параметрів роботи досліджуваних ансамблів нейронних мереж із різними підходами до формування остаточного результату їх роботи шляхом порівняння поверхонь їх похибок роботи на основі MAPE (рис. 2.5).

Слід зазначити, що поверхні отримано шляхом повного перебору параметрів роботи обох GRNN (σ_1 та σ_2) на проміжках [0.01, 1.49] з кроком 0.01 [43] для кожного із ансамблів. У [18], виявлено, що з ростом значення σ похибка роботи GRNN під час розв'язання поставленої задачі монотонно зростала. Саме тому вищі значення $\sigma > 0.5$ не слід брати до уваги.

З рисунка рис. 2.5 а) видно, що значення MAPE, % суттєво зростало при рості параметрів σ_1 та σ_2 обох мереж ансамблю з [11]. Під час застосування зваженого сумування зі зміщенням з [44] (рис. 2.5 б)) уся поверхня похибки суттєво згладжувалася. Найкращі результати отримано під час використання третього ансамблю (рис. 2.5 в). Функціональне розширення входів за допомогою членів полінома Колмогорова-Габора реалізує нелінійну апроксимацію поверхні відгуку вихідного параметра, що підвищило точність апроксимації, а незначне збільшення підсумкової кількості входів (з двох до п'яти) третьої компоненти ансамблю практично не погіршило властивість генералізації.

Загалом, коректуюча нейроподібна структура неітеративного типу нівелювала суттєвий ріс похибки при рості σ_1 та σ_2 (рис. 2.5 б і в). Не зважаючи на це, локальні мінімуми, залишилися в усіх досліджуваних ансамблів. Шляхом пошуку оптимальних параметрів повним перебором на зазначених проміжках для трьох методів, встановлено мінімальні значення похибок MAPE, % та RMSE. Ці результати, а також часові затримки роботи досліджуваних ансамблів зведено у таблиці 2.2.

З табл. 2.2 видно, що результати щодо точності роботи досліджуваних ансамблевих методів варіюються дуже мало. Зокрема значення MAPE, % між найбільш та найменш точним методами становить близько 0,8%. Проте, найбільш точним виявився метод на основі третього ансамблю.

Оскільки різниця в похибках MAPE результатів роботи досліджуваних ансамблів не є великою, до уваги також бралася похибка RMSE. Перший ансамбль лише з двох мереж GRNN [43] демонструє найменш точні результати серед розглянутих методів на основі обох похибок. Ансамбль із використанням третьої компоненти– лінійної коректуючої нейроподібної структури МПГП [44] показує хороші результати на основі RMSE, проте кардинально інші на основі MAPE. Лише метод прогнозування на основі третього ансамблю [46] забезпечує найменшу похибку на основі обох показників серед усіх розглянутих методів.

Щодо тривалості застосування досліджуваних методів, то, як видно з табл. 2.2 найменший час роботи демонструє метод на основі першого ансамблю. Метод з [44] показує дещо більші часові затримки в порівнянні із базовим. Зокрема за рахунок заміщення процедури сумування на процедуру зваженого сумування зі зміщенням на основі нейроподібної структури МПГП.

Табл. 2.2. Експериментальні результати роботи досліжуваних бустингових ансамблів на основі GRNN

№	Назва ансамблевого методу	Оптимальні параметри роботи методу	MAPE (в режимі	MAPE (в режимі	RMSE (в режимі	RMSE (в режимі	Час роботи, секунд

			навчання)	застосування)	навчання)	застосування)	
1	Бустинговий ансамбль на основі двох мереж GRNN (рис. 2.2)	$\sigma_1 = 0.12,\ \sigma_2 = 0.02$	-	18,104	-	0,492	23.61
2	Бустинговий ансамбль на основі двох мереж GRNN та нейроподібної структури МПГП (рис. 2.3)	$\sigma_1 = 0.23,\ \sigma_2 = 0.05$ два нейрони у вхідному та прихованому шарах нейропо-дібної структури МПГП	20,267	18,828	0,494	0,459	33.29
3	Бустинговий ансамбль на основі двох мереж GRNN на нейроподібної структури МПГП з нелінійним розширенням входів (рис. 2.4)	$\sigma_1 = 0.10,\ \sigma_2 = 0.02$ другий степінь полінома Колмогорова-Габора; пять нейронів у вхідному та прихованому шарах нейроподібної структури МПГП	20.03	18.051	0.502	0.4495	42.15

символ «-» у таблиці 3.2 означає відсутність процедури навчання

Найбільший час застосування демонструє третій ансамблевий метод. Це пояснюється необхідністю нелінійного розширення входів (виходів обох GRNN) поліномом Колмогорова–Габора, яке покладено в основу роботи методу. Необхідність виконання цієї процедури, а також опрацювання п’яти замість двох входів нейроподібною структурою і робить таку затримку в часі роботи методу. Незважаючи на це, третій ансамблевий метод із удосконаленою процедурою зваженого сумування зі зміщенням забезпечує найменші похибки (на основі як MAPE так і RMSE) в режимах як навчання так і застосування.

2.1 Висновки до розділу

У розділі розглянуто застосування бустингового підходу до підвищення точності розв’язання задач апроксимації на основі мереж GRNN. Проведено ряд експериментальних досліджень та порівняння розроблених ансамблів цього типу. В основі їх покладено можливість часткового усунення похибки роботи мережі GRNN. Основною відмінністю трьох ансамблів є процедури формування вихідного результату. Оцінка ефективності роботи трьох ансамблевих методів відбувалася шляхом розв’язання задачі регресії на основі використання реального набору даних.

Експериментальним чином встановлено, що найвищі показники точності (на основі MAPE and RMSE) роботи демонструє третій ансамблевий метод. Експериментальним шляхом встановлено часові ресурси, необхідні для реалізації досліджуваних методів. Серед недоліків методів цього класу слід зазначити суттєві часові затримки, необхідні для їх роботи. Для їх усунення необхідно застосування ряд оптимізаційних методів підбору оптимальних параметрів роботи штучних нейронних мереж цього типу.

РОЗДІЛ 3 ПОШУК ЙМОВІРНІСНИХ ПРОДУКЦІЙНИХ ЗАЛЕЖНОСТЕЙ НА ОСНОВІ МОДЕЛІ ВЕЛИКИХ ДАНИХ

Відсутні значення викликають серйозні проблеми при аналізі даних. Різні алгоритми аналізу даних вимагають матриці показників без прогалин. Імпутація відсутнього значення є важливою річчю для рішення IoT, а також для аналізу та попередньої обробки великих даних. Наприклад, для підтримки медичних рішень нам потрібно визначити фізичні та психоемоційні характеристики пацієнта, а також показники навколишнього середовища (температура повітря в приміщенні та на вулиці, вологість, тиск тощо). Інший приклад - платформа розумного міста, яка складається з інформації про місцезнаходження транспорту, рівень CO_2, споживання енергії, а також інформацію про параметри здоров'я громадян. Більше того, колосальний ріст датчиків та їх придбання, збір та зберігання викликають нові проблеми з відсутніми даними. Окрім відсутніх значень, як і будь-який метод аналізу даних, потрібно розуміння їх природи. Ключовим питанням при аналізі неповних даних є механізм прогалин у них.

Метою цього розділу є покращення відновлення відсутніх даних за допомогою імовірнісної продукційних залежності.

Група методів усунення пропусків містить два типи методів, пов’язаних із використанням деяких моделей, а саме, немодельні або модельні підходи. Що стосується кількості значень, які є результатом використання методів імпутації, існують підходи, засновані на методах одноразової імпутації та методів множинної імпутації [51]. Алгоритми одноразового заповнення дають єдиний повний набір даних, де кожен простір замінюється значенням. Перевагами цього підходу є використання методів повного аналізу даних на наступних етапах обробки та скорочення роботи для заповнення кожного прогалини. Алгоритми багаторазового заповнення генерують кілька повних наборів даних, які аналізуються окремо і отримані результати об'єднуються. Це дозволяє

використовувати методи повного аналізу даних на наступних етапах обробки даних та забезпечує покращені оцінки, а також мінімізує стандартні помилки. Однак ці методи вимагають багато ресурсів для створення більшої кількості наборів даних і потребують більше часу для проведення аналізу, а також більше пам'яті для зберігання результатів [52].

Найбільшою проблемою методів, проаналізованих вище, є обробка структурованих даних, але насправді напівструктуровані та неструктуровані дані домінують у реальних ситуаціях [53]. Крім того, дуже важливо проаналізувати приховані залежності в наборі даних, а також врахувати характер набору даних та передбачити відсутність даних для кожного джерела даних окремо. Ось чому парадигма великих даних використовується для попередньої обробки та обробки інформації з різних джерел, що складається з безперервних числових даних та категоріальних даних. Тим не менше, природа відсутніх даних у різних джерелах даних теж різна. Отже, основною ідеєю є аналіз даних з різних джерел на основі специфіки та природи цих джерел.

Правила асоціації та генерація правил широко використовуються, і вони стикаються з багатьма проблемами, головним з яких є наявність великих даних та багатовимірних наборів даних [53]. Одно процесорні системи з нормальною швидкістю процесора не можуть впоратися з такими великими даними, що робить алгоритм неефективним для використання. В останніх розробках зростання мережевих технологій та особливо хмарних платформ дало нові ідеї щодо генерації мультиатрибутних правил, використовуючи паралельне середовище, таке як Hadoop [54]. MapReduce був популярним і більше використовуваним для обчислення великої кількості даних з тих пір, як компанія Google її запустила на своїй платформі. Розподілена файлова система Google (GFS) та веб-служба Amazon (AWS) використовують платформу Hadoop та MapReduce для надання своїх послуг.

У випадку генерації асоціативних правил в MapReduce, Mapper відповідає за завдання отримати різні комбінації елементів як "ключ", а "значення" ви-

користовується для відстеження кількості входів або підрахунку підтримки. Задачею Редуктора є зменшити отриманий від Mapper набір для кожного ключового значення та обчислити остаточну підтримку для всіх наборів даних кандидатів.

Таким чином асоціативні правила можуть бути створені з максимальною підтримкою та впевненістю.

Алгоритм Apriori має великі проблеми з великими обсягами Big Data, оскільки він сканує всю базу даних кілька разів. Це означає, що час виконання збільшуєть-ся відповідно до кількості транзакцій. У нашому дослі-дженні ми використовували Spark та ієрархічний метод формування правил для покращення алгоритму Apriori.

Введемо поняття асоціативної залежності. Шукатимемо залежність на відношенні *r*. Це відношення може бути сформоване як для реляційних джерел даних, так і для нереляційних (NoSQL) шляхом формування пари значень - назви об’єкта та його характеристик. Для різних об’єктів кількість характеристик може бути різною. У такому випадку відношення *r* формуватиметься як CROSS(*r*). Далі по тексту вживатимемо тільки позначення *r*.

Асоціативна залежність (АЗ) – це продукційне правило в селекції відношення *r*, яке справджується для значущої кількості об’єктів цієї селекції. Поріг значущості повинен визначатись експертним шляхом, або виходячи з розрахунків імовірності помилкового виділення цієї залежності.

$$\begin{gathered} F_I(S;T):(s;t)=\left(r_S(R);r_T(R)\right), \\ i=\overline{1..n_s},\forall i:A_i\in R, \\ j=\overline{1..n_T},\forall j:A_j\in R, \\ s=r_S(R)=\sigma_{S(\{A_i\})}\left(r(R)\right), \\ t=r_T(R)=\sigma_{T\left(\{A_j\}\right)}(s), \end{gathered} \tag{3.1}$$

де S, T – предикати селекції умовної та результуючої частини відповідно, s, t – результати операцій селекції за цими предикатами з основного відношення.

Позначення $S(\{A_i\})$ означає, що предикат використовує множину атрибутів $\{A_i\}$

.

Аналіз великих обсягів даних вимагає виявлення груп атрибутів, які утворюють функціональні залежності (FD). Однак у реальному світі набори даних набагато частіше, важливі залежності визначаються лише на підмножині значень групи ключових атрибутів. Більше того, залежності можуть з'являтися не лише для кортежів у реляційних джерелах даних, але і між різними частинами різних кортежів. Ми будемо називати їх імовірнісною продукційною залежністю (PPD).

Імовірнісна продукційна залежність - це правило продукції при виборі основного співвідношення, яке діє для значної кількості суб'єктів у цьому відборі. Поріг значущості повинен визначатися експертно або на основі розрахунків ймовірності помилкового вибору цієї залежності. Основна відмінність асоціативних правил від PPD полягає в тому, що PPD генерується із існуючих FD у наборі даних.

Основними параметрами традиційних асоціативних правил є рівень довіри та рівень підтримки.

Рівень довіри – відношення кількості об'єктів, для яких має місце така АЗ до кількості об'єктів в селекції:

$$Conf(S \to T) = P(S \to T) = \frac{|\sigma_{S \wedge T}(r)|}{|\sigma_S(r)|}. \qquad (3.2)$$

Рівень підтримки – характеристика предиката селекції на відношенні, що обчислюється як відношення кількості об'єктів, які задовольняють предикат P до загальної кількості об'єктів у відношенні:

$$Supp(P) = \frac{|\sigma_P(r)|}{|r|}. \qquad (3.3)$$

У випадку обчислення рівня підтримки для АЗ умовний та результуючий предикат залежності об'єднуються знаком кон'юнкції:

$$Supp(S \to T) = Supp(S \wedge T) = \frac{|\sigma_{S \wedge T}(r)|}{|r|}. \quad (3.4)$$

З використанням цього поняття рівень довіри можна обчислити, як:

$$Conf(S \to T) = \frac{Supp(S \to T)}{Supp(S)}. \quad (3.5)$$

Рівень покращення обчислюється, як відношення рівнів довіри та підтримки АЗ:

$$\mathrm{Imp}(S \to T) = \frac{Conf(S \to T)}{Supp(T)} = \frac{Supp(S \wedge T)}{Supp(S) \cdot Supp(T)}. \quad (3.6)$$

Рівень підтримки та рівень покращення є "симетричними", тобто, $Sup(X \to Y) = Sup(Y \to X)$ та $Imp(X \to Y) = Imp(Y \to X)$. Те ж справедливо й для повної взаємної інформації. Рівень довіри є "направленим" параметром, тобто, $Conf(X \to Y) \neq Conf(Y \to X)$. Цим пояснюється непряма залежність інформативності від рівня довіри.

Метод пошуку ймовірнісної продукційної залежності

Побудуємо метод пошуку PPD.

Вхідні дані

1. Відношення CROSS(r), схема R визначена лише на аналізованій селекції.
2. Хеш-функції для кожного атрибуту відношення R: $h_j(A_j)$.
3. Порогове значення рівня довіри залежностей, що шукаються – p_0. Замість даного параметра рівноцінно може використовуватись кількість кортежів, на яких повинна бути визначена шукана залежність – minSupport.
4. Порогове значення рівня довіри залежностей, що враховуються при утворенні нових залежностей $= p^*$.

Вихідні дані

Множина асоціативних залежностей, що відповідає вказаним критеріям: $\{S_i \to T_i\}, \forall i : Conf(S_i \to T_i) \geq p_0$

І етап. Аналіз вхідних даних.

1. Створити дерево хеш-таблиць статистики значень атрибутів відношення.Хеш-таблиця кожного атрибута A_j міститиме відповідність значень цього атрибута структурі (кількість повторів значення; масив описів умовних значень інших атрибутів). Опис умовних значень – хеш-таблиця значень атрибута A_k та кількості їх повторень на множині кортежів $X = R_{A_j = x_i[A_j]}$.

   Вершини парних рівнів дерева розгалужуються по імені атрибута, на який здійснюється проекція; непарних рівнів – значення атрибута батьківської вершини.

   Хеш-таблиці непарних рівнів дерева використовують наперед задані хеш-функції $h_j(A_j)$; для парних рівнів використовується внутрішня хеш-функція порівняння атрибутів на рівність.

2. Заповнити структуру, описану в п.1. Заповнюється не все дерево повністю, а лише гілки, що відповідають наперед заданим критеріям якості залежностей. Додаткове уточнення статистики можливе, використовуючи принцип відкладених обчислень:

a) Для кожного кортежу $x \in r(R)$:

b) Ініціалізувати поточну вершину дерева статистики v коренем дерева: $v = a$

.

c) Ініціалізувати $startAttr = 0$ – номер атрибуту, з якого починати розгалуження дерева.

d) Для кожного атрибуту $A_j \in R, j \geq startAttr$:

e) Якщо не існує гілка дерева статистики $v[j][x[A_j]]$,

f) Створити вершини $v[j]$ та $v[j][x[A_j]]$;

g) Ініціалізувати атрибути

$v[j]\left[x\left[A_j\right]\right].count = 0$,

$v[j]\left[x\left[A_j\right]\right].childs = \varnothing$,

$v[j]\left[x\left[A_j\right]\right].nextAttrId = startAttr + 1$,

Якщо $v.count > splitThreshold$,

h) Рекурсивно перейти до кроку 2d.

Таким чином, у структурі даних a містяться елементарні залежності виду $A_j = x_i\left[A_j\right] \to A_k = x_i\left[A_k\right]$, а також забезпечується можливість обчислення кількості кортежів довільних проекцій $\left|\sigma_{A_{i_1}=v_{i_1} \wedge A_{i_2}=v_{i_2} \wedge \ldots \wedge A_{i_l}=v_{i_l}}(r)\right|$.

3. Оголосити список залежностей $z[l]\{S:set;T:set;NS:N;NT:N\}$. Кожна залежність представляє собою структуру з атрибутами: S – множина значень атрибутів умовної частини залежності, на яких вона визначена; T – множина значень атрибутів результуючої частини залежності, на яких вона визначена; NS – кількість кортежів, для яких залежність справджується; NT – кількість кортежів, для яких виконується умовна частина залежності.

4. Заповнити список Z залежностями $Conf\left(A_j = x_i\left[A_j\right] \to A_k = x_i\left[A_k\right]\right) \geq p_0$ проходженням по структурі даних a.

II етап. Агрегування залежностей.

1. Скопіювати список Z у окремий список залежностей Aggr;
2. Оголосити хеш-таблицю h_{aggr} множин значень результуючих частин КАЗ. Це дасть змогу ефективно шукати АЗ, з якими можна агрегувати кожну конкретну залежність d_{aggr};
3. Для кожної залежності $z_1 \in Aggr$:
4. Для кожної залежності:
5. Утворити залежність $z_3 = z_1 + z_2$;

6. Якщо $Conf(z_3) \geq p^*$, додати z_3 в кінець списку Aggr для подальшого агрегування з іншими залежностями.
7. Занести залежність z_1 у хеш-таблицю h_{aggr} з ключем $h(z_1[\Pr T])$.

ІІІ етап. Побудова АЗ.

1. Оголосити список залежностей y;
2. Ініціалузувати $Y = Aggr$;
3. Оголосити хеш-таблицю результуючих частин предикатів pr для ефективного пошуку множини залежностей, що мають однакову результуючу частину предикату;
4. Для кожної залежності $F_I \in Y$ занести АЗ F_I у відповідний список хеш-таблиці: $pr[h(F_I[\Pr T])]$..
5. Побудований алгоритм пошуку асоціативних залежностей дає змогу проводити ефективий аналіз однотипних даних на наявність АЗ, сумарна складність якого по часу складає

$$t = O\left(minSupport \cdot \left(\frac{n}{minSupport} \right)^{1+\log_{avgD}(m)} + \right.$$
$$\left. + \frac{Z_d^2}{m \cdot D(A)} + \frac{Z_{aggr}^2 \cdot \log(sz_{aggr})}{m \cdot D(A)} \right) =$$
$$= O\left(minSupport \cdot \left(\frac{n}{minSupport} \right)^{1+\log_{avgD}(m)} + \right.$$
$$\left. + \frac{Z_{aggr}^2 \cdot \log(sz_{aggr})}{m \cdot D(A)} \right).$$

Складність алгоритму по пам’яті рівна

$$M = O\left(M_{stat} + M_{aggr} + M_{ma}\right) =$$
$$= O\left(\left(\frac{n}{minSupport} \right)^{1+\log_{avgD}(m)} + Z_{aggr} \cdot sz_{aggr} + Z_{ma} \cdot sz_{ma} \right) =$$
$$= O\left(\left(\frac{n}{minSupport} \right)^{1+\log_{avgD}(m)} + Z_{ma} \cdot sz_{ma} \right).$$

Друга компонента суми домінується третьою як у випадку часу аналізу, так і для використаної пам’яті. Порівнювати ж перший і третій член суми неможливо

в загальному випадку: їхній розмір по-різному залежить від обсягу даних та структури даних.

Низька асимптотична складність розробленого алгоритму та широкий набір типів даних, що підтримуються для аналізу, дають змогу застосовувати розроблений алгоритм практично у всіх предметних галузях, що працюють із асоціативними залежностями в даних.

Результат аналізу подано на рис. 3.1.

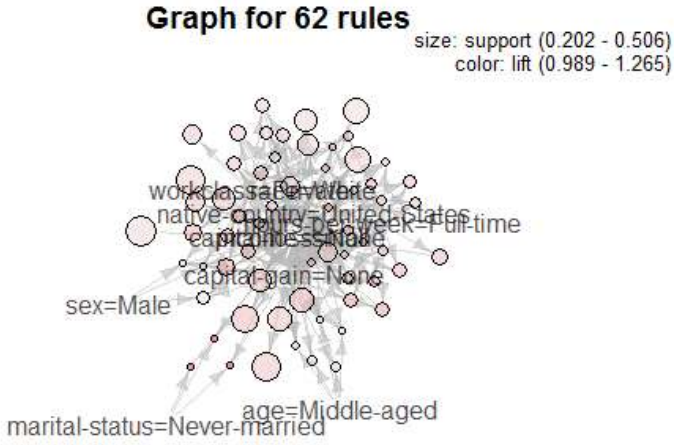


Рис. 3.1. Візуалізація знайдених PPD

Для запуску експерименту використовується набір даних з платформи даних інвентаризації охорони здоров’я великих міст („Платформа даних ВСНС”). Платформа містить понад 18 000 елементів даних у більш ніж 50 сферах охорони здоров’я, соціально-економічних, міських (інформація за допомогою інтелектуальних датчиків) та демографічних показників у 11 категоріях США. Це приклад напівструктурованих даних із прихованими залежностями всередині сутностей та між сутностями.

Відсутні дані моделюються як випадкове видалення точних даних.

Запропоновані та існуючі методи апробовані на тому самому обладнанні: Intel Core 5 Quad E6600 2,4 ГГц, 16 ГБ оперативної пам'яті, жорсткий диск 2 ТБ 7200 об / хв. Критеріями створення PPD є Conf (F1)> 0,7 та minSupport = 100. RStudio використовується для моделювання та аналізу даних.

Запропонований метод було порівняно з існуючими: AR, RF, SVM, MLP, EM та KNN (рис. 3.2), де помилка відновлення представлена нормалізованою середньоквадратичною помилкою (NRMSE).

Як видно з рис. 3.2, запропонований метод PPD є найбільш точним серед усіх методів, і класифікатор EM дає найближчі результати.

Як видно з результатів аналізу (рис. 3.3), у межах приблизно 1% -40% пропущених даних запропонований метод PPD виглядає краще, ніж більшість існуючих методів. Наприклад, запропонований метод PPD працює на 12% краще, ніж EM та RF методи для 30% відсутніх даних. Для більшої частки пропущених даних (діапазон близько 40% -50% пропущених даних) метод EM виглядає найкращим чином, а PPD має порівнянні результати з методом SVM.

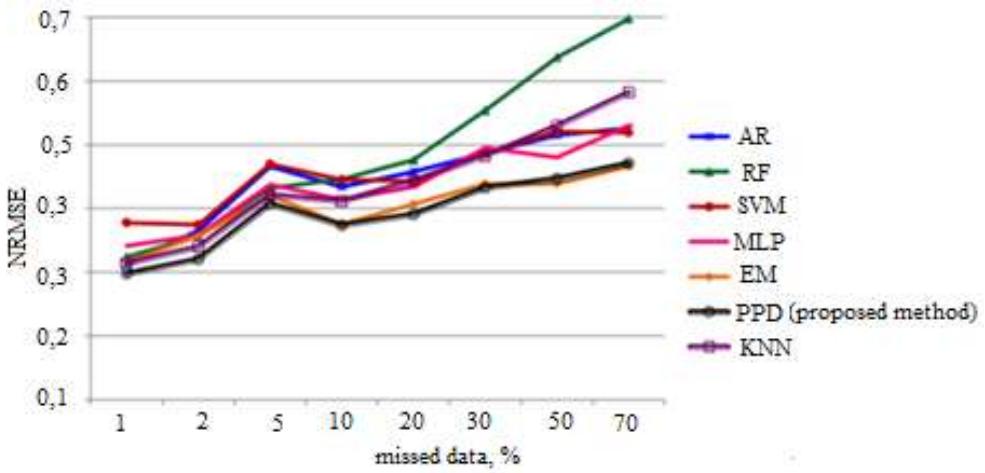


Рис. 3.2. NRMSE помилка відновлення даних

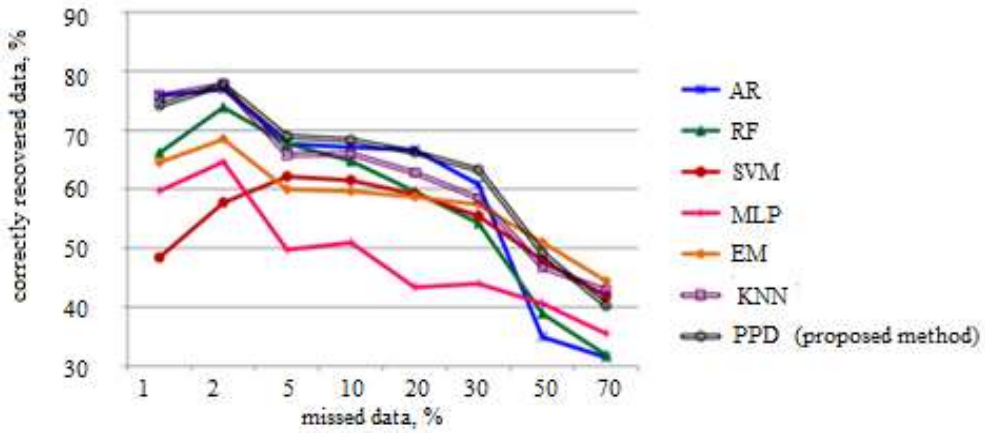


Рис. 3.3. Аналіз відновлення даних за рівнями зашумленості

Результати порівняння часової складності алгоритмів подано у табл. 3.1.

Табл. 3.1. Часова складність в мс

Кількість рядків	**SVM**	**AR**	**EM**	**PPD (пропонований метод)**
2 000	31	190	9	8
4 000	49	362	23	19
6 000	95	437	37	31
8 000	144	639	49	42
10 000	210	827	62	51
18 000	286	1019	77	65

Складність часу розраховується для етапу генерації моделі разом із етапом імпутації відсутніх даних. Як видно з табл. 3.1, часова складність методу PPD є найменшою, і алгоритм EM показав найближчий результат до PPD.

У міру зростання кількості атрибутів кількість знайдених залежностей зростає набагато швидше, ніж коли обсяг аналізованих даних зростає, і цей висновок не обмежується описаним вище.

Проаналізуємо складність алгоритму в паралельному режимі з урахуванням кортежів на кожен з k серверів (процесорів). У такому випадку складність алгоритму слід оцінювати за кількістю ознак m = Card (Bd) у наборі даних Bd та його складністю в часі.

Функція часової складності має субполіноміальний характер. Розроблений алгоритм дозволяє стверджувати, що проблема виявлення PPD у великих даних належить до класу P-проблем з підставою для використання високопродуктивних обчислень. Автори дослідили шляхи прискорення алгоритму з використанням різної кількості серверів (рис. 3.4).

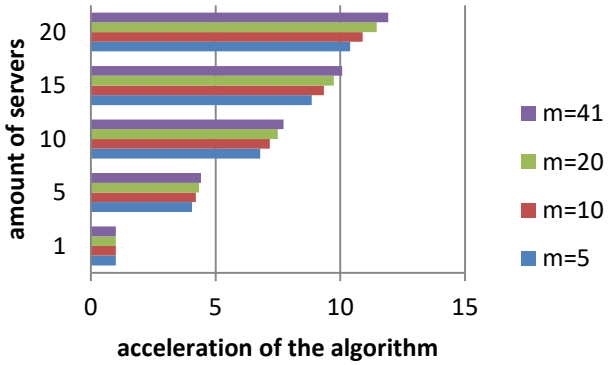


Рис. 3.4. Графік прискорення розробленого алгоритму

Як видно з рисунка 3.4, прискорення для атрибутів m = 41 більше у 12,5 разів для 20 серверів (процесорів) порівняно з непаралельним режимом. Прискорення алгоритму також залежить від кількості атрибутів. Наприклад, прискорення для 20 серверів (процесорів) дорівнює 12,5 для 41 атрибута і 10,5 для 5 атрибутів (див. рис. 3.4). Таким чином, прискорення алгоритму PPD вище приблизно у 1,2 рази для великих наборів даних (41 атрибут), ніж для малих наборів даних (5 атрибутів).

3.1 Висновки до розділу

Розроблено триступеневий підхід до імпутації відсутніх даних, який включає: (i) проектування моделі великих даних у завданні відновлення відсутніх даних, що дозволяє обробляти напівструктуровані дані; (ii) розробка методу відновлення відсутніх даних на основі функціональних залежностей та правил асоціації та (iii) оцінка складності алгоритму відновлення відсутніх даних.

Запропонований спосіб відновлення відсутніх даних створює додаткові значення даних із використанням залежності домену та функціональних залежностей та додає ці значення до наявних навчальних даних. Правильність обчислених значень перевіряється на класифікаторі, побудованому на вихідному наборі даних. Запропонований метод PPD працює на 12% краще, ніж EM та RF методи для 30% відсутніх даних. Для більшої частки відсутніх даних (діапазон близько 40% -50% відсутніх даних) найкраще виглядає метод EM, а PPD має порівнянні результати з методом SVM. Порівняння методів аналізу та моделювання статистичних процесів за якісними критеріями підтвердило, що розроблений метод має наступні переваги: володіє характеристиками стійкості до помилок у даних, що забезпечує паралельне виконання в розподілених базах даних, автоматизує та виконує аналіз різних типів даних . Прискорення для атрибутів m = 41 більше у 12,5 разів для 20 серверів (процесорів) порівняно з непаралельним режимом.

РОЗДІЛ 4 РЕКОМЕНДАЦІЙНА СИСТЕМА ДЛЯ ЗАПОБІГАННЯ ПОШИРЕННЯ ІНФЕКЦІЇ: ПРЕВЕНТИВНІ ЗАХОДИ НА РІВНІ ДЕРЖАВИ

4.1 Огляд предметної області

Новий вид коронавірусу (COVID-19) був вперше виявлений наприкінці 2019 року в Ухані, Китай, і з тих пір поширився майже по всьому світу. В даний час загальна кількість повідомлень про поширення цієї вірусної хвороби охоплює 215 країн та територій [55]. Через швидке поширення випадків захворювання, їх широке територіальне охоплення та потенційно серйозні наслідки, що загрожують життю людей, Всесвітня організація охорони здоров’я оголосила поширення COVID-19 пандемією [56]. Станом на 26 жовтня 2020 року було виявлено 43 694 475 підтверджених випадків зараження, що призвело до 1162 999 смертей [57].

Пандемія має безпрецедентний вплив на суспільство та економіку, зачіпаючи практично всі аспекти повсякденного життя людей та всі сектори економіки. За деякими оцінками, близько 4 мільярдів людей, або близько половини світового населення, страждають від суворих обмежень мобільності та соціальних комунікацій. Пандемія COVID-19 спричинила глобальну економічну кризу, яка відрізняється від глобальної економічної кризи попередніх десятиліть і може мати більш серйозні та тривалі наслідки. Епідеміологічні обмеження призвели до значних змін у повсякденному житті людей, їх звичках, соціальних та економічних зв'язках, методах і формах спілкування, виробничих процесах, соціальному та політичному житті.

Через такий значний вплив на повсякденне життя людей та світову економіку, найважливіші питання, що хвилюють суспільство та уряди, включають: (i) коли поширення COVID-19 досягне свого максимуму як на національному, так і на національному рівні. світові ваги; (ii) як довго триватиме

пандемія; (iii) скільки людей взагалі постраждає від хвороби [58]. Найбільш надійні та аргументовані відповіді на такі питання допоможуть урядам вибрати та реалізувати адекватну стратегію боротьби з розповсюдженням хвороби шляхом балансування між безпекою людей та обмеженнями соціального та економічного життя.

Метою роботи є побудова рекомендаційних правил щодо адекватної державної політики щодо зменшення поширення нових справ Covid-19. Система цих правил базується на новому ансамблі методів машинного навчання, таких як регресійне дерево та кластеризація.

4.2 Огляд аналогів

Широкомасштабне дослідження та аналіз публікацій про COVID-19 з використанням методів машинного навчання А. Доанво та ін. [59] показали, що основний обсяг статей припадає на дослідження в галузі охорони здоров'я, пандемії, клінічної діагностики та лікування коронавірусу. У той же час опубліковано невелику кількість робіт щодо мікробіологічних деталей цього вірусу, включаючи його патогенез та шляхи передачі. Детальний огляд літератури щодо застосування алгоритмів машинного та поглибленого навчання при обробці медичних зображень для діагностики коронавірусної хвороби проведено в [60] разом із порівняльним аналізом застосування кожного алгоритму та його ефективності.

Розробка нових ефективних моделей прогнозування кількості та динаміки нових інфекцій COVID-19, а також епідеміологічне прогнозування в цілому надзвичайно важлива для системи охорони здоров’я, оскільки дозволяє ефективно планувати усунення або зменшення можливих епідемій. Головною вимогою до таких моделей є максимально можлива точність і надійність прогнозу епідеміологічних часових рядів. Для задоволення цих вимог протягом ряду років використовується ряд моделей на основі ШІ [61-63]. Методи, що використовуються для вирішення проблеми побудови моделі прогнозування

COVID-19, варіюються від статистичної авторегресії, як ARIMA, до більш надійних методів машинного навчання [57, 53–65, 67–73].

Так, Ribiero та ін. [68] розглянемо різні моделі, що базуються на регресії такі як cubist, ridge, reference vector regressions, ARIMA разом з random forest та ансамблями для тренування одного, трьох та шести днів передбачення захворюваності у Бразилії. Вони демонструють, що в більшості випадків регресія опорних векторів і навчання ансамблю стеків мають кращі результати за прийнятими критеріями, ніж інші вивчені моделі. Загалом, моделі, розроблені в [68], можуть генерувати точні прогнози з похибкою менше 6,90%. У роботі [68] було зроблено висновок, що подальшого вдосконалення у розробці моделей прогнозування COVID-19 можна досягти шляхом поєднання глибокого навчання та накопичення ансамблевого навчання, використовуючи багатоцільову оптимізацію для налаштування гіперпараметрів моделей прогнозування та прийняття набору особливості, що дозволяють пояснити залежності майбутніх випадків COVID-19.

Чжан та ін. [57] використовував сегментовану модель Пуассона для прогнозування та аналізу часових рядів випадків COVID-19 у шести країнах G7. У розробленій моделі вони включали втручання уряду (поради / накази щодо сидіння вдома, соціального дистанціювання, блокування та карантину) як фактори, що впливають на спалах COVID-19. Аналіз дозволяє [57] статистично прогнозувати переломний момент, тривалість та інтенсивність нападів для досліджуваних країн. У статті [57] випливає, що якби не було серйозних примусових заходів проти стратегій контролю, таких як блокування, дистанціювання, накази залишатися вдома, тоді вірус поширювався б у геометричній прогресії. Це вказує на те, що втручання / дії значно зменшили масштаби спалахів і згладили криві епідемії.

Динаміка спалаху COVID-2019 у Китаї, Італії та Франції у січні – березні 2020 р. Проаналізована в [64]. Деяке загальне поширення епідемії було виявлено на основі аналізу простих добових карт відставання, що дозволяє авторам [64]

припустити, що такі прості моделі можуть бути успішно використані для кількісної оцінки розповсюдження епідемії, зокрема висоти та часу піку заражених осіб.

Руза та ін. [65] використовував феноменологічні моделі, протестовані під час попередніх спалахів, для оцінки короткострокових прогнозів загальної кількості підтверджених випадків COVID-19 у провінції Хубей та окремо для всієї іншої території Китаю. Вони надають прогнози на 5, 10 і 15 днів протягом п'яти днів поспіль, з 5 по 9 лютого 2020 року, з кількісною невизначеністю на основі узагальненої моделі логістичного зростання, моделі зростання Річардса та субепідемічної хвильової моделі.

Вайш'я та ін. [66] провів детальний аналіз публікацій щодо використання ШІ для пандемії COVID-19. Показано, що основні програми ШІ в пандемії COVID-19 включають, але не обмежуючись ними: діагностику та раннє виявлення, моніторинг лікування пацієнтів, відстеження заражених, прогнозування захворювань та смертності, розробку наркотиків та вакцин, зменшення навантаження на медичні працівники і, нарешті, профілактика захворювань. Таким чином, у [66] зроблено висновок, що системи, засновані на ШІ, можуть ефективно використовуватися для ранньої діагностики та лікування пацієнтів, у системах підтримки прийняття рішень у медичній галузі, а також для навчання лікарів та студентів-медиків щодо хвороби COVID-19, що матиме вплив на стратегію та тактику лікування наступних пацієнтів, а отже, зменшить навантаження на медичний персонал у лікарнях під час пандемії.

Басу та Кемпбелл [67] пропонують модель, засновану на довготривалій короткочасній пам’яті, яка була навчена на кількості випадків COVID-19 та смертей окремо. Вони навчають розроблену модель, використовуючи набір даних, що містить кількість підтверджених випадків COVID-19 протягом періоду більше чотирьох місяців. Параметри запропонованої моделі можуть бути скориговані таким чином, щоб забезпечити необхідну точність прогнозування. У статті [67] з використанням розробленої моделі LSTM проведена кількісна

оцінка та аналіз впливу заходів, що вживаються різними округами США на поширення хвороби.

Neher та співавт. [68] розглянемо сезонну модель, засновану на підході до відновлення сприйнятливо-інфекційних захворювань для аналізу та прогнозування поширення реінфекції вірусом COVID-19 у світі у найближчі роки. У своїй моделі вони враховують такі фактори, як рівень зараження, обсяг та швидкість еміграції та переміщення населення тощо. Оновлена версія їх моделі [69] спрямована на аналіз та прогнозування необхідних лікарняних ресурсів для боротьби з пандеміями такого масштабу, як COVID-19.

Ху та ін. [70] побудував та дослідив модель машинного навчання на основі композитних автокодерів для опису поширення коронавірусу в Китаї у квітні 2020 року на основі даних попередніх періодів. Вони кластеризували міста та провінції на основі характеристик, отриманих із розробленої моделі автоматичного кодування. Подібні моделі, засновані на LSTM, використовувались для вивчення статистичних даних щодо пацієнтів [71], а також для виявлення тенденцій поширення епідемії в Китаї [72].

Дослідження S. Tuli та співавт. [73] базується на вдосконаленій моделі машинного навчання для аналізу та прогнозування поширення епідемії COVID-19 у різних країнах. Автори [73] розробили та вивчили міцну модель Вейбулла з повторним зважуванням. Вони прийшли до висновку, що розроблена модель Вейбулла статистично краща за базову модель Гауса для прогнозування спалаху COVID-19. Було показано [73], що проста гауссова модель призводить до надто оптимістичного сценарію поширення COVID-19. Розроблена модель була розгорнута на платформі хмарних обчислень із використанням системи FogBus для більш точного та реалістичного прогнозування динаміки зростання епідемії.

4.3 Методи та засоби дослідження

Опис датасету

У цьому дослідженні ми використовували дані Оксфордського урядового відстежувача COVID-19, а також Європейського центру з профілактики та контролю захворювань (ECDC) Covid-19 [74].

Набір даних Oxford COVID-19 урядового відстеження (OxCGRT) надає інформацію про те, які уряди вживали заходів, які дії вони вживали та коли вони вживали. OxCGRT систематично збирає інформацію про низку спільних політичних відповідей певного уряду та визначає, наскільки уряд реалізує ці заходи. Відповідні бали об'єднані у набір показників політики.

Другий набір даних містить нові публічні дані про географію випадків COVID-19 у всьому світі від Європейського центру з профілактики та контролю захворювань. Кожен рядок або запис містить кількість нових випадків на день за країною чи регіоном.

Для проведеного аналізу ми об’єднали згадані набори даних та розглянули дані з 01 січня по 04 жовтня.

Використані методи

Посилення стохастичного градієнта - це рішення проблеми регресії шляхом побудови ансамблю "слабких" дерев прогнозних рішень. На першій ітерації будується дерево рішень з обмеженою кількістю вузлів. Потім обчислюється різниця між тим прогнозованим деревом, помноженим на швидкість навчання, та бажаною змінною на цьому етапі. І наступна ітерація ґрунтується на цій різниці. Це триває до тих пір, поки результат не перестане покращуватися. На кожному кроці ми намагаємось виправити помилки попереднього дерева.

Обробка даних

Набір даних складається з атрибутів різного характеру. Також представлені дані про шум та викиди. Ось чому на першому етапі необхідні наступні кроки:

• вибір функцій,

• аналіз порожніх даних,

• нормалізація та масштабування даних.

Вибір ознак здійснюється на основі теорії інформації. Спільна взаємна інформація між кожною ознакою та цільовим атрибутом ConfirmedCases обчислюється як:

$$JI(f_i) = \sum_{f_j \in S}\left(I(f_i; c) + I(f_i; c|f_j)\right) \quad (4.1)$$

де c - цільовий клас, , *I(f,c)* - взаємна інформація, f_j - це вже обрана функція, f_i обробляється.

Вже вибрані ознаки були обрані вручну. Список уже вибраних змінних виглядає так: countryName, loaddate. В результаті було обрано 19 із 38 змінних.

Наступним кроком є аналіз порожніх даних. Набір даних складається з 66 998 рядків, 26 003 з яких мають порожні значення у вибраних атрибутах. Через невелику кількість рядків з порожніми даними ці рядки були ліквідовані.

Бажано привести всі вхідні змінні до одного діапазону та нормалізувати (максимальне абсолютне значення вхідних змінних не повинно перевищувати одного). В іншому випадку помилки через змінні, що змінюються в широкому діапазоні, будуть більш впливовими, ніж помилки через змінні, що змінюються у вузькому діапазоні. Переконавшись, що кожна функція змінюється в одному і тому ж діапазоні, ми гарантуємо, що кожна з них має однаковий ефект. Тому вхідні змінні, як правило, масштабуються так, що змінні змінюються в діапазоні функції, як правило, [0,1] або [−1,1]. Використовується масштабування Softmax. Розподіл набору даних для кожної функції показано на рис. 4.1.

Результати

Спочатку розробляється регресійне дерево рішень. Дерева рішень ділять простір об’єктів відповідно до набору правил розділення. Ці правила є логічними твердженнями про змінну і можуть бути істинними або хибними. Тут є три обставини:

- правила дозволяють здійснювати послідовну дихотомічну сегментацію даних,
- два об'єкти вважаються подібними, якщо вони з'являються в одному сегменті розділу,
- на кожному кроці розділу кількість інформації про досліджувану змінну (відповідь) збільшується.

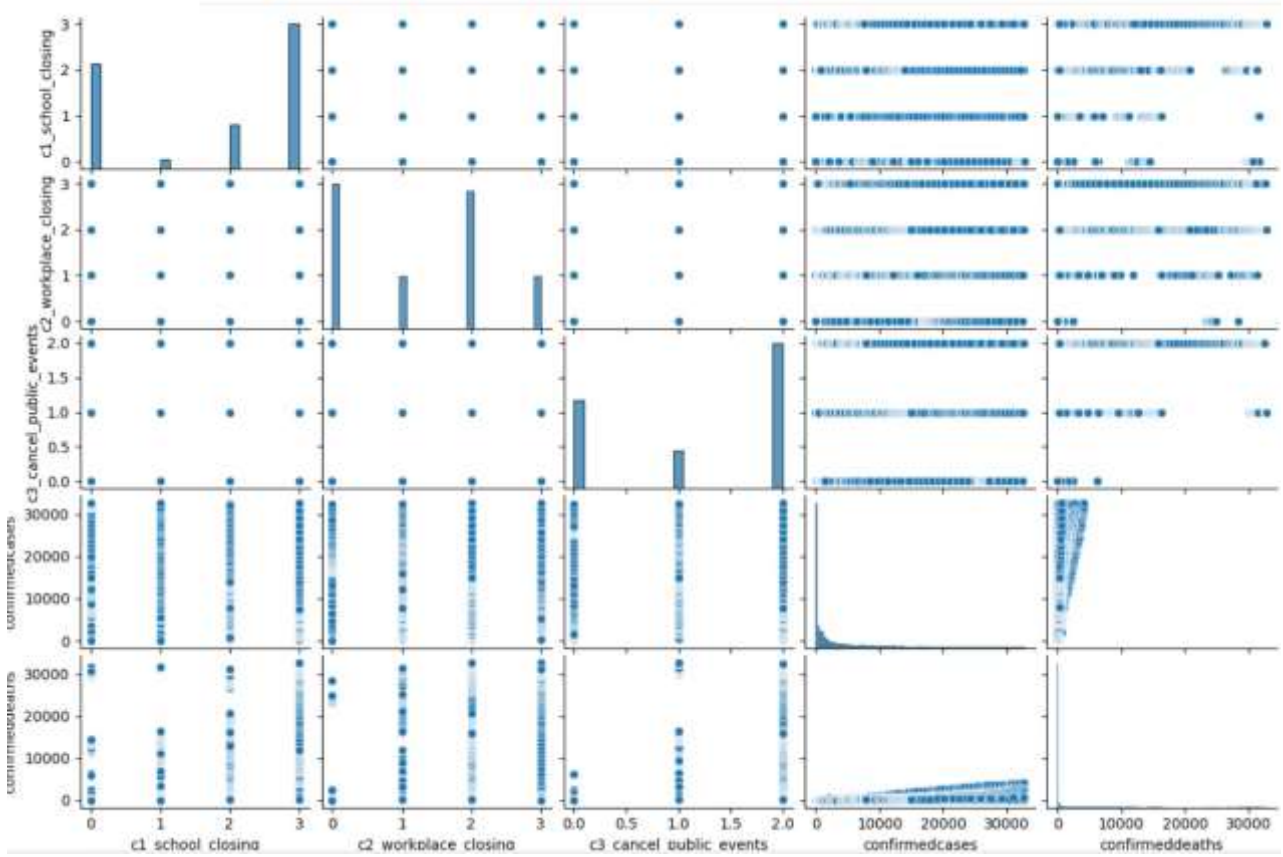
Головною особливістю запропонованого алгоритму є його k-дугова структура. Розгалуження за обраною ознакою x розділяє навчальні об'єкти на k підпроби, де k - кількість різних характеристичних значень.

Не втрачаючи загальності, будемо вважати, що ознака x має значення від $\{0, 1,\ldots, k\text{-}1\}$, $k \geq 2$. У цьому випадку при побудові дерева рішень з вершини are є k дуги позначені цифрами з $\{0, 1,\ldots, k\text{-}1\}$. Нехай σ - мітка однієї з дуг, що залишають вершину x, $\sigma \in \{0, 1,\ldots, k\text{-}1\}$. Щоб сформувати нову поточну підмножину об’єктів та новий поточний набір ознак, ті об’єкти з T' видаляються, для яких значення ознаки x не дорівнює σ, а також сама функція x видаляється з набору ознак.

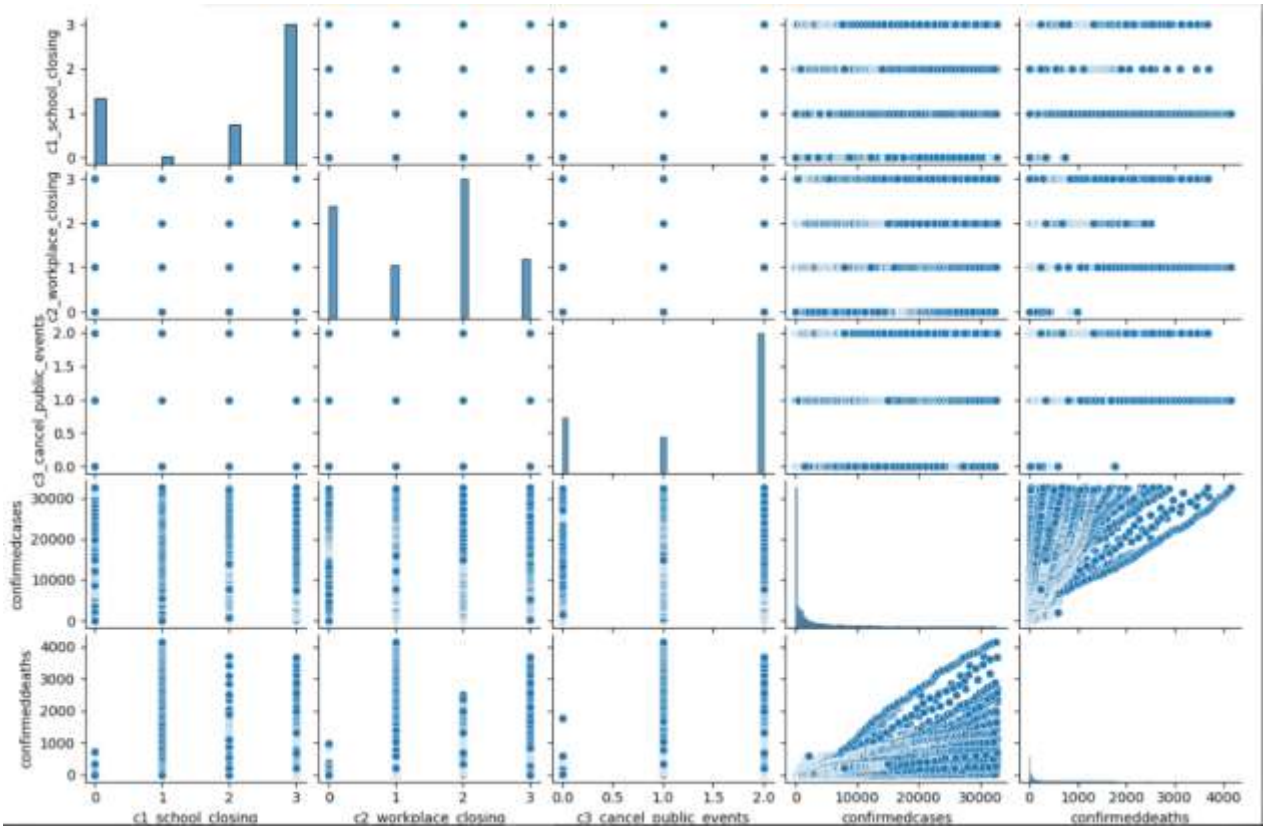
Нехай v - звисаюча вершина, породжена гілкою дерева з внутрішніми вершинами $x_{j_1},\ldots, x_{j_r}$ і нехай дуга, що виходить з вершини x_{j_i}, $i \in \{1,\ldots, r\}$, позначити як σi. Далі нехай $T(v)$ - поточний набір об’єктів, які потрапили у вершину v. Вершина v асоційована з парою $(B, w(v))$, де $w(v)$ дорівнює середньому значенню цільової змінної для всіх об’єктів з $T(v)$, а B - елементарний зв’язок форми $x_{j_1}^{\sigma_1}\ldots x_{j_r}^{\sigma_r}$. Якщо вершина v не є підвісною, тоді ми призначаємо їй сполучнику $B = xx_{j_1}^{\sigma_1}\ldots x_{j_r}^{\sigma_r}$ ... Інтервал істини елементарного сполучення B позначаємо через NB. Нехай S - впізнаваний об’єкт. Для кожної

висячої вершини (B, w (v)) виконується перевірка належності опису об’єкта тесту до інтервалу істини NB. Якщо опис S належить NB, то об’єкт S пов’язаний зі значенням цільової змінної w (v). Об'єкту S присвоюється значення цільової змінної.

$$W = \frac{\sum_{i=1}^{l} w(v_i) * I_{B_i}}{\sum_{i=1}^{l} I_{B_i}}, \quad \text{де} \quad I_{B_i} = \begin{cases} 1, if\ S \in N_{B_i}, \\ 0, \text{інакше.} \end{cases}$$



(*a*)



(*b*)

Рис. 4.1. Розподіл даних до (а) та після (б) попередньої обробки

Наступним кроком є кластеризація. Використовується метод k-середніх значень. Статистика ліктя дозволяє знайти відповідну кількість кластерів.

Кластерні центроїди дозволяють знаходити “середній” об’єкт у кожній групі та створювати правила регуляризації. Кластер №3 показує країни, на які поширюється обмеження, кластер №1 складається з країн без майже будь-яких обмежень (див. табл. 4.1).

Табл. 4.1. Зважені характеристики для класів

Cluster	School closing	Workplace closing	Cancel public events	Close public transport
1	0.04368254	0.02120635	0.07822222	0.005587302
2	2.89865433	2.26230025	1.94790791	1.109703532
3	1.88113534	1.02194531	1.30058844	0.256905504

Cluster	Stay at home req.	Restrictions on internal movement	International travel controls
1	0.0351746	0.01853968	1.137778
2	1.7232443	1.68355761	3.438709
3	0.5112496	0.50931118	2.850190

Cluster	Income support	Public information campaigns	Testing policy	Contact tracing

1	0.02933333	0.7869206	0.6651429	0.6673016
2	1.00667578	1.9615749	1.5702271	1.3380467
3	1.17223953	1.9540325	1.8985808	1.6339218

Система рекомендацій на основі розподілу об'єктів у кожному кластері зображена на рис. 4.2.

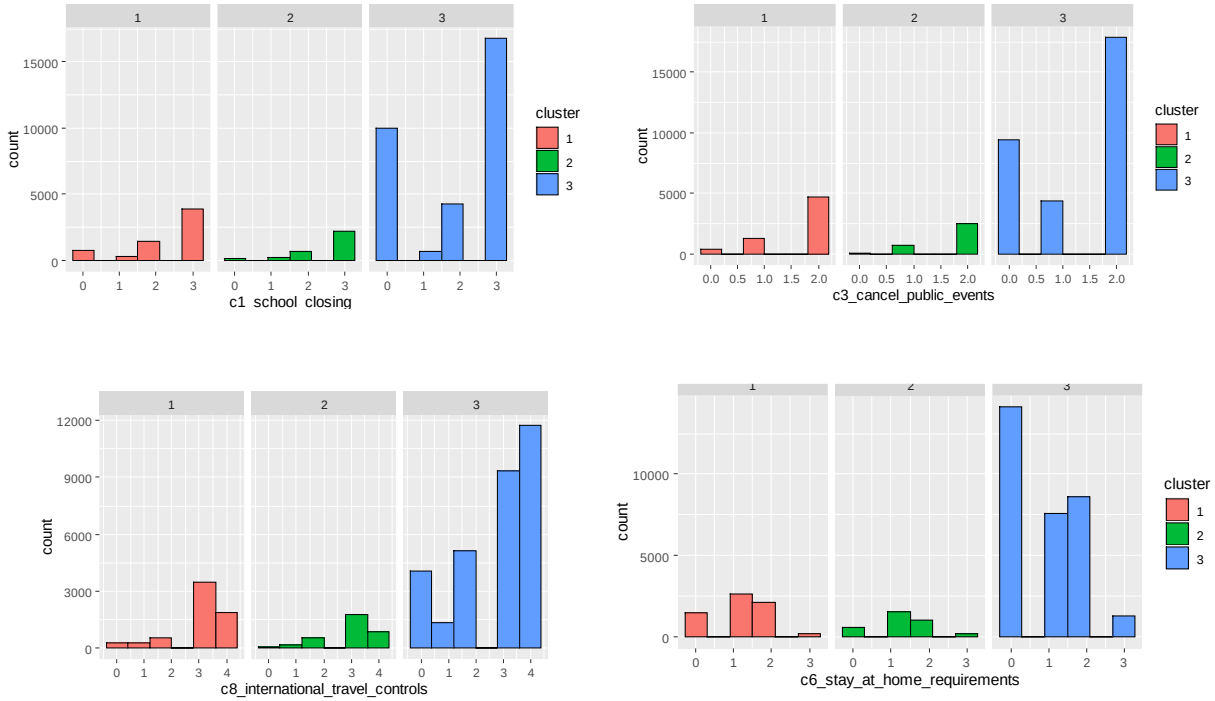


Рис. 4.2. Розподіл даних між кластерами за рекомендацією на основі розподілу об’єктів

Як видно з рис. 4.2, кластери відрізняються найбільш чіткими рекомендаціями, які можна узагальнити наступним чином:

• Кластер №1: рекомендується закривати школи та контролювати міжнародні поїздки,

• Кластер №2: рекомендується залишатися вдома,

• Кластер №3: рекомендується залишатися вдома та скасовувати публічні заходи.

Вплив такого скупчення на розповсюдження COVID-19, положення та тривалість піку, а також рівень смертності буде предметом нашого подальшого дослідження.

Приклад країн з найбільш частим числом кластерів наведено в табл. 4.2. Одна і та ж країна може бути приєднана до різних кластерів у різні часові інтервали. Отже, кластеризація за країнами не буде настільки однозначною. Ось чому часові ряди для відокремленої країни можуть бути цікавими і будуть предметом подальшого вивчення. На цьому етапі ми використовували процедуру класифікації набору даних.

Табл. 4.2. Список країн (фрагментів), що належать до різних кластерів

Country	Cluster
Afghanistan	1
Albania	1
Algeria	2
Andorra	1
Austria	2
Poland	2
Portugal	2
Puerto Rico	1
United Kingdom	3
United States	1
Uruguay	1
Uzbekistan	1
Venezuela	1
Vietnam	1
Yemen	2
Zambia	1
Zimbabwe	1

Проаналізовано три класифікатори: Random Forest (500 дерев), логістична регресія та XGBOOST (алгоритм навчання дерева). Бали моделей наведені в .

Табл. 4.3. Оцінка різних класифікаторів

Random forest	Logistic regression	XGBOOST
0.96	0.91	0.89

Правила, наведені нижче, представляють стратегію, засновану на дереві рішень, і можуть бути використані для стратегічного планування в системі охорони здоров'я, щоб уникнути смертей та тяжких наслідків епідемії.

if (confirmedcases <= 11577.5) {

if (confirmedcases <= 3806.5) { [[confirmeddeaths=20.21931171]] }

else { [[confirmeddeaths =206.84760845]] }

```
 } else {
  if ( confirmedcases <= 17153.5 ) {
   if ( e1_income_support <= 1.5 ) {      [[confirmeddeaths=370.77655172]] }
    else {
     if ( c8_international_travel_controls <= 2.5 ) {
      if ( c6_stay_at_home_requirements <= 0.5 )
{   [[confirmeddeaths=452.20754717]] }
       else {    [[confirmeddeaths=1410.8028169]]      }
     } else {    [[confirmeddeaths=501.40616622]]     }
    }
  } else {
   if ( e1_income_support <= 1.5 ) {
    if ( confirmedcases <= 24529.0 ) {  [confirmeddeaths= [503.03383459]] }
     else {        [[confirmeddeaths=780.60655738]] }
   } else {
    if ( c8_international_travel_controls <= 2.5 ) {
     if ( c5_close_public_transport <= 0.5 )
{      [[confirmeddeaths=898.76146789]] }
      else {     [[confirmeddeaths=1899.2962963]]         }
     else {
      if ( c6_stay_at_home_requirements <= 0.5 ) {
       if ( c7_flag <= 0.5 ) {
        if ( c7_restrictions_on_internal_movement <= 0.5 )
{         [[confirmeddeaths=1346.375]]  }
         else {            [[confirmeddeaths=413.04]]  }
        } else {            [[confirmeddeaths=3256.]] }
      } else {
       if ( confirmedcases <= 29128.0 ) {
        if ( c2_workplace_closing <= 2.5 )
{           [[confirmeddeaths=665.88839286]]  }
         else {
          if ( c5_flag <= 0.5 ) {
           if ( c1_school_closing <= 2.5 )
{            [[confirmeddeaths=2504.]] }
```

```
            else {              [[confirmeddeaths=815.2952381]]  }
          } else {        [[confirmeddeaths=1627.93548387]] }
        }
      } else {
        if ( h2_testing_policy <= 2.5 ) {
          if ( c8_international_travel_controls <= 3.5 )
{            [[confirmeddeaths=1457.]] }
            else {          [[confirmeddeaths=530.2173913]]  }
        } else {            [[confirmeddeaths=708.86764706]]       }
      }
    }
  }
 }
}
```

4.4 Висновки до розділу

Цей розділ присвячено побудові рекомендаційних правил щодо адекватної державної політики щодо зменшення розповсюдження нових справ Covid-19. Система цих правил базується на новому ансамблі методів машинного навчання, таких як регресійне дерево та кластеризація. У цьому дослідженні використовувались об’єднані дані з урядового відстежувача відповідей COVID-19 та ECDC Covid-19.

Кластеризацію проводили методом k-середніх. Статистика розривів дозволяє знайти відповідну кількість кластерів, і у випадку дослідження було обрано три кластери. Кластери відрізняються рекомендаціями та діями урядів-кореспондентів. Таким чином, перші країни кластеру обрали закриття шкіл та контроль міжнародних поїздок як головні рекомендації; країни другого кластеру рекомендували залишатися вдома, тоді як головні рекомендації урядів країн, що входять до третього кластеру, полягали в тому, щоб залишатися вдома та скасовувати публічні заходи.

Одна і та ж країна може бути приєднана до різних кластерів у різні часові інтервали. Отже, кластеризація за країнами не буде настільки однозначною. Ось чому часові ряди для відокремленої країни можуть бути цікавими і будуть предметом подальшого вивчення. Крім того, вплив такого скупчення на поширення COVID-19, положення та тривалість піку, а також рівень смертності також буде предметом нашого подальшого дослідження.

Було побудовано регресійне дерево прийнятя рішень, набір правил було отримано з дерева рішень і може бути використано для стратегічного планування в системі охорони здоров’я.

РОЗДІЛ 5 МЕТОД ВИЗНАЧЕННЯ ПРОГНОСТИЧНИХ ЧИННИКІВ НЕСПРИЯТЛИВОГО ПЕРЕБІГУ SARS-COV-2-АСОЦІЙОВАНОЇ ПНЕВМОНІЇ

Пандемія SARS-CoV-2-інфекції, що розпочалася в грудні 2019 року, за півтора року вразила всі країни світу і станом на 31.05.21 призвела до смерті 3,5 мільйонів людей, в тому числі більше 50 тисяч громадян України. Спалах коронавірусної інфекції в Україні розпочався з Чернівців, які протягом року стійко утримувалися на сумному першому місці по рівню захворюваності. Надзвичайна ситуація в медицині змусила лікарів різних спеціальностей надавати допомогу пацієнтам з SARS-CoV-2-асоціойваними пневмоніями, на власному практичному досвіді вивчати особливості перебігу SARS-CoV-2-інфекції.

Не дивлячись на величезний накопичений клінічний, лабораторний матеріал, надзвичайну увагу медичної громадськості до лікування пацієнтів з SARS-CoV-2-асоціойваними пневмоніями, досі достеменно не відомо чому у деяких пацієнтів захворювання набувало фатального перебігу (Priya S, Selva Meena M, Sangumani J, Rathinam P, Brinda Priyadharshini C, Vijay Anand V. "Factors influencing the outcome of COVID-19 patients admitted in a tertiary care hospital, Madurai.- a cross-sectional study". *Clin Epidemiol Glob Health*. 2021;10:100705. doi:10.1016/j.cegh.2021.100705).

Мета дослідження – визначити прогностичні чинники несприятливого перебігу SARS-CoV-2-асоційованої пневмонії.

5.1 Опис набору даних

Проведений ретроспективний аналіз результатів лікування SARS-CoV-2-асоційованої пневмонії 121 пацієнта, що перебували в Чернівецькій міській лікарні №1 (з березня 2020 р. – опірна клінічна база по наданню медичної допомоги хворим з SARS-CoV-2). Критерієм включення в дослідження був

середній чи важкий варіанти перебігу SARS-CoV-2-асоційованої пневмонії, критерієм виключення – смерть в лікарні раніше п'ятої доби від поступлення. Згідно результатів сформовано дві групи: 60 пацієнтів (перша група), чиє захворювання закінчилося летально, і 61 пацієнтів (друга група), котрі в задовільному стані були виписані з лікарні.

Кожного пацієнта можна описати великою кількістю параметрів. В якості параметрів, серед котрих намагалися виокремити чинники несприятливого прогнозу, розглядали 77 показників, які умовно можна поділити на 9 блоків відповідно від робочих гіпотез. Так, включення віково-гендерних показників в аналіз мало на меті перевірку робочої гіпотези щодо вищого рівня смертності в старшій віковій групі. Аналіз даних анамнезу хвороби (тривалість захворювання до госпіталізації, важкість перебігу на момент поступлення, тривалість гіпертермії на момент поступлення) проведений для підтвердження/спростування робочої гіпотези щодо зростання смертності при запізнілому зверненні хворого за спеціалізованою стаціонарною допомогою. Показники, що описували стан пацієнта при поступленні (температура, артеріальний тиск, сатурація без дотації кисню і з дотацією, ступінь враження легень згідно даних КТ і УЗД, ускладнення і супутній діагноз), включені в дослідження для визначення ймовірних несприятливих чинників, які б могли допомогти лікарю в прийнятті рішень вже при першому огляді хворого. Окремо як прогностичні чинники оцінювали динаміку показників температури, тиску, сатурації на 5-7 день від поступлення.

Враховуючи патофізіологічні особливості розвитку SARS-CoV-2-інфекції з всього різноманіття лабораторних показників обрані параметри периферійної крові: гемоглобін, гематокрит, лейкоцити, паличко-ядерні і сегменто-ядерні лейкоцити, лімфоцити, ШОЕ; показники гострої фази запалення: С-реактивний білок, інтерлейкін-6, прокальцитонін; біохімічні параметри: білок і його фракції, білірубіни (загальний, прямий, непрямий), сечовина і креатинін, трансамінази крові (АЛАТ і АСАТ), глюкоза, феритин; а також параметри згортання крові:

протромбіновий час, фібрин, МНВ (міжнародне нормалізоване відношення), Д-димер. Аналіз лабораторних показників проводили двічі: на момент поступлення і на 5-7 день лікування.

Вважаємо доцільним зауважити, що обрані параметри є загальнопоширеними, простими і абсолютно доступними в рутинній практиці надання спеціалізованої медичної допомоги пацієнту з SARS-CoV-2-асоційованою пневмонією.

Достеменно не має однозначної гіпотези залежності імовірності смерті пацієнта із цими параметрами.

Задачу можна віднести до задачі класифікації, де необхідно визначити належність до одного з класів (помре чи житиме) на основі множини різних факторів.

Попередня підготовка даних.

На цьому етапі слід виділити декілька кроків, що зумовлені особливістю отримання та збереження даних. Для реалізації кожного кроку був написаний скрипт на мові python.

Вилучення полів

З розрахунку вилучаються поля, що містять персоналізовану інформацію та ті, що однозначно не впливають на діагноз. Зокрема: номер і/х, ПІП, телефон, діагноз, ускладнення, КТ (залишається ступінь КТ), КТ опис, УЗД.

Верифікація

Особливістю наявних даних є те, що всі вони внесені людьми, а це призводить до наявних технічних посилок. Отже першою процедурою є верифікація даних та виправлення їх в автоматичному та ручному режимі. Для цього був створений скрипт, що визначає, та по можливості виправляє людські помилки.

Трансформація та зміна значень полів

Значна кількість полів не є придатна для цифрового аналізу, так як вона містить інформацію у текстовому форматі, що не відповідає жодному із придатних для аналізу типів даних. Зокрема були проведені такі зміни в даних:

Стать – уніфіковані позначення статі людини, та приведено до категорійного типу даних ['ч', 'ж'].

Перебіг – приведено до категорій ['важкий', 'серед важ', 'вкрай важкий ']

ЛН – проміжний діагноз перетворено до верхнього рівня (II-III –> III) та приведено до категорійного поля: ['IV', 'III', 'II', 'I']

Супутній – значення цього поля змінюється на True, якщо в супутньому діагнозі присутній цукровий діабет. В інакшому випадку значення поля – False.

ступінь КТ – перетворюється на категорійне поле

Т вище38, Т вище 37,5, Т 37-37,5 – фіксується температура, яка була до віднесення пацієнта у стаціонар.

АТ пост та АТ стац – виділяється верхня та нижня границі артеріального тиску та розташовуються в нових полях

Дотація – виділяються числові значення та вилучається будь-яка текстова інформація

Дата розтину – фактично представляє вихідне поле та перетворюється в бінарне, що є індикатором помер пацієнт чи ні.

Всі інші поля є числовими та залишаються без змін. В результаті отримані наступні вхідні поля: ['Вік', 'Стать', 'Перебіг', 'ЛН', 'Супутній', 'ступінь КТ', 'spO2 без дот', 'spO2 з дот', 'spO2 випис', 'Днів до поступ', 'Поступлення', 'Виписка', 'Ліжко днів', 'Дата розтину', 'Т вище38', 'Т вище 37,5', 'Т 37-37,5', 'Дотація', 'Гемогл1', 'Гемогл2', 'Л1', 'Л2', 'П/Я 1', 'П/Я2', 'С/Я1', 'С/Я2', 'Лім1', 'Лім2', 'ШОЕ1', 'ШОЕ2', 'СРБ1', 'СРБ2', 'ІЛ6-1', 'ІЛ6-2', 'ФЕРИТ1', 'Д-дим1', 'Д-дим2', 'ФЕРИТ2', 'Прокальцитонін', 'Гематокрит1', 'Гематокрит2', 'Прот час1', 'Прот час2', 'Фібрин1', 'Фібрин2', 'МНВ1', 'МНВ2', 'Білок1', 'Білок2', 'Сечов1', 'Сечов2', 'Креат1', 'Креат2', 'БР заг1', 'БРзаг2', 'БРпр1', 'БРпр2', 'БРнепр1', 'БРнепр2',

'АЛАТ1', 'АЛАТ2', 'АСАТ1', 'АСАТ2', 'Цукор1', 'Цукор2', 'АТ постН', 'АТ постL', 'АТ стацН', 'АТ стацL']. Вихідним є бінарне поле помер/жив.

Видалення полів з відсутніми даними

Наступним етапом є видалення записів, що містять пропущені значення. Велика кількість полів призводить до того, що спроба вилучити записи, які містять хоча б одне пропущене значення призводить до суттєвого зменшення набору даних – що унеможливить застосування методів класифікації. Для розв'язання цієї проблеми порожні значення були заповнені значенням за замовчуванням (якщо це можливо). Далі були визначені поля, що містять найбільше пропусків. До них були віднесені:

Поле	**Відсоток пропусків**
ФЕРИТ2	99%
ФЕРИТ1	98%
ІЛ6-2	95%
ІЛ6-1	83%
Д-дим2	75%
ступінь КТ	65%
Прокальцитонін	62%
Д-дим1	51%
Т 37-37,5	45%
spO2 випис	44%
МНВ2	37%
Т вище38	36%
Т вище 37,5	36%

Було прийняте рішення усунути ці поля з подальшого розрахунку, адже їх наявність унеможливить подальший аналіз. Подальше вилучення записів з пропущеними даними зменшило навчальний набір на 19%. (із 121 до 99 записів). Загальна кількість полів склало 54 – 53 вхідне та одне вихідне поле. З яких 49 – цифрових поля, 3 категорійних та одне логічне.

Визначення важливих факторів

Для визначення важливості факторів для класифікації пацієнтів був використаний метод сортування критерію xi^2 та взаємної інформації (MI). Величина фактора визначає значущість поля при класифікації. Результати приведені в таблиці:

Табл. 5.1. Визначення важливих факторів для класифікації

Фактори	xi^2	Фактори	MI
Л2	434	Лім2	0.3
П/Я2	352	Л2	0.28
Лім2	352	П/Я2	0.25
Гематокрит2	250	spO2 без дот	0.23
Креат2	226	Ліжко днів	0.20
spO2 без дот	22	Гематокрит2	0.19
Ліжко днів	183	Креат2	0.17
СРБ2	146	ГЕМОГЛ1	0.15
ЛН	68	Вік	0.13
Стать	50	Перебіг	0.13

Примітка: Л2 – кількість лейкоцитів в периферійній крові через 7 днів від поступлення, П/Я2 – кількість паличко-ядерних нейтрофілів через 7 днів від поступлення, Лім2 – кількість лімфоцитів в периферійній крові через 7 днів від поступленння, Гематокрит2 – гематокрит через 7 днів від поступлення, Креат2 – креатинін через 7 днів від поступлення, spO2 без дот – сатурація крові без дотації кисню, ліжко днів – кількість дні, проведених в лікарні до виписки/смерті, СРБ2 – С-реактивний білок в сироватці крові через 7 днів від поступлення, ЛН – легенева недостатність, ГЕМОГЛ1 – рівень гемоглобіну при поступленні.

Як видно з табл. 5.1, перші сім факторів у двох методів співпадають. Відрізняється тільки їх важливість. Тому, набір даних був скорочений до перших семи факторів.

Наступним була перевірка наявності взаємокореляції, для цього був проведений автокореляційний аналіз, який показав відсутність факторів, що мають сильну взаємну кореляцію.

Рис. 5.1. Автокореляційна матриця вхідних параметрів

5.2 Побудова класифікаційної моделі

Для визначення класифікаційних правил та візуалізації результатів був використаний метод дерев рішень. Основною причиною вибору цього методу є можливість візуалізувати результат класифікаційного аналізу у вигляді дерева прийняття рішень. Однак точність цього методу є не найкращою.

Це непараметричний метод навчання з вчителем, який використовується для класифікації та регресії. Мета полягає в тому, щоб створити модель, яка прогнозує значення цільової змінної, вивчаючи прості правила прийняття рішень, виведені з характеристик даних. Дерево можна розглядати як кусково-постійне наближення.

Критерієм вимірювання порогу розщеплення був вибраний коефіцієнт Джіні (gini) – показник нерівності розподілу деякої величини чисел, що приймає значення між 0 і 1, де 0 означає абсолютну рівність (величина приймає лише одне значення), а 1 позначає повну нерівність. Стратегія, яка використовується для вибору розщеплення в кожному вузлі – це пошук найкращого розподілу.

Для навчання та визначення точності алгоритму набір даних був поділений на навчальний та тестовий в пропорції 70/30. В результаті отримане дерево рішень представлене на рис. 5.2:

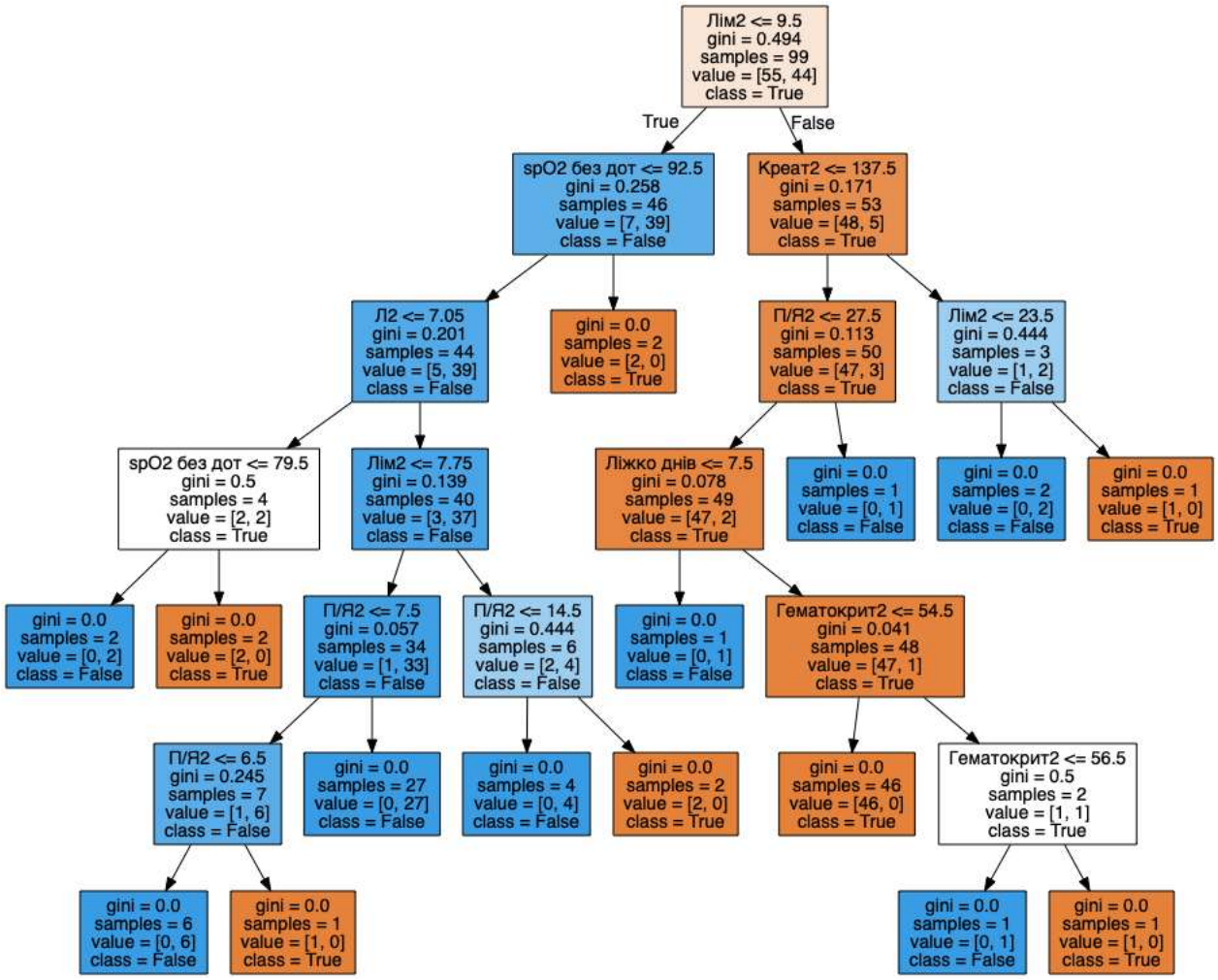


Рис. 5.2. Дерево рішень визначення чи помре (False) пацієнт, чи залишиться живим (True). Gini – нерівність розподілу, samples – кількість випадків, value – значення функції класифікатора, class – належність до класу.

Побудова дерева рішень дала змогу встановити важливість факторів саме для цього класифікатора:

Табл. 5.2. Важливість факторів дерева рішення

Фактор	Важливість
Лім2 (кількість лімфоцитів в периферійній крові через 7 днів)	0.62
П/Я2 (кількість паличко-ядерних нейтрофілів в периферійній крові через 7 днів)	0.13
spO2 без дот (сатурація крові без дотації кисню при поступленні)	0.12
Креат2 (рівень креатиніну через 7 днів)	0.04
Ліжко днів	0.04
Л2 (кількість лейкоцитів в периферійній крові через 7 днів)	0.03
Гематокрит2 (гематокрит через 7 днів від поступлення)	0.02

Як видно з таблиці, найважливішим фактором дерева рішень є кількість лімфоцитів через тиждень від поступлення (Лім2). Лімфоцитопенія як маркер важкості перебігу SARS-CoV-2-інфекції описана в науковій літературі останніх років, наше дослідження доводить важливість врахування показника як маркера несприятливого прогнозу [50,51]. Наростання депресії лімфоцитів протягом тижня інтенсивної терапії пацієнта з SARS-CoV-2-асоційованою пневмонією свідчить про виснаження імунного захисту і збільшує ймовірність фатального результату.

Наступним за важливістю фактором є загальновизнаний показник активності запального процесу - кількість паличко-ядерних нейтрофілів [53] на сьомий день від початку стаціонарного лікування SARS-CoV-2-асоційованої пневмонії. Так, прогностично несприятливим виявилося поєднання збільшення вмісту паличко-ядерних нейтрофілів на фоні граничної лімфоцитопенії. Тобто шанси вижити у пацієнта із важким перебігом SARS-CoV-2-асоційованої пневмонії зменшуються при різкій активації запалення на тлі пригнічення специфічної імунної відповіді.

Третій важливий фактор дерева рішень – сатурація крові без дотації кисню на момент поступлення пацієнта. Низька сатурація на момент поступлення опосередковано віддзеркалює важкість стану пацієнта і масивність враження легень, вказує на виснаження захисних і компенсаторних можливостей організму, декомпенсацію роботи серцево-судинної системи, важку гіпоксію [54]. Цілком логічною виявилася інформативність цього показника як предиктора несприятливого прогнозу захворювання.

Точність навчання та тестування склала відповідно 100% та 79%. Висока точність на навчальній вибірці та низька на тестовій (велика дисперсія) свідчить про перенавчання цього методу. Однак він дозволяє лікарю власноруч провести пацієнта по цьому дереву та швидко визначити клас, до якого він належить. Створення ж автоматизованої програми прийняття рішень на основі дерев не складає жодної проблеми.

Наведемо декілька прикладів використання дерева рішень тестового набору. Пацієнт 1 поступив в стаціонар з сатурацією 85%, через тиждень його аналізи засвідчили: кількість лейкоцитів 34,9 Г/л, паличко-ядерних нейтрофілів – 24, лімфоцитів -3, гематокрит – 43,1, креатинін – 142. Проведемо пацієнта по дереву рішень: Лім2 <= 9.5 (так) –> spO2 без дот <=92.5 (так) –> Л2<=7.05 (ні) –> Лім2 <= 7.75 (так) –> П/Я2 <= 7.5 (ні) –> Class False, тобто прогнозуємо летальний результат. Дійсно, на 12-ту добу від поступлення зафіксовано смерть пацієнта.

Пацієнт 2 поступив в стаціонар з сатурацією 91%, через тиждень його аналізи засвідчили: кількість лейкоцитів 6,1 Г/л, паличко-ядерних нейтрофілів – 2, лімфоцитів -9, гематокрит – 46, креатинін – 117. Проведемо пацієнта по дереву рішень: Лім2 <= 9.5 (так) –> spO2 без дот <=92.5 (так) –> Л2<=7.05 (так) –> spO2 без дот <= 79.5 (ні) –> Class True, що дає можливість зробити сприятливий прогноз. Дійсно, на 13-ту добу від поступлення пацієнт виписаний з лікарні.

5.3 Ансамбль класифікаційних моделей

Для покращення якості навчання та усунення перенавчання було прийнято рішення збільшити навчальний набір та використати ансамбль класифікаційних моделей. Для цього в якості навчального набору було використано всі наявні записи. А для отримання тестового було проведено додаткове дослідження пацієнтів. Тестовий набір склав 83 записи.

Для ансамблю моделей були використані наступні класифікатори:

1. Логістична регресія
2. Квадратичний класифікатор
3. Наївний Баєсів класифікатор
4. Класифікатор на основі випадкового лісу
5. Adaptive Boosting класифікатор

Нові дані отримувались в тому самому відділені тому і розподіл тестової вибірки був тим самим.

Комбіноване рішення цих методів знаходилось за допомогою Voting Classifier. Основна ідея класифікатора голосування полягає в тому, щоб об’єднати концептуально різні класифікатори машинного навчання та використовувати більшість голосів або середні прогнозовані ймовірності (м’яке голосування) для прогнозування міток класів. В нашому випадку було використано «жорстке» голосування, тобто вибір класу визначався більшістю «голосів» класифікаторів. В результаті точність навчання склала: 93%. А точність на тестовому наборі даних зросла до 81%.

Отже використання ансамблю моделей дало змогу уникнути перенавчання та підвищити точність прогнозу. Помилка прогнозу на навчальному наборі становить 7% а, дисперсія тестового набору від навчального 12%. Отже можна прийти до висновку, що для зменшення дисперсії (зменшення помилки тестового набору) достатньо просто збільшити навчальний набір. Це призведе до незначного зменшення точності навчального набору та суттєвого зростання точності тестового. В ідеалі збільшення навчального набору даних може призвести до зростання точності орієнтовно до 85% (при неминучому зменшенні точності на навчальному наборі даних).

Подальше збільшення точності двох показників можливе при умові одночасного зростання навчальної вибірки та включення в розрахунок нових факторів, або ускладнення класифікаційних моделей, наприклад долучення до ансамблю класифікаторів на основі нейронних мереж.

Таким чином, єдиним прогностичним маркером несприятливого перебігу SARS-CoV-2-асоційованої пневмонії на момент поступлення пацієнта в стаціонар виявився рівень сатурації крові без дотації кисню менше 92,5%. Це базовий показник, який першим перевіряється при поступленні таких пацієнтів, визначає необхідність і об’єм кисневої дотації. Натомість ні важкість стану, ні супутній цукровий діабет, ні тривалість хвороби до поступлення не свідчили про підвищений ризик летального результату. Масивність враження легень згідно даних комп’ютерної томографії чи УЗД не збільшували і не зменшували шанси пацієнта померти від SARS-CoV-2-асоційованої пневмонії. Так само, жоден з лабораторних показників не давав можливість робити фатальний прогноз.

Натомість вже через 5-7 днів після початку стаціонарного лікування з’являються інформативні прогностичні лабораторні маркери, зокрема кількість лімфоцитів і паличко-ядерних нейтрофілів в периферійній крові. Наростання активності запального процесу, що віддзеркалюється в рості паличко-ядерних нетрофілів і загального пулу лейкоцитів на тлі депресії лімфоцитарної ланки, свідчить про виснаження специфічного імунного захисту організму, втрату

імунологічного контролю запалення і вказує на високу ймовірність фатального прогнозу.

Використання загальновживаних параметрів опису стану пацієнтів, котрі щодня рутинно використовуються в клінічній практиці, доступне і зрозуміле дерево рішень дозволить лікарю за декілька хвилин визначити прогноз і, відповідно, зрозуміти необхідність корекції лікування, переведення пацієнта в відділення невідкладної терапії тощо.

РОЗДІЛ 6 МОДЕЛЬ ПЕРЕДБАЧЕННЯ СКЛАДНОСТІ ПЕРЕБІГУ ЗАХВОРЮВАННЯ НА ОСНОВІ ІЄРАРХІЧНОГО ПРЕДИКТОРА

6.1 Побудова експерименту та опис набору даних

Експеримент організований таким чином:

- Дослідницький аналіз даних (нормалізація та кодування);
- Побудова простору ознак на основі визначення важливих атрибутів;
- Відбір слабких предикторів;
- Розробка ієрархічного предиктора;
- Оцінка результатів.

Усі розрахунки були зроблені за допомогою RStudio. Дані передавались через Sampler Data для балансування. В результаті всі випадки враховуються. Це означає, що зібраний набір даних збалансований.

Ми обробляємо персоналізовані дані пацієнтів зі зібраного набору даних (https://doi.org/10.6084/m9.figshare.14865411.v1). Цей набір даних зібрано у хірургічному відділенні Львівської державної лікарні (Україна). Пацієнти пройшли клінічне лікування післяопераційних ускладнень в черевній порожнині.

Набір даних складається з таких характеристик:

- Вік - ціле число,
- Стать – булева змінна,
- Вага –категорійна змінна,
- Дата прийому - дата,
- Діагноз – категорійна змінна,
- Пов'язана діагностика –категорійна змінна,
- Флора – категорійна змінна,
- Медикамент –категорійна змінна,

- Діюча речовина –категорійна змінна,
- Time_in_hospital (кількість ліжко-днів у лікарні, цільовий параметр)-ціле число.

Завдання полягає в тому, щоб передбачити кількість днів перебування в лікарні (тривалість лікування) на основі медикаментозного лікування та особистих параметрів пацієнтів.

Набір даних складається з 51 екземпляра та 10 параметрів. Time_in_hospital - це цільова змінна. Після етапу попередньої обробки та одноразового гарячого кодування використання набору даних категорійних змінних складається з 39 функцій. Пропущені значення знаходяться у функції Flora та у параметрі Пов'язана діагностика. Процедура внесення відсутніх даних не використовується, оскільки природа відсутніх даних - це дійсно порожнє значення.

Для побудови простору умов виконуються наступні кроки:

- вибір найважливіших ознак;
- екземпляри розбиваються на кластери з подібними залежними від часу та незалежними від часу параметрами.

Початковий вибір ознак здійснюється кореляційною матрицею, Борутою та деревом регресії. Жорстке голосування використовується для остаточного відбору функцій.

Співвідношення між параметрами наведено на рис. 6.1

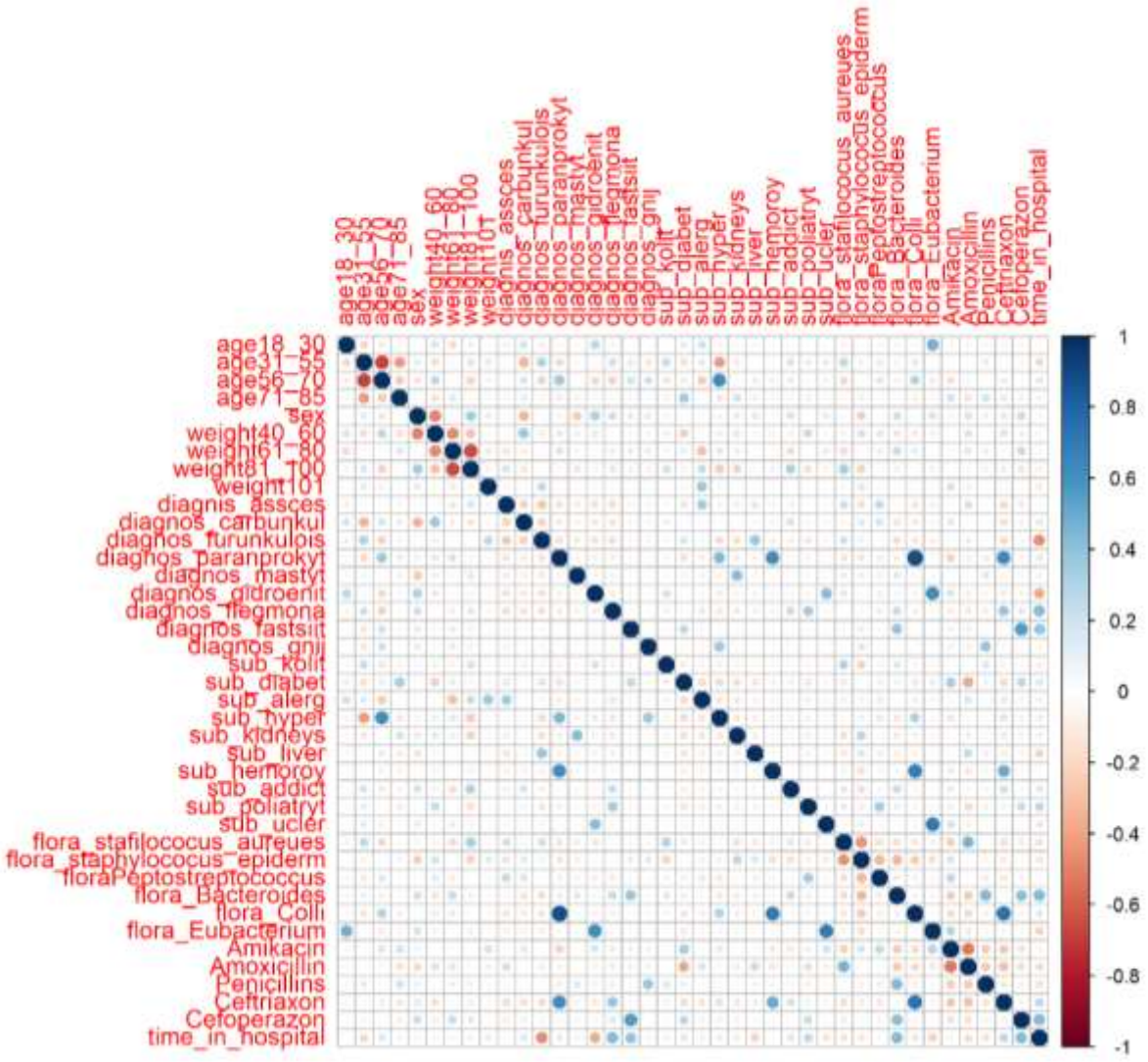


Рис. 6.1. Кореляційна матриця

З рис. 6.1 видно, що значна кореляція відсутня.

Використовується наступний алгоритм Boruta [21] (рис. 6.2).

Алгоритм Boruta - це обгортка, побудована навколо алгоритму класифікації випадкового лісу. На рис. 6.2 показано важливі (зелений) та попередньо важливі (жовтий) ознаки. Ділянки синього кольору відповідають мінімальній, середній та максимальній оцінці Z атрибута тіні.

Нижче наведені важливі змінні.

Нижче наведені важливі змінні.

	meanImp	decision
diagnos_furunkulois	18.631571	Confirmed

diagnos_gidroenit	9.901603	Confirmed
diagnos_flegmona	8.915723	Confirmed
sub_diabet	7.557117	Confirmed

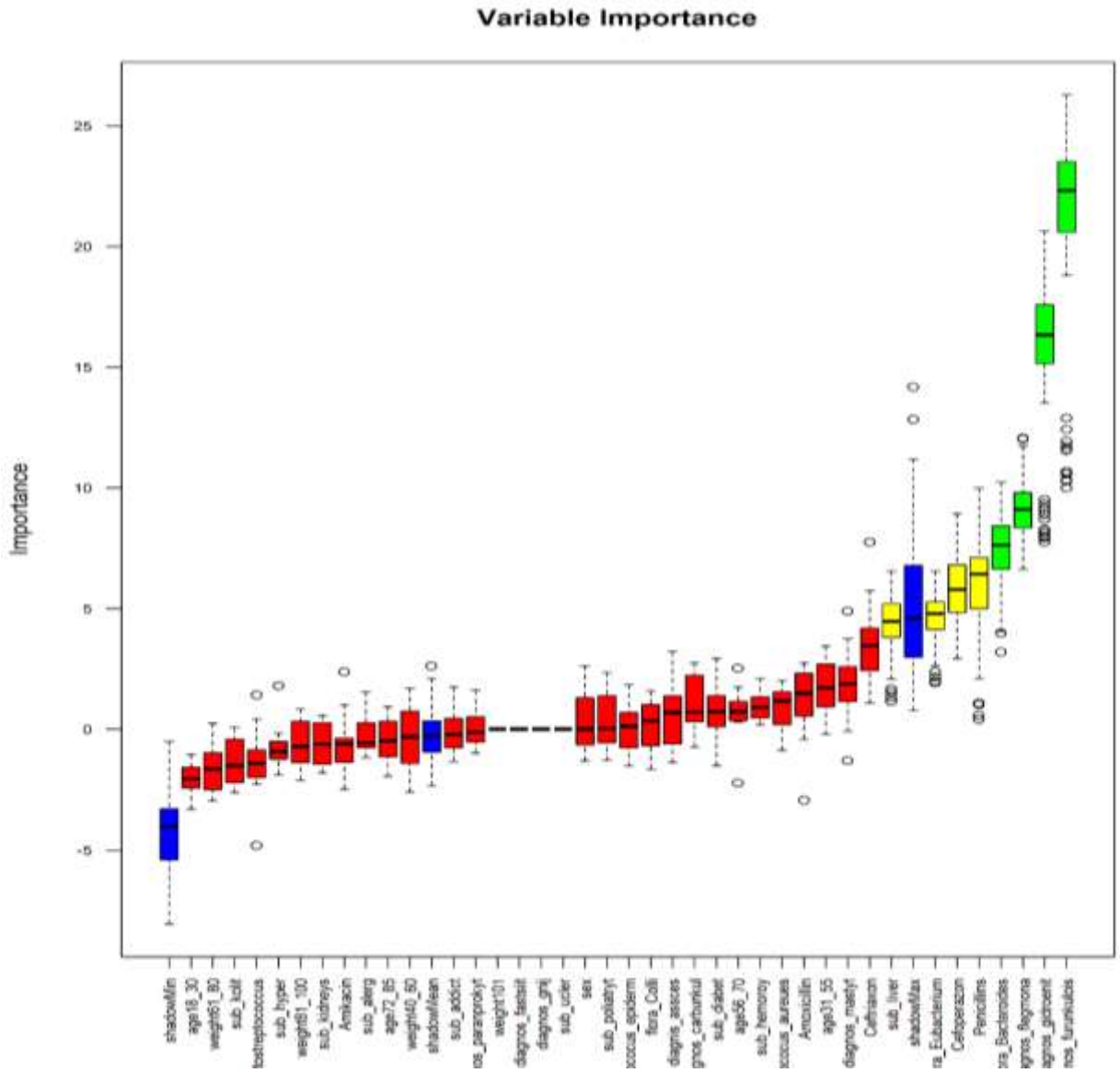


Рис. 6.2. Результат Boruta

Далі для відбору ознак використано регресійне дерево.

Регресійне дерево подано нижче:

1) root 51 1981.92200 10.372550

2) diagnos_furunkulois>=0.5 11 17.63636 4.818182 *

3) diagnos_furunkulois< 0.5 40 1531.60000 11.900000

6) sub_diabet< 0.5 33 1075.87900 11.060610

12) Ceftriaxon< 0.5 26 818.65380 10.115380

24) sex>=0.5 12 602.66670 8.666667 *

25) sex< 0.5 14 169.21430 11.357140 *

13) Ceftriaxon>=0.5 7 147.71430 14.571430 *

7) sub_diabet>=0.5 7 322.85710 15.857140 *

Подане вище регресійне дерево дає змогу знайти важливі атрибути та їх значення, які є умовою поділу на гілки. Отже, відібрано такі ознаки:

- diagnos_furunkulois,
- sub_diabet,
- Ceftriaxon,
- sex.

Далі порівнюємо відібрані ознаки різними методами. Для цього застосовуємо метод hard voting для трьох методів відбору ознак: кореляційної матриці, Boruta та регресійного дерева. Результат подано нижче.

- diagnos_furunkulois,
- sub_diabet,
- diagnos_gidroenit,
- Ceftriaxon,
- diagnos_flegmona,
- sex.

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

Далі використовується кластеризація. Візуальна оцінка (Кластерної) тенденції (VAT) використовується для аналізу можливості поділу об'єктів. ПДВ демонструє погану тенденцію до кластеризації (рис.6.3). Невеликі відмінності представлені темними відтінками, а великі - світлими [22].

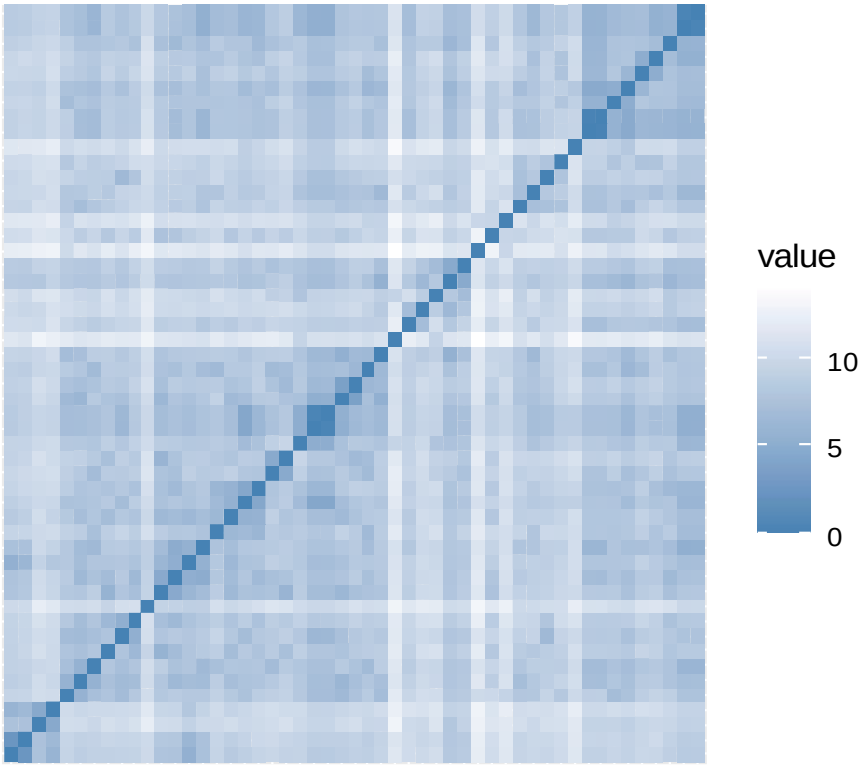


Рис. 6.3. Результат VAT-аналізу

Такий самий результат ми отримаємо, застосувавши k-means (Рис. 6.4). Ми бачимо явне перекриття кластерів, особливо першого та третього. Це свідчить про низьку тенденцію до групування.

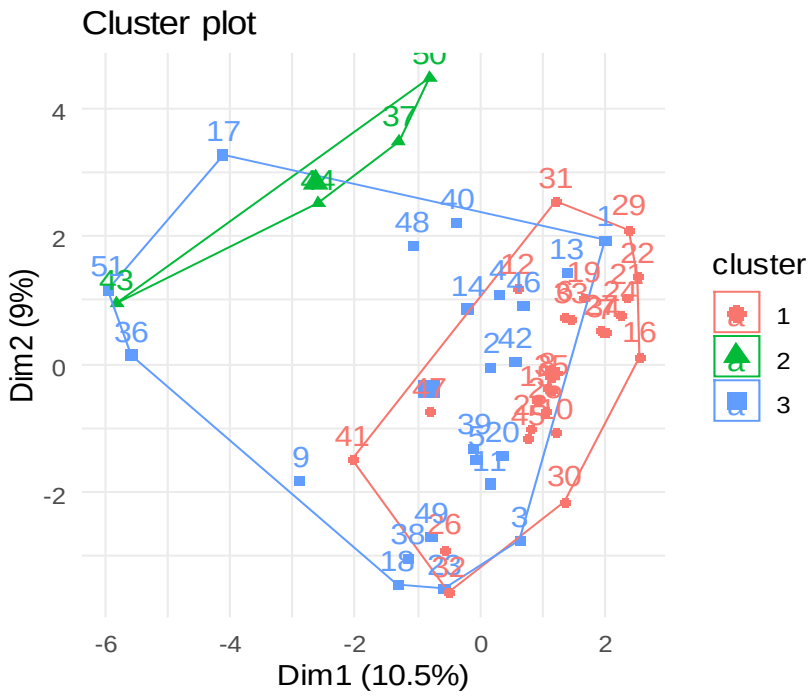


Рис. 6.4. Результат кластеризації га основі kmeans

Статистика Хопкінса [23, 42, 43] дорівнює 0.71. Це також свідчить про те, що набір даних не надається до кластеризації.

Для підтвердження цього скористаємося методом Fuzzy c-mean. Таблиця 6.1 демонструє результати кластеризації. Усі елементи, які відносяться до одного кластеру, позначені потовщенням.

Табл. 6.1. Результат методу fuzzy c-mean

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

Отже, набір даних розділимо на 3 кластери. Об’єкти 7, 13 17, 20, 26, 34, 39, 42, 45, 47, 50, аналізуватимуться окремо, оскільки значення функції приналежності для цих об’єктів менше, ніж 0.65. Отже, між ними не можна знайти значної відмінності.

6.2 Вибір слабких предикторів

На наступному кроці ми будуємо слабкі предиктори та порівнюємо їх результат по точності. Для цього обрано дві метрики: Root Mean Squared Error (RMSE) and Mean Absolute Percentage Error (MAPE). Результати для повного набору даних (усі ознаки) подано в табл. 6.2.

Табл. 6.2. Результати предикторів

Модель	RMSE	MAPE
Linear regression	6.109753	0.4946111
Regression tree	4.970676	0.4763212
Random forest (500 trees, mtree-3)	3.560431	0.3753441
knn	3.360431	0.3753441
SVM with Radial Basis kernel	3.194611	0.2923931
SVM with Polynomial kernel	2.262171	0.2670496
ANN with 12 units in single hidden layer	2.0972937	0.2025539

Результат предикторів для відібраних ознак подано у рис. 6.3.

Табл. 6.3. Точність передбачення для моделей, побудованих на відібраних ознаках.

Model based on selected variables	RMSE	MAPE
Linear regression	3.972548	0.3240777
Regression tree	4.970676	0.4763212
Random forest (500 trees, mtree-3)	3.546447	0.2730968
knn	3.346447	0.2730968
SVM with Radial Basis kernel	3.095239	0.2386008
SVM with Polynomial kernel	2.226906	0.1766799
ANN with 12 units in single hidden layer	2.059849	0.1513523

6.3 Розроблення ієрархічного предиктора

На наступному етапі для кожного відокремленого кластера будуть використовуватися слабкі предиктори. До четвертого кластера додано екземпляри з невідомим кластером. Всього враховано 4 кластери.

Ієрархічний предиктор будується за допомогою таких кроків:

1. Fuzzy c-means ділить об’єкти на чотири кластери (табл. 6);

2. Для кожного кластера окремо використовуються випадковий ліс лінійної регресії, SVM з ядром Radial Basis та SVM з поліноміальним ядром;

3. Забезпечується середнє голосування за отримані результати. На основі цього буде вибрано середнє значення.

Точність прогнозування ієрархічного предиктора представлена в табл. 6.4 Як бачимо, точність передбачення є гірша для всього набору даних.

Повторна перехресна перевірка K-кратними методами використовується для перевірки невеликих наборів даних. Перевагою цієї техніки є здатність до паралелізації. Результат представлений у табл. 6.4.

Табл. 6.4. Точність передбачення розроблених предикторів.

Model based on whole variables	RMSE	MAPE
Hierarchical predictor	1.401258	0.137792
Model based on selected variables	**RMSE**	**MAPE**
Hierarchical predictor	1.401257	0.102961
Hierarchical predictor with repeated K-fold cross-validation, 5 fold, repeated 3 times	1.401125	0.102753

РОЗДІЛ 7 РОЗПАРАЛЕЛЕННЯ ПРОПОНОВАНИХ МОДЕЛЕЙ

7.1 Розпаралелення ієрархічного предиктора

Розпаралелення здійснюється на основі задач, тобто кожен предиктор працює на окремому ядрі.

Використано 4-ядерний процесор.

Ми бачимо, що RMSE для ієрархічної бази прогнозування у всьому наборі даних не набагато вище, ніж для вибраних функцій. З іншого боку, точність прогнозування для розділених кластерів краща, ніж для найкращого слабкого предиктора Штучна нейронна мережа (ANN) з 12 одиницями в одному прихованому шарі. Точність ієрархічного предиктора з повторною перехресною перевіркою у K разів не настільки вища, ніж для неперехресно перевіреного методу.

Усі тривалі операції розробленого алгоритму можуть бути виконані паралельно. Єдина операція, що повинна бути виконана частково послідовно – hard voting. Оскільки операція hard voting є комутативною і асоціативною, то її реалізовано з використанням алгоритму здвоювання, що дало змогу отримати часову оцінку складності $t_{merge} = O\left(\log_2(k) \cdot V_{stat}\right) = O\left(\log_2(k) \cdot \left(\frac{n}{k \cdot minSupport}\right)^{1+\log_{avgD}(m)}\right)$. У випадку послідовного виконання алгоритму цією операцією можна знехтувати, оскільки $t_{merge} = o\left(t_{stat}\right)$, проте у випадку паралельного виконання її необхідно враховувати.

На кожен сервер даних у системі із k серверів припадає $\frac{n}{k}$ кортежів даних у випадку рівномірного розподілення по серверах.

Таким чином, розроблений алгоритм на системі із k процесорами має часову складність

$$t_k = O\left(minSupport \cdot \left(\frac{n}{k \cdot minSupport} \right)^{1+\log_{avgD}(m)} + \log_2(k) \cdot \left(\frac{n}{k \cdot minSupport} \right)^{1+\log_{avgD}(m)} \right) +$$

$$+O\left(\frac{Z_{aggr}^2 \cdot \log\left(sz_{aggr} \right)}{k \cdot m \cdot D(A)} \right) =$$

$$= O\left(\left(minSupport + \log_2(k) \right) \cdot \left(\frac{n}{k \cdot minSupport} \right)^{1+\log_{avgD}(m)} + \frac{Z_{aggr}^2 \cdot \log\left(sz_{aggr} \right)}{k \cdot m \cdot D(A)} \right)$$

Якщо $k = O(n)$, то

$$t_n = O\left(\left(minSupport + \log_2(n) \right) \cdot minSupport^{-1-\log_{avgD}(m)} + \frac{Z_{aggr}^2 \cdot \log\left(sz_{aggr} \right)}{n \cdot m \cdot D(A)} \right)$$

Для даних, що мають практичну цінність, кількість залежностей Z_{aggr} складає не більше декількох тисяч у той час, як кількість кортежів n може обчислюватися мільярдами. Тому у випадку паралельних обчислень на значній кількості комп'ютерів другим членом функції t_n можна знехтувати. З тих же міркувань $\log_2(n) = o(minSupport)$. Таким чином, асимптотична оцінка часу виконання алгоритму на системі із k комп'ютерів складає:

$$t_n = O\left(minSupport^{-\log_{avgD}(m)} \right) \qquad (7.1)$$

Отримана оцінка часу виконання алгоритму є субполіноміальною, а отже, розроблений алгоритм є ефективним паралельним алгоритмом.

Розроблений алгоритм дає змогу стверджувати, що задача виявлення кон'юнктивних асоціативних залежностей у розподілених базах даних належить до класу P-задач.

На практиці досягнення теоретичної ефективності паралелізму неможливе через затримки передачі даних, синхронізації та ін. Результати тестів показані на графіках:

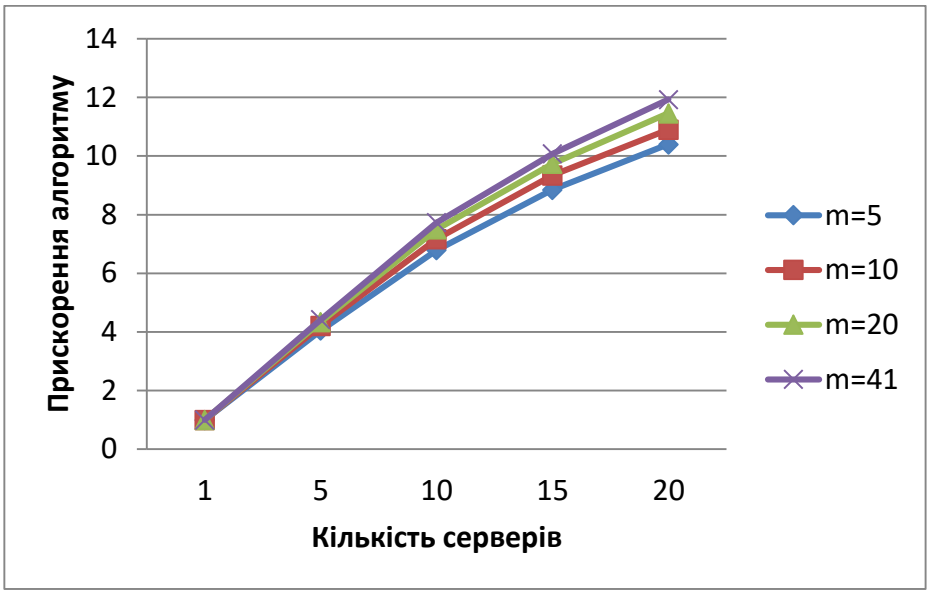


Рис. 7.1. Графік прискорення розробленого алгоритму

Залежність прискорення розробленого алгоритму від кількості серверів, як і було передбачено, є нелінійним. Але ефективність паралелізації залишається високою навіть для десятків серверів.

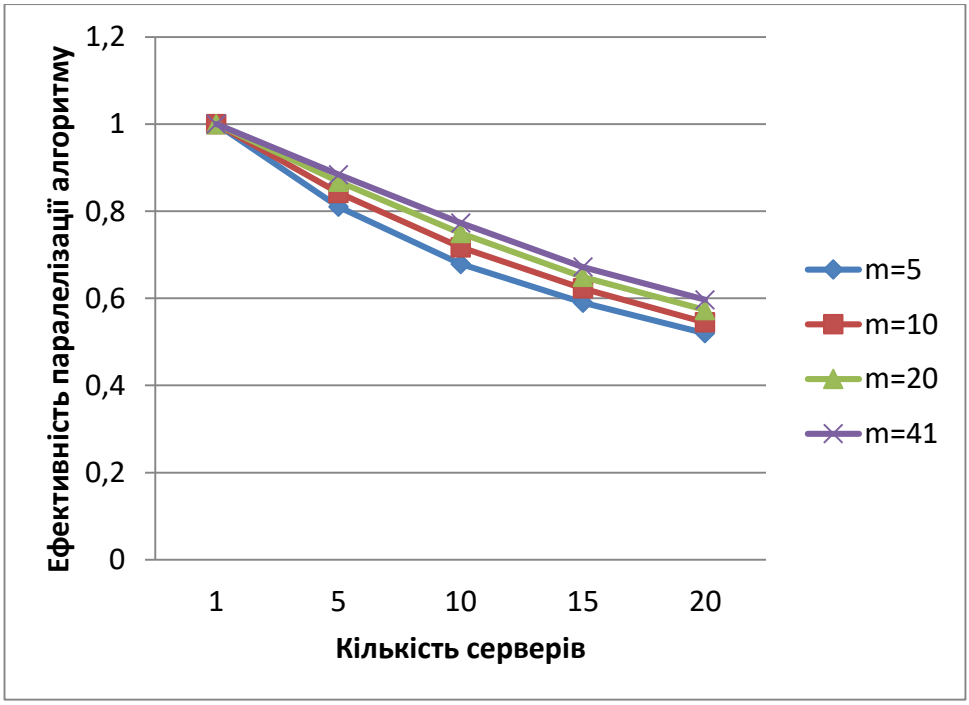


Рис. 7.2. Графік ефективності паралелізації розробленого алгоритму

Отримані результати підтверджують теоретичну оцінку складності алгоритму при паралельному виконанні на декількох серверах даних. Як і очікувалось, ефективність алгоритму зменшується із зростанням кількості

серверів через збільшення затрат на передачу даних та об’єднання результатів. Проте зменшення ефективності достатньо повільне і дає змогу виконувати алгоритм на розподілених сховищах даних із десятків серверів.

7.2 Прискорення паралельного алгоритму

Прискорення паралельного алгоритму є його найінформативнішою характеристикою, яка показує в скільки раз застосування паралельного алгоритму зменшує час розв'язку завдання в порівнянні з послідовним алгоритмом. Прискорення паралельного алгоритму визначається величиною

$$S_N = \frac{T_1}{T_N}, \qquad (7.2)$$

де T_1 – час виконання алгоритму на одному процесорі, T_N – час виконання того ж алгоритму на N процесорах.

Ідеальним, мабуть, є прискорення $S_N = N$. У реальності це прискорення недосяжне. Перерахуємо причини неможливості досягнення ідеального прискорення:

- відсутність максимального паралелізму в алгоритмі;
- незбалансованість завантаження процесорів (якщо, наприклад, необхідно додати два вектори з 9 чисел кожний на восьмипроцесорній системі);
- тимчасові витрати на обмін даними, конфлікти в пам'яті й на синхронізацію.

Розглянемо як приклад прискорення, яке може забезпечити алгоритм подвоєння при множенні n дійсних чисел на $N = \frac{n}{2}$–процесорній обчислювальній системі. Припустимо, що числа a_1, a_2 зберігаються на першому процесорі, числа a_3, a_4 – на другому процесорі, і т.д. Нехай t – час виконання додавання на одному процесорі, а αt – час обміну одним дійсним числом між сусідніми процесорами. Тоді шукане прискорення

$$S_N \approx \frac{nt}{\log_2 n(1+\alpha)t}. \qquad (7.3)$$

Помітимо, що більш строго в чисельнику формули, поданої вище, варто було б написати $(n-1)$, а в знаменнику – $(\log_2 n-1)(1+\alpha t)$. Однак, при досить великих n одиницею в чисельнику й у першому зі співмножників знаменника можна знехтувати.

Оскільки $\log_2 n = \log_2(2N) = 1+\log_2 N$, з рівності (7.3) випливає

$$S_N \approx \frac{1}{1+\alpha}\frac{2N}{1+\log_2 N}. \qquad (7.4)$$

Очевидно, ідеальною ситуацією є ситуація, коли можна знехтувати часом обміну даними, тобто коли можна присвоїти $\alpha = 0$. Порівняємо цей випадок з випадком, коли час обміну дорівнює часу виконання множення $\alpha = 1$. З формули (7.3) випливає, що при цьому прискорення алгоритму подвоєння падає у два рази (рис. 7.3).

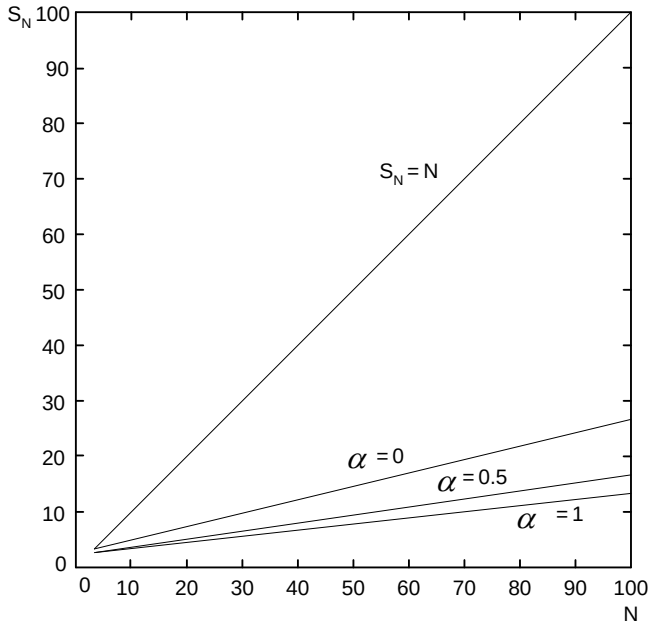


Рис. 7.3. Залежність прискорення алгоритму здвоювання від числа процесорів у системі. Пряма $S_N = N$ відповідає ідеальному випадку.

Як уже зазначалось, у формулі (7.4) величина T_1 – час виконання алгоритму на одному процесорі. Але, як ми бачили, паралельний алгоритм може містити зайві операції в порівнянні з відповідним послідовним алгоритмом. Тобто, загалом кажучи, паралельний алгоритм не є «гарним» для послідовної

реалізації. Тому, більш строго, у формулі (7.4) у якості величини T_1 слід використовувати час виконання найшвидшого послідовного алгоритму на одному процесорі.

7.3 Ефективність паралельного алгоритму.

Ефективний паралельний алгоритм – алгоритм, що може бути виконаний на PRAM з кількістю процесорів $O(n)$ за поліноміальний час. Ефективність паралельного алгоритму визначається величиною

$$E_N = \frac{S_N}{N}, \tag{7.5}$$

де S_N – прискорення паралельного алгоритму, N – кількість процесорів у системі.

З того факту, що $S_N \le N$, випливає обмеження зверху на величину ефективності: $E_N \le 1$.

Розглянемо ефективність, яку може забезпечити алгоритм подвоєння. Маємо

$$E_N \approx \frac{1}{1+\alpha}\frac{2}{1+\log_2 N}, \tag{7.6}$$

що проілюстровано на рис. 7.4.

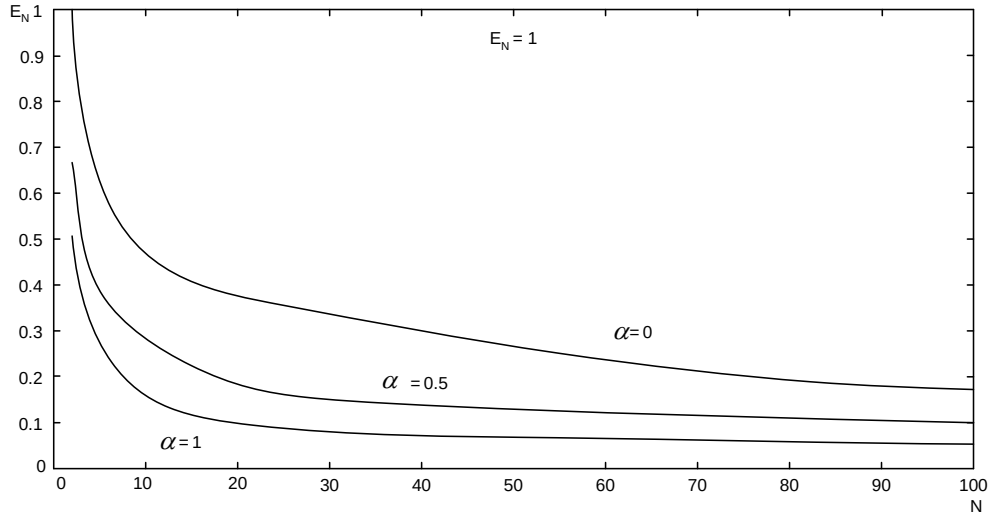


Рис. 7.4. Залежність ефективності алгоритму здвоювання від числа процесорів у системі. Пряма $E_N=1$ відповідає ідеальному випадку.

Можна виділити такі основні причини втрати ефективності паралельних обчислень:

- час ініціалізації паралельної програми (startup);
- незбалансованість завантаження процесорів (load imbalance);
- витрати на комунікації (communication costs);
- наявність у програмі послідовних частин (serial part of the code).

Час ініціалізації паралельної програми – це час генерації початкових даних (часто на одному процесорі) і розсилання результатів по всіх процесорах.

Незбалансованість завантаження процесорів виникає тому, що частина процесорів змушена якийсь час простоювати.

Витрати на комунікації визначаються пропускною здатністю комунікаційної мережі, латентністю комунікаційної мережі, діаметром комунікаційної мережі, який значною мірою визначається топологією комунікаційної мережі. Комунікаційні витрати можуть бути скорочені за рахунок використання зустрічних обмінів даними й асинхронної передачі даних. Сучасні комунікаційні засоби паралельних обчислювальних систем, а також комунікаційні бібліотеки допускають використання зустрічних обмінів, коли дані передаються одночасно в обидва боки. Асинхронні обміни даними (обміни даними, що виконуються паралельно з виконанням обчислень) також підтримуються більшістю сучасних паралельних комп’ютерів, а також і програмному рівні.

РОЗДІЛ 8 ОПИС СИСТЕМИ МОДЕЛЮВАННЯ ПОШИРЕННЯ ЗАХВОРЮВАНЬ, СПРИЧИНЕНИХ ВІРУСАМИ ТА ІНФЕКЦІЯМИ

8.1 Архітектура системи

На рис. 8.1 представлено UML діаграму структури інформаційної системи моделювання та прогнозування поширення інфекційний різного роду.

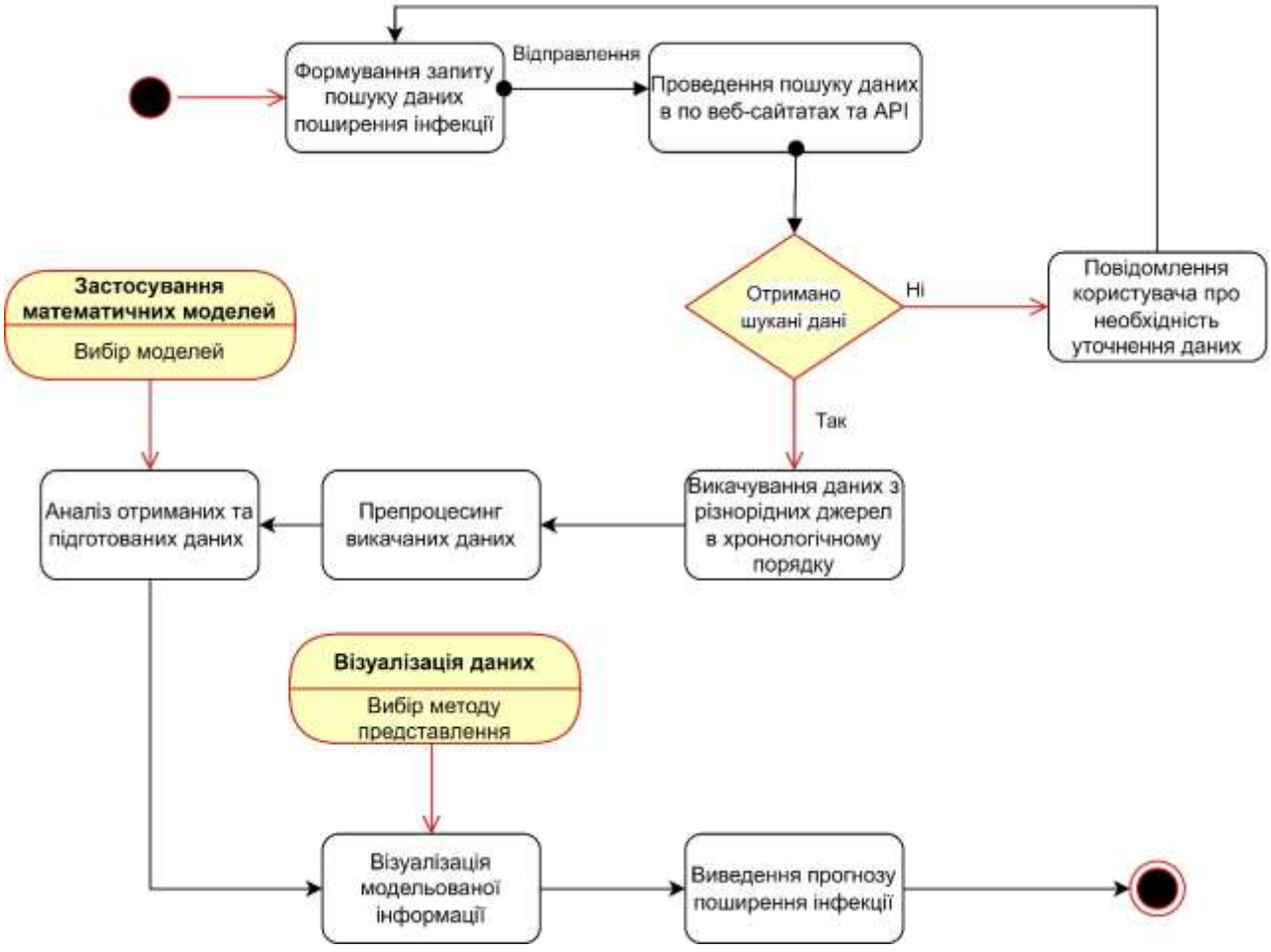


Рис. 8.1. Модель інформаційної системи моделювання та прогнозування поширення інфекцій різного роду представлена у виді UML діаграми

Згідно з структурою інформаційної системи (рис. 8.1), користувач формує поля запиту, встановлює часові рамки та ініціює запит на отримання даних про поширення певної інфекції із веб-сайтів та API. Система проводить пошук, та в

разі не виявлення даних повідомляє оператора про необхідність коригування запиту, якщо пошук пройшов успішно, розпочинається викачування даних.

Наступним кроком є препроцесинг даних, оскільки вони отримані з різних джерел, в різних форматах, а як наслідок використовувати їх для проведення прогнозування неможливо. В нашому випадку дані потрібно збирати власноруч. Для цього процес збирання даних буде в більшій мірі складатися з їх аналізу.

Просіювання через масивні набори даних може бути трудомістким завданням, навіть для автоматизованих систем. Ось чому етап зменшення даних є настільки важливим, оскільки він обмежує набір даних найважливішою інформацією, тим самим збільшуючи ефективність зберігання, зменшуючи при цьому затрати та час, пов'язані з роботою з такими наборами [2].

Рис. 8.2. UML діаграма діяльності препроцесингу.

Підготовані та очищені дані вже можна використати для проведення аналізу та прогнозування поширення інфекцій. Проте, для проведення аналізу поширення інфекції необхідно мати моделі, які будуть натреновані для прогнозування поширення інфекційних захворювань. Цей процес включає в себе такі етапи:

- підготовка набору даних;
- тренування моделей;
- тестування;
- збереження.

Рис. 8.3. UML діаграма тренування моделей для прогнозування поширення інфекційних захворювань.

Всі натреновані моделі перевіряються на адекватність, та в подальшому будуть використовуватися для прогнозування поширення інфекційних захворювань. Вихідні дані розробленої системи надають користувачу розуміння ситуації відносно епідеміологічних основних показників: нових захворювань, одужання, смерті тощо. Дані обчислюються в якості передбачення на основі вихідних даних і створюють прогноз для даного набору характеристик.

У вигляді вихідних даних користувач отримуватиме візуалізацію пробачення у вигляді графіку для певного набору даних, а також карту з передбаченням для певного регіону для простішого візуального сприйняття.

Блок візуалізації (рис. 8.1) дозволяє представити дані в необхідному вигляді, чи то представлення на карті поширення інфекції по різних територія, чи то динаміка зміни смертності, захворювання і т.д..

Прогнозування поширення інфекційних хвороб передбачає використання нейронних мереж та моделей закладених для прогнозування. Детальний огляд архітектури штучних нейронних мереж дає розуміння, як організована алгоритмічна складова системи. Демонстрація проектної системи в основному починається з контекстної діаграми. На рис. 8.4 видно, що сутність “Користувач” надає системі початковий відлік та вказує часові рамки, відколи зробити прогнозування. Після цього система може зробити певний прогноз від даної точки відліку.

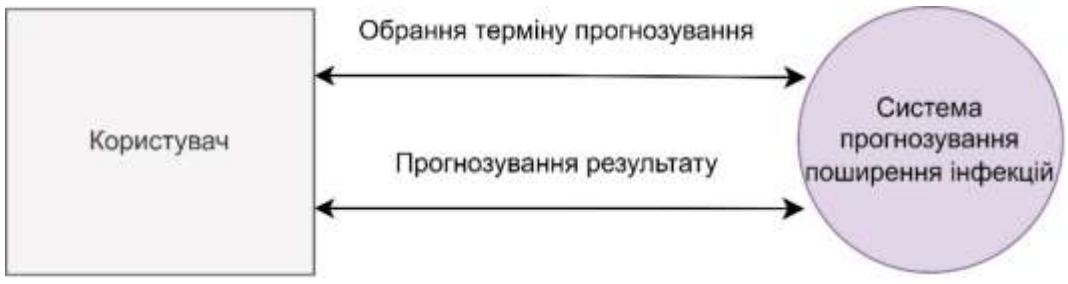


Рис. 8.4. Контекстна діаграма проектованої системи

Для того, щоб розроблена система була здатна прогнозувати приналежність зображення до класу, попередню потрібно пройти етап

формування набору даних, його підготовку (препроцесинг). Коли набір даних готовий, наступний етап – навчання моделі на тренувальних зразках.

Досягнувши достатньої точності на тренувальній вибірці, система може передбачати дані відносно поширення інфекції певного типу (рис. 8.5).

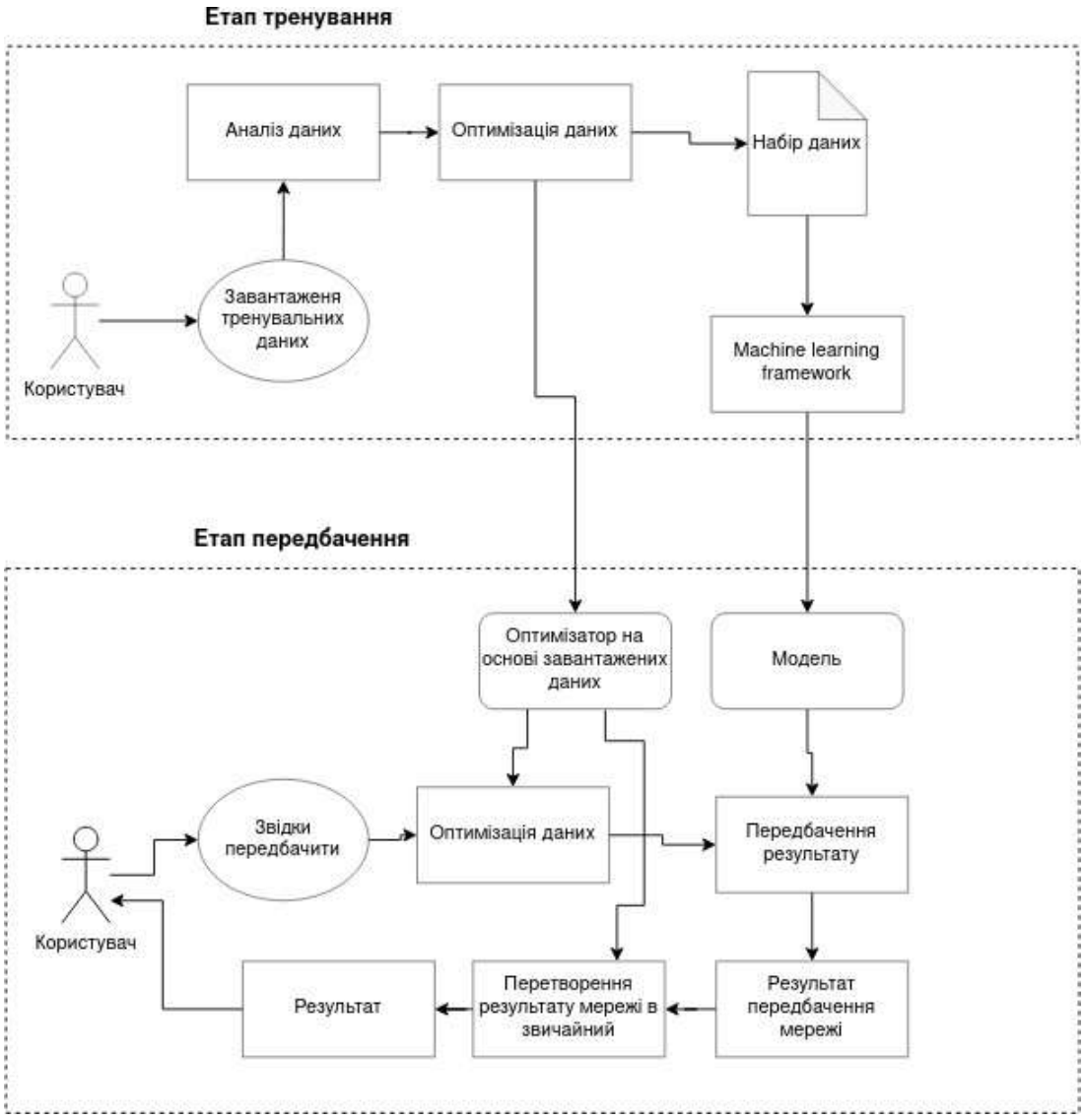


Рис. 8.5. Алгоритм побудови системи

У верхній частині рис. 8.5 зображено етап навчання. Спочатку підготований набір даних проходить попередню підготовку. Даний етап являється доволі важливим, адже при не нормуванні даних результат може мати дуже погані результати. Після цього дані оптимізуються, в результаті чого ми отримуємо оптимізатор, який є створеним на основі вихідних даних для подальшого декодування результатів, а також ми отримаємо оптимізовані

результати функції. Після цього створити модель для машинного навчання і натренувати її.

Після етапу навчання можна переходити на етап прогнозування, де дані спочатку оптимізуються на основі нетренованих даних, а після цього передаються на прогнозування мережі. Результат роботи декодується тим самим оптимізатором з поверненням результату юзеру.

Основний функціонал розробленої системи наведений в табл. 8.1 на основі проаналізованих аналогічних рішень.

Табл. 8.1. Функціональні вимоги

Розділ	**Функціональні вимоги**
Робота з даними	Зчитування даних
	Нормалізація даних для подальшого використання
	Аналіз даних на різних етапах його використання(зчитування, підготовка, зміна даних тощо)
	Вибір моделі навчання та проведення навчання
Машинне навчання	Виокремлення основних ознак
Нейронна мережа	Мережа утворює передбачення на основі попередніх даних

Для проектування системи було використано наступні технології:

• Anaconda – вільний, безкоштовний дистрибутив для деяких програмних продуктів.

• Jupyter Notebook – це певна некомерційна організація, яка застосовується для розроблення програмного забезпечення.

• Python – це високорівнева, інтерпретована, об’єктно-орієнтована мова програмування із строго-динамічною типізацією.

• Scikit-learn – відкрита безкоштовна бібліотека для такої мови програмування як Python, вона надає деяку функціональність для створення, тренування різних алгоритмів регресії, класифікації, кластеризації.

• Pandas – безкоштовна програмна бібліотека для мови програмування Python, вона дозволяє маніпулювати даними та аналізувати їх.

• Google Colaboratory – це хмарний сервіс, який спрямований на певне спрощення досліджень в галузі глибинного та машинного навчання, з ним ми можемо запускати Jupyter Notebook із доступом до графічного процесора.

Розроблену систему можна, також використовувати не тільки для прогнозування поширення інфекційних захворювань. Нижче можна побачити діаграму варіантів використання системи, на якій актором є власне система оскільки, розроблювана система не потребує користувацьких чи адміністраторських функцій.

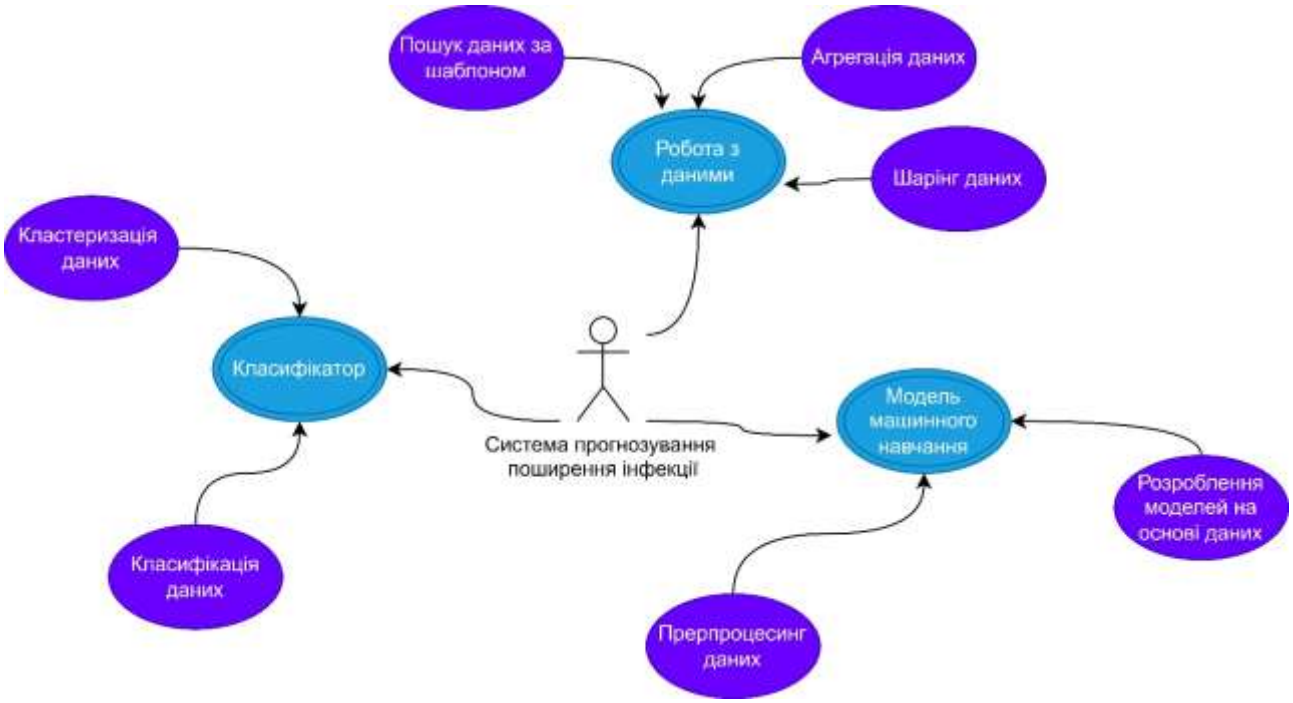


Рис. 8.6. Діаграма варіантів використання системи

8.2 Візуалізація поширення інфекції

Основними полями для аналізу епідемій являються географічна ідентифікація, а також характеристики, які виступають як нові значення наступних критеріїв: підозра, підтвердження, смерть, відновлення.

Також набір характеристик має прив’язку до певної локації, наприклад місто чи регіон відносно якого будуть здійснюватися передбачення.

В якості вхідних даних був використаний набір даних, поданий як часовий ряд з різними метриками, які можуть бути використані для аналізу чи передбачення. Зазначентй датасет містить інформацію про вірус COVID-19 відносно міст україни. Приклад набору даних до їхньої нормалізації подано на рис. 8.7.

zvit_date	registration_area	new_susp	new_confirm	new_death	new_recover
2020-01-27	Львівська	1	0	0	0
2020-01-29	Львівська	1	0	0	0
2020-03-13	Львівська	1	0	0	0
2020-03-16	Львівська	9	0	0	0
2020-03-17	Львівська	5	0	0	0
...	...	...	...	...	...
2021-04-25	Львівська	295	22	11	154
2021-04-26	Львівська	378	301	16	346
2021-04-27	Львівська	469	645	13	365
2021-04-28	Львівська	669	458	10	301
2021-04-29	Львівська	459	73	3	229

Рис. 8.7. Приклад вхідних даних для створення передбачення

Вихідні дані розроблювальої системи повинні надавати розуміння ситуації відносно епідеміологічних основних показників: нових захворювань, одужання, смерті тощо. Дані обчислюються в якості передбачення на основі вихідних даних і створюють прогноз для даного набору характеристик.

У вигляді вихідних даних користувач отримуватиме візуалізацію пробачення у вигляді графіку для певного набору даних, а також карту з передбаченням для певного регіону для простішого візуального сприйняття.

Система передбачення поширення епідемій повинна бути кросплатформенною і працювати на таких операційних системах як Windows, Linux та MacOS. Отже для забезпечення роботи системи необхідний інтерпретатор Python версії 3.8.8 та ряд бібліотек Python встановлених на пристрої.

Для створення віртуального середовища для інтерпретації програми рекомендовано використовувати відкритий інсталятор дистрибутивів Anaconda. Віртуальне середовище повинне містити такі бібліотеки:

− tensorflow_gpu або tensorflow. Під час розробки програми було використано tensorflow_gpu версії 2.4.1;

− pandas − для маніпулювання даними та їхнього аналізу версії 1.2.3;

− numpy − для роботи з багатовимірними масивами і матрицями версії 1.19.2. Також бібліотека має підримку великої кількості операції над масивам ;

− statsmodels − для аналізу даних в графічному вигляді версії 0.12.2;

− matplotlib − для візуалізації результатів роботи версії 3.3.4;

− keras − для проектування нейронної мережі версії 2.4.3;

− sklearn − для допомоги з реалізацією алгоритмів, зокрема LSTM версії 0.24.1;

− plotity − графічна бібліотека Python Plotly робить інтерактивні графіки якості публікації. Приклади того, як робити лінійні графіки, діаграми розсіювання, діаграми площ тощо. Використовується для зображення географічної карти з використанням версії 7.6.3;

− ipywidgets − для графічної зміни параметрів даних для подальшої графічної візуалізації версії 4.14.3;

Дана система складається з функцій, які виконуються в “Jupyter Notebook” для покрокового і візуального відображення роботи системи. Завдяки цьому є можливість слідкування за роботою системи.

Для початку виконується імпортування усіх бібліотек і зчитування даних, на основі яких буде виконуватися передбачення. Після цього дані перевіряються на нульові значення, адже їхня наявність може перешкодити якосному подальшому навчанню моделі. Результати роботи даного кроку наведені в рис. 8.8.

	zvit_date	registration_area	registration_region	registration_settlement	registration_settlement_lng	registration_settlement_lat	new_susp	new_confirm	a
0	2020-01-22	Донецька	Українськ	Українськ	37.362549	48.096803	0	0	
1	2020-01-22	Луганська	Сєвєродонецьк	Сєвєродонецьк	38.487877	48.948177	0	0	
2	2020-01-22	Рівненська	Рівне	Рівне	26.251617	50.619900	1	0	
3	2020-01-27	Донецька	Костянтинівка	Костянтинівка	37.707581	48.523943	0	0	
4	2020-01-27	Львівська	Львів	Львів	24.029717	49.839683	1	0	

Рис. 8.8. Попередній огляд даних для створення моделі

Після цього виконується аналіз даних чи не потрібна попередня обробка даних. На рисунку 8.9 подано кількість випадків відносно різних даних.

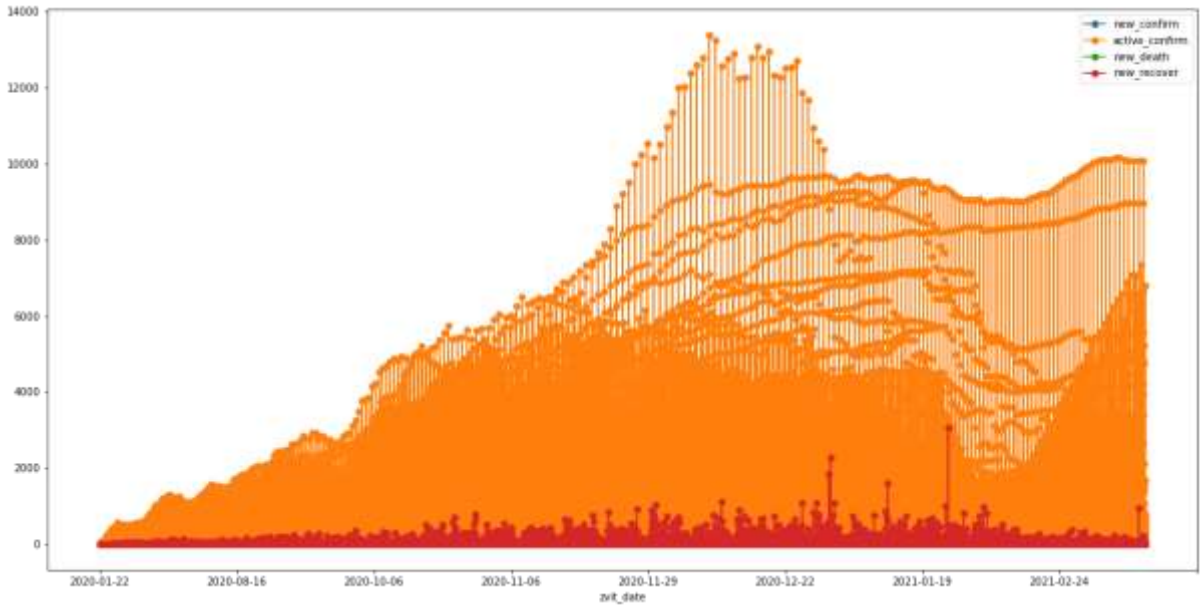


Рис. 8.9. Кількість випадків відносно різних характеристик

З рис. 8.9 випливає, що дані не достатньо підходять для подальшого застосування. Причиною цього є неоднорідність даних завдяки присутності декількох локацій. Подальшим кроком є перевірка чи можна усунути цю проблему за допомогою групування даних відносно певного періоду часу. Цей етап відображено на рис. 8.10 для нових підтверджених випадків.

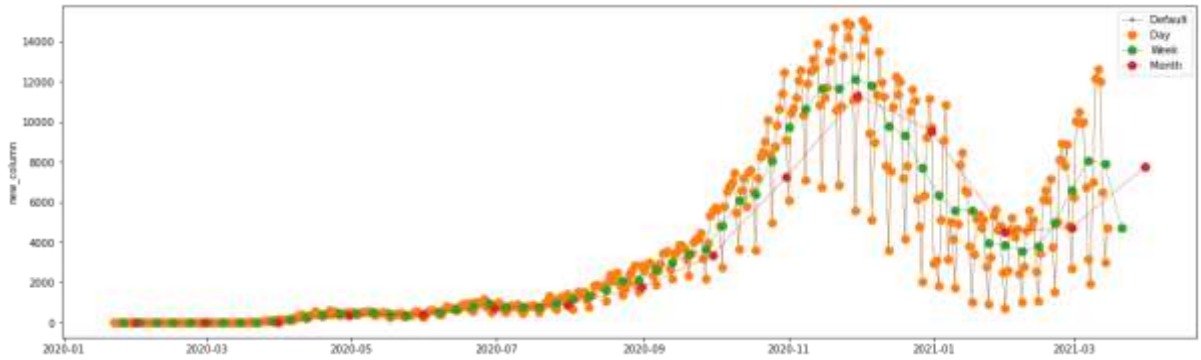


Рис. 8.10. Кількість нових випадків відносно різних проміжків часу

На цьому рисунку можна побачити, що усунулися розриви в даних, отже групування даних відносно певних періодів є доцільним. У подальшому буде створена модель для передбачення відносно регіонів, отже наступним кроком відбувається групування даних з подальшим виділенням певної області. Обрано Львівську область (рис. 8.11).

zvit_date	registration_area	new_susp	new_confirm	new_death	new_recover
2020-01-27	Львівська	1	0	0	0
2020-01-29	Львівська	1	0	0	0
2020-03-13	Львівська	1	0	0	0
2020-03-16	Львівська	9	0	0	0
2020-03-17	Львівська	5	0	0	0
...	...	...	...	...	...
2021-03-11	Львівська	715	915	12	65
2021-03-12	Львівська	876	1345	8	226
2021-03-13	Львівська	582	442	5	12
2021-03-14	Львівська	559	3	10	7

Рис. 8.11. Результати групування даних відносно Львівської області з інтервалом в один день

Далі здійснюється нормалізація даних. Для цього було перетворено область в числовий еквівалент для подальшого опрацювання, а також було приведено усі характеристики в діапазон від 0 до 1, оскільки алгоритм являється чутливим до великих перепадів значень, а також прискорення їхнього навчання. Для цього було використано функцію MinMaxScaler з бібліотеки sklearn.

Для початку перед тренуванням моделі потрібно розділити дані на навчальні та валідаційні дані, де було використано функцію train_test_split з бібліотеки sklearn без перемішування даних, адже в даному випадку важливим являється послідовність даних.

Після цього утворюємо модель, з урахуванням трьох основних критеріїв:

a. Кількість минулих днів, які ми враховуємо для передбачення,
b. Кількість днів, які потрібно передбачити,
c. Кількість критеріїв, які потрібно врахувати.

Для нашого випадку було враховано один день для передбачення наступного дня з використанням чотирьох характеристик аналізуючи попередні сім днів. структуру моделі виконуємо команду model.summary() (рис. 8.12).

```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
lstm (LSTM)                  (7, 4, 4)                 144
_________________________________________________________________
lstm_1 (LSTM)                (7, 4, 4)                 144
_________________________________________________________________
lstm_2 (LSTM)                (7, 4, 4)                 144
_________________________________________________________________
lstm_3 (LSTM)                (7, 4, 4)                 144
_________________________________________________________________
lstm_4 (LSTM)                (7, 4, 4)                 144
_________________________________________________________________
lstm_5 (LSTM)                (7, 4)                    144
=================================================================
Total params: 864
Trainable params: 864
Non-trainable params: 0
_________________________________________________________________
```

Рис. 8.12. Структура моделі

Далі було здійснено тренування раніше створеної моделі з умовою виходу, якщо точність являтиметься більшою 92% чи не будуть пройдені усі епохи. Також зберігається кожна модель з найменшим результатом втрат для валідаційних даних. Процес тренування моделі подано на рис. 8.13, а історію тренування моделі на рис. 8.14.

```
- accuracy: 0.7806 - val_loss: 0.1430 - val_accuracy: 0.6765
Epoch 95/100
363/363 [==============================] - 2s 6ms/step - loss: 0.0503
- accuracy: 0.7993 - val_loss: 0.1471 - val_accuracy: 0.6765
Epoch 96/100
363/363 [==============================] - 2s 6ms/step - loss: 0.0527
- accuracy: 0.7709 - val_loss: 0.1418 - val_accuracy: 0.6765
Epoch 97/100
363/363 [==============================] - 2s 6ms/step - loss: 0.0511
- accuracy: 0.7530 - val_loss: 0.1475 - val_accuracy: 0.6765
Epoch 98/100
363/363 [==============================] - 2s 6ms/step - loss: 0.0490
- accuracy: 0.7476 - val_loss: 0.1560 - val_accuracy: 0.5882
Epoch 99/100
363/363 [==============================] - 2s 6ms/step - loss: 0.0507
- accuracy: 0.7062 - val_loss: 0.1563 - val_accuracy: 0.5294
Epoch 100/100
363/363 [==============================] - 2s 6ms/step - loss: 0.0544
- accuracy: 0.7204 - val_loss: 0.1466 - val_accuracy: 0.6765
```

Рис. 8.13. Тренування моделі

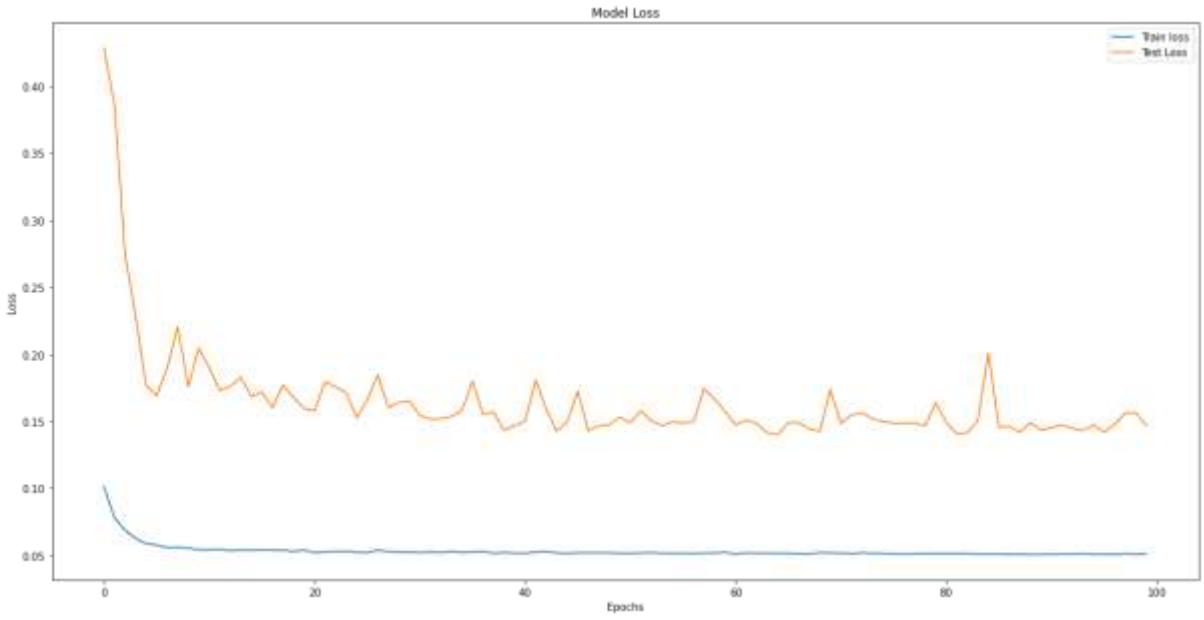
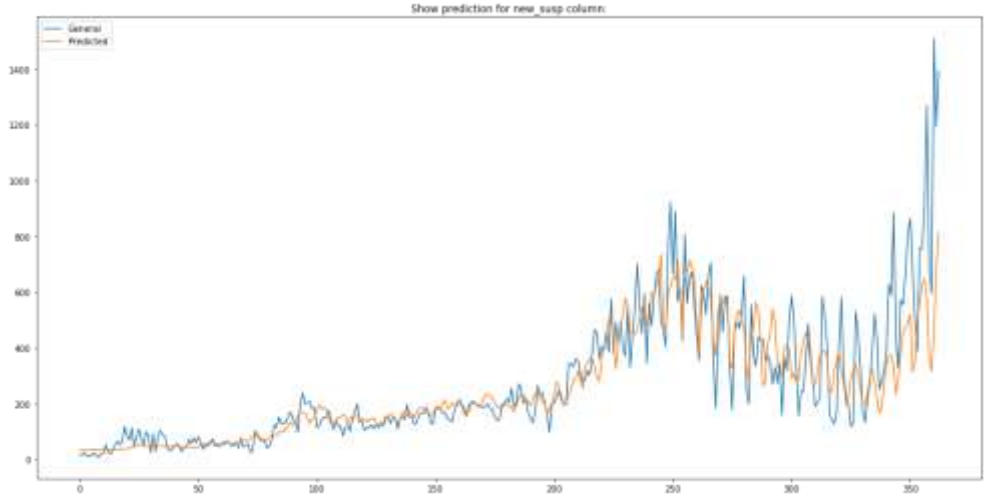


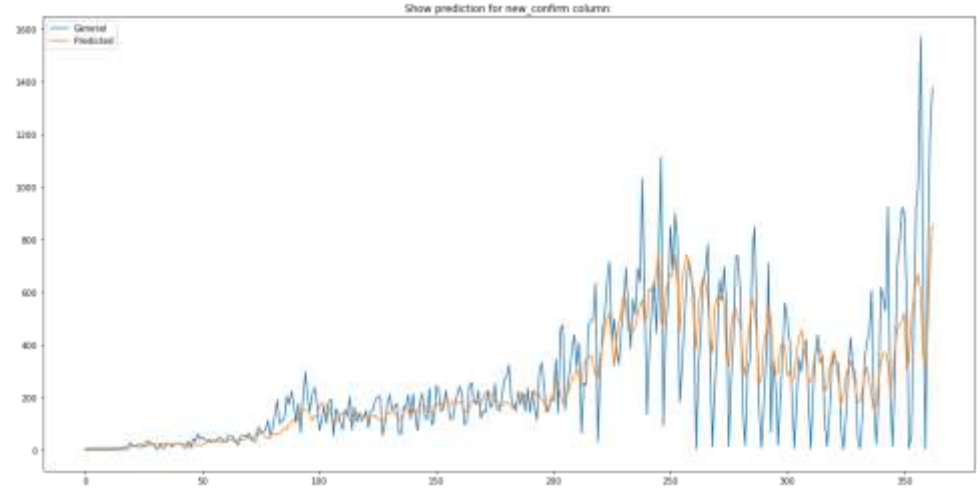
Рис. 8.14. Оцінка втрат моделі

Далі відображаються результати роботи функції відносно тренувальних даних для кожного з параметрів. Приклад даного відображення представлено на рис. 8.15. На графіку представлено відображення звичайних тренувальних даних,

які представлені синьою лінією, а також передбачення, яке представлене оранжевою лінією. З даних графіків можна говорити, що точність передбачення відносно тестувальних даних є доволі високою.



а)



б)

Рис. 8.15. Передбачення для тренувальних даних відносно:

а) – нових спостережень;

б) – нових підтверджень.

Після цього здійснюється передбачення на основі тестувальних даних (рис. 8.16).

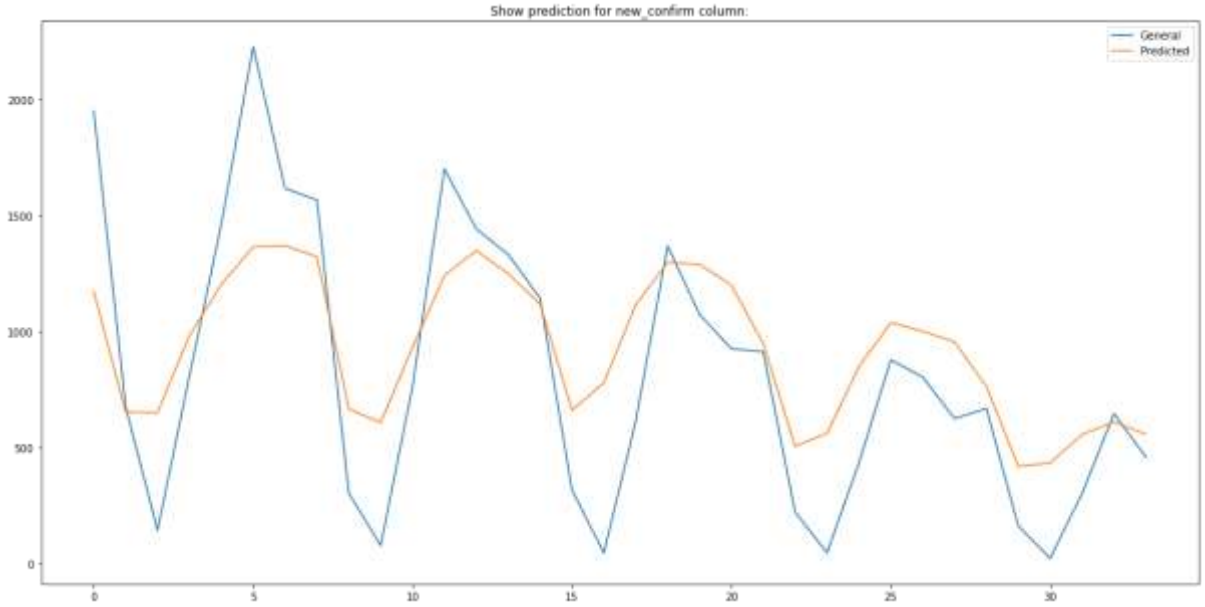


Рис. 8.16. Передбачення для тестувальних даних відносно нових підтверджень

Для роботи з новими даними потрібно мати готову модель даних і готовий MinMaxScaler для подальшого правильного приведення змінних відносно попередніх даних. Завершивши етап нормалізації даних, модель тренується після чого викликається функція plt_result() для графічного відображення точності навчання відносно кожного параметру, які знаходяться в окремих колонках.

Для відображення графічних даних карті потрібно виконати певну попередню обробку даних. Для цього використовується функція update_df_to_plt(), де важливим параметром являється “registration_area”. Даний параметр відповідає за область, яка буде пізніше відображена.

Після цього виконується функція show_map() для відображення даних на карті. Результат роботи зображений на рис. 8.17.

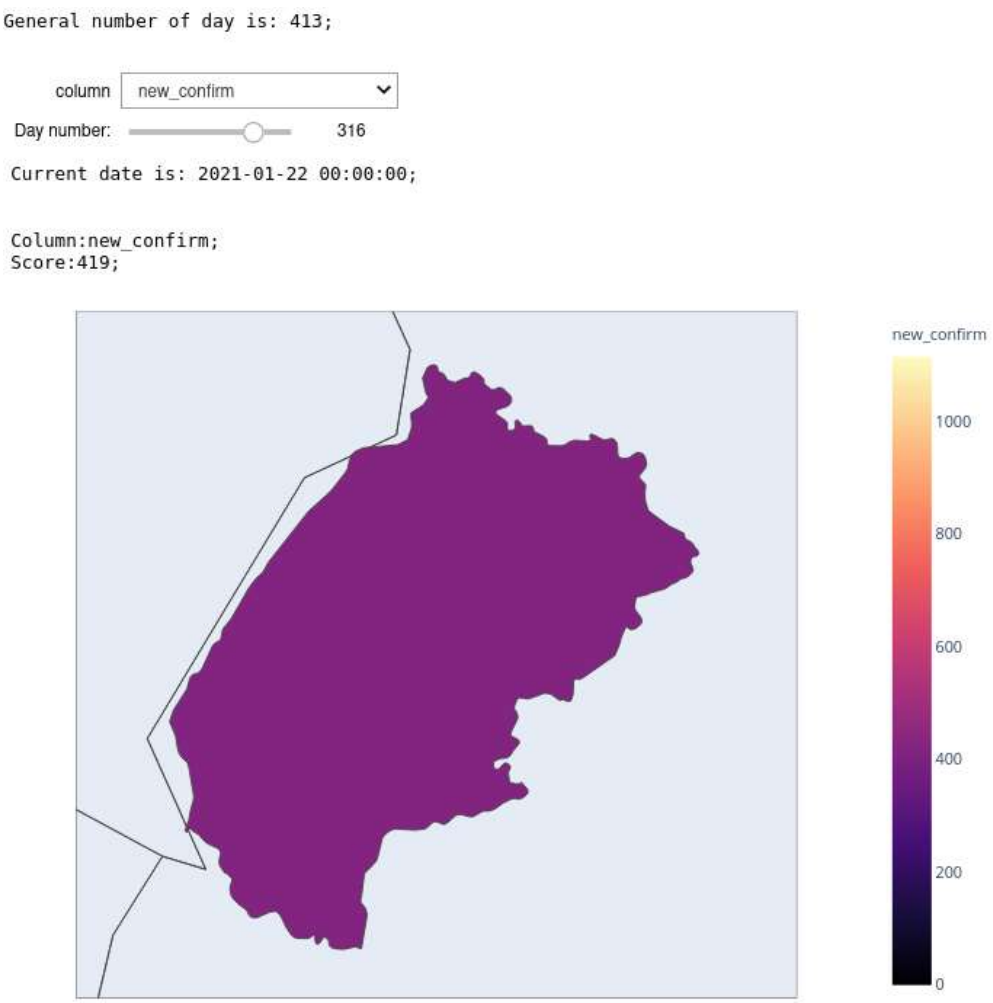


Рис. 8.17. Відображення нових підтверджених випадків для Львівської області

ВИСНОВКИ

В результаті дослідження запропоновано методи співставлення комп'ютерого час розрахунку, динамічних параметрів мультиагентної SIR моделі з реальним часом та такими показниками як кількість активних хворих, рівень смертності, ефективне репродуктивне число. Це дало змогу максимально реалістично побудавати імітаційну модель, що показала достатній рівень адекватності і точності прогнозування поширення пандемії.

Запропоновано модифікацію класичної мульти-агентної SIR моделі, що дало змогу врахувати час інкубаційного періоду, дотримання людьми безпечної віддалі під час пересування, змоделювати карантин, ізоляцію, відвідування громадських місць, таких як супермаркети, парки, церкви, спортзали, змоделювати транспорт, будівництво, спортзали, тощо.

Запропоновано метод моделювання великих регіонів, що складається з декількох адміністративних одиниць шляхом симуляції множини автономних мульти-агентних систем. Запропоновано шляхи моделювання транспортного потоку між регіонами та врахування поширення вірусу між ними.

Був проведений аналіз чутливості моделі до різних заходів запобігання інфекції, визначені як абсолютна так і комплексна ефективність цих заходів. Оцінено ефективність їх використання на різних стадіях поширення вірусної інфекції.

Вдосконалена SIR модель дала змогу змоделювати та провести прогнозні симуляції поширення COVID-19 в Чернівцях та Чернівецькій області та оцінити ефективність різних заходів обласної державної адміністрації для запобігання його поширенню, зокрема відкриттю будівництва та введення жорсткого карантину на вихідні дні. Була проведена симуляція наслідків масового відвідування церков на Пасху та Вербну неділю.

Розроблена модель дала змогу спрогнозувати поширення COVID-19 в таких країнах як Туреччина, Словаччина, Сербія. Отримані результати показали добре

узгодження з експериментом та дали змогу передбачити другу хвилю поширення хвороби.

Досліджено застосування бустингового підходу до підвищення точності розвязання задач апроксимації на основі мереж GRNN. Розроблено ряд ансамблів на основі сумісного використання мереж GRNN те нейроподібних структур Моделі Послідовних Геометричних Перетворень. В основі цього покладено можливість часткового усунення похибки роботи мережі GRNN. Основною їх відмінністю є процедура формування вихідного результату. Оцінка ефективності роботи усіх досліджуваних ансамблевих методів відбувалася шляхом розв'язання задачі регресії на основі використання реального набору даних. Експериментальним чином встановлено, що найвищі показники точності (на основі MAPE and RMSE) роботи демонструє розроблений ансамблевий метод. Експериментальним шляхом встановлено часові ресурси, необхідні для реалізації досліджуваних методів. Серед недоліків методів цього класу слід зазначити суттєві часові затримки, необхідні для їх роботи. Для їх усунення необхідно застосування ряд оптимізаційних методів підбору оптимальних параметрів роботи штучних нейронних мереж цього типу.

Розроблено триступеневий підхід до імпутації відсутніх даних, який включає: (i) проектування моделі великих даних у завданні відновлення відсутніх даних, що дозволяє обробляти напівструктуровані дані; (ii) розробка методу відновлення відсутніх даних на основі функціональних залежностей та правил асоціації та (iii) оцінка складності алгоритму відновлення відсутніх даних.

Запропонований спосіб відновлення відсутніх даних створює додаткові значення даних із використанням залежності домену та функціональних залежностей та додає ці значення до наявних навчальних даних. Правильність обчислених значень перевіряється на класифікаторі, побудованому на вихідному наборі даних. Запропонований метод PPD працює на 12% краще, ніж EM та RF методи для 30% відсутніх даних. Для більшої частки відсутніх даних (діапазон

близько 40% -50% відсутніх даних) найкраще виглядає метод EM, а PPD має порівнянні результати з методом SVM. Порівняння методів аналізу та моделювання статистичних процесів за якісними критеріями підтвердило, що розроблений метод має наступні переваги: володіє характеристиками стійкості до помилок у даних, що забезпечує паралельне виконання в розподілених базах даних, автоматизує та виконує аналіз різних типів даних . Прискорення для атрибутів m = 41 більше у 12,5 разів для 20 серверів (процесорів) порівняно з непаралельним режимом.

Побудовано рекомендаційні правила щодо адекватної державної політики щодо зменшення розповсюдження нових справ Covid-19. Система цих правил базується на новому ансамблі методів машинного навчання, таких як регресійне дерево та кластеризація. У цьому дослідженні використовувались об’єднані дані з урядового відстежувача відповідей COVID-19 та ECDC Covid-19.

Кластеризацію проводили методом k-середніх. Статистика розривів дозволяє знайти відповідну кількість кластерів, і у випадку дослідження було обрано три кластери. Кластери відрізняються рекомендаціями та діями урядів-кореспондентів. Таким чином, перші країни кластеру обрали закриття шкіл та контроль міжнародних поїздок як головні рекомендації; країни другого кластеру рекомендували залишатися вдома, тоді як головні рекомендації урядів країн, що входять до третього кластеру, полягали в тому, щоб залишатися вдома та скасовувати публічні заходи.

Одна і та ж країна може бути приєднана до різних кластерів у різні часові інтервали. Отже, кластеризація за країнами не буде настільки однозначною. Ось чому часові ряди для відокремленої країни можуть бути цікавими і будуть предметом подальшого вивчення. Крім того, вплив такого скупчення на поширення COVID-19, положення та тривалість піку, а також рівень смертності також буде предметом нашого подальшого дослідження.

Було побудовано регресійне дерево прийнятя рішень, набір правил було отримано з дерева рішень і може бути використано для стратегічного планування в системі охорони здоров'я.

Отримані результати дослідження запропонованих моделей для зібраного, загалом, підтверджують гіпотезу про відмінності в точності прогнозування для всього набору даних та умовного простору, побудованого на кластерному аналізі та відборі ознак. Однак слід зробити відразу два зауваження:

• Кількість екземплярів у зібраному наборі даних занадто мала. Дану гіпотезу слід довести великим набором даних. Ми працюємо над розширенням набору даних. Однак подібний підхід був використаний у нашому попередньому дослідженні для іншого набору даних [39, 40] з такими ж результатами;

• Автори припускають, що точність прогнозу буде сильно залежати від кількості порожніх значень.

Різниця між найкращим слабким предиктором (персептрон) у 1,01 краще для вибраних ознак (табл. 2,3). Якість розробленого ієрархічного предиктора для метрики RMSE на 1,47 краща, ніж для найкращого слабкого предиктора. Дерево регресії показує однаковий результат для всього набору даних та вибраних змінних. Лінійна регресія показує 1,53 кращі показники RMSE для вибраних функцій у порівнянні з усім набором даних. Решта слабких предикторів показують кращі результати щодо вибраних функцій [42].

Вибір характеристик важливий не тільки для підвищення точності. Час навчання за архівним предиктором Гієра для вибраних функцій у 1,2 рази менше, ніж для всього набору даних [44].

Запропонований метод використовується для невеликих наборів даних, він подібний до [41]. Тут автори пропонують удосконалити метод подвоєння введення на основі RBF шляхом введення додаткових елементів у формулу для розрахунку вихідного сигналу методу. Ми збільшуємо точність на 4%, виходячи з MAE та RMSE порівняно з основним методом [46, 47 - 49].

Результатом етапу є вибір функцій та кількість ліжко-днів у лікарняному прогнозуванні з використанням умовного простору та розроблених ієрархічних предикторів. Набір персональних медичних параметрів збирається в державній лікарні. Цей набір даних використовується для прогнозування кількості днів перебування в лікарні. Пацієнти пройшли клінічне лікування післяопераційних ускладнень у животі. Наше дослідження показує низьку попарну кореляцію між величезною підмножиною параметрів, перерахованих у наборі даних. Однак для підвищення якості моделі прогнозування необхідний правильний вибір ознак.

Попередня обробка даних дозволяє нам підвищити якість аналізу. Борута, дерево регресії та кореляція використовуються для вибору ознак. Результати відбору формуються на основі жорсткого голосування всіх відбірників функцій. Для поділу об'єкта на розділені кластери використовується кілька алгоритмів кластеризації. Таке розщеплення дозволяє розробляти кондиціонований простір на основі залежних від часу та незалежних від часу даних та медичного протоколу.

У цій роботі було розроблено ієрархічний предиктор, що ґрунтується на поєднанні результатів кластеризації та чотирьох слабких предикторів для кожного кластера окремо [43]. Тому запропонований алгоритм демонструє більш високу точність прогнозування порівняно з найкращим перцептроном прогнозування. Також запропоновано новий метод візуалізації дендрограм з можливістю обчислення площі та периметру кластера в 3D [45].

ЛІТЕРАТУРА

1. Odeloui AE, Edoh TO, A Context-Aware Machine-to-Machine-Enabled Pervasive Cardiac Telemetry for Personalizing Health Care Delivery. IREHI: 1-8, 2018
2. Djulbegovic, B., & Guyatt, G. H. Progress in evidence-based medicine: a quarter century on. The Lancet, 390(10092), 415-423., 2017.
3. Danhof, M., Klein, K., Stolk, P., Aitken, M., & Leufkens, H. The future of drug development: the paradigm shift towards systems therapeutics. Drug discovery today, 23(12), 1990-1995, 2018.
4. Kuznietsova N. V., Bidyuk P. I. Business intelligence techniques for missing data imputation, Scientific news of the National Technical University of Ukraine, Kyiv Polytechnic Institute, 5:.47-56, 2015.
5. Mishyna, M., Volokh, O., Danilova, Y., Gerasimova, N., Pechnikova, E., & Sokolova, O. S. Effects of radiation damage in studies of protein-DNA complexes by cryo-EM. Micron, 96, 57-64, 2017.
6. Khanmohammadi, S. An improved synchronization likelihood method for quantifying neuronal synchrony. Computers in biology and medicine, 91, 80-95, 2017.
7. Perov, Y., Graham, L., Gourgoulias, K., Richens, J., Lee, C., Baker, A., & Johri, S. Multiverse: causal reasoning using importance sampling in probabilistic programming. In Symposium on advances in approximate bayesian inference (pp. 1-36). PMLR, 2020
8. Tang Y., Wang Y., Cooper K. M., Li L., Towards big data Bayesian network learning-an ensemble learning based approach, 2014 IEEE International Congress on Big Data. IEEE, pp.355-357, 2014.
9. Lakho, Shamshad, et al. Decision Support System for Hepatitis Disease Diagnosis using Bayesian Network. Sukkur IBA Journal of Computing and Mathematical Sciences, 1.2: 11-19, 2017.

10. Seixas, Flávio L., et al. A Bayesian network decision model for supporting the diagnosis of dementia, Alzheimer′ s disease and mild cognitive impairment. Computers in biology and medicine, 51: 140-158, 2014.
11. Perova I, Bodyanskiy Y. Fast Medical Diagnostics Using Autoassociative Neuro-Fuzzy Memory / International Journal of Computing, 16(1), 34-40, 2017.
12. Bhatt, Chintan, et al., editors. Internet of Things and Big Data Technologies for Next Generation Healthcare. Springer International Publishing, 2017. DOI.org (Crossref), doi:10.1007/978-3-319-49736-5.
13. Podletskaya NI, Divak MP Information technology for the identification of the reverse laryngeal nerve during thyroid surgery. Measuring and computing technology in technological processes, 1: 151-157, 2015.
14. Silva-Ramírez E. L., Pino-Mejías R., López-Coello M., Single imputation with multilayer perceptron and multiple imputation combining multilayer perceptron and k-nearest neighbours for monotone patterns, Applied Soft Computing, vol.29, pp.65-74, 2015.
15. Chiacchio F., Aizpurua J.I., D'Urso D., Compagno L., Coherence region of the Priority-AND gate: 5 Analytical and numerical examples, Quality and Reliability Engineering International, vol.34, no.1, pp.107-115, 2018.
16. Tsai C.W., Wu N.-K., Huang C.-H., A multiple-state discrete-time Markov chain model for estimating suspended sediment concentrations in open channel flow, Applied Mathematical Modelling, vol. 40, no. 23-24, pp. 10002-10019, 2016.
17. Kadri, H., Ahmed, S. B., & Collart-Dutilleul, S. Formal approach to control design of complex and dynamical systems. Procedia Computer Science, 108, 2512-2516, 2017.
18. Masic, I., Miokovic, M., & Muhamedagic, B. Evidence based medicine–new approaches and challenges. Acta Informatica Medica, 16(4), 219, 2008.
19. Sobrinho, A., Perkusich, A., da Silva, L. D., Cordeiro, T., Rêgo, J., & Cunha, P. Towards medical device certification: A colored petri nets model of a surface

electrocardiography device. In *IECON 2014-40th Annual Conference of the IEEE Industrial Electronics Society,*. 2645-2651, 2014.

20. Boubeta-Puig, J., Díaz, G., Macià, H., Valero, V., & Ortiz, G. MEdit4CEP-CPN: An approach for complex event processing modeling by prioritized colored Petri nets. *Information Systems, 81,* 267-289, 2019.
21. Anand, N., Sehgal, R., Anand, S., & Kaushik, A. Feature selection on educational data using Boruta algorithm. International Journal of Computational Intelligence Studies, 10(1), 27-35, 2021.
22. Pakhira, M. K. Finding number of clusters before finding clusters. Procedia Technology, 4, 27-37, 2012.
23. Li, Z., Tian, Z., Zhou, M., Zhang, Z., & Jin, Y. Awareness of line-of-sight propagation for indoor localization using Hopkins statistic. IEEE Sensors Journal, 18(9), 3864-3874, 2018.
24. N. Boyko, M. Kuba, L. Mochurad, S. Montenegro, Fractal Distribution of Medical Data in Neural Network, CEUR-WS.Org. 2488 (2019) 307–318.
25. N. Melnykova, V. Melnykov, E. Vasilevskis, The personalized approach to the processing and analysis of patients' medical data, Proceedings of the 1st International Workshop on Informatics & Data-Driven Medicine (IDDM 2018). 2255 (2018) 103–112.
26. T. Shaikhina, N.A. Khovanova, Handling limited datasets with neural networks in medical applications: A small-data approach, Artificial Intelligence in Medicine. 75 (2017) 51–63. https://doi.org/10.1016/j.artmed.2016.12.003.
27. Y.V. Bodyanskiy, O.K. Tyshchenko, A Hybrid Cascade Neural Network with Ensembles of Extended Neo-Fuzzy Neurons and Its Deep Learning, in: P. Kulczycki, J. Kacprzyk, L.T. Kóczy, R. Mesiar, R. Wisniewski (Eds.), Information Technology, Systems Research, and Computational Physics, Springer International Publishing, Cham, 2020: pp. 164–174. https://doi.org/10.1007/978-3-030-18058-4_13.

28. Z. Hu, Y. Bodyanskiy, O.K. Tyshchenko, Self-learning procedures for a kernel fuzzy clustering system, Advances in Intelligent Systems and Computing. 754 (2019) 487–497. https://doi.org/10.1007/978-3-319-91008-6_49.
29. A. Pasini, Artificial neural networks for small dataset analysis, Journal of Thoracic Disease. 7 (2015). https://doi.org/10.3978/j.issn.2072-1439.2015.04.61.
30. D.F. Andrews, A.M. Herzberg, Data: A Collection of Problems from Many Fields for the Student and Research Worker, Springer-Verlag, New York, 1985. https://doi.org/10.1007/978-1-4612-5098-2.
31. R: Urine Analysis Data, (n.d.). https://vincentarelbundock.github.io/Rdatasets/doc/boot/urine.html (accessed December 12, 2020).
32. A.C. Davison, D.V. Hinkley, Bootstrap Methods and their Application, Cambridge University Press, Cambridge, 1997. https://doi.org/10.1017/CBO9780511802843.
33. Р.О. Ткаченко, П.Р. Ткаченко, І.В. Ізонін, Нейромережеві засоби штучного інтелекту, Видавництво Львівської політехніки, Львів, 2017. https://vlp.com.ua/node/17380.
34. S. Fedushko, T. Ustyianovych, Predicting Pupil's Successfulness Factors Using Machine Learning Algorithms and Mathematical Modelling Methods, in: Z. Hu, S. Petoukhov, I. Dychka, M. He (Eds.), Advances in Computer Science for Engineering and Education II, Springer International Publishing, 2020: pp. 625–636.
35. R. Tkachenko, H. Kutucu, I. Izonin, A. Doroshenko, Y. Tsymbal, Non-iterative Neural-like Predictor for Solar Energy in Libya, in: V. Ermolayev, M.C. Suárez-Figueroa, A. Lawrynowicz, R. Palma, V. Yakovyna, H.C. Mayr, M. Nikitchenko, A. Spivakovsky (Eds.), Proceedings of the 14th International Conference on ICT in Education, Research and Industrial Applications. Integration, Harmonization and Knowledge Transfer. Volume I: Main Conference, Kyiv, Ukraine, May 14-

17, 2018, CEUR-WS.org, 2018: pp. 35–45. http://ceur-ws.org/Vol-2105/10000035.pdf (accessed February 17, 2019).

36. I. Izonin, R. Tkachenko, N. Kryvinska, M. Gregus, P. Tkachenko, P. Vitynskyi, Committee of SGTM Neural-Like Structures with RBF kernel for Insurance Cost Prediction Task, in: 2019 IEEE 2nd Ukraine Conference on Electrical and Computer Engineering (UKRCON), 2019: pp. 1037–1040. https://doi.org/10.1109/UKRCON.2019.8879905.
37. R. Tkachenko, P. Tkachenko, I. Izonin, P. Vitynskyi, N. Kryvinska, Y. Tsymbal, Committee of the Combined RBF-SGTM Neural-Like Structures for Prediction Tasks, in: I. Awan, M. Younas, P. Ünal, M. Aleksy (Eds.), Mobile Web and Intelligent Information Systems, Springer International Publishing, Cham, 2019: pp. 267–277. https://doi.org/10.1007/978-3-030-27192-3_21.
38. A. Doroshenko, Piecewise-Linear Approach to Classification Based on Geometrical Transformation Model for Imbalanced Dataset, in: 2018 IEEE Second International Conference on Data Stream Mining Processing (DSMP), 2018: pp. 231–235. https://doi.org/10.1109/DSMP.2018.8478537.
39. Melnykova N, Shakhovska N, Greguš M., Melnykov V., Using Big Data for Formalization the Patient's Personalized Data. Procedia Computer Science, vol. 155, 2019, pp. 624–29. DOI.org (Crossref), doi:10.1016/j.procs.2019.08.088.
40. Shakhovska, N., Izonin, I., & Melnykova, N. The hierarchical classifier for COVID-19 resistance evaluation. Data, 6(1), 6, 2021.
41. Izonin, I., Tkachenko, R., Dronyuk, I., Tkachenko, P., Gregus, M., & Rashkevych, M. Predictive modeling based on small data in clinical medicine: RBF-based additive input-doubling method. Math. Biosci. Eng, 18, 2599-2613, 2021.
42. Melnykova, N.; Shakhovska, N.; Melnykov, V.; Melnykova, K.; Lishchuk-Yakymovych, K. Personalized Data Analysis Approach for Assessing Necessary Hospital Bed-Days Built on Condition Space and Hierarchical Predictor. *Big Data Cogn. Comput.* 2021, *5*, 37. https://doi.org/10.3390/bdcc5030037

43. Yakovyna, V., & Shakhovska, N. (2021). Modelling and predicting the spread of COVID-19 cases depending on restriction policy based on mined recommendation rules. Mathematical Biosciences and Engineering: MBE, 18(3), 2789-2812
44. Izonin, I., Tkachenko, R., Shakhovska, N., & Lotoshynska, N. (2021). The Additive Input-Doubling Method Based on the SVR with Nonlinear Kernels: Small Data Approach. Symmetry, 13(4), 612
45. Kaminskyy, R., Shakhovska, N., Kryvinska, N., & Younas, M. (2021). Dendrograms-based disclosure method for evaluating cluster analysis in the IoT domain. Computers & Industrial Engineering, 158, 107402.
46. Ivan Izonin, Roman Tkachenko, Ivanna Dronyuk, Pavlo Tkachenko, Michal Gregus, Mariia Rashkevych. Predictive modeling based on small data in clinical medicine: RBF-based additive input-doubling method. Mathematical Biosciences and Engineering, 2021, 18(3): 2599-2613. doi: 10.3934/mbe.2021132
47. Izonin I., Tkachenko R., Fedushko S., Koziy D., Zub K., Vovk O. (2021) RBF-Based Input Doubling Method for Small Medical Data Processing. In: Hu Z., Zhang Q., Petoukhov S., He M. (eds) Advances in Artificial Systems for Logistics Engineering. ICAILE 2021. Lecture Notes on Data Engineering and Communications Technologies, vol 82. Springer, Cham, pp. 23-31. https://doi.org/10.1007/978-3-030-80475-6_3
48. Ivan Izonin, Roman Tkachenko, Michal Gregus, Khrystyna Zub, Nataliia Lotoshynska. “Input Doubling Method based on SVR with RBF kernel in Clinical Practice: Focus on Small Data”, Elsevier Journal Procedia Computer Science, Special Issue on “The 12th International Conference on Ambient Systems, Networks and Technologies (ANT) / The 4th International Conference on Emerging Data and Industry 4.0 (EDI40) / Affiliated Workshops”, vol. 184, pp. 606–613, 2021. ISSN 1877-0509, DOI: 10.1016/j.procs.2021.03.075
49. Ivan Izonin, Roman Tkachenko, Michal Gregus ml., Khrystyna Zub, Pavlo Tkachenko. “A GRNN-based Approach towards Prediction from Small Datasets

in Medical Application", Elsevier Journal Procedia Computer Science, Special Issue on "The 12th International Conference on Ambient Systems, Networks and Technologies (ANT) / The 4th International Conference on Emerging Data and Industry 4.0 (EDI40) / Affiliated Workshops", vol. 184, pp. 242–249, 2021. ISSN 1877-0509, DOI: 10.1016/j.procs.2021.03.075

50. Wagner J, DuPont A, Larson S, Cash B, Farooq A. Absolute lymphocyte count is a prognostic marker in Covid-19: A retrospective cohort review. Int J Lab Hematol. 2020;42(6):761-765. doi:10.1111/ijlh.13288;
51. Mazzoni A, Salvati L, Maggi L, Annunziato F, Cosmi L. Hallmarks of immune response in COVID-19: Exploring dysregulation and exhaustion [published online ahead of print, 2021 Oct 26]. Semin Immunol. 2021;101508. doi:10.1016/j.smim.2021.101508
52. Wang J, Jiang M, Chen X, Montaner LJ. Cytokine storm and leukocyte changes in mild versus severe SARS-CoV-2 infection: Review of 3939 COVID-19 patients in China and emerging pathogenesis and therapy concepts. J Leukoc Biol. 2020;108(1):17-41. doi:10.1002/JLB.3COVR0520-272R
53. Böning D, Kuebler WM, Bloch W. The oxygen dissociation curve of blood in COVID-19. Am J Physiol Lung Cell Mol Physiol. 2021;321(2):L349-L357. doi:10.1152/ajplung.00079.2021
54. V. Piven, VasiaPiven/covid19_ua. 2021. Accessed: Apr. 26, 2021. [Online]. Available: https://github.com/VasiaPiven/covid19_u

ДОДАТКИ

Додаток 1. Основні публікації за результатами проекту

- **Shakhovska, N., Melnykova**, N., & Chopiyak, V. (2022). An Ensemble Methods for Medical Insurance Costs Prediction Task. CMC-COMPUTERS MATERIALS & CONTINUA, 70(2), 3969-3984. **Q=1**
- **Melnykova, N.; Shakhovska**, N.; Melnykov, V.; Melnykova, K.; Lishchuk-Yakymovych, K. Personalized Data Analysis Approach for Assessing Necessary Hospital Bed-Days Built on Condition Space and Hierarchical Predictor. *Big Data Cogn. Comput.* 2021, *5*, 37. https://doi.org/10.3390/bdcc5030037 **Q=2**
- **I. Izonin**, R. Tkachenko, V. Verhun et al., An approach towards missing data management using improved GRNN-SGTM ensemble method, https://doi.org/10.1016/j.jestch.2020.10.005 **Q=1**
- Kaminskyy, R., **Shakhovska, N**., Kryvinska, N., & Younas, M. (2021). Dendrograms-based disclosure method for evaluating cluster analysis in the IoT domain. *Computers & Industrial Engineering*, *158*, 107402. **Q=1**
- **Yaroslav Vyklyuk**, Mykhailo Manylich, Miroslav Škoda, Milan M. Radovanović, Marko D. Petrović, Modeling and analysis of different scenarios for the spread of COVID-19 by using the modified multi-agent systems – Evidence from the selected countries, Results in Physics, 2020, 103662, ISSN 2211-3797, https://doi.org/10.1016/j.rinp.2020.103662 **Q=2**
- **Shakhovska, N., Izonin, I., & Melnykova, N**. (2021). The hierarchical classifier for COVID-19 resistance evaluation. Data, 6(1), 6. Data **Q=2**
- Yakovyna, V., & **Shakhovska, N**. (2021). Modelling and predicting the spread of COVID-19 cases depending on restriction policy based on mined recommendation rules. Mathematical Biosciences and Engineering: MBE, 18(3), 2789-2812. **Q=2**

- **Izonin, I.,** Tkachenko, R., **Shakhovska, N**., & Lotoshynska, N. (2021). The Additive Input-Doubling Method Based on the SVR with Nonlinear Kernels: Small Data Approach. *Symmetry*, *13*(4), 612 **Q=2**
- **Ivan Izonin,** Roman Tkachenko, Ivanna Dronyuk, Pavlo Tkachenko, Michal Gregus, Mariia Rashkevych. Predictive modeling based on small data in clinical medicine: RBF-based additive input-doubling method. Mathematical Biosciences and Engineering, 2021, 18(3): 2599-2613. doi: 10.3934/mbe.2021132 **Q=2**

- **Ivan Izonin**, Roman Tkachenko, Pavlo Vitynskyi, Khrystyna Zub, Pavlo Tkachenko, Ivanna Dronyuk Stacking-based GRNN-SGTM Ensemble Model for Prediction Tasks. Proceedings of the International Conference on Decision Aid Sciences and Applications, 8 – 9 November 2020, Kingdom of Bahrain, pp. 326-330.
- **Ivan Izonin**, Roman Tkachenko, Liliia Ryvak, Khrystyna Zub, Mariia Rashkevych, Olena Pavliuk Addressing Medical Diagnostics Issues: Essential Aspects of the PNN-based Approach Ceur, Vol-2753 , 2020, 209-218
- **Nataliia Melnykova, Natalya Shakhovska**, Volodymyr Melnykov, Mykola Logoyda, Yulia Peleshchak The Problem of Analysing the Relationships between Individual Characteristics of Individuals with COVID`19 Ceur, Vol-2753 , 2020 473-482
- Vitaliy Yakovyna, **Natalya Shakhovska**, Khrystyna Shakhovska, Jaime Campos. Recommendation Rules Mining for Reducing the Spread of COVID-19 Cases. *Ceur*, Vol 2753, p.219-229.
- **Ізонін І. В**., Ткаченко Р. О., Зуб Х. В. Застосування бустингу в ансамблях нейронних мереж узагальненої регресії для підвищення точності розв’язання задач апроксимації. На шляху до індустрії 4.0: інформаційні технології, моделювання, штучний інтелект, автоматизація : монографія / кол. авт. : В .яб. Артеменко, Л. В. Артеменко, О. В. Артеменко [та ін.]; за

заг. ред. В. С. Котлика. - Одеса : Астропринт, 2021. - с. 394-409. ISBN 978-966-702-2

- **Nataliya Shakhovska**. Method for Missing Data Imputation in Big Data technologies // Інформаційні технології, економіка та право: стан і перспективи розвитку : матеріали Міжнародної науково-практичної конференції ІТЕП-2020 (Чернівці, 10 листопада 2020 р.). – 2020. – С. 67–68.
- Вергун В. Р., Вітинський П. Б., **Ізонін І. В.,** Ткаченко Р. О., Кобзар* Н. О. Дослідження ефективності використання бустингових ансамблів штучних нейронних мереж узагальненої регресії для розвязання задач прогнозування // Інформаційні технології і автоматизація – 2020 : збірник тез доповідей XIII Міжнародної науково-практичної конференції, Одеса, 22–23 жовтня 2020 р. – 2020. – С. 282–283.
- **Vyklyuk Y.,** Nevinskyi D. *Modelling And Analysis of Different Scenarios for the Spread of Covid-19 by Using bhe Modified Multi-Agent Systems - Evidence from the Selected Countries* // Міжнародна наукова конференція, присвячена 75-річчю кафедри диференціальних рівнянь та 85-річчю від дня народження Михайла Павловича Ленюка, 28 - 30 жовтня 2021 р., Чернівці: матеріали конференції. - Чернівці, 2021. - с. 193-194
- **Ізонін І. В**., Ткаченко Р. О., Ривак Л. А. Використання імовірнісних нейронних мереж в задачах медичної діагностики // Інформаційні технології, економіка та право: стан і перспективи розвитку : матеріали Міжнародної науково-практичної конференції ІТЕП-2020 (Чернівці, 10 листопада 2020 р.). – 2020. – С. 106–107.
- **Nataliya Shakhovska**. Method for Missing Data Imputation in Big Data technologies // Інформаційні технології, економіка та право: стан і перспективи розвитку : матеріали Міжнародної науково-практичної

конференції ІТЕП-2020 (Чернівці, 10 листопада 2020 р.). – 2020. – С. 67–68.

- Kunanets N., Pasichnyk V., Kunanets O., **Kryvenchuk Y., Vyklyuk Y**. The modern information technologies in the process of raising the virus pandemics. Інформаційні технології, економіка та право: стан і перспективи розвитку: матеріали Міжнародної науково-практичної конференції ІТЕП-2020 (Чернівці, 10 листопада 2020 р.). – 2020. – С. 67.
- **Ізонін І. В**., Ткаченко Р. О., Підкостельний Р. Р. Методи штучного розширення коротких вибірок даних // Інформаційні технології, економіка та право: стан і перспективи розвитку : матеріали Міжнародної науково-практичної конференції ІТЕП-2020 (Чернівці, 10 листопада 2020 р.). – 2020. – С. 104–105.
- **Ivan Izonin,** Roman Tkachenko, Michal Gregus, Khrystyna Zub, Nataliia Lotoshynska. “Input Doubling Method based on SVR with RBF kernel in Clinical Practice: Focus on Small Data”, Elsevier Journal Procedia Computer Science, Special Issue on “The 12th International Conference on Ambient Systems, Networks and Technologies (ANT) / The 4th International Conference on Emerging Data and Industry 4.0 (EDI40) / Affiliated Workshops”, vol. 184, pp. 606–613, 2021. ISSN 1877-0509, DOI: 10.1016/j.procs.2021.03.075
- **Ivan Izonin,** Roman Tkachenko, Michal Gregus ml., Khrystyna Zub, Pavlo Tkachenko. “A GRNN-based Approach towards Prediction from Small Datasets in Medical Application”, Elsevier Journal Procedia Computer Science, Special Issue on “The 12th International Conference on Ambient Systems, Networks and Technologies (ANT) / The 4th International Conference on Emerging Data and Industry 4.0 (EDI40) / Affiliated Workshops”, vol. 184, pp. 242–249, 2021. ISSN 1877-0509, DOI: 10.1016/j.procs.2021.03.075
- **Natalya Shakhovska**, Ihor Darmoriz, **Yaroslav Vyklyuk, Yurii Kryvenchuk** and Pavlo Pukach Visualization of the Epidemics Forecasting Results. IDDM 2021

Свідоцтво про авторське право:

- А. с. 103422 "Комп’ютерна програма “Гібридна система обчислювального інтелекту для підвищення точності розвязання задач класифікації” / **І. В. Ізонін,** Л. А. Ривак, Р. О. Ткаченко; заявн. І. В. Ізонін. – с202101616; заявл. 19.03.2021; опубл. 24.03.2021

  А. с. 103447 "Комп’ютерна програма “Бустинговий ансамбль з двох нейронних мереж узагальненої регресії” / **І. В. Ізонін**, Н. О. Кобзар, Р. О. Ткаченко, О. М. Павлюк; заявн. І. В. Ізонін. – с202101202; заявл. 01.03.2021; опубл. 25.03.2021

Додаток 2. Публікації, прийняті до друку

11th International Conference on Current and Future Trends of Information and Communication Technologies in Healthcare,
ICTH-2021, November 1-4, 2021, Leuven, Belgium

AUTHOR NAME: Ivan Izonin
AFFLIATION: Lviv Polytechnic National University, S. Bandery str., 12, Lviv, 79013, Ukraine
PAPER ID: SBDaH-2021_paper_10

Please be so kind and make the appropriate payment for the Conference Fee for the International workshop on Small and Big Data Approaches in Healthcare, SBDaH-2021 (http://science.lpnu.ua/sbdah2021), that will be held in conjunction with: the 11th International Conference on Current and Future Trends of Information and Communication Technologies in Healthcare, ICTH-2021 (http://cs-conferences.acadiau.ca/icth-21/) on behalf of Dr. Ivan Izonin, Lviv Polytechnic National University, Stepan Bandera str., 12, 79013 Lviv, UKRAINE

Bank Sender:
JOINT STOCK COMPANY 'THE STATE EXPORT-IMPORT BANK OF UKRAINE'Ukraine,
03150 Kyiv, 127 Antonovycha Str.
S.W.I.F.T.: EXBSUAUX
Acc. № 26005000028040
IBAN: UA943223130000026005000028040

Payment should be made as per following invoice:

Pro Forma Invoice	**No. 2021/ICTH/WS010**
Sum:	**EUR 405.00**
In writing:	Conference Participant - 405 EUR (four hundred and five euros)
Date:	**August 10, 2021**

INSTRUCTIONS FOR CUSTOMER (RECEIVER) TRANSFERS IN EUR
Beneficiary:
Title: A. YASAR
Recipient's address: Wetenschapspark 5/6, 3590 Diepenbeek, Belgium
IBAN: BE27377002642073 (ING Bank Belgium)
Swift: BBRUBEBB
Beneficiary's bank address: ING Belgium, Bondgenotenlaan 31, 3000 Leuven
Message: ** INVOICE NUMBER **

Dr. Elhadi Shakshuki
Steering Chair

Secretariat • ICTH Conference

11th International Conference on Current and Future Trends of Information and Communication Technologies in Healthcare,
ICTH-2021, November 1-4, 2021, Leuven, Belgium

AUTHOR NAME: Natalia Shakhovska
AFFLIATION: Lviv Polytechnic National University, S. Bandery str., 12, Lviv, 79013, Ukraine
PAPER ID: SBDaH-2021_paper_1

Please be so kind and make the appropriate payment for the Conference Fee for the International workshop on Small and Big Data Approaches in Healthcare, SBDaH-2021 (http://science.lpnu.ua/sbdah2021), that will be held in conjunction with: the 11th International Conference on Current and Future Trends of Information and Communication Technologies in Healthcare, ICTH-2021 (http://cs-conferences.acadiau.ca/icth-21/) on behalf of Prof. Natalia Shakhovska, Lviv Polytechnic National University, Stepan Bandera str., 12, 79013 Lviv, UKRAINE

Bank Sender:
JOINT STOCK COMPANY 'THE STATE EXPORT-IMPORT BANK OF UKRAINE'Ukraine,
03150 Kyiv, 127 Antonovycha Str.
S.W.I.F.T.: EXBSUAUX
Acc. № 26005000028040
IBAN: UA943223130000026005000028040

Payment should be made as per following invoice:

Pro Forma Invoice	**No. 2021/ICTH/WS001**
Sum:	**EUR 405.00**
In writing:	Conference Participant - 405 EUR (four hundred and five euros)
Date:	**August 10, 2021**

INSTRUCTIONS FOR CUSTOMER (RECEIVER) TRANSFERS IN EURBeneficiary:

Title: A. YASAR

Recipient's address: Wetenschapspark 5/6, 3590 Diepenbeek, Belgium

IBAN: BE27377002642073 (ING Bank Belgium)

Swift: BBRUBEBB

Beneficiary's bank address: ING Belgium, Bondgenotenlaan 31, 3000 Leuven

Message: ** INVOICE NUMBER **

Dr. Elhadi Shakshuki
Steering Chair

Secretariat ICTH Conference

11th International Conference on Current and Future Trends of Information and Communication Technologies in Healthcare, ICTH-2021, November 1-4, 2021, Leuven, Belgium

AUTHOR NAME: Nataliia Melnykova
AFFLIATION: Lviv Polytechnic National University, S. Bandery str., 12, Lviv, 79013, Ukraine
PAPER ID: SBDaH-2021_paper_8

Please be so kind and make the appropriate payment for the Conference Fee for the International workshop on Small and Big Data Approaches in Healthcare, SBDaH-2021 (http://science.lpnu.ua/sbdah2021), that will be held in conjunction with: the 11th International Conference on Current and Future Trends of Information and Communication Technologies in Healthcare, ICTH-2021 (http://cs-conferences.acadiau.ca/icth-21/) on behalf of Dr. Nataliia Melnykova, Lviv Polytechnic National University, Stepan Bandera str., 12, 79013 Lviv, UKRAINE

Bank Sender:
JOINT STOCK COMPANY 'THE STATE EXPORT-IMPORT BANK OF UKRAINE'Ukraine,
03150 Kyiv, 127 Antonovycha Str.
S.W.I.F.T.: EXBSUAUX
Acc. № 26005000028040
IBAN: UA943223130000026005000028040

Payment should be made as per following invoice:

Pro Forma Invoice	**No. 2021/ICTH/WS008**
Sum:	**EUR 405.00**
In writing:	Conference Participant - 405 EUR (four hundred and five euros)
Date:	**August 10, 2021**

INSTRUCTIONS FOR CUSTOMER (RECEIVER) TRANSFERS IN EUR
Beneficiary:
Title: A. YASAR
Recipient's address: Wetenschapspark 5/6, 3590 Diepenbeek, Belgium
IBAN: BE27377002642073 (ING Bank Belgium)
Swift: BBRUBEBB
Beneficiary's bank address: ING Belgium, Bondgenotenlaan 31, 3000 Leuven
Message: ** INVOICE NUMBER **

Dr. Elhadi Shakshuki
Steering Chair

Secretariat • ICTH Conference

11th International Conference on Current and Future Trends of Information and Communication Technologies in Healthcare,
ICTH-2021, November 1-4, 2021, Leuven, Belgium

AUTHOR NAME: Nataliia Melnykova
AFFLIATION: Lviv Polytechnic National University, S. Bandery str., 12, Lviv, 79013, Ukraine
PAPER ID: SBDaH-2021_paper_9

Please be so kind and make the appropriate payment for the Conference Fee for the International workshop on Small and Big Data Approaches in Healthcare, SBDaH-2021 (http://science.lpnu.ua/sbdah2021), that will be held in conjunction with: the 11th International Conference on Current and Future Trends of Information and Communication Technologies in Healthcare, ICTH-2021 (http://cs-conferences.acadiau.ca/icth-21/) on behalf of Dr. Nataliia Melnykova, Lviv Polytechnic National University, Stepan Bandera str., 12, 79013 Lviv, UKRAINE

Bank Sender:
JOINT STOCK COMPANY 'THE STATE ECPORT-IMPORT BANK OF UKRAINE'
Ukraine, 03150 Kyiv, 127 Antonovycha Str.
S.W.I.F.T.: EXBSUAUX
Acc. № 26005000028040
IBAN: UA943223130000026005000028040

Payment should be made as per following invoice:

Pro Forma Invoice	**No. 2021/ICTH/WS009**
Sum:	**EUR 405.00**
In writing:	Conference Participant - 405 EUR (four hundred and five euros)
Date:	**August 10, 2021**

INSTRUCTIONS FOR CUSTOMER (RECEIVER) TRANSFERS IN EUR
Beneficiary:
Title: A. YASAR
Recipient's address: Wetenschapspark 5/6, 3590 Diepenbeek, Belgium
IBAN: BE27377002642073 (ING Bank Belgium)
Swift: BBRUBEBB
Beneficiary's bank address: ING Belgium, Bondgenotenlaan 31, 3000 Leuven
Message: ** INVOICE NUMBER **

Dr. Elhadi Shakshuki
Steering Chair

Secretariat • ICTH Conference