ЗАТВЕРДЖУЮ
Проректор з наукової роботи
Національного університету
«Львівська політехніка»
Демидов І.В.
(підпис)
м.п.

АНОТОВАНИЙ ЗВІТ
про виконану роботу в рамках реалізації проєкту
із виконання наукових досліджень і розробок
«Система підтримки прийняття рішень моделювання поширення вірусних інфекцій»

Назва конкурсу: _«Наука для безпеки людини та суспільства»
Реєстраційний номер Проєкту: 2020.01/0025

Підстава для реалізації Проєкту з виконання наукових досліджень і розробок (реєстраційний номер та назва Проєкту) 2020.01/0025 «Система підтримки прийняття рішень моделювання поширення вірусних інфекцій»

Рішення Наукової ради Національного фонду досліджень України щодо визначення переможця конкурсу «Наука для безпеки людини та суспільства»

протокол № 21 від 16-17.09.2020

1. ЗАГАЛЬНА ІНФОРМАЦІЯ ПРО ПРОЄКТ

Початок: 11.2020

Закінчення: 12.2021 рік

Загальна вартість Проєкту, грн.
1 810 932 грн.

Вартість Проєкту по роках, грн.:
1-й рік 604 632 грн.
2-й рік 1 206 300 грн.

2. ІНФОРМАЦІЯ ПРО ВИКОНАВЦІВ ПРОЄКТУ

до виконання Проєкту було залучено 5 виконавців, з них:

доктори наук	2
кандидати наук	3

3. ІНФОРМАЦІЯ ПРО ГРАНТООТРИМУВАЧА ТА ОРГАНІЗАЦІЮ(Ї) СУБВИКОНАВЦЯ(ІВ) ПРОЄКТУ

Інформація про виконавців (авторів) Проєкту (в тому числі особи, які залучені до виконання Проєкту за трудовим договором або угодою цивільно-правового характеру: ПІБ, основне місце роботи, посада, науковий ступінь).
Завідувач кафедри систем штучного інтелекту, д.т.н., проф. Шаховська Н.Б.,
Професор кафедри СШІ, д.т.н., проф. Виклюк Ярослав Ігорович,
Доцент кафедри СШІ, к.т.н., Мельникова Наталія Іванівна,
Доцент кафедри СШІ, к.т.н., Ізонін Іван Вікторович,
Доцент кафедри СШІ, к.т.н., Кривенчук Юрій Павлович.

Львівський національний медичний університет імені Данила Галицького як субвиконавець залучений до формування бази даних по Covid`19- резистентності, а також для підтвердження адекватності отриманих у процесі моделювання нових залежностей.

Повна назва підприємства: Львівський національний медичний університет імені Данила Галицького

Організаційно-правова форма підприємства /установи/організації: Державна організація (установа, заклад, підприємство)

Підпорядкованість підприємства/установи /організації: Міністерство охорони здоров'я України

Код ЄДРПОУ: 02010793

Код(и) КВЕД: 85.42 - вища освіта, 86.22 Спеціалізована медична практика

Стратегічні напрями наукової діяльності: навчання, лікарська практика

До проекту залучені кандидати медичних наук:

1. Мельников В.А. к.м.н.,доцент кафедри загальної хірургії Львівського Національного Медичного Університету ім. Данила Галицького;
2. Магльований В.А к.м.н.,доцент кафедри загальної хірургії Львівського Національного Медичного Університету ім. Данила Галицького

4. ОПИС ПРОЄКТУ

4.1. Мета Проєкту

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

4.2. Основні завдання Проєкту

Пошук прихованих залежностей даних для визначення характеру перебігу хвороби індивіда. Аналіз великих обсягів даних вимагає визначення груп атрибутів, які утворюють функціональні залежності. Однак, у реальних наборах даних, отриманих з різних джерел, важливі залежності визначені лише для підмножини значень групи атрибутів (існують залежності, наприклад між перенесеними раніше захворюваннями та характером перебігу хвороби зараз - така залежність встановлена між підмножинами значень різних кортежів і не може бути знайдена існуючими методами пошуку прихованих даних); ми будемо називати такі залежності частковими функціональними залежностями. Відповідно, рівень підтримки таких залежностей є низьким, що не дає змоги використовувати їх для подальшого аналізу даних. Водночас, часткові функціональні залежності є модифікованими асоціативними правилами, але такими, що виконуються лише для

частини даних, і залежить від фактору часу. Тому пропонується розробити апарат ймовірнісних секвенційних залежностей як розширення асоціативних правил та секвенційних залежностей. Метод пошуку таких залежностей базуватиметься на основі відкладених обчислень (модифікація FP-дерева). Це дозволяє зменшити часову складність та використовувати паралельний і розподілений режим для розрахунку. Отже, алгоритм пошуку залежностей може бути реалізований на MapReduce.

4.3. Детальний зміст Проєкту:

- *Сучасний стан проблеми*

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.

Проаналізуємо існуючі програмні засоби моделювання інфекційних захворювань.

Epigrass - програмний засіб для мереж епідеміологічного моделювання та аналізу, дозволяє дослідникам здійснювати просторово-часові моделювання. Включає в собі епідеміологічні дані та моделі передачі хвороби і контролю. Epigrass розрахований на побудову популяційних моделей на основі теорії графів.

HealthMapper - це розробка ВООЗ для спостереження та роботи з картографічною інформацією на національному та глобальному рівнях. Опрацьовує географічну, демографічну та медико-санітарну інформацію, у тому числі розташування громади, охорони здоров'я та освіти, доступності дороги, доступ до безпечної води і демографії. В даний час система розрахована на підтримку цілої низки інфекційних захворювань в більш ніж 60 країнах в усіх регіонах ВОЗ.

Model-Builder – це графічна утиліта для дизайну, симуляції та аналізу математичних моделей на основі диференційних рівнянь.

AnyLogic – це інструмент імітаційного моделювання, що об'єднав методи системної динаміки, "процесного" дискретно-дійового і агентного моделювання. AnyLogic підтримує різноманітні типи експериментів з моделями: простий прогон, порівняння прогонів, варіювання параметрів, Монте-Карло, аналіз чутливості, оптимізація, калібрування, а також експеримент по призначеному для користувача сценарієм.

Для математичного моделювання поширення інфекцій використовують 2 види моделювання епідемій - стохастичний та детермінований.

Стохастичні моделі спираються на міжіндивідуальні ризики зміни впливу, хвороби та інших факторів. Вони часто використовуються, коли ймовірність коливання або знання різнорідності є важливим як в малій так в ізольованій популяції. Стохастичні моделі дозволяють перевірку

кожного індивідууму в популяцій. Стохастичні моделі, крім того, можуть бути дуже трудомісткі і потребують багато симуляцій для того щоб отримати корисні прогнози.

Детерміновані математичні моделі, також відомі як моделі з відсіками, описують, що в загалом відбувається у популяції. Ці моделі поділяють індивідууми у різні підгрупи (відсіки). Модель SEIR (Susceptible - вразливий, Exposed – заражений, Infected - хворий та Recovered - видужали), наприклад, включає в себе чотири відсіки представлені як вразливі, заражені, хворі і здорові (ті, що перехворіли хворобу і мають вже імунітет до неї). Крім того, моделі включать швидкість переходу між відсіками, як вразливий може стати зараженим, заражений хворим, і так далі. Найвідоміший швидкістю переходу є сила інфекції або швидкість атаки яка вимірює швидкість з якою вразливі стають хворими.

Більшість моделей інфекційних захворювань, які використовуються в наш час є детермінованими, оскільки вони вимагають менше даних і їх легше розробляти засобами обчислювальних машин та систем. Поширення SIR-моделі (Susceptible - вразливий, Infected - хворий та Recovered - видужали) в даний час добре вивчена, тому детерміновані моделі, як правило, використовуватися для вивчення ефективності тієї чи іншої стратегії управління. На жаль, такі моделі, побудовані на основі диференційних рівнянь, важко застосовувати в умовах, де для ситуації необхідно вказувати якісь специфічні параметри або у випадку швидкого переналаштування моделі.

- *Новизна Проєкту*

Створено інформаційну систему моделювання та прогнозування поширення інфекцій різного роду, що має здатність масштабування (країна, регіон, місто). Дана система базується на двох підходах до моделювання поширення інфекцій та визначення причин відхилень даних. Перший з них зосереджено на імітації в режимі реального часу просторового поширення інфекції на основі агентного підходу та клітинних алгоритмів. Другий - на моделюванні перебігу хвороби та визначенні основних факторів впливу на конкретного індивіда на основі модифікованих асоціативних та секвенційних залежностей.

Розроблено методу пошуку прихованих залежностей даних для визначення характеру перебігу хвороби індивіда.

Аналіз великих обсягів даних вимагає визначення груп атрибутів, які утворюють функціональні залежності. Однак, у реальних наборах даних, отриманих з різних джерел, важливі залежності визначені лише для підмножини значень групи атрибутів (існують залежності, наприклад між перенесеними раніше захворюваннями та характером перебігу хвороби зараз - така залежність встановлена між підмножинами значень різних кортежів і не може бути знайдена існуючими методами пошуку прихованих даних); ми будемо називати такі залежності частковими функціональними залежностями. Відповідно, рівень підтримки таких залежностей є низьким, що не дає змоги використовувати їх для подальшого аналізу даних. Водночас, часткові функціональні залежності є модифікованими асоціативними правилами, але такими, що виконуються лише для частини даних, і залежить від фактору часу. Тому розроблено апарат ймовірнісних секвенційних залежностей як розширення асоціативних правил та секвенційних залежностей. Метод пошуку таких залежностей базується на основі відкладених обчислень (модифікація FP-дерева). Це дозволяє зменшити часову складність та використовувати паралельний і розподілений режим для розрахунку. Отже, алгоритм пошуку залежностей реалізований на MapReduce.

Розроблено модель поширення інфекції в режимі реального часу. Як результат отримується динаміка (просторова та статистична) таких показників, як кількість інфікованих, здорових, тих, що перехворіли, летальних випадків та інші показники. Також можливим є імітувати різні сценарії запобігання поширенню інфекції на кшталт карантину чи інших заходів, а також оцінювати їх ефективність та науково-обгрунтовано приймати стратегічні рішення.

- *Методологія дослідження*

Задача пошуку залежностей в даних потребує аналізу залежностей між десятками параметрів досліджуваного процесу та сотнями можливих джерел впливу на цей процес. Залежності носять недетермінований характер і тому моделювання потребує застосування статистичних методів

аналізу випадкових процесів. Значна частина інформації часто є прихована від спостереження або ж спостереження за нею не ведеться. Це вносить багато труднощів у процес аналізу зібраної інформації.

На сьогодні розроблені методи статистичного аналізу дають змогу працювати з частково невизначеними чи розмитими процесами. Проте наявні методи мають суттєві обмеження в області застосування та типах даних, що можуть аналізуватись цими методами.

Іншою особливістю медичних даних є їх ієрархічність та мережевість. До мережевих даних належить інформація про супутні патології, алергічні реакції тощо, що також є прямим або непрямим чинником, який визначає характер захворюваності індивіда. Отже, необхідним є відшукання не тільки лінійних залежностей у даних.

Таким чином, усі вищезазначені фактори можуть негативно впливати на проведення, інтерпретацію і узагальнення результатів досліджень, і на розуміння й тлумачення досліджуваного феномену. У роботі розроблено підхід до моделювання характеру захворюваності індивіда на основі підходу Великих даних. Метод складається з двох частин:

1) Пошук ймовірнісних продукційних залежностей на основі моделі великих даних;

2) Використання ймовірнісних продукційних залежностей для моделювання характеру захворюваності.

В основі цього методу покладено розробку спеціальних методів формування навчальної множини даних і попередньої обробки атрибутів з урахуванням специфіки контенту медичних даних та даних навколишнього оточення і розроблення ансамблів моделей імпутації даних на основі базових моделей різнобічної природи у складі спеціалізованої інформаційної технології відновлення пропущених даних для автоматизованого опрацювання інформації.

На етапі 2 розроблено метод генерації ймовірнісних продукційних залежностей на основі асоціативних правил секвенційних залежностей, що дає змогу визначити приховані залежності даних не тільки на рівні кортежів (залежності між значеннями атрибутів), а також на рівні підмножин кортежів (врахування часового фактору).

Оптимізація відомих методів полягає в тому, що для кожної залежності через хеш-таблицю визначається багато залежностей з однаковою частиною результату або однією і тією ж умовною частиною. Комбінація відбувається не з усіма іншими елементарними залежностями, а лише з відповідним злиттям стану.

На наступному етапі здійснено валідацію даних, порівняння отриманих результатів з результатами відомих методів.

Наступним етапом є прогнозування динаміки поширення інфекції та моделювання різних сценаріїв впливу з боку держави. Ця модель базується на теорії агентних систем, та може бути реалізована методами імітаційного моделювання. Ці системи передбачають, що кожна окремо взята людина представляє агента, який переміщується, наприклад, в межах обмеженої області та за певними правилами взаємодіє з іншими агентами. Перевагою цих моделей теж є легка адаптованість та здатність вдосконалювати систему моделювання без великого втручання в код програми. Для моделювання необхідні наступні дані, що можуть бути отримані з першої частини розробки та відомих статистичних даних: Кількість населення центру та всіх районів області, Густина населення обласного центру та районів, Соціальна дистанція між людьми, Тривалість хвороби, Імовірність захворювання при контактах людей, Рівень смертності, Наявність місць скупчення людей (супермаркети, церкви, аптеки, ринки, об’єкти будівництва, спортзали), Відсоток людей, що переносять хворобу безсимптомно, Наявність процедури ізоляції хворих людей, Можливість переміщення людини з району до обласного центру та назад, Дотримання людьми необхідної дистанції, інкубаційний період та інші. Для стійкості результату використано ансамблі моделей, які легко розпаралелити. Тому реалізація розрахункового ядра може виконуватись на комп’ютерних кластерах.

- *Інформація про наявну матеріально-технічну базу, обладнання та устаткування, необхідні для виконання Проєкту*

Усе обладнання необхідне для виконання проєкту є в наявності.

- *Очікувані результати виконання Проєкту:*

а) Опис наукової або науково-технічної продукції (за її наявності), яка буде створена в результаті виконання Проєкту (із зазначенням її очікуваних якісних та кількісних (технічних) характеристик) у 2021 році.

У результаті виконання проєкту у 2021 р. 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) Інформаційна технологія для моделювання поширення захворюваності із візуалізацією на карті, а також рекомендаційна лікарська система моделювання перебігу захворюваності індивідом.

Особливістю розробленої інформаційної технології є можливістю симуляції поширення захворюваності за заданих початкових умов та зміна цих умов у динаміці. Адекватність моделі підтверджена шляхом симуляції поширення Covid`19 та порівняння зі статистичними даними не тільки для України, а також Туреччини, Чехії та Румунії.

Впродовж виконання проєкту авторами підготовлено та опубліковано **9 статей** у рейтингових наукових **журналах** з **Q1/2**, опубліковано **15 тез** у міжнародних конференціях, з них **7 індексуються в наукометричні базі даних Scopus**, отримано **2 свідоцтва про авторське право на комп'ютерну програму**.

б) Обґрунтування переваг очікуваної наукової або науково-технічної продукції (за її наявності) у порівнянні з існуючими аналогами на підставі порівняльного аналізу.

Розроблені засоби можуть використовуватися для опрацювання даних по різних типах вірусів та інфекцій. Порівняно з аналогами, перевагами наукової та науково-технічної продукції є:

1. універсальність - можливість моделювати поширення вірусів та інфекцій різного роду, задавши лише початкові умови агентів та характер їх взаємодії, без внесення коректівів у моделі даних;
2. застосовність до інших цілей - знайдені приховані залежності даних можуть бути використані не тільки для моделювання перебігу складності захворювання, але й для оцінювання загалом стани індивіда та підтримки показника якості життя на належному рівні;
3. візуальний аналіз - візуалізація поширення захворюваності на карті з накладанням різних шарів дасть змогу корегувати управлінські рішення як а рівні малих громад, так і на рівні держав.

5. ОТРИМАНІ НАУКОВІ АБО НАУКОВО-ТЕХНІЧНІ РЕЗУЛЬТАТИ (до 2 сторінок) в поточному році/ в рамках реалізації Проєкту, зокрема:

5.1. Опис наукових або науково-технічних результатів, отриманих в рамках виконання Проєкту (із зазначенням їх якісних та кількісних (технічних) характеристик)

У результаті виконання проєкту досягнуто таких цілей:

1. Розроблено імітаційну модель на основі теорії агентних систем
2. Розроблено метод прогнозування поширення інфекцій на основі ансамблів моделей
3. Розпаралелено обчислення для пришвидшення отримання результатів моделювання.
4. Розроблено метод пошуку ймовірнісних продукційних залежностей на основі моделі великих даних;

5. Розроблено метод використання ймовірнісних продукційних залежностей для моделювання характеру захворюваності.
6. Розроблено інформаційну систему моделювання та прогнозування поширення інфекцій різного роду.

5.2. За наявності науково-технічної продукції обґрунтування її переваг у порівнянні з існуючими аналогами

Порівняно з аналогами, перевагами наукової та науково-технічної продукції є:

1. універсальність - можливість моделювати поширення вірусів та інфекцій різного роду, задавши лише початкові умови агентів та характер їх взаємодії, без внесення коректівів у моделі даних;
2. застосовність до інших цілей - знайдені приховані залежності даних можуть бути використані не тільки для моделювання перебігу складності захворювання, але й для оцінювання загалом стани індивіда та підтримки показника якості життя на належному рівні.

5.3. Практична цінність отриманих результатів реалізації Проєкту для економіки та суспільства (стосується проєктів, що передбачають проведення прикладних наукових досліджень і науково-технічних розробок)

Загалом, розроблена інформаційна технологія може використовуватися користувачами різних категорій. Користувачами науково-практичної продукції можуть як медики, так і органи державного управління, соціальні служби тощо. Очікувані результати проєкту матимуть широке застосування і за кордоном, оскільки відповідають європейським трендам щодо диджилітизації медицини і покращення рівнів безпеки та якості життя. Передбачення характеру поширення захворюваності сприятиме зменшенню видатків та лікування, а аналіз персональних даних та пошук прихованих залежностей стане основою для запобігання рецидивів та покращить якість життя індивідів.

5.4. Опис шляхів та способів подальшого використання результатів виконання Проєкту в суспільній практиці.

Інформаційна система, розроблена як результат виконання проекту в цілому (2020-2021 рр.), а також розроблені моделі та методи зокрема, можуть бути використані кількома способами:

- як система/компонент системи підтримки прийняття рішень лікарями - для передбачення стану індивіда та визначення оптимальних лікувальних схем;
- як навчальна система/компонент системи для студентів-медиків та молодих фахівців - для імітаційного моделювання перебігу хвороби,
- органами місцевого та державного управління - для запобігання поширення захворюваності у неконтрольованих обсягах,
- соціальними службами - для визначення груп населення, які можуть бути найбільш вразливими для тих чи інших захворювань;
- студентами математичних та технічних спеціальностей - для навчання в курсах імітаційного моделювання, математичної статистики тощо.

Науковий керівник Проєкту
Завідувач кафедри систем штучного інтелекту
Національного університету «Львівська політехніка»
Шаховська Н.Б.

(підпис)