КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
ІМЕНІ ТАРАСА ШЕВЧЕНКА
Факультет інформаційних технологій
Кафедра інтелектуальних технологій

КВАЛІФІКАЦІЙНА РОБОТА
на здобуття освітнього ступеня «магістр»

НА ТЕМУ:

«Нейромережна технологія прогнозування розвитку пандемії COVID-19 на території України»

Галузь знань: 12 «Інформаційні технології»
Спеціальність: 122 «Комп'ютерні науки»
Освітньо-наукова програма «Технології штучного інтелекту»

Виконав:
студент(ка) 2 курсу магістратури, групи ТШІ-21
Дідик А.А.
(ПІБ)

Науковий керівник: Тмєнова Н.П.
(ПІБ)
к.ф.- м.н., доцент
(науковий ступінь, вчене звання)

Засвідчую, що в цій кваліфікаційній роботі немає запозичень з праць інших авторів без відповідних посилань

Студент(ка) _____
підпис

Кваліфікаційна робота допущена до захисту
рішенням кафедри *інтелектуальних технологій*

Протокол № 8 від «5» травня 2022 р.

Зав. кафедри _____ доц. Іларіонов О.Є.
підпис

Київ – 2022

РЕФЕРАТ

Кваліфікаційна робота складається зі вступу, 3 розділів, висновків, списку використаної літератури із 18 джерел та 1 додатку. Загальний обсяг роботи 66 сторінок. Робота містить 5 таблиці та 38 рисунки.

Тема кваліфікаційної роботи є актуальною, адже прогнозування розвитку та швидкості поширення захворювань за допомогою нейромережних технологій – це новий перспективний напрямок розвитку медичної сфери і штучного інтелекту.

Моделі, що призначені для прогнозування коронавірусної хвороби, можуть використовуватись для передбачення можливих варіантів розвитку інших епідемій в майбутньому.

Метою роботи - розробка нейромережевої технології прогнозування розвитку епідемії COVID-19 в Україні.

Об’єкт дослідження – процеси прогнозування розвитку епідемії на основі статистичних даних.

Предмет дослідження – методи прогнозування розвитку епідемії COVID-19 на території України на базі використання нейромережних технологій.

Методи дослідження – методи нейромережних технологій. Результати експериментальних досліджень візуалізовано за допомогою програмних засобів мови Python.

У результаті дослідження було проаналізовано моделі прогнозування, їх ефективність та доцільність застосування для даної задачі та обрано алгоритм, який дозволить отримувати точний середньостроковий прогноз. Побудовано модель, яка дозволить ефективно розв’язувати задачу прогнозування.

Ключові слова: LSTM, COVID-19, нейронна мережа, прогнозування, часові ряди.

ABSTRACT

Qualification work consists of an introduction, 3 chapters, conclusions, a list of references from 18 sources and 1 appendix. The total volume of the work is 66 pages. The work contains 5 tables and 38 figures.

The topic of qualification work is relevant, because forecasting the development and rate of spread of diseases using neural network technologies is a new promising area of development of the medical field and artificial intelligence.

Models designed to predict coronavirus disease can be used to predict possible future epidemics.

The aim of the work is to develop a neural network technology for predicting the development of the COVID-19 epidemic in Ukraine.

The object of research is the processes of forecasting the development of the epidemic on the basis of statistical data.

The subject of the research is methods of forecasting the development of the COVID-19 epidemic on the territory of Ukraine on the basis of the use of neural network technologies.

Research methods - methods of neural network technologies. The results of experimental research are visualized using Python software.

As a result of the research, forecasting models, their efficiency and expediency of application for this task were analyzed and an algorithm was chosen that will allow to obtain an accurate medium-term forecast. A model has been built that will allow to effectively solve the problem of forecasting.

Key words: LSTM, COVID-19, neural network, forecasting, time series.

ЗМІСТ

ВСТУП 6
РОЗДІЛ 1 АНАЛІТИЧНИЙ ОГЛЯД, ПОСТАНОВКА ЗАВДАННЯ 7
1.1 Актуальність досліджуваної проблеми 7
1.2. Аналіз відомих рішень в області медичного прогнозування 15
1.3. Постановка задачі 20
Висновки до розділу 20
РОЗДІЛ 2 ТЕОРЕТИЧНЕ ТА МЕТОДИЧНЕ ДОСЛІДЖЕННЯ ТЕМИ 21
2.1. Особливості задачі прогнозування 21
2.2. Огляд методів прогнозування 24
2.3. Аналіз архітектури моделі LSTM 28
2.4. Методичне забезпечення організації проведення досліджень 35
2.4.1. Проектування абстрактної моделі прогнозування 35
2.4.2. Опис навчального набору даних 36
Висновки до розділу 43
РОЗДІЛ 3 РЕАЛІЗАЦІЯ, ТЕСТОВИЙ ПРИКЛАД 44
3.1 Обґрунтування вибору середовища програмування та допоміжних бібліотек 44
3.1.1. Використання бібліотеки Keras 45
3.1.2. Використання бібліотеки Pandas 46
3.1.3. Використання бібліотеки Matplotlib 47
3.1.4. Використання бібліотеки NumPy 48
3.2. Опис результатів моделювання 49

3.3. Оцінка адекватності експериментального дослідження..... 58
ВИСНОВКИ..... 60
СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ..... 61
ДОДАТОК А Лістинг програми навчання LSTM мережі..... 64

ВСТУП

Епідемії з давніх часів створювали смертельну небезпеку для людства. Не є винятком і коронавірусна хвороба 2019 (COVID-19) — інфекційна хвороба, що була вперше виявлена в грудні 2019 року в місті Ухань (Центральний Китай). Хвороба почалася як спалах, що розвинувся у пандемію. Хоча багато країн вже запустили програми масової вакцинації COVID-19 для швидкого контролю спалаху захворювання, багато країн по всьому світу борються з безпрецедентним сплеском нових випадків COVID-19 через більш заразні і небезпечні штами. Втома від пандемії і апатія громадськості до різних стратегій втручання ставлять перед урядовцями нові завдання по боротьбі з пандемією.

Сучасна епідеміологія в повній мірі базується на системному підході, тому приділяє велику увагу передбаченню можливих варіантів розвитку епідемій. На сьогоднішній день спостерігається широке впровадження сучасних комп'ютерних технологій в сферу медицини, і для задач прогнозування у тому числі. Це відкриває нові можливості у сфері медицини, адже прогнозування дає змогу передбачати можливі варіанти розвитку епідемій для вчасного проведення адекватних протиепідемічних заходів.

У першому розділі роботи розглянуто вже відомі дослідження у цій галузі, описано актуальність досліджуваної теми, основні проблеми, що виникають під час її розв’язання, а також описано основні задачі дослідження.

У другому розділі проведено теоритичне та методичне дослідження теми, визначено основні напрямки вирішення поставленної задачі, розглянуто структуру навчального набору даних.

У третьому розділі описано програмну реалізацію процесу прогнозування розвитку епідемії, проаналізовано отримані результати, а також графічно проілюстровано отримані результати.

РОЗДІЛ 1 АНАЛІТИЧНИЙ ОГЛЯД, ПОСТАНОВКА ЗАВДАННЯ

1.1 Актуальність досліджуваної проблеми

Коронавірусне захворювання або COVID-19 – це інфекційна хвороба, причиною якої є новий штам коронавірусу. Коронавірус SARS-CoV-2 спричиняє виникнення цього захворювання, його циркуляція в людській популяції раніше була невідомою.

Всесвітня організація охорони здоров'я (ВООЗ) офіційно визнала COVID-19 пандемією, тобто надзвичайною ситуацією в галузі міжнародної охорони здоров'я та внесла заходи з боротьби проти неї до тимчасових рекомендацій згідно з Міжнародними медико-санітарними правилами 2005 року [1].

Курці відносяться до основних груп ризику, у них за результатами численних проведених мета аналізів передбачається частий тяжкий перебіг [2]; особи чоловічої статі, люди азійської етногрупи, похилого віку [3] й особи з тяжкими хронічними хворобами (есенціальна артеріальна гіпертензія, цукровий діабет тощо) і значними зрушеннями в імунній системі, люди похилого віку 60-65 років. Захист людей з цих категорії ризику є завданням профілактики, адже їх кількість досягає багатьох сотень мільйонів осіб на Землі [4]. Найбільший ризик мають медичні працівники, адже для лікарів, що безпосередньо і тривалий час контактують із хворими, у них він у 8 разів більший, ніж у звичайного населення. Тривалість імунітету після перенесеної хвороби невідома. ВООЗ заявили, що хоча керівництво деяких країн роблять припущення, що виявлення антитіл до SARS-CoV-2 може послужити основою для «сертифіката відсутності ризику» або «імунного паспорта», на даний момент не можна стверджувати, що люди, які одужали від коронавірусної хвороби 2019 і мають антитіла до коронавірусу, захищені від повторної інфекції [5].

Розповсюдження коронавірусу 2019 територією України зафіксовано 3 березня 2020 року, саме тоді вперше було підтверджено випадок хвороби в одному з населених пунктів Чернівецької області.

Перший летальний випадок було зафіксовано 13 березня 2020 року внаслідок інфекції коронавірусу. Станом на 8 травня 2022 в Україні налічувалося близько 5002870 хворих, з них летальних випадків — 108411.

Відповідно до статистики 18 квітня 2021 року Україна займала 16 місце у світі (8-е в Європі) за кількістю зафіксованих випадків інфікування і 18-е у світі (8-е в Європі) — за кількістю летальних випадків. Кількість інфікованих на 1 мільйон населення складає 44722 осіб (67-е місці у світі і 36-е в Європі), а кількість летальних випадків з COVID-19 на мільйон населення — 915 осіб (46-е місце у світі і 33-є в Європі) [6].

У таблиці 1.1 наведено поточну статистику по коронавірусу на 8.05.2022 (Україна) за даними офіційного сайту МінфінМедіа [7].

Таблиця 1.1 Поточна статистика по коронавірусу на 8.05.2022 (Україна)

Населення	41 130 тис.	
Всього інфіковано	5 002 870	12,2 %
Смертельні випадки	108 411	2,2 %
Видужали	4 848 175	96,9 %
Наразі хворіють	46 284	0,9 %
Зроблено тестів	19 521 252	
тестів на 1 млн.	451 363	

Далі буде наведено візуалізації розвитку епідемії COVID-19 в Україні за різними показниками (рис. 1.1- 1.8) за даними джерела [7].

Нижче надається зведений графік по всіх категоріях інфікованих (рис. 1.1).

Діаграма, що зображена на рисунку 1.3, показує кількість підтверджених випадків захворювання на COVID-19 на день. Через обмежене тестування кількість підтверджених випадків нижча за справжню кількість інфекцій.

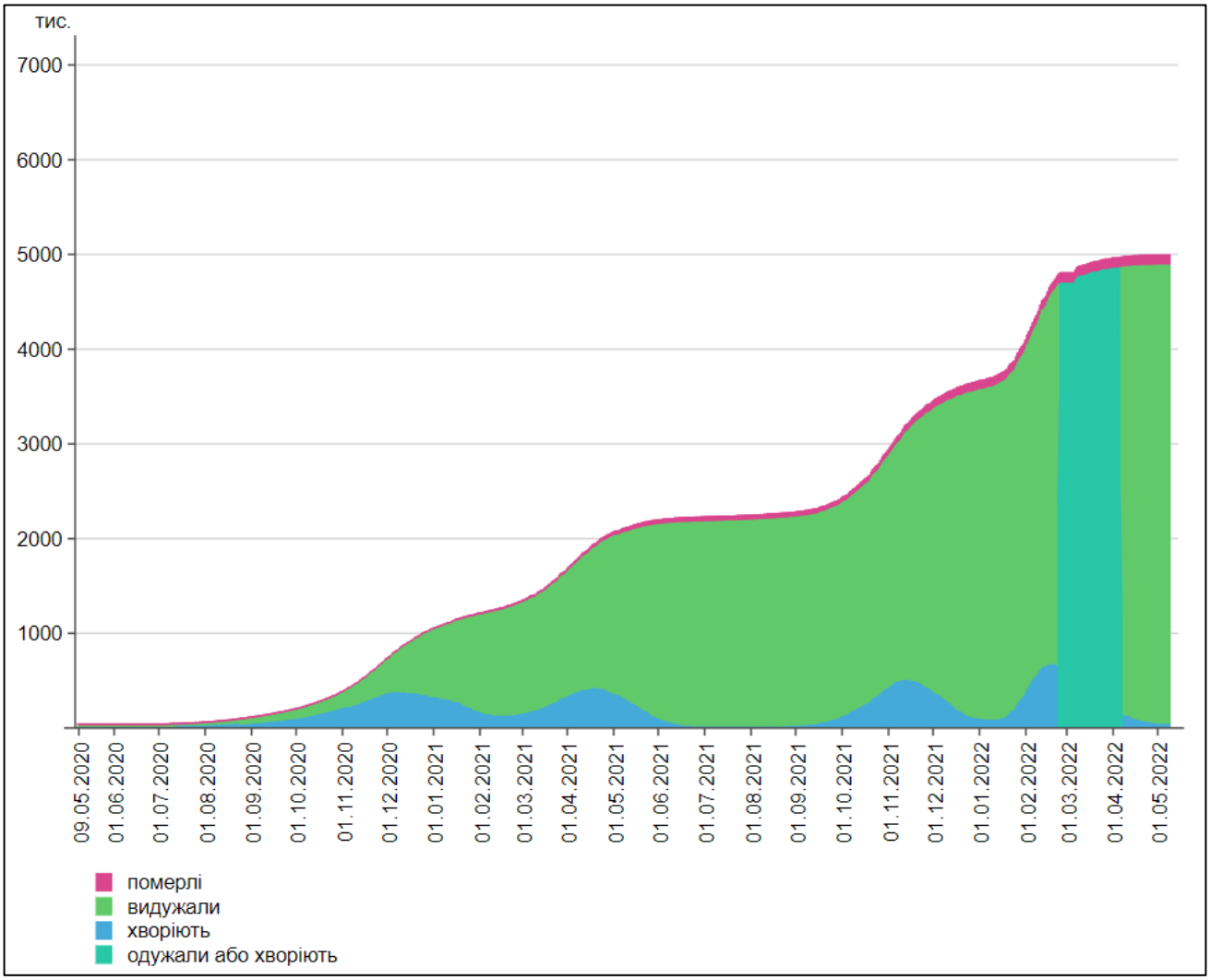

Рисунок 1.1 Зведений графік по всіх категоріях інфікованих [7]

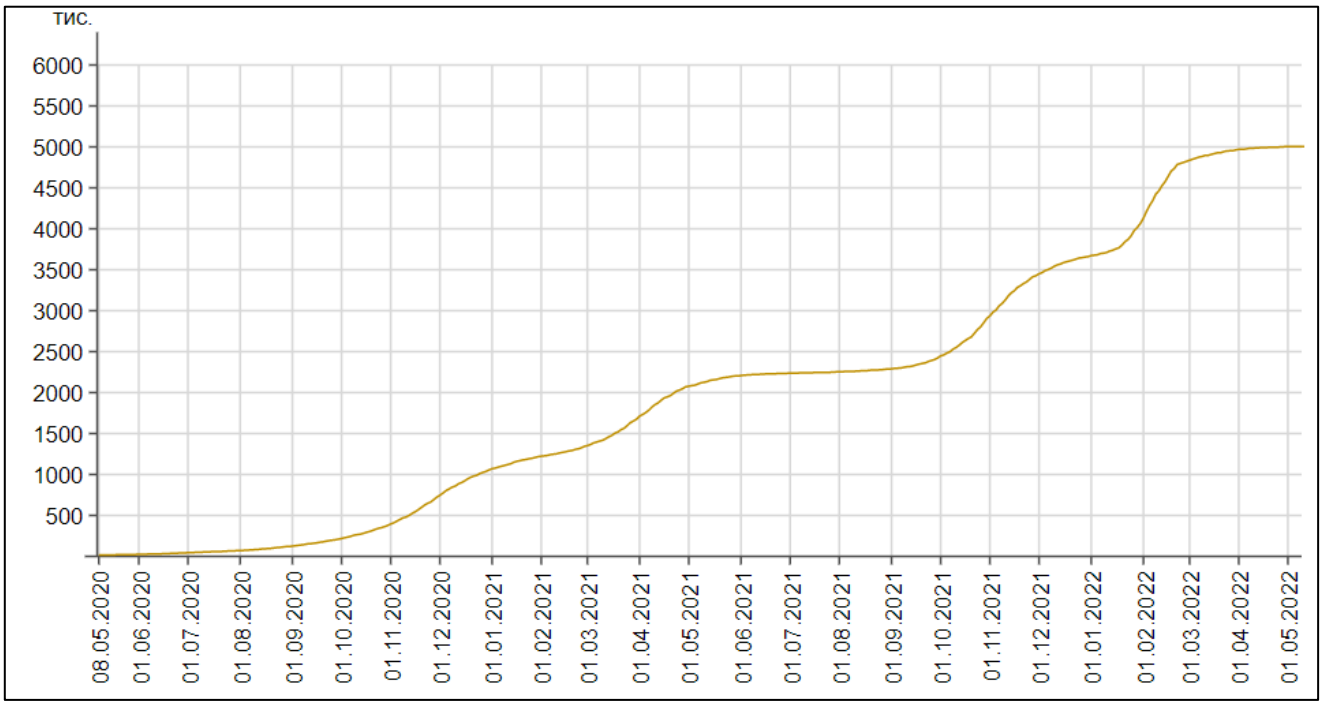

Рисунок 1.2 Загальна кількість інфікованих [7]

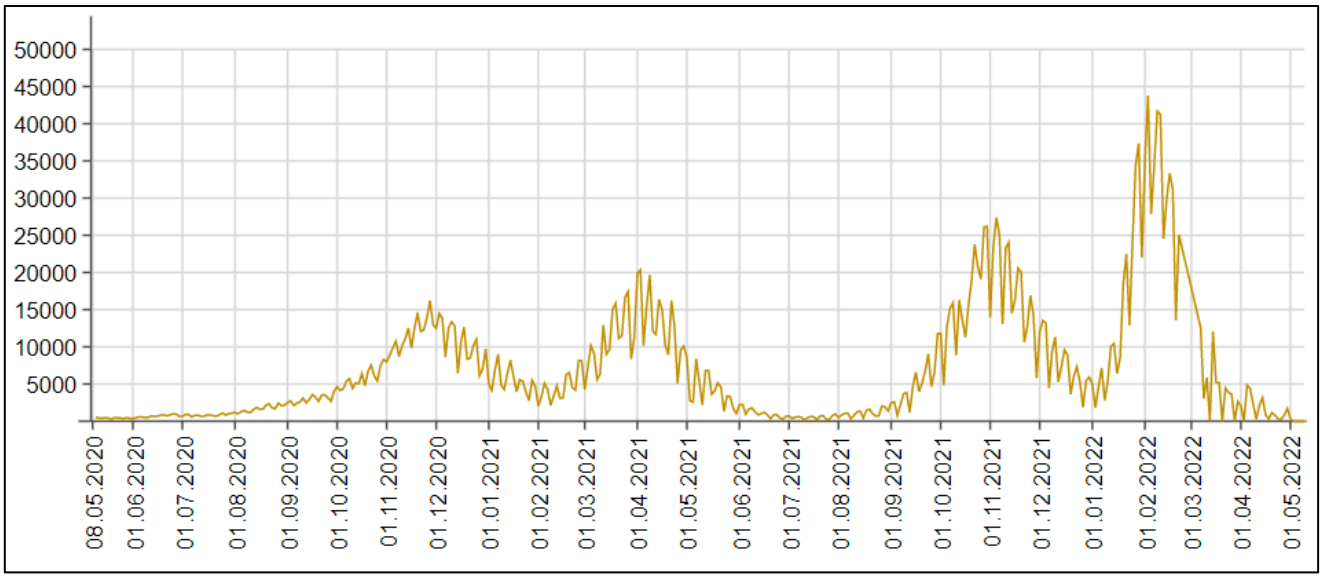

Рисунок 1.3 Загальна кількість інфікованих, по днях [7]

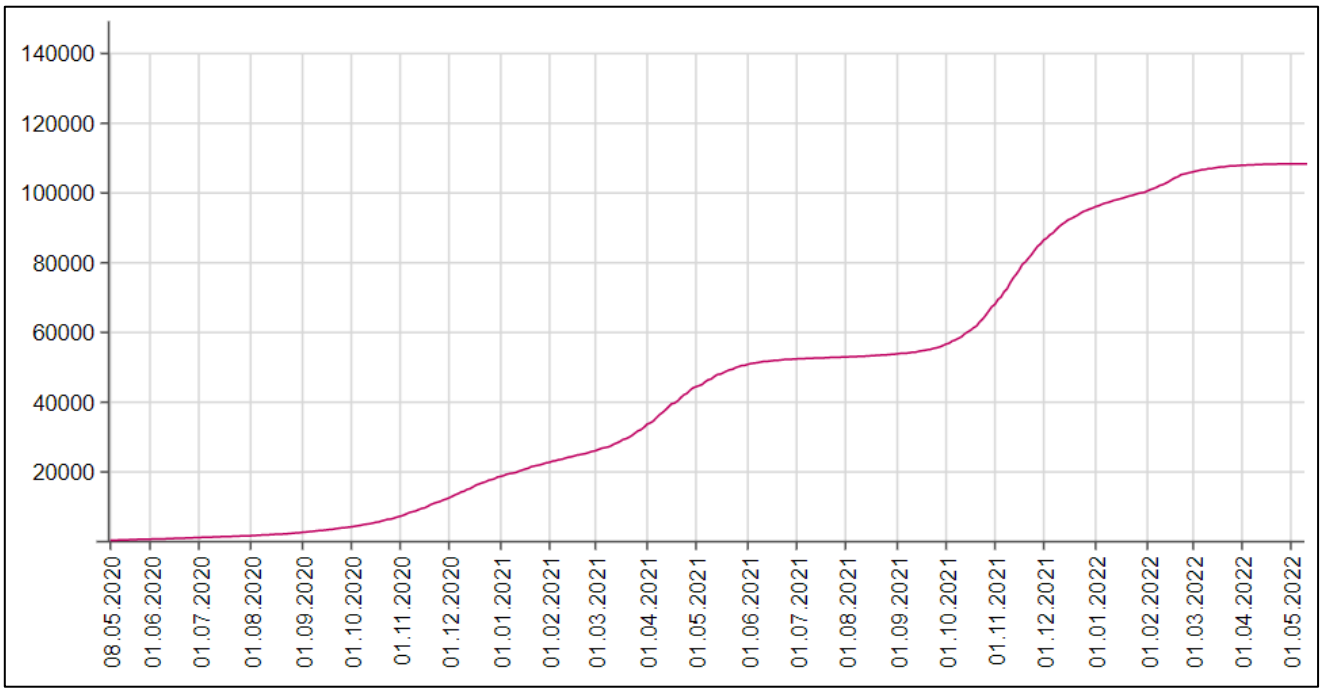

Рисунок 1.4 Загальна кількість летальних випадків [7]

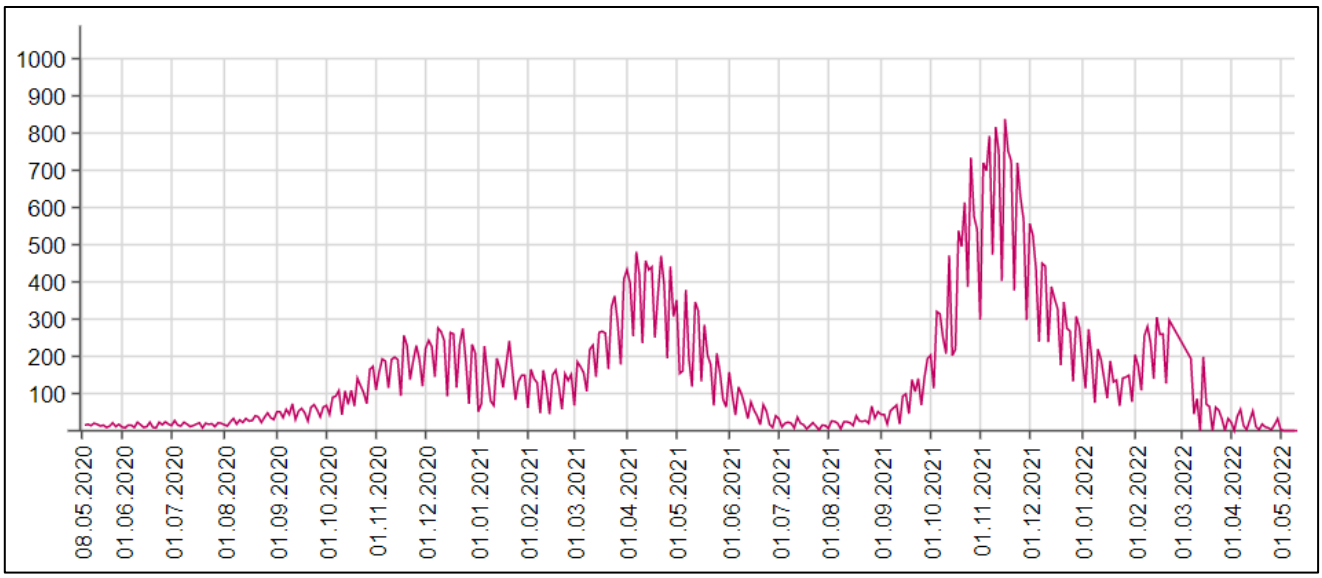

Рисунок 1.5 Загальна кількість летальних випадків, по днях [7]

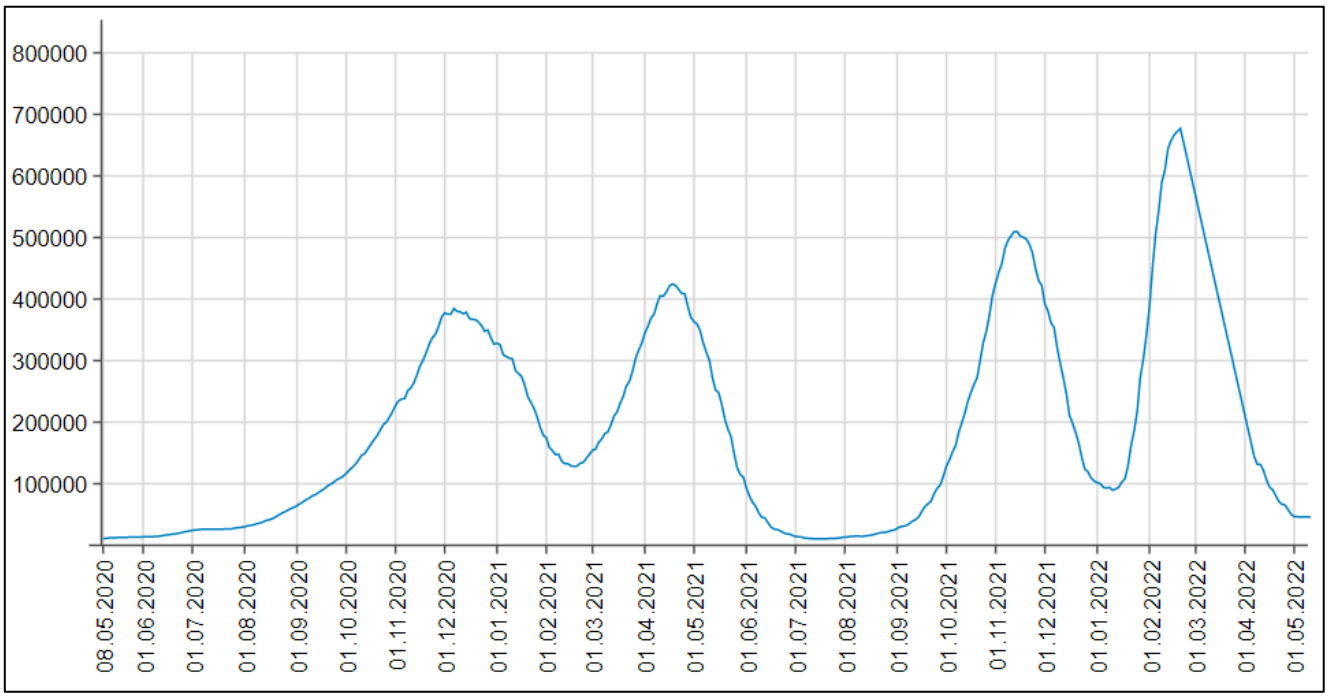

Рисунок 1.6 Кількість хворих [7]

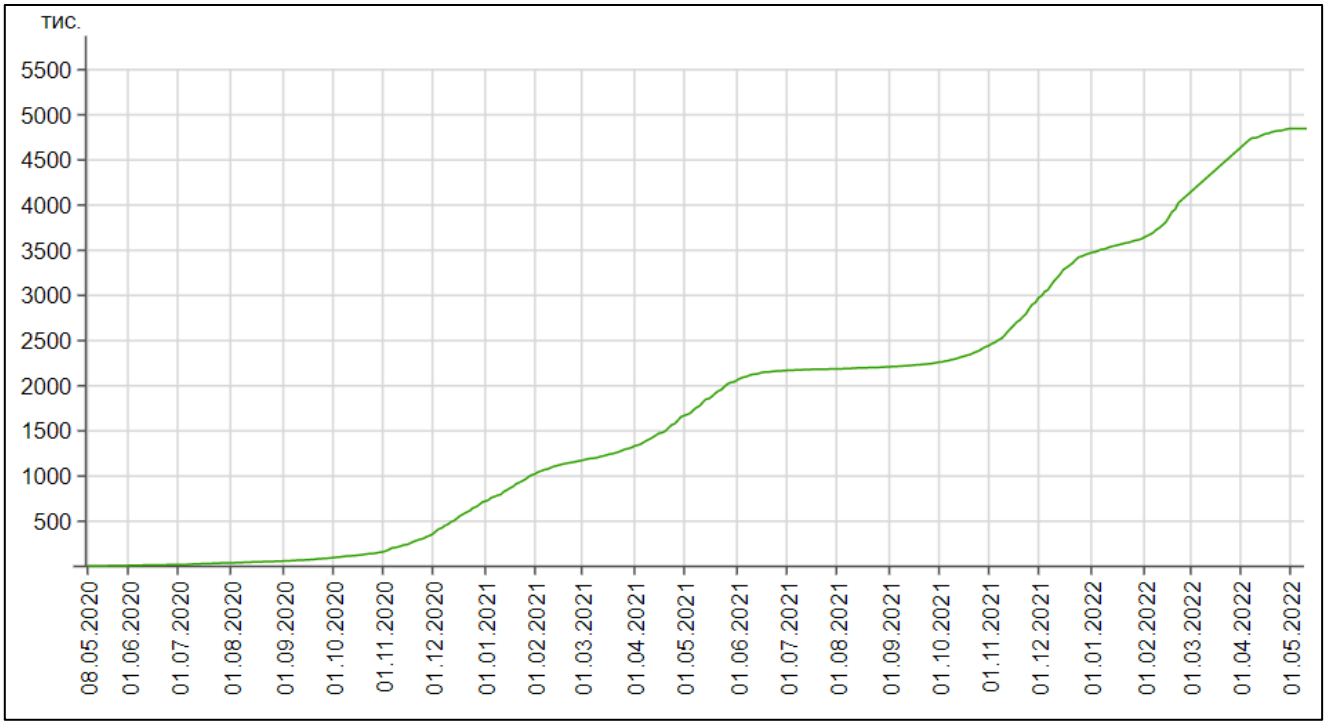

Рисунок 1.7 Загальна кількість тих, що одужали [7]

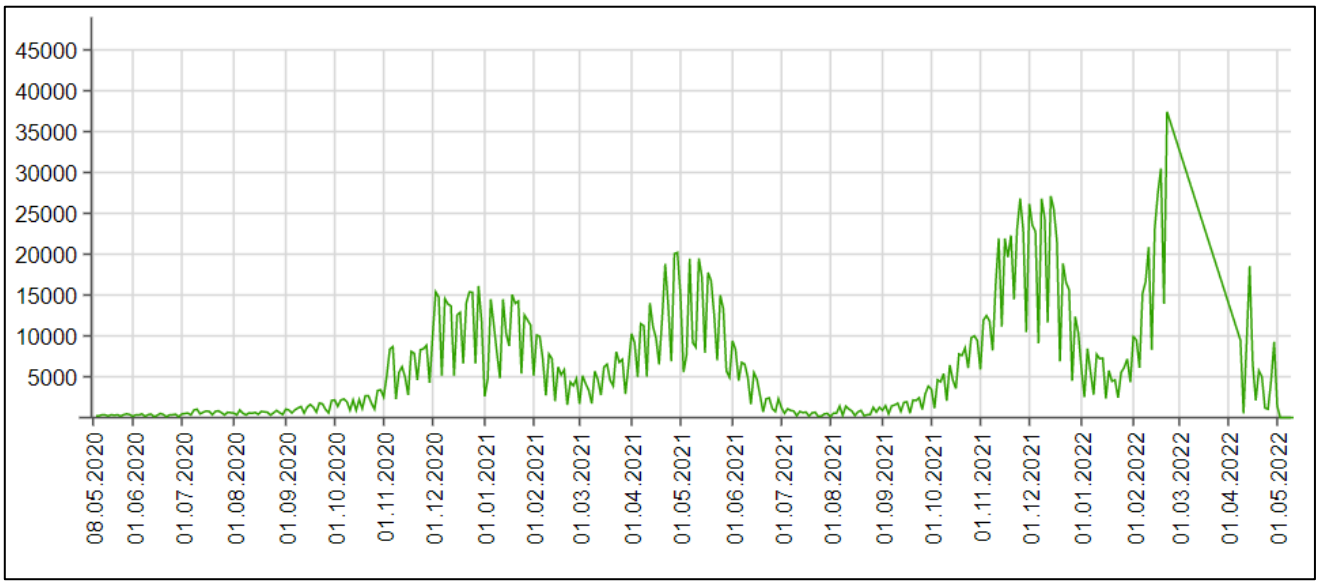

Рисунок 1.8 Кількість тих, що одужали, по днях [7]

Різниця в чисельності населення між різними країнами часто велика. Для порівняння захворюваності у різних країнах важливо подивитися на кількість підтверджених випадків на мільйон людей (рис. 1.9) [8].

Варто відмітити, що в країнах, які проводять дуже мало тестів, фактична кількість випадків може бути набагато більшою, ніж кількість підтверджених випадків.

Діаграма 1.11 демонструє показник смертельних випадків від COVID-19 за ковзним середнім. Коефіцієнт летальності (CFR) — це співвідношення між підтвердженими смертями та підтвердженими випадками. Дане ковзне середнє CFR розраховується як відношення між середньою кількістю смертей за 7 днів і середньою кількістю випадків за 7 днів 10 днів раніше.

Отже, проаналізувавши статистичні дані, видно, що є велика необхідність у передбаченні розвитку коронавірусної хвороби в Україні, у першу чергу, для попередження розвитку подібних вірусів у майбутньому.

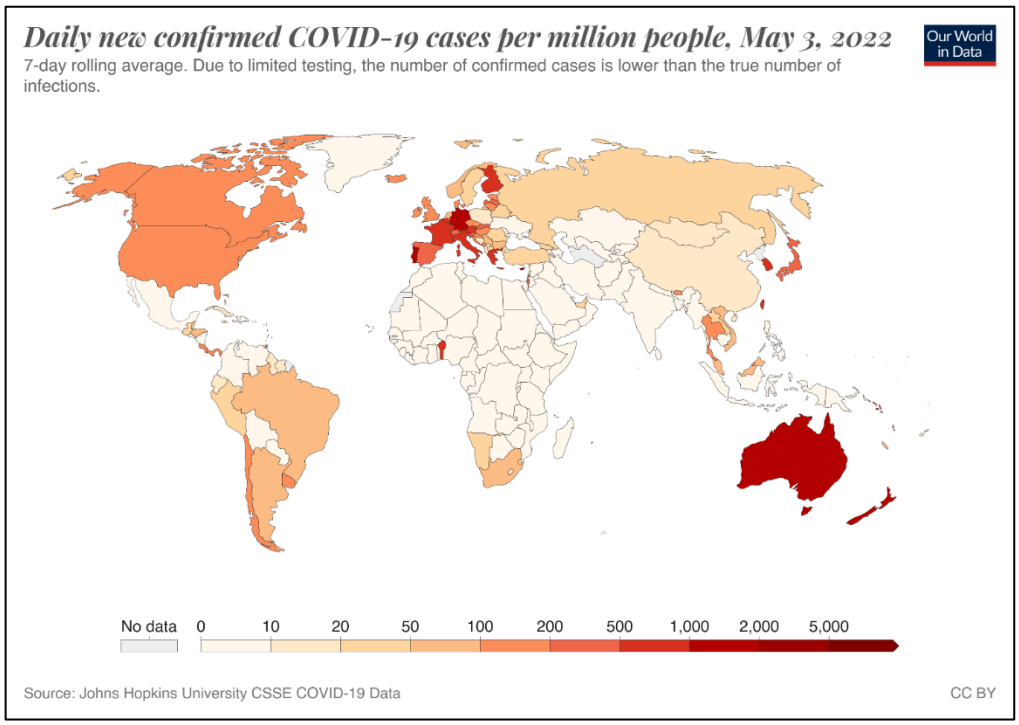

Рисунок 1.9 Щоденна кількість доз вакцини проти COVID-19, що вводиться щодня [8]

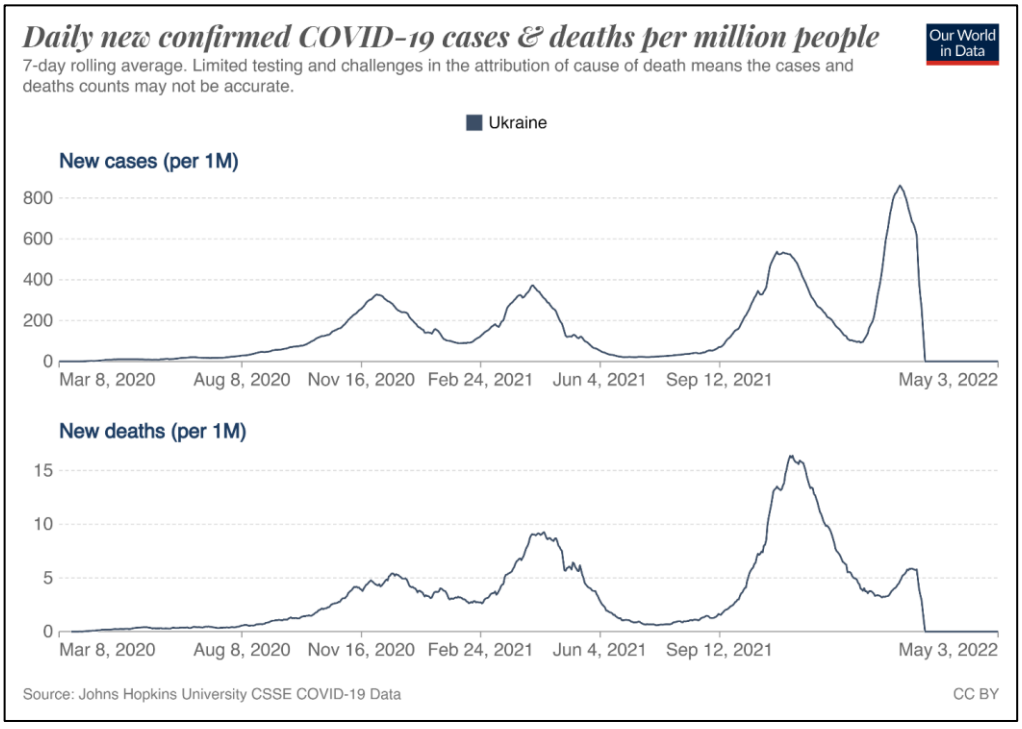

Рисунок 1.10 Щоденна кількість нових випадків та летальних випадків [8]

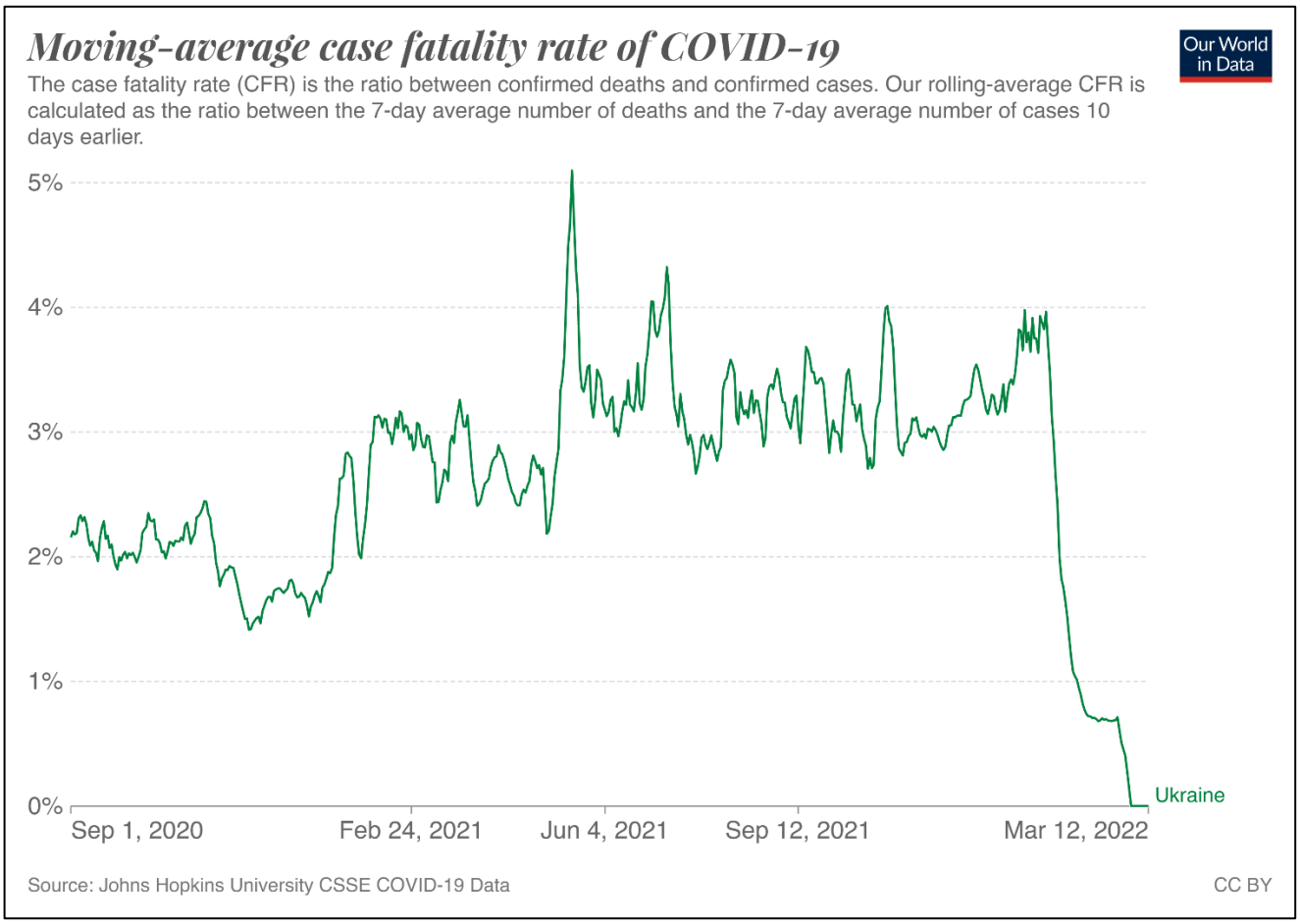

Рисунок 1.11 Показник смертельних випадків від COVID-19 (ковзне середнє) [8]

Для вибору архітектури нейронної мережі для вирішення задачі прогнозування розвитку COVID-19 в Україні далі наведено короткий аналіз відомих досліджень у цій галузі.

1.2. Аналіз відомих рішень в області медичного прогнозування

Вивченням застосування нейронних мереж з метою прогнозування в медичній сфері в своїх роботах розглядали такі вчені як А.Г. Хасанов в роботі [9], М. А. Кондратьев [10], KhondokerNazmoonNabi [11], Rahele Kafieh [12], K.E.ArunKumar [13] та багато інших.

У роботі [11] для прогнозування розвитку епідемії коронавірусу в Бразилії, Росії та Сполученому королівстві використовувалося чотири моделі глибокого навчання: закритий рекурентний блок (GRU), згорткова нейронна мережа (CNN), довгострокова короткострокова пам'ять (LSTM) і багатовимірна згорткова нейронна мережа (MCNN).

У дослідженні виявлено, що CNN може забезпечити надійні результати довгострокового прогнозування в аналізі часових рядів завдяки своїй здатності до навчання основних функцій, інваріантності спотворень і навчання тимчасової залежності.

LSTM добре себе зарекомендувала при горизонті прогнозування всього три дні. Це говорить про те, що архітектура LSTM не дає істотних відомостей про довгострокове прогнозування часових рядів з достатньою точністю. З іншого боку, алгоритм LSTM добре працює з великим набором даних, коли потрібно знайти сезонність при аналізі часових рядів. Однак при відносно меншій кількості точок даних він не може знайти ніякої закономірності.У роботі [12] наведено дослідження прогнозування спалаху COVID-19 в дев'яти країнах (Іран, Німеччина, Італія, Японія, Корея, Швейцарія, Іспанія, Китай і США).

Найкраща продуктивність була виявлена у модифікованої версії LSTM, званої M-LSTM, для прогнозування майбутньої траєкторії пандемії в згаданих країнах. Для цього було зібрано дані з 22 січня по 30 липня 2020 роки для навчання і з 1 серпня 2020 року по 31 серпня 2020 роки для фази тестування.

LSTM була порівняна з багатошаровим перцептроном (MLP), а також з Random Forest, що показали гірші результати.

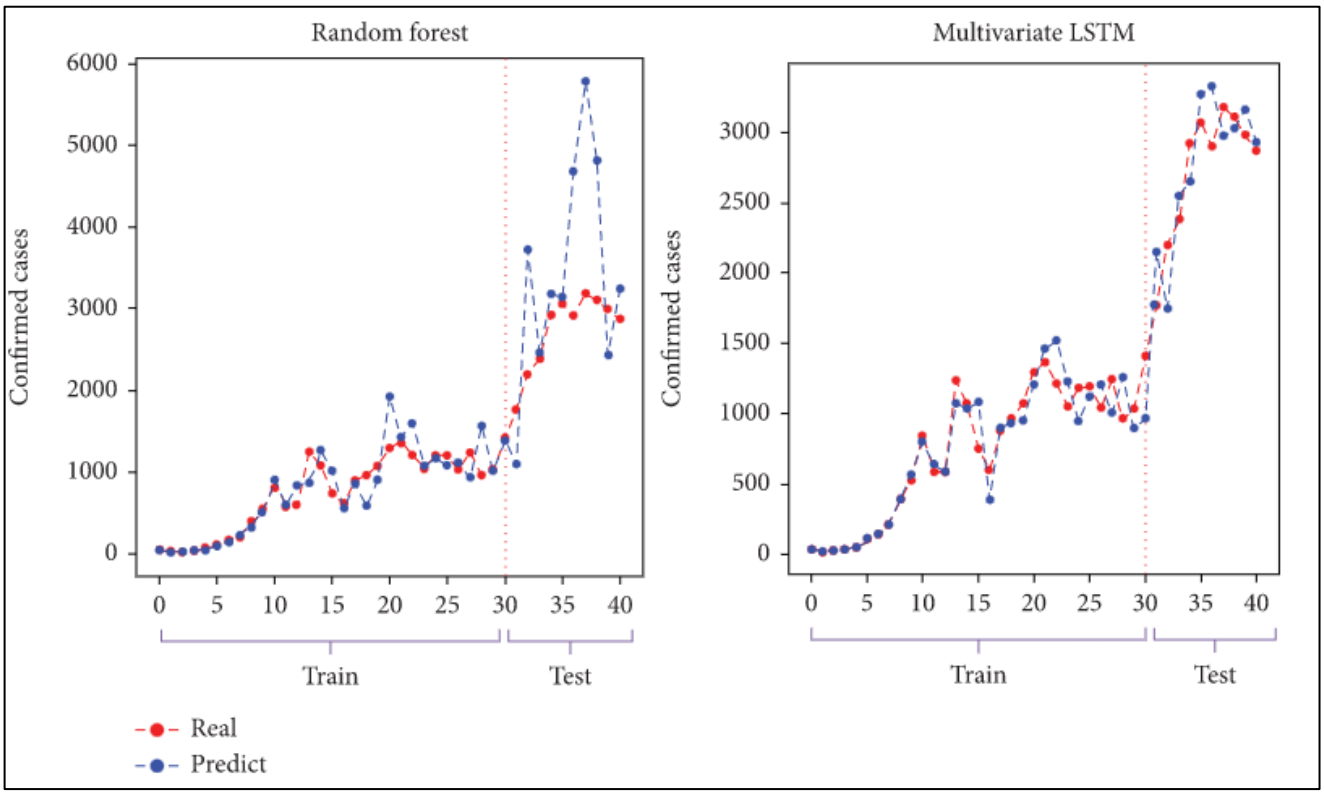

Рисунок 1.12 Порівняння найгірших і найкращих моделей у правильному дотриманні навчального набору та точному прогнозуванні значень тесту

У роботі [13] наведено 60-денний прогноз випадків COVID-19 і їх тенденцій для топ-10 країн, а також наведено порівняння моделей глибокого навчання RNN на основі LSTM і GRU.

Для аналізу використано загальнодоступні дані з бази даних COVID-19 Університету Джона Хопкінса. Також у роботі враховано важливість різних факторів, таких як вік, профілактичні заходи та медичні установи, щільність населення і т. д., які відіграють життєво важливу роль в швидкому поширенні пандемії COVID-19.

Що стосується прогнозу даних про загиблих, модель LSTM перевершила модель GRU для Бразилії, Індії, Росії, Південної Африки, Мексики і Ірану.

Деякі дослідження проводились і на території України. У дослідженні [14] науковці КПІ та Інституту епідеміології та інфекційних хвороб ім. Л.В.

Громашевського Національна академія медичних наук України відзначили, що наступна хвиля COVID-19, пов'язана з появою нового штаму Omicron, стрімкими темпами розповсюджується у світі. Її рух спосперігається переважно у напрямку Східної Європи, в тому числі і у напрямку України.

У порівнянні з штамом Delta передання вірусу Omicron відбувається значно швидше, ураховуючи останні дані щодо різкого зростання кількості заіксованих випадків у Західній Європі і Сполучених Штатах Америки (станом на 29.12.2021 маємо: Велика Британія — 183 тис., Франція — 208 тис., Іспанія — 100,7 тис., Італія — 98 тис., США — 465,5 тис. випадків). Проте варто відмітити, що кількість щоденних летальних випадків у наведених вище країнах значно зменшилася (відповідно 57, 184, 78, 136 і 1777 випадків). Можна припустити, що й на території України перебіг нової хвилі COVID-19 буде легшим, ніж попередні.

Аналізуючи попередню хвилю хвороби, вчені вказують, що найвищий рівень згладженої за тиждень кривої нових випадків COVID-19, викликаних Delta-штамом, припав в Україні на 27.10.2021—2.11.2021 і становив 23 333 осіб. Загалом у ті дні хворіло близько 509 тис. осіб. Коли рівень повністю вакцинованих осіб досяг 20,2% (15.11.2021р.), хвиля захворюваності почала спадати, а під кінець 2021 року середнє значення захворюваності за тиждень вже становило 4,5 тис. випадків на добу. Наразі повністю вакциновано 2 дозами 31,2% громадян України, що передбачає спадну тенденцію і далі. До 2022 року в січні цього року щоденна кількість нових випадків зараження становила 3-5 тисяч, досягнувши найнижчого рівня.

Крім того, в Україні почав циркулювати ще один варіант – вірус SARS-CoV-2 та штам вірусу грипу H3N2 Omicron. Це вплине на захворюваність наступною хвилею, що вимагає диференційної діагностики між COVID-19 та грипом. Щоб перевірити цю гіпотезу, були зроблені прогнози на 100 днів вперед (довгострокові) на основі використання нейронних мереж, таких як «прямі персептрони». За

отриманими результатами можна вважати, що протягом січня-лютого 2022 року може початися поширення нового штаму Omicron, що спричинить нову хвилю захворюваності. Ця хвиля, за даними прогнозу, досягне свого піку в середині лютого 2022 року, а кількість інфікованих на добу становитиме 23—25 тисячі. В березні або квітні 2022 року можна очікувати її спад. Проте слід враховувати той факт, що дані офіційної захворюваності та смертності у більшій мірі залежатимуть від обсягу лабораторної діагностики, а тривалість наступної хвилі від темпів вакцинації та обсягу протиепідемічних заходів.

У роботі [15] алгоритм машинного навчання без учителя використовується для прогнозування майбутніх випадків COVID-19 з точки зору підтверджених випадків, випадків одужання та летальних виадків. Для цього дослідження використовуються моделі регресії лінійної регресії ((LR), поліноміальної регресії (PR), регресії Ріджа і Лассо (RLR), регресії Риджа (RR) і опорного вектора (SVM).

У процесі роботи були проведені комплексні експерименти, і кожен тест оцінювався за параметрами середньої квадратичної помилки (MSE), середньої абсолютної похибки (MAE), середньої квадратичної помилки (RMSE).

Результати представлені в таблиці 1.2. Результати моделей лінійної та регресії Ріджа дуже схожі один на одного. Модель регресії SVM показала найкращі результати.

Таблиця 1.2 Результати дослідження [15]

Назва моделі	**MSE**	**MAE**	**RMSE**
Linear Regression	3230371601512722.5	56457125.4	56836358.0
Polynomial Regression	1795330053135026.2	37454555.9	42371335.2
Ridge	3233647448477261.5	56485978.2	56865169.0
Polynomial ridge	737383324668728.4	24617677.8	27154802.9
SVM	3645218671704054.5	56956658.6	60375646.3

Отже, проаналізувавши наведену інформацію, було прийнято рішення для прогнозування розвитку епідемії COVID-19 в Україні використовувати модель LSTM, яка є однією з найкращих для короткострокових та середньострокових прогнозів і активно використовується для прогнозування часових рядів.

1.3. Постановка задачі

Завданням даної роботи є розробка нейромережевої технології прогнозування розвитку епідемії COVID-19 на території України на основі статистичних даних 2020-2022 років.

Для досягнення даної мети були поставлені наступні задачі:

- розглянути існуючі нейромережеві алгоритми в контексті застосування до завдань прогнозування;
- знайти і підготувати набір даних, необхідний для навчання алгоритму;
- використовуючи комбінації та покращення існуючих алгоритмів реалізувати алгоритм прогнозування розвитку епідемії;
- провести тестування та оцінити точність отриманого прогнозу;
- проаналізувати отримані результати.

Вхідна інформація:

- текстовий набір даних, створений у середовищі MS Excel на базі офіційної статистики захворюваності на COVID-19 на території України;

Вихідною інформацією є набір прогнозованих значень на певний період.

Висновки до розділу 1

У даному розділі проаналізовано актуальність досліджуваної теми, розглянуто вже відомі дослідження у цій галузі, основні проблеми, що виникають під час її розв’язання, описано основні задачі дослідження та обрано модель, за допомогою якої буде вирішуватись задача прогнозування.

РОЗДІЛ 2 ТЕОРЕТИЧНЕ ТА МЕТОДИЧНЕ ДОСЛІДЖЕННЯ ТЕМИ

2.1. Особливості задачі прогнозування

Наведемо деякі основні поняття з прогностики – науки, предметом вивчення якої є принципи побудови прогнозів.

Ось кілька основних понять прогнозування - науки, предметом якої є принципи прогнозування.

Насправді, формалізувати таке поняття, як прогнозування, важко, оскільки прогнозування є досить новою науковою дисципліною, яка остаточно сформувалася наприкінці XX століття [16]. Одне з найточніших визначень наразі:

Прогноз – це імовірнісне судження про стан об’єкта (процесу чи явища) у певний момент у майбутньому та/або альтернативний метод досягнення будь-якого результату.

Теорія оцінювання дає дещо інше визначення: передбачення — це умовне математичне очікування функції, яке дає змогу обчислити майбутнє значення цільової (первинної) змінної.

Надалі ми розглядатимемо прогнози лише з наукової точки зору, оскільки всі припущення мають бути науково обґрунтованими. Зрозуміло, що наведене визначення є досить широким і тому вимагає певної класифікації методів і методів прогнозування. По-перше, необхідно розрізняти кількісні та якісні методи прогнозування. Перший ґрунтується на математичних моделях та статистичних методах, а другий — на дослідженнях аналітиків предметної галузі та формуванні певних експертних оцінок.

До першої категорії належать, наприклад, моделі регресії, моделі ковзних середніх, нейронні мережі, до другої — метод Delphi, морфологічний аналіз тощо. Принципова відмінність між цими методами полягає в тому, що перша група методів передбачає вивчення поведінки процесу в минулому і передбачає, що певна

закономірність процесу триватиме, а друга група методів передбачає врахування значного впливу змін якості процесу. У цій роботі ми обмежимося вивченням кількісних методів, які ми будемо мати на увазі, якщо прямо не вказано інше.

Важливим параметром, який може класифікувати прогнози, є горизонт прогнозу - період часу, протягом якого робляться прогнози. Різні джерела надають різні поділки класів, наприклад, за діапазоном передбачення можна виділити наступні типи [17]:

а) короткострокові - до 2 місяців;

б) середньострокові - від 2 місяців до 6 місяців;

в) довгострокові - шість місяців і більше.

Тому для оперативного управління та виявлення спалахів епідемії використовуються короткострокові прогнози на найближчі тижні.

Найкориснішими можна вважати середньострокові прогнози від двох до шести місяців для тактичного управління. Звичайно, це не так точно, як короткостроковий період, але залишає достатньо часу для можливих надзвичайних ситуацій та вжиття запобіжних заходів. Приймаючи стратегічні рішення, не залишайтеся без довгострокових прогнозів на наступний рік і далі. Досягти якісних прогнозів у більшості випадків неможливо, але вони необхідні при оцінці виробництва необхідних ліків і вакцин, медичних закладів та підготовки кадрів.

Час відіграє особливо важливу роль у проблемах прогнозування; щоб проаналізувати динаміку процесу, нам потрібні не тільки розміри певного набору змінних, але й тимчасове впорядкування цих вимірів. Тому для опису вимірювання значень змінних використовуються часові ряди.

Часовий ряд – набір послідовних вимірів значень певної змінної, як правило, часові інтервали між вимірами є однаковими. Окремі спостереження у часовому ряді також називають рівнями цього ряду. Таким чином фактично часовий ряд складається з двох компонент – час та рівень ряду в цей момент часу. Слід

пам’ятати, що між різними вимірами ряду існує статистична, або ж іншими словами кореляційна, залежність. В аналізі часових рядів виділяють такі основні компоненти:

а) тренд – це фактично середнє значення або математичне сподівання процесу в даний момент часу, тобто тренд фактично відображає тенденції змін значень показника. Тренди можуть бути як детермінованими, тобто описуватись рівняннями різних видів, наприклад, лінійними або кубічними, чи бути випадковими, які описують комбінаціями випадкових процесів. Детермінований тренд відображає більш довгострокові тенденції у порівнянні з випадковим. 13 За означенням з теорії оцінювання тренд: умовне математичне сподівання процесу, які вказує на його довгострокові зміни;

б) випадкова компонента – це випадкові коливання значень вимірів змінної, що накладається на тренд, якщо він існує, та зумовлені різними факторами, наприклад, неточність вимірів, певні зовнішні впливи, окремо врахувати всі ці фактори практично неможливо. Можлива також класифікація цих компонент на поточні та раптового впливу, поточні відповідно враховують безліч факторів, кожен з яких не має сильного впливу на процес, але їх сукупний вплив необхідно врахувати, фактори раптового впливу враховують випадкові сильні збурення, наприклад, стихійні лиха;

в) сезонність – це компонента, що пояснює певні періодичні коливання процесу (період не надто великий, менше року), що часто пов’язані зі змінами днів тижня, місяців, пір року тощо. Наприклад, якщо часовий ряд описує динаміку зміни кількості респіраторних захворювань, то з великою ймовірністю ця кількість буде більшою в зимові пори року;

г) циклічність – компонента, що описує періодичні коливання процесу, період яких є досить великим (більше року);

д) нелінійність – це компонента, що визначає характер поведінки процесу і відповідно часового ряду, нелінійний процес не можна досить точно описати лінійними рівняннями, приклад нелінійного процесу – процес зі змінною дисперсією або асиметричними циклами [17].

2.2. Огляд методів прогнозування

Прогнозування розвитку епідемії COVID-19 дасть змогу оцінити кількість необхідних тестів для виявлення хвороби в майбутньому, кількість вакцин, необхідну кількість лікарняних ліжок та обсяг місць в спеціальних відділеннях лікарень, оцінити кількість необхідних апаратів штучної вентиляції легень для кожної лікарні. Тобто завдяки прогнозуванню є можливість завчасно підготуватися і зменшити кількість летальних випадків. Також прогноз розвитку одного вірусу дасть змогу застосувати отримані навички для прогнозування подібних явищ в майбутньому.

Через те, що прогнозування набуло актуальності в різних сферах людської діяльності, було розроблено велику кількість різних методів побудови прогнозів. На сьогодні активно досліджуються та модифікуються існуючі методи, розробляються нові, тому не існує однозначної класифікації методів прогнозування. В даній роботі наведемо таку схему класифікації [18]:

- формалізовані та інтуїтивні методи;

Інтуїтивні методи не передбачають побудову моделей та базуються на експертних оцінках, досвіді, інтуїції. Зокрема подібний підхід застосовується при вирішенні задач, які складно або неможливо формалізувати та побудувати адекватну модель. Формалізовані методи передбачають побудову моделі процесу. В свою чергу ці методи можна класифікувати за видами моделей, які використовуються для побудови прогнозів.

- статистичні моделі та структурні моделі;

Статистичні моделі прогнозування передбачають встановлення аналітичної залежності між теперішніми та минулими рівнями часового ряду та майбутніми, які ми хочемо спрогнозувати. До цього класу входять такі групи методів:

- регресійні моделі;
- авторегресійні моделі;
- методи експоненціального згладжування.

Структурні моделі передбачають побудови певних структур для побудови прогнозу, наприклад вибір кількості шарів та штучних нейронів у нейронній мережі, до цієї категорії відносяться методи машинного навчання:

Решту цього підрозділу присвятимо короткому огляду основних методів цих категорій.

- регресійні моделі:
    а)моделі лінійної регресії;
    б)моделі нелінійної регресії;
    в)моделі множинної регресії;
    г) метод групового врахування аргументів (МГУА);
- авторегресійні моделі:
    а)моделі авторегресії (АР);
    б)моделі ковзного середнього (КС);
    в)моделі авторегресії та ковзного середнього (АРКС);
    г)інтегровані моделі авторегресії та ковзного середнього(АРІКС);
    д)авторегресійні/узагальнені авторегресійні умовно гетероскедастичні моделі (АРУГ/УАРУГ);
- моделі експоненційного згладжування:
    а)просте експоненційне згладжування(експоненційне згладжування першого поряку);
    б)експоненційне згладжування другого порядку;

в)експоненційне згладжування третього порядку;

- структурні моделі:

а)моделі на основі нейронних мереж:

- LSTM мережа;

- рекурентна нейронна мережа (RNN);

б)моделі на основі ланцюгів Маркова;

в)моделі опорних векторів;

г) моделі на основі класифікаційно-регресійних дерев;

д)генетичні алгоритми.

Перераховані вище групи методів пропонують різні підходи та застосовуються в залежності від специфіки задачі та властивостей процесу. Однак можна виділити деякі спільні недоліки та переваги різних класів методів, основні з них для найбільш популярних груп методів наведемо у таблиці 2.1.

Слід зауважити, що досить важко відповісти на питання, методи якої категорії або який конкретний метод дозволить будувати більш точні прогнози, оскільки ефективність прогнозування залежить не лише від специфіки розвитку процесу, поставленої задачі та вибору методу, а й від багатьох інших факторів, зокрема від навичок та досвіду фахівця, достовірності та обсягу даних, помилок, допущених на будь-якому етапі вирішення задачі. Також важливими є питання, які в нас є обмеження на обчислювальні ресурси, скільки часу ми можемо витрати на побудову прогнозів, що зобов'язує нас знаходити раціональний компроміс.

Таблиця 2.1 Порівняння різних класів методів прогнозування [17]

Клас моделей	**Переваги**	**Недоліки**
Регресійні та авторегресійні моделі	Простота реалізації моделей, можливість аналізу проміжних обчислень, швидка побудова прогнозу на основі розробленої моделі, гнучкість.	Трудомісткість обчислень для ідентифікації параметрів. Необхідність аналізу процесу, вибору моделей з множини альтернативних. Низька ефективність прогнозування нелінійних процесів
Моделі експоненційного згладжування	Простота реалізації, ефективність побудови довгострокових прогнозів	Негнучкість, ефект затримки прогнозу при тендеціях зростання або спадання значень вибірки Складність ідентифікації параметрів
Моделі на основі нейромереж	Висока адаптивність, нелінійність моделей, можливість застосування без поглибленого аналізу природи процесу.	Скадність вибору структури моделі, методів навчання. Чутливість до якості даних. Необхідінсть ідентифікації великої кількості параметрів Високі вимоги до обчислювальних потужностей, навчання мережі часто потребує багато часу.

2.3. Аналіз архітектури моделі LSTM

Ми обрали модель LSTM у попередньому розділі для вирішення поставленої задачі, тому далі буде розглянуто детально теоретичні засади цієї архітектури.

Рекурентні нейронні мережі (рис. 2.1) вирішують цю проблему. Вони представляють мережі з циклами, що дозволяють зберігати інформацію.

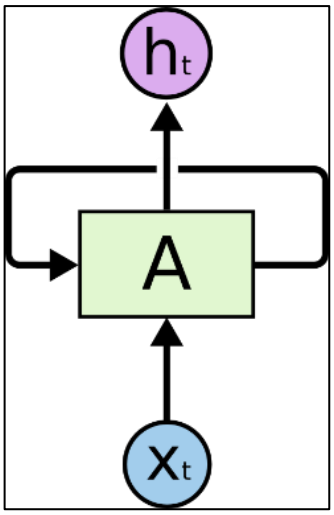

Рисунок 2.1 Спрощена модель рекурентної нейронної мережі

На наведеній вище схемі фрагмент нейронної мережі, A, приймає на вхід значення x_t і повертає значення h_t. Цикл дозволяє передавати інформацію від одного кроку мережі до наступного.

Рекурентну нейронну мережу можна розглядати як кілька копій однієї і тієї ж мережі, кожна з яких передає повідомлення наступнику.

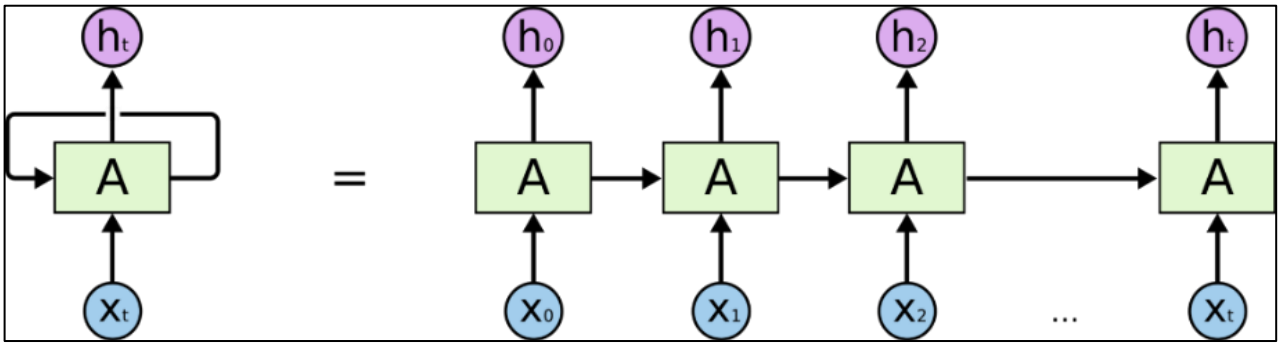

Рисунок 2.2 Розгорнута модель рекурентної нейронної мережі

За останні кілька років був неймовірний успіх у застосуванні RNN для різноманітних задач: розпізнавання мови, моделювання мови, переклад, субтитри до зображень , тощо.

Значну роль відіграє «LSTM» - одна з модифікацій рекурентної нейронної мережі, яка на багатьох задачах значно перевершує стандартну версію.

Одна з ідей RNN полягає в тому, що вони потенційно вміють пов'язувати попередню інформацію з поточним завданням, так, наприклад, знання про попередній кадр відео можуть допомогти в розумінні поточного кадру. Якби RNN мали таку здатність, вони були б надзвичайно корисні.

Бувають випадки, коли для вирішення поточного завдання ми потребуємо лише недавню інформацію. Але також є випадки, коли нам необхідно більше контексту, коли розрив між актуальною інформацією і точкою її застосування стає дуже великим. На жаль, у міру зростання цієї відстані, RNN втрачають здатність зв'язувати інформацію [13].

Довга короткострокова пам'ять (LSTM) - особливий різновид архітектури рекурентних нейронних мереж, здатна до навчання довготривалим залежностям. Вони були представлені Зеппом Хохрайтер і Юргеном Шмідхубер в 1997 році, а потім вдосконалені і популярно викладені в роботах багатьох інших дослідників. Вони прекрасно вирішують цілий ряд різноманітних завдань і в даний час широко використовуються [14].

LSTM розроблені спеціально, щоб уникнути проблеми довготривалої залежності.

Будь-яка рекурентна нейронна мережа має форму ланцюжка повторюваних модулів нейронної мережі. У класичній RNN структура такого модуля максимально спрощена, наприклад, він може представляти собою один шар з функцією активації tanh - гіперболічний тангенс.

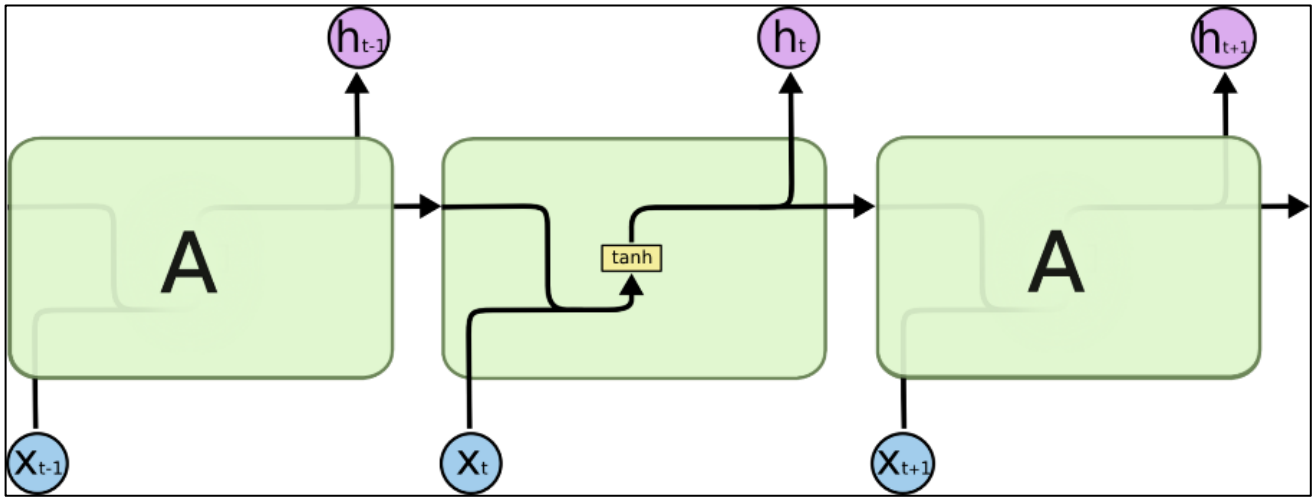

Рисунок 2.3 Повторюваний модуль в стандартній RNN

Структура LSTM також нагадує ланцюжок, але модулі виглядають інакше. Ч отири шари нейронної мережі взаємодіють особливим чином, натомість стандартн а RNN містить лише один шар.

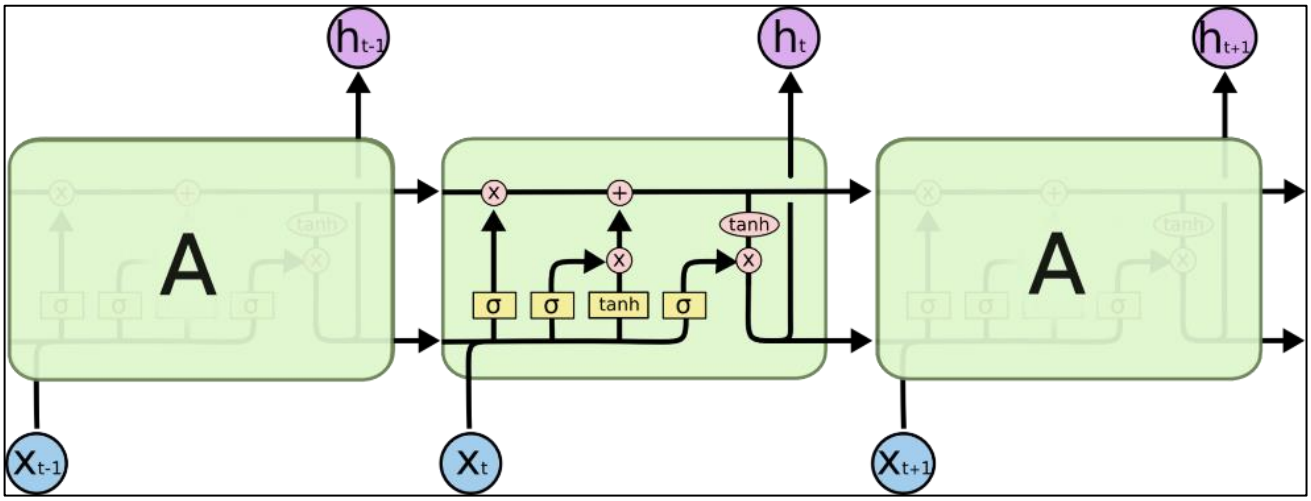

Рисунок 2.4 Повторюваний модуль в LSTM мережі

Ключовий компонент LSTM - це стан комірки (cell state) - горизонтальна лінія, що проходить по верхній частині схеми.

Стан комірки нагадує конвеєрну стрічку. Вона проходить безпосередньо через весь ланцюжок, і бере участь тільки в декількох лінійних перетвореннях. Інформація може легко текти по ній, не піддаючись змінам.

Проте, LSTM може видаляти інформацію зі стану комірки; цей процес регулюється структурами, так званими фільтрами (gates).

Фільтри дозволяють пропускати інформацію на підставі деяких умов. Вони складаються з шару сигмоїдальної нейронної мережі і операції поточкового множення.

Сигмоїдальний шар повертає числа на проміжку від нуля до одиниці, які позначають, яку частку кожного блоку інформації, яку можна пропустити далі по мережі. Нуль відповідає значенню «не пропускати нічого», одиниця – «пропустити все».

У LSTM три таких фільтри, що дозволяють захищати і контролювати стан комірки.

Перший крок в LSTM - визначити, яку інформацію можна викинути зі стану комірки. Це рішення приймає сигмоїдальний шар, названий «шаром фільтра забування» (forget gate layer). Він дивиться на h_{t-1} і повертає число x_t від 0 до 1 для кожного числа зі стану комірки C_{t-1}. Одиниця означає «повністю зберегти», а 0 – «повністю викинути».

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f \tag{1}$$

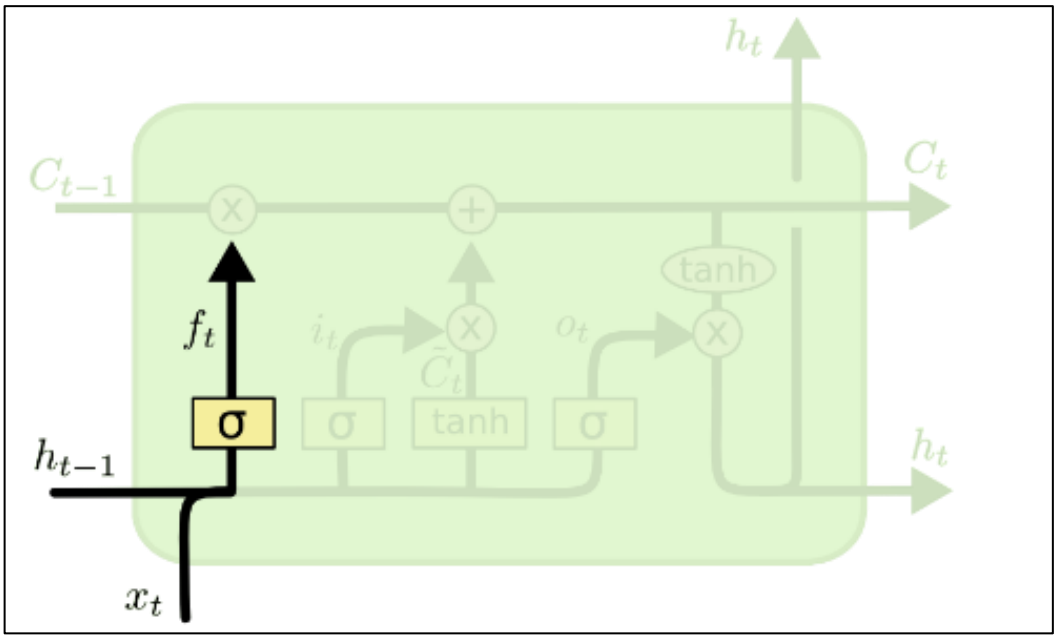

Рисунок 2.5 Forget gate layer в LSTM мережі

Наступний крок - вирішити, яка нова інформація буде зберігатися в стані комірки. Цей етап складається з двох частин. Спочатку сигмоїдальний шар під назвою «шар вхідного фільтра» (input layer gate) визначає, які значення слід оновити. Потім tanh-шар будує вектор нових значень-кандидатів C_t, які можна додати в стан комірки.

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i \tag{2}$$

$$C_t = tanh(W_C \cdot [h_{t-1}, x_t] + b_C \quad (3)$$

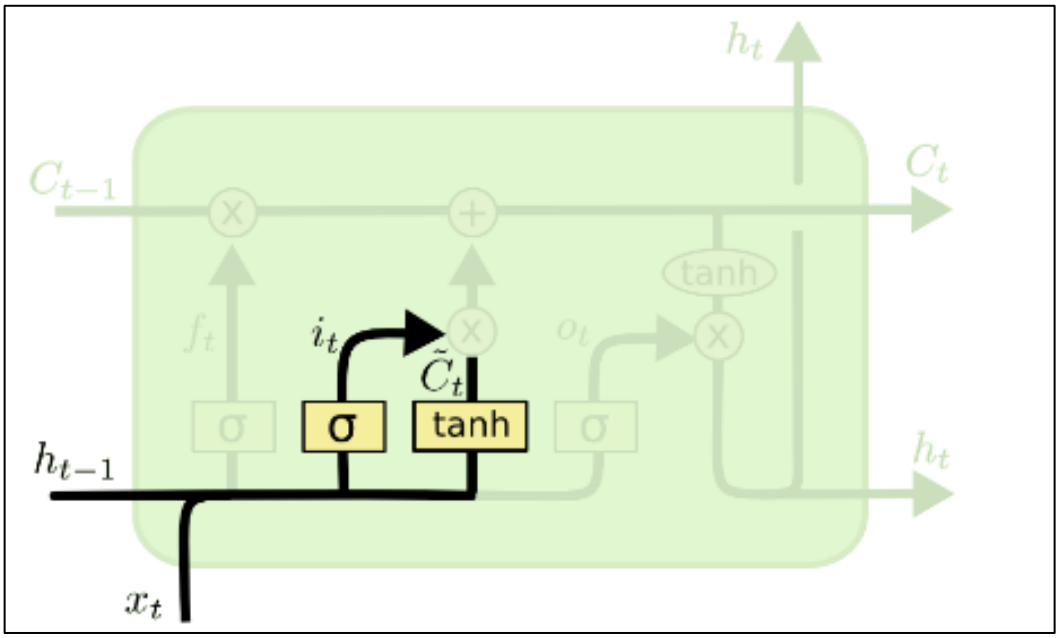

Рисунок 2.5 Збереження інформації в LSTM мережі

Настав час замінити старий стан комірки C_{t-1} на новий стан C_t.

Ми множимо старий стан на f_t, забуваючи те, що ми вирішили забути. Потім додаємо $i_t \cdot C_t$. Це нові значення-кандидати, помножені на t - на скільки ми хочемо оновити кожне з значень стану.

$$C_t = f_t \times C_{t-1} + i_t \times C_t \quad (4)$$

Отже, сновна відмінність між RNN і LSTM полягає в тому, що зберігати інформацію в пам'яті протягом тривалого періоду часу. LSTM може обробляти інформацію в пам'яті протягом тривалого періоду часу, завдяки цій особливості вона має значну перевагу порівняно з RNN. Але питання полягає в тому, чим LSTM відрізняється від RNN, завдяки чому LSTM здатні підтримувати довгострокові тимчасові залежності.

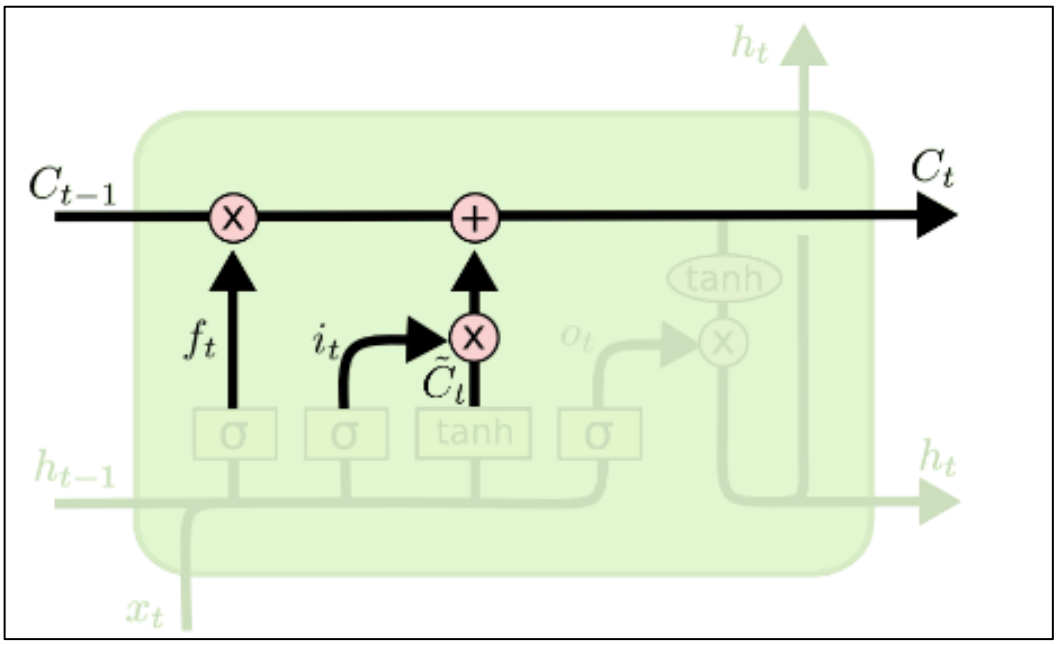

Рисунок 2.6 Зміна стану в LSTM мережі

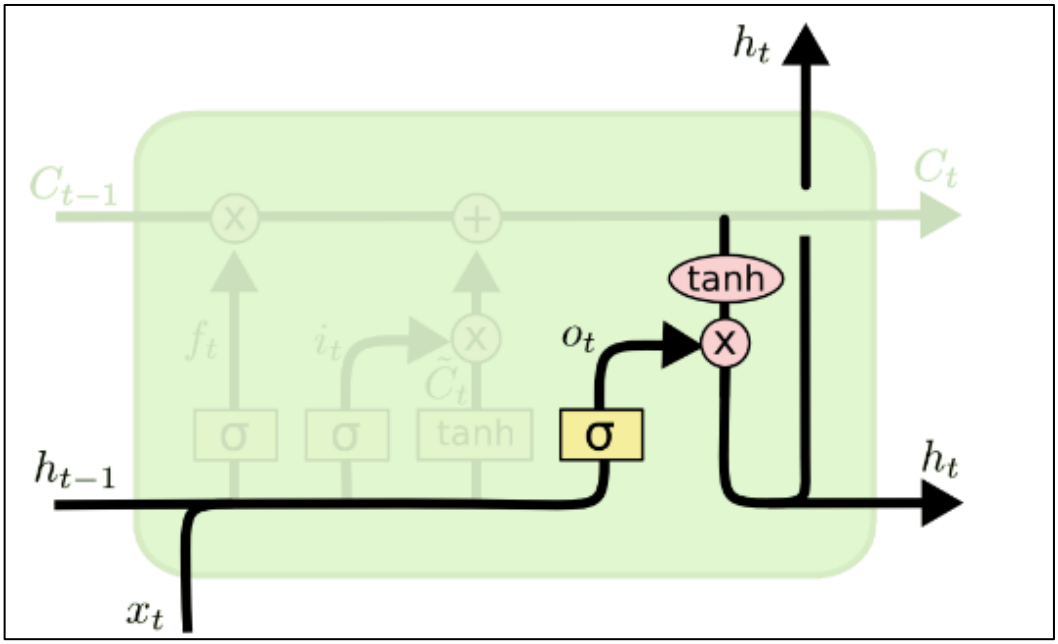

Рисунок 2.7 Отримання вихідної інформації в LSTM мережі

RNN мають петлі зворотного зв'язку в повторюваному шарі, завдяки цьому вони здатні зберігати інформацію в «пам’яті» протягом тривалого часу. Але може бути важко навчити стандартні RNN вирішувати проблеми, які вимагають вивчення

довгострокових тимчасових залежностей. Це пов’язано з тим, що градієнт функції втрат експоненціально має тенденцію спадати з часом (називається проблемою градієнта зникнення).

LSTM справляються з проблемою градієнта, вводячи нові фільтри, які дозволяють тримати під контролем градієнтний потік і забезпечують краще збереження «довгострокових залежностей».

2.4. Методичне забезпечення організації проведення досліджень

2.4.1. Проектування абстрактної моделі прогнозування

Для розуміння етапів створення моделі прогнозування та їх послідовності була розроблена діаграма діяльностей за допомогою уніфікованої мови моделювання.

Уніфікована мова моделювання (UML), що є графічною мовою для візуалізації, специфікації, документування та проектування систем. UML дозволяє розробити детальну модель створюваної інтелектуальної системи. UML є мовою широкого профілю, це відкритий стандарт, що використовує графічні позначення для створення абстрактної моделі системи.

У рамках UML-моделі всі основні спеціальні графічні конструкції представляються у вигляді діаграм. Далі розглянемо один із видів діаграми поведінки.

Діаграма діяльності - візуальне представлення графу діяльностей в UML, де вершини графу – це певні дії, які поєднані низкою переходів.

На рисунку 2.8 зображено діаграму діяльності «Навчання LSTM мережі», де представлені етапи завантаження датасету, попередньої обробки даних, утворення навчальної і тестової вибірок та безпосередньо процеси навчання та тестування мережі.

Рисунок 2.8 Діаграма діяльностей «Навчання LSTM мережі»

2.4.2. Опис навчального набору даних

На підготовчому етапі було обрано навчальний набір даних та проведена попередня обробка даних.

У якості навчального набору даних використана статистика розвитку епідемії COVID-19 в Україні в період з 3.03.2020 по 7.05.2022. В обраному датасеті представлена загальносвітова статистика, що включає наступні дані (таблиця 2.2):

Таблиця 2.2 Дані, що представлені у датасеті [8]

Критерій	**Джерело**	**Оновлення**	**Країни**
Вакцинації	Офіційні дані зібрано командою «Our World in Data»	Щоденно	218
Тести	Офіційні дані зібрано командою «Our World in Data»	Щотижнево	187
Госпіталізації та госпіталізації у відділення інтенсивної терапії	Офіційні дані зібрано командою «Our World in Data»	Щоденно	47
Підтверджені випадки	JHU CSSE COVID-19 Data	Щоденно	216
Підтверджені летальні випадки	JHU CSSE COVID-19 Data	Щоденно	216
Швидкість розмноження	Arroyo-Marioli F, Bullano F, Kucinskas S, Rondón-Moreno C	Щоденно	192
Відповідь політики	Оксфордський інструмент відстеження реагування уряду на COVID-19	Щоденно	187
Інше	Міжнародні організації (ООН, Світовий банк, ОЕСР, IHME…)	Фіксовано	241

Підтверджені випадки та летальні випадки: дані надходять із сховища даних COVID-19 Центру системних наук та інженерії (CSSE) Університету Джона Хопкінса (JHU). Набір даних про випадки і смертність оновлюється щодня.

Примітка: підтверджені випадки та летальні випадки збираються Університетом Джона Хопкінса за датою звіту, а не за датою тесту/смерті. Таким чином, кількість, яку вони повідомляють у певний день, не обов'язково відповідає дійсному числу на цю дату через довгий ланцюжок звітності, який існує між новим випадком/смертю та його включенням до статистики. Це також означає, що часові ряди можуть показувати раптові зміни (негативні чи позитивні), коли країна виправляє історичні дані, оскільки раніше вона занижувала чи завищувала кількість випадків/смертей.

Госпіталізації та госпіталізації у відділення інтенсивної терапії (ICU): наші дані зібрані з офіційних джерел і зіставлені Our World in Data.

Тестування на COVID-19 та вакцинація проти COVID-19: ці дані збирає команда «Our World in Data» з офіційних звітів.

Інші змінні: ці дані зібрані з різних джерел (Організація Об'єднаних Націй, Світовий банк тощо).

Дані зберігаються в форматі .xlsx (рис. 2.10). Дані необхідно попередньо відфільтрувати та привести всі значення до потрібного формату.

Фільтр - це один з інструментів MS Excel, що допомагає відобразити на екрані лише обрані записи, які задовольняють певній умові. У нашому випадку в датасеті необхідно відфільтрувати записи, що стосуються саме України. Важливо, щоб кожному стовпчику відповідав заголовок, тому це слід перевірити завчасно. Фільтрація за простими критеріями реалізується за допомогою автофільтру.

Виділивши всі дані таблиці, включно з заголовками стовпців, виконується команда Дані / Фільтр / Автофільтр. Біля заголовка кожного з стовпців з'явиться стрілка.

Усі порожні значення в таблиці було замінено на значення 0.

Детальний опис критеріїв датасету, що використовується у роботі, представлений у таблиці 2.3.

Таблиця 2.2 Опис критеріїв, що представлені у датасеті [8]

Критерій	**Опис**
total_cases	Загальна кількість підтверджених випадків COVID-19. Підрахунок може включати ймовірні випадки, якщо повідомлялося.
new_cases	Нові підтверджені випадки COVID-19. Підрахунок може включати ймовірні випадки, якщо повідомлялося. У рідкісних випадках, коли джерело повідомляє про негативну щоденну зміну через виправлення даних, значення цього показника NA.
new_cases_smoothed	Нові підтверджені випадки COVID-19 (згладжено за 7 днів). Підрахунок може включати ймовірні випадки, якщо повідомлялося.
total_cases_per_million	Загальна кількість підтверджених випадків COVID-19 на 1 000 000 осіб. Підрахунок може включати ймовірні випадки, якщо повідомлялося.
new_cases_per_million	Нові підтверджені випадки COVID-19 на 1 000 000 осіб. Підрахунок може включати ймовірні випадки, якщо повідомлялося.

new_cases_smoothed_per_million	Нові підтверджені випадки COVID-19 (згладжено 7 днів) на 1 000 000 осіб. Підрахунок може включати ймовірні випадки, якщо повідомлялося.
total_deaths	Загальна кількість смертей від COVID-19. Підрахунок може включати ймовірні смерті, якщо повідомлялося.
new_deaths	Нові випадки смерті від COVID-19. Підрахунок може включати ймовірні смерті, якщо повідомлялося. У рідкісних випадках, коли джерело повідомляє про негативну щоденну зміну через виправлення даних, значення цього показника NA.
new_deaths_smoothed	Нові випадки смерті від COVID-19 (згладжено за 7 днів). Підрахунок може включати ймовірні смерті, якщо повідомлялося.
total_deaths_per_million	Загальна кількість смертей від COVID-19 на 1 000 000 осіб. Підрахунок може включати ймовірні смерті, якщо повідомлялося.
new_deaths_per_million	Нові випадки смерті від COVID-19 на 1 000 000 людей. Підрахунок може включати ймовірні смерті, якщо повідомлялося.
new_deaths_smoothed_per_million	Нові випадки смерті від COVID-19 (згладжені 7 днів) на 1 000 000 людей. Підрахунок може включати ймовірні смерті, якщо повідомлялося.

stringency_index	Індекс жорсткості реагування уряду: зведений показник на основі 9 індикаторів реагування, включаючи закриття шкіл, закриття робочих місць та заборони на поїздки, змінений до значення від 0 до 100.
reproduction_rate	Оцінка в режимі реального часу ефективного рівня відтворення (R) COVID-19.
iso_code	ISO код країни
continent	Материк географічного положення
location	Географічне положення
date	Дата спостереження
population	Розмір популяції
population_density	Кількість людей, поділена на площу землі, виміряна в квадратних кілометрах, за останній рік.
median_age	Середній вік населення, прогноз ООН на 2020 рік.
aged_65_older	Частка населення віком 65 років і старше, за останній рік.
aged_70_older	Частка населення віком від 70 років у 2015 році
gdp_per_capita	Валовий внутрішні.й продукт за паритетом купівельної спроможності (постійні міжнародні долари 2011 року), доступний останній рік.
extreme_poverty	Частка населення, що живе в умовах крайньої бідності, останній рік доступний з 2010 року.

cardiovasc_death_rate	Смертність від серцево-судинних захворювань у 2017 р. (річна кількість смертей на 100 тис. осіб).
diabetes_prevalence	Поширеність цукрового діабету (% населення у віці від 20 до 79 років) у 2017 році.
female_smokers	Частка жінок, які палять за останній рік.
male_smokers	Частка чоловіків, які палять за останній рік.
hospital_beds_per_thousand	Лікарняні ліжка на 1000 осіб, останній рік доступний з 2010 року.
life_expectancy	Очікувана тривалість життя.
human_development_index	Зведений індекс, що вимірює середні досягнення в трьох основних вимірах людського розвитку — довгому і здоровому житті, знаннях і гідному рівні життя. Значення за 2019 рік.

LSTM чутливі до масштабу вхідних даних, особливо коли використовуються сигмоїдні (за замовчуванням) або tanh функції активації. Тому необхідно нормалізувати дані в діапазоні від 0 до 1.

Для дослідження також необхідно поділити датасет на навчальну та тестову вибірки у відношенні 70% до 30% відповідно.

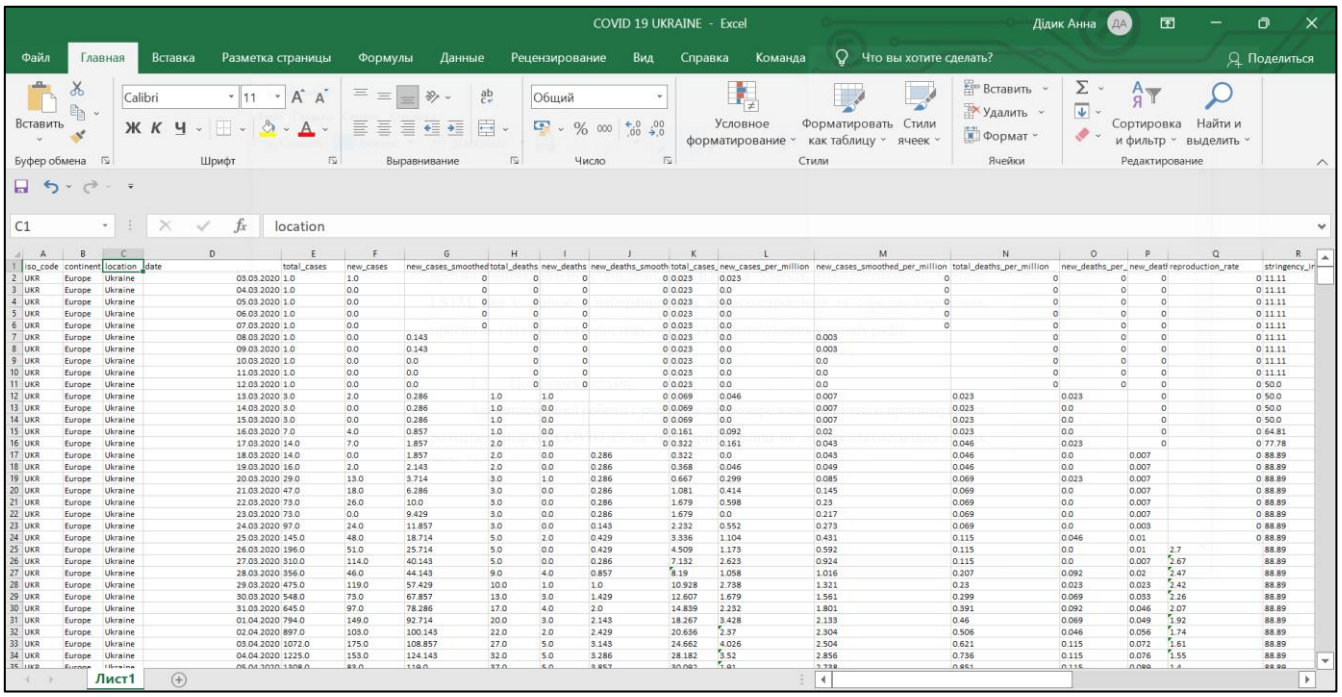

Рисунок 2.10 Представлення даних у датасеті

Висновки до розділу 2

У другому розділі проведено теоритичне та методичне дослідження теми, визначено основні напрямки вирішення поставленної задачі, проаналізовано архітектуру обраної моделі та розглянуто структуру навчального набору даних.

РОЗДІЛ 3 РЕАЛІЗАЦІЯ, ТЕСТОВИЙ ПРИКЛАД

3.1 Обґрунтування вибору середовища програмування та допоміжних бібліотек

Для реалізації нейронних мереж, візуалізації отриманих результатів, а також для підключення до Microsoft Excel необхідно обрати таку мову програмування, яка б містила весь необхідний функціонал.

Python - це скриптова мова програмування. Він універсальний, тому підходить для вирішення різноманітних завдань. Серед основних її переваг можна назвати такі:

- чистий синтаксис;

- переносність програм (що властиве більшості інтерпретованих мов);

- стандартний дистрибутив має велику кількість корисних модулів (у тому числі для розробки графічних інтерфейсів);

- можливість використання Python в режимі розмови (корисно для експериментів і вирішення простих задач);

- стандартний дистрибутив має просте, але в той же час дуже потужне середовище розробки під назвою IDLE, яке написане на Python;

- легко розв'язувати математичні задачі (має метод роботи з комплексними числами, може обробляти цілі числа будь-якого розміру, може використовуватися як потужний калькулятор в режимі розмови);

- відкритий код (можливість редагувати його іншими користувачами).

Python має ефективні високорівневі структури даних і простий, але ефективний підхід до об’єктно-орієнтованого програмування. Вільний синтаксис Python, динамічна обробка типів і інтерпретована мова роблять його ідеальним для написання сценаріїв і швидкої розробки додатків у багатьох галузях на більшості платформ.

Розглянемо детально бібліотеки, які використовуються для вирішення даної задачі.

3.1.1. Використання бібліотеки Keras

Однією з головних переваг Python – це Keras. Бібліотека високорівневих нейронних мереж, написана на Python і здатна працювати поверх TensorFlow - відкритої програмної бібліотеки для машинного навчання, розроблена компанією Google. Він був розроблений з акцентом на швидке експериментування. Уміння переходити від ідеї до результату якомога швидше - це ключове значення для проведення хороших досліджень.

Основні переваги Keras:

- дозволяє легко і швидко створювати прототипи (через загальну модульність, мінімалізм і розширюваність);
- підтримує як згорткові мережі, так і рекурсивні мережі, а також комбінації цих двох;
- підтримує довільні схеми підключення (включаючи навчання з декількома входами й кількома виходами);
- легко працює на процесорі та графічному процесорі.

На офіційному сайті Keras ми маємо доступ до Keras Applications - це моделі глибокого навчання, які доступні разом з попередньо підготовленими вагами (pre-trained). Ці моделі можна використовувати для прогнозування, вилучення функцій та тонкої настройки. А також важливою складовою є Keras Preprocessing - модуль попередньої обробки та доповнення даних бібліотеки глибокого навчання Keras.

3.1.2. Використання бібліотеки Pandas

Оскільки раніше було сказано, що статистичні дані зберігаються в таблиці Microsoft Excel, то необхідно забезпечити швидкий доступ до них, а також зберігати їх у зручному форматі. Для цього використовується бібліотека Pandas, а саме структура даних Pandas DataFrame.

Pandas — це програмна бібліотека для маніпуляції та аналізу даних, написана для мови програмування Python. Зокрема, він надає структури даних та операції для маніпулювання числовими таблицями та часовими рядами.

Можливості бібліотеки:

- об'єкт DataFrame із вбудованим індексуванням для маніпулювання даними;
- інструменти для читання та запису даних між структурами даних в пам'яті та різними форматами файлів;
- вирівнювання даних та вбудована підтримка пропущених даних;
- переформатовування для отримання зведених наборів даних;
- отримання зрізів за мітками, індексування з розширеними можливостями та отримання піднаборів з великих наборів даних;
- вставляння та вилучення стовпчиків у структурах даних;
- рушій групування, що дозволяє робити з наборами даних операції розділення-зміни-об'єднання (англ. split-apply-combine);
- злиття та з'єднання наборів даних;
- ієрархічне індексування осей для роботи з даними високої вимірності в структурі даних нижчої вимірності;
- функціональність для часових рядів: породження діапазонів дат та перетворення частоти, статистики рухливого вікна, лінійні регресії рухливого вікна, зсування дат та запізнювання.

Об'єкт DataFrame візуалізовано у вигляді звичайних таблиць, тому DataFrame є табличною структурою даних. Колонки в об'єкті DataFrame представляють собою об'єкти Series, рядки яких являються їх безпосередніми елементами.

3.1.3. Використання бібліотеки Matplotlib

Matplotlib — це бібліотека на мові програмування Python для візуалізації даних за допомогою 2D-графіки (також підтримується 3D-графіка). Отримані зображення можна використовувати як ілюстрації в публікаціях.

Matplotlib в основному написаний і підтримується Джоном Хантером і поширюється за ліцензією, схожою на BSD. Зображення, створені в різних форматах, можна використовувати в інтерактивній графіці, наукових публікаціях, графічних інтерфейсах користувача, веб-додатках, де потрібно побудувати діаграми. У документації автор визнає, що Matplotlib починався з імітації графічних команд MATLAB, але це окремий проект.

Matplotlib — це гнучкий, легко настроюваний пакет, який забезпечує функціональність, подібну до MATLAB, а також NumPy, SciPy та IPython. Зараз цей пакет працює з кількома графічними бібліотеками, включаючи wxWindows і PyGTK.

Пакет підтримує багато видів графіків і діаграм:

- Графіки (line plot);
- Діаграми розсіювання (scatter plot);
- Стовпчасті діаграми (bar chart) і гістограми (histogram);
- Секторні діаграми (pie chart);
- Діаграми «Стовбур-листя» (stem plot);
- Контурні графіки (contour plot);
- Поля градієнтів (quiver);
- Спектральні діаграми (spectrogram).

Користувачі можуть вказувати осі, сітки, додавати заголовки та пояснення, використовувати логарифмічний масштаб або полярні координати. Просту 3D-графіку можна створити за допомогою набору інструментів mplot3d. Є й інші набори інструментів: для створення діаграм, роботи з Excel, утилітами GTK тощо. Набір підтримуваних форматів зображень, векторних і растрових, можна знайти в словнику FigureCanvasBase.filetypes. Типові підтримувані формати:

- Encapsulated PostScript (EPS);
- Enhanced Metafile (EMF);
- JPEG;
- PDF;
- PNG;
- PostScript;
- RGBA;
- SVG;
- SVGZ;
- TIFF.

3.1.4. Використання бібліотеки NumPy

Numpy — це розширення мови Python, яке додає підтримку великих багатовимірних масивів і матриць, а також велику бібліотеку високорівневих математичних функцій для маніпулювання цими масивами. Numeric, попередник Numpy, спочатку був створений Джимом Хугуніном. Numpy — це програмне забезпечення з відкритим вихідним кодом і численними розробниками.

Оскільки Python — інтерпретована мова, математичні алгоритми, часто працюють в ньому набагато повільніше ніж у компільованих мовах, таких як C або навіть Java. NumPy намагається вирішити цю проблему для великої кількості обчислювальних алгоритмів забезпечуючи підтримку багатовимірних масивів і

безліч функцій і операторів для роботи з ними. Таким чином будь-який алгоритм, який може бути виражений в основному як послідовність операцій над масивами і матрицями, працює так само швидко, як еквівалентний код, написаний на C.

NumPy можна розглядати як гарну вільну альтернативу MATLAB, оскільки мова програмування MATLAB зовні нагадує NumPy: обидві вони інтерпретовані, і обидві дозволяють користувачам писати швидкі програми поки більшість операцій проводяться над масивами або матрицями, а не над скалярами. Перевага MATLAB у великій кількості доступних додаткових тулбоксів, включаючи такі як пакет Simulink. Основними пакетами, які доповнюють NumPy, є: SciPy - бібліотека, яка додає більше функціональних можливостей MATLAB; Matplotlib - пакет для створення графіків у стилі MATLAB. Внутрішньо і MATLAB, і NumPy засновані на бібліотеці LAPACK, розробленій для вирішення фундаментальних проблем лінійної алгебри.

3.2. Опис результатів моделювання

Спочатку протестуємо модель мережі для прогнозування деяких окремих критеріїв, не враховуючи вплив інших критеріїв, а саме: total_cases, new_cases, total_deaths, new_deaths, total_cases_per_million, new_cases_per_million.

Перед навчанням моделі слід звернути увагу на налаштування параметрів навчання. Отже, навчання проводиться за наступних умов:

- кількість епох: 100;
- look back (кількість даних, вікно для прогнозування) для навчальної вибірки: 30 (усереднена кількість днів у місяці).

Було написано функцію для перетворення нашого одного стовпця даних у набір даних із n-стовпців: перший стовпець матиме кількість зареєстрованих випадків, а наступні стовпці також матимуть кількість випадків, але в наступні $(n - 1)$ періоди (рис. 3.1). Функція приймає два аргументи: початковий набір даних, який

є масивом NumPy, який ми хочемо перетворити на новий набір даних, і look_back, тобто кількість попередніх кроків часу, які використовуються як вхідні змінні для прогнозування наступного періоду часу.

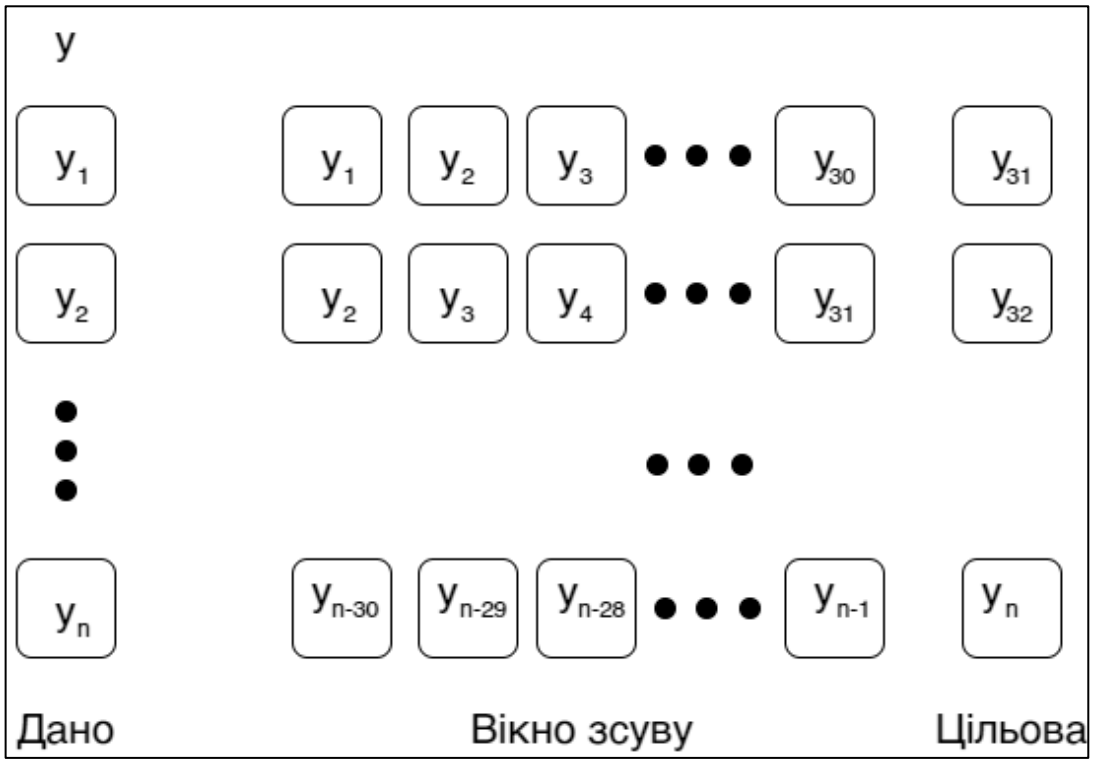

Рисунок 3.1 Схема перетворення даних у набір даних для прогнозування

Далі використовуємо цю функцію для підготовки наборів даних для тренувань та тестування нейронної мережі та перетворення даних у структуру, що відповідає входу нейронної мережі.

LSTM має 4 блоки всередині та 1 вихід зовні. Крім того, функція втрат обчислюється як середня квадратична помилка (MSE) [18]. Mean Square Error — середній квадрат помилки визначення будь-якої величини, квадратний корінь із MSE є середньоквадратичним відхиленням визначеної величини від її математичного очікування.

Після навчання моделі ми можемо оцінити якість моделі на навчальних та на тестових наборах даних для критеріїв, описаних вище (рис. 3.2-3.13).

Оригінальний набір даних відображається чорним кольором, прогнози для навчального набору - червоним, а прогнози для тестового набору - зеленим.

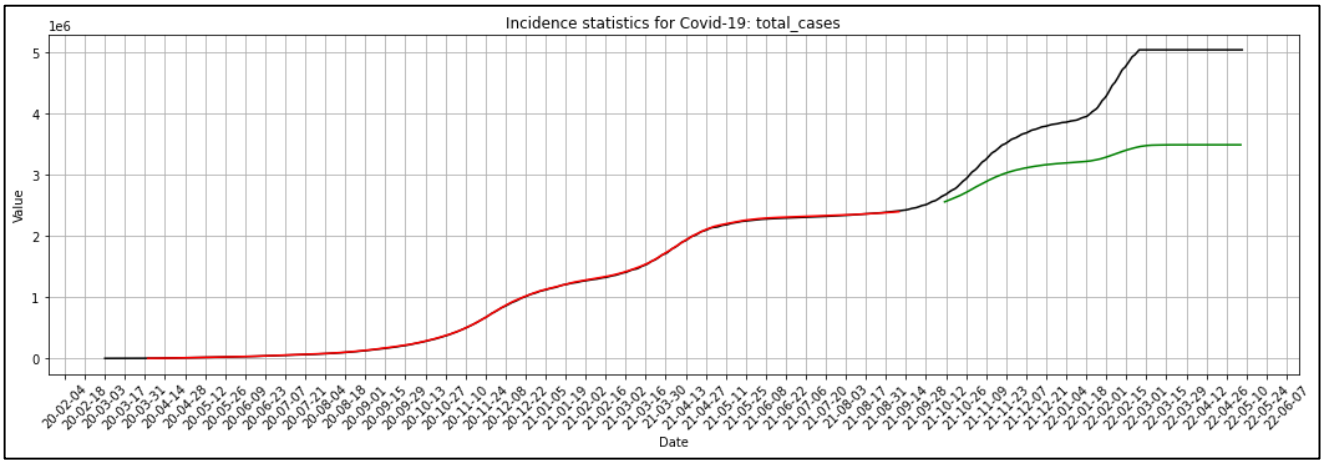

Рисунок 3.2 Прогнозування загальної кількості виявлених випадків

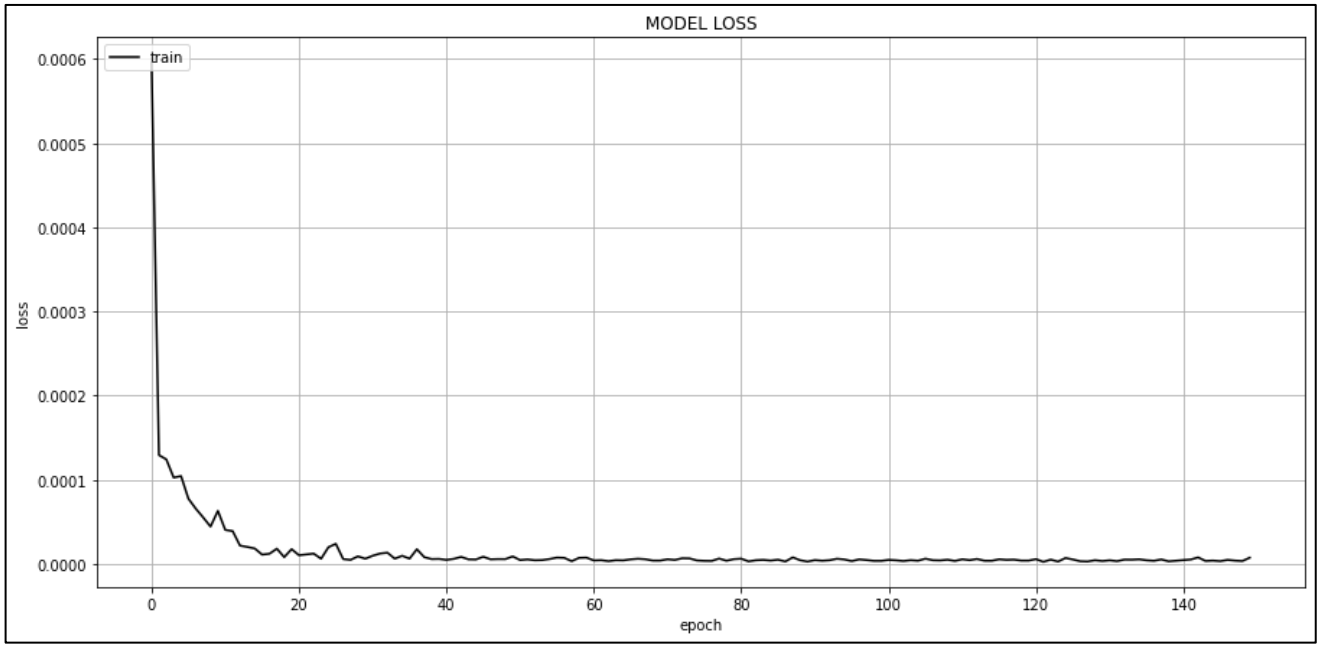

Рисунок 3.3 Графік функції втрат для прогнозу загальної кількості виявлених випадків

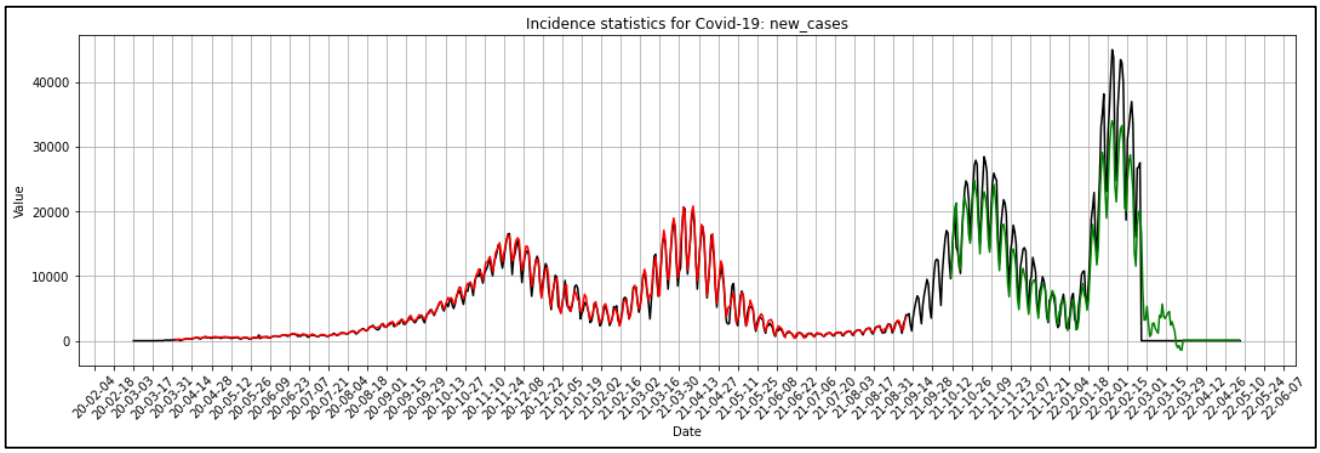

Рисунок 3.4 Прогнозування кількості нових виявлених випадків щоденно

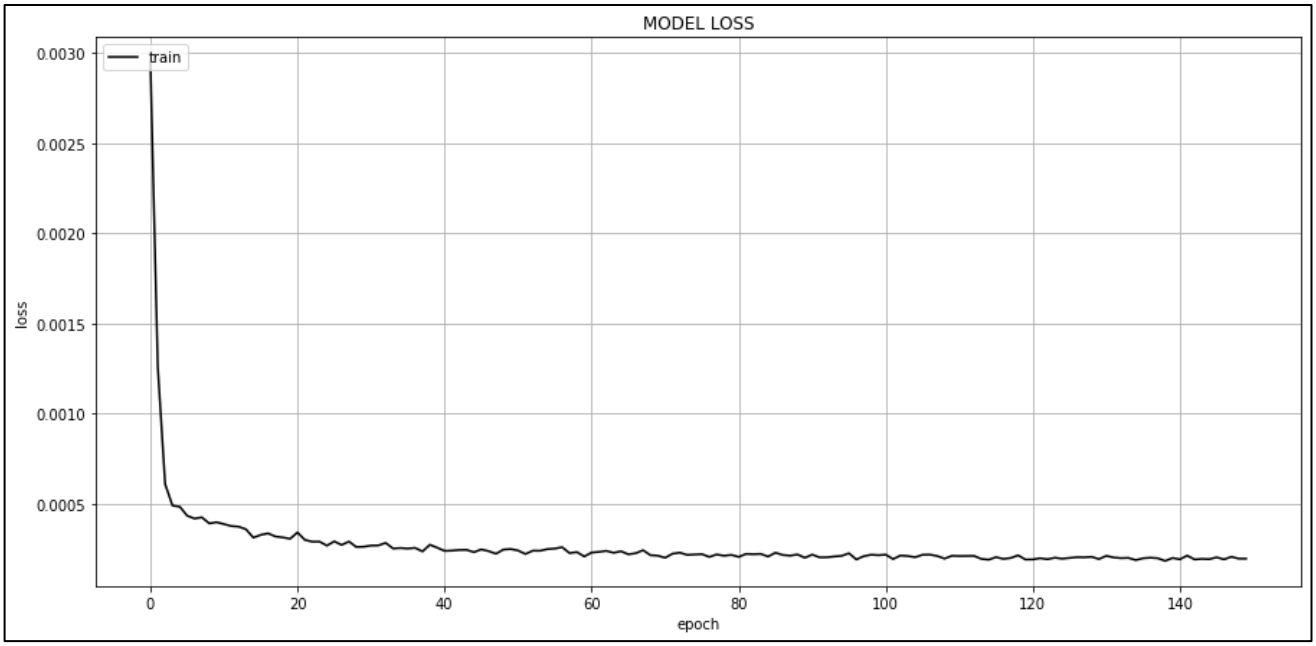

Рисунок 3.5 Графік функції втрат для прогнозу
кількості нових виявлених випадків

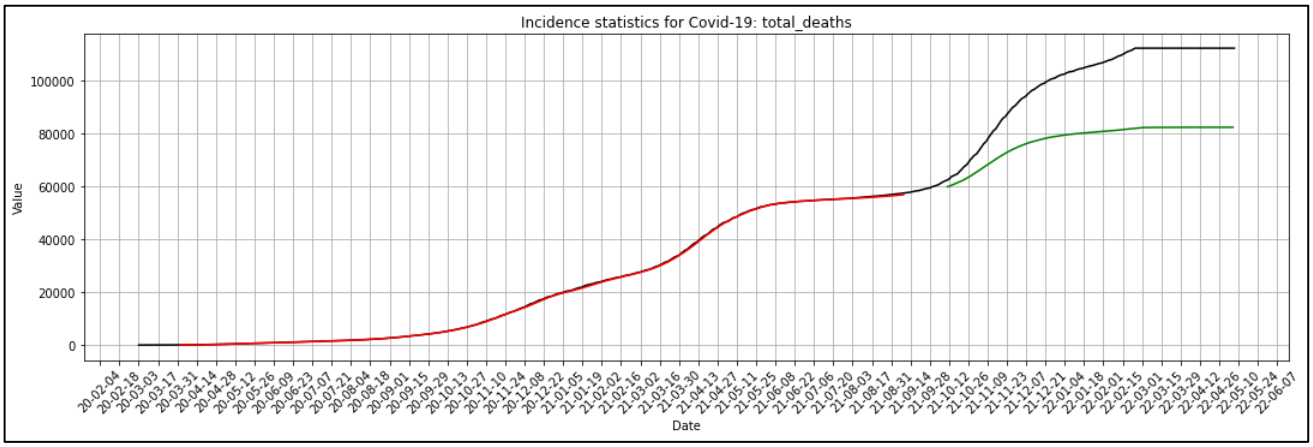

Рисунок 3.6 Прогнозування загальної кількості летальних випадків

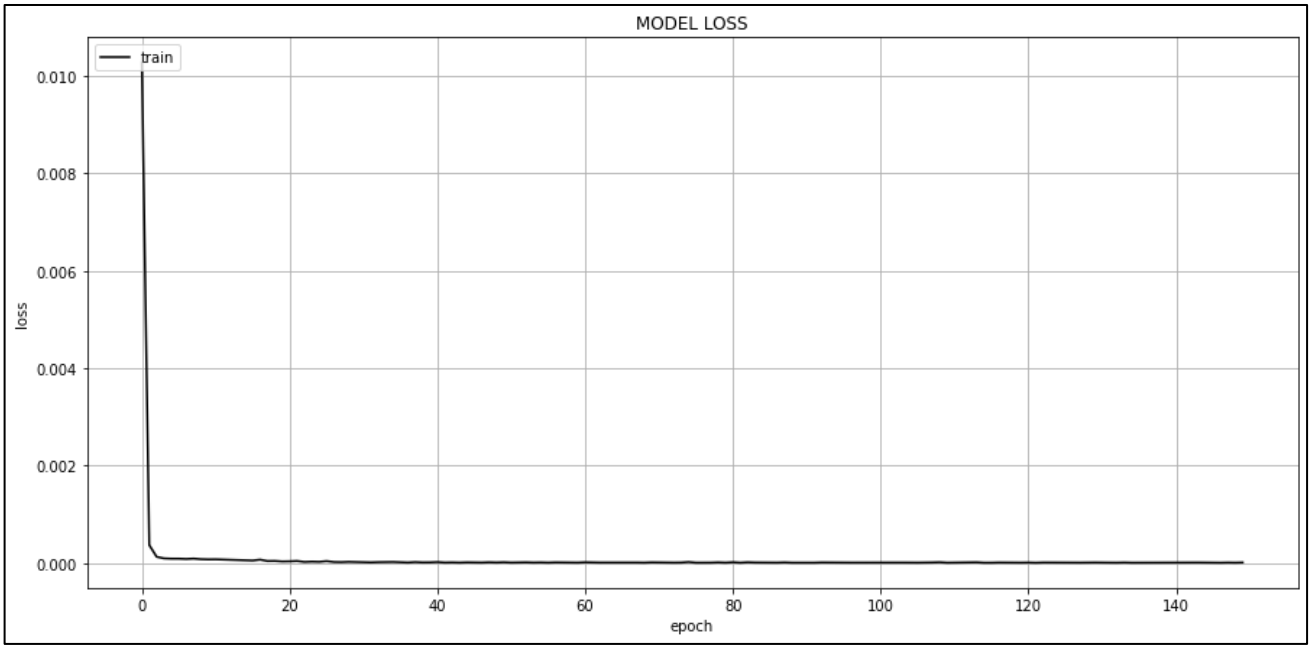

Рисунок 3.7 Графік функції втрат для прогнозу загальної кількості летальних випадків

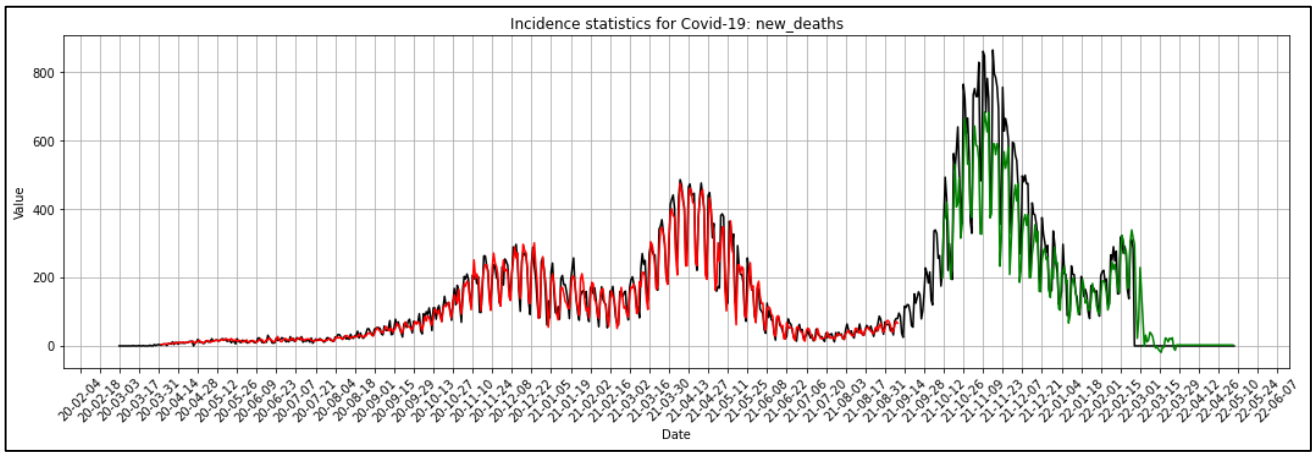

Рисунок 3.8 Прогнозування кількості нових летальних випадків щоденно

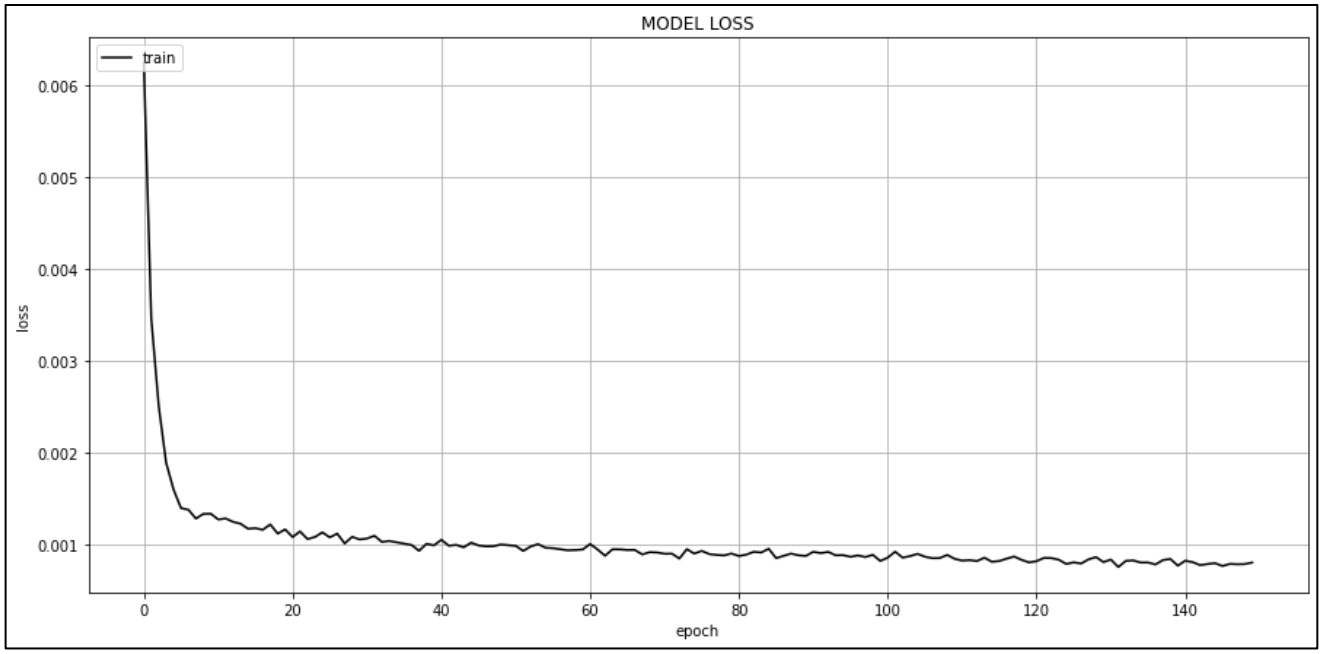

Рисунок 3.9 Графік функції втрат для прогнозу
кількості нових летальних випадків

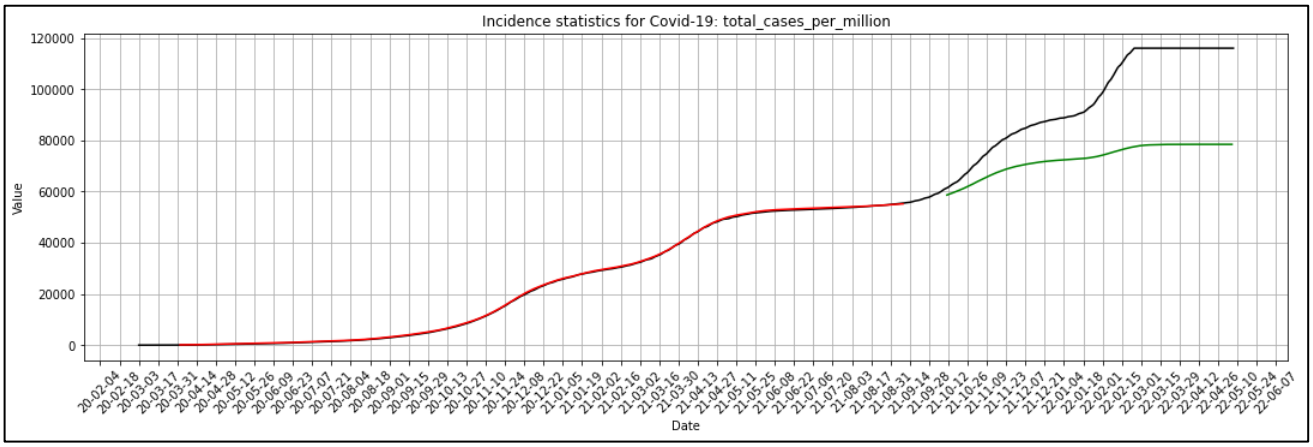

Рисунок 3.10 Прогнозування загальної кількості підтверджених випадків COVID-19 на 1 000 000 осіб

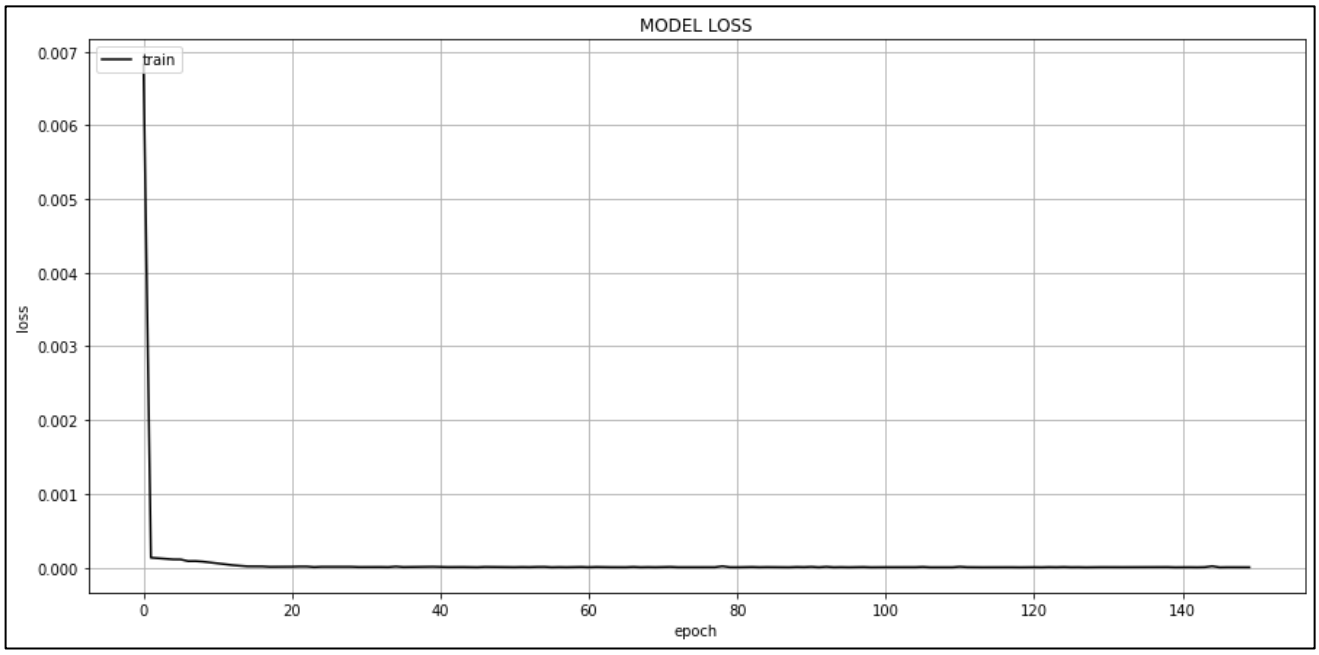

Рисунок 3.11 Графік функції втрат для прогнозу загальної кількості підтверджених випадків COVID-19 на 1 000 000 осіб

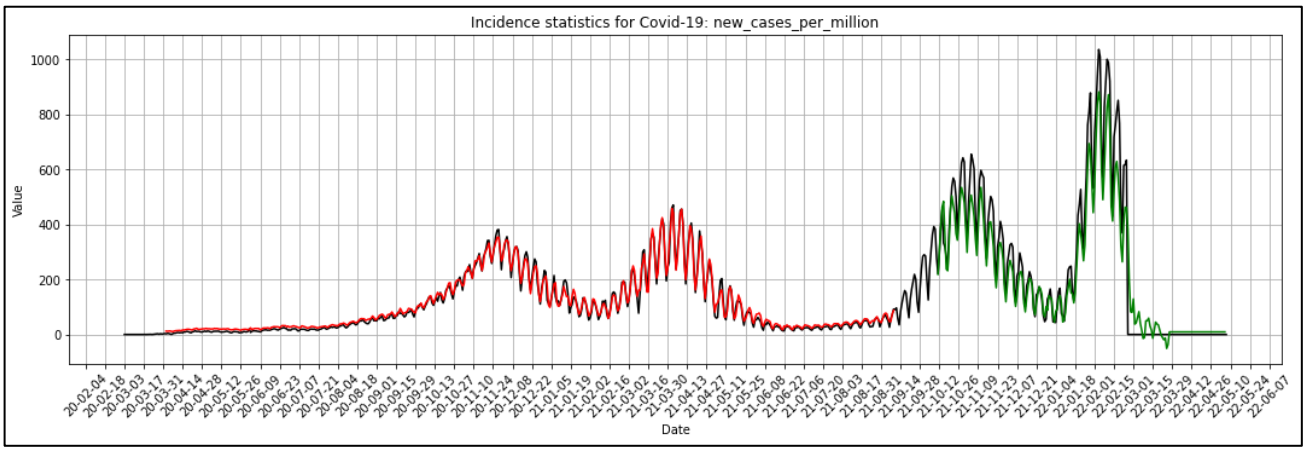

Рисунок 3.12 Прогнозування кількості
підтверджених випадків COVID-19 на 1 000 000 осіб щоденно

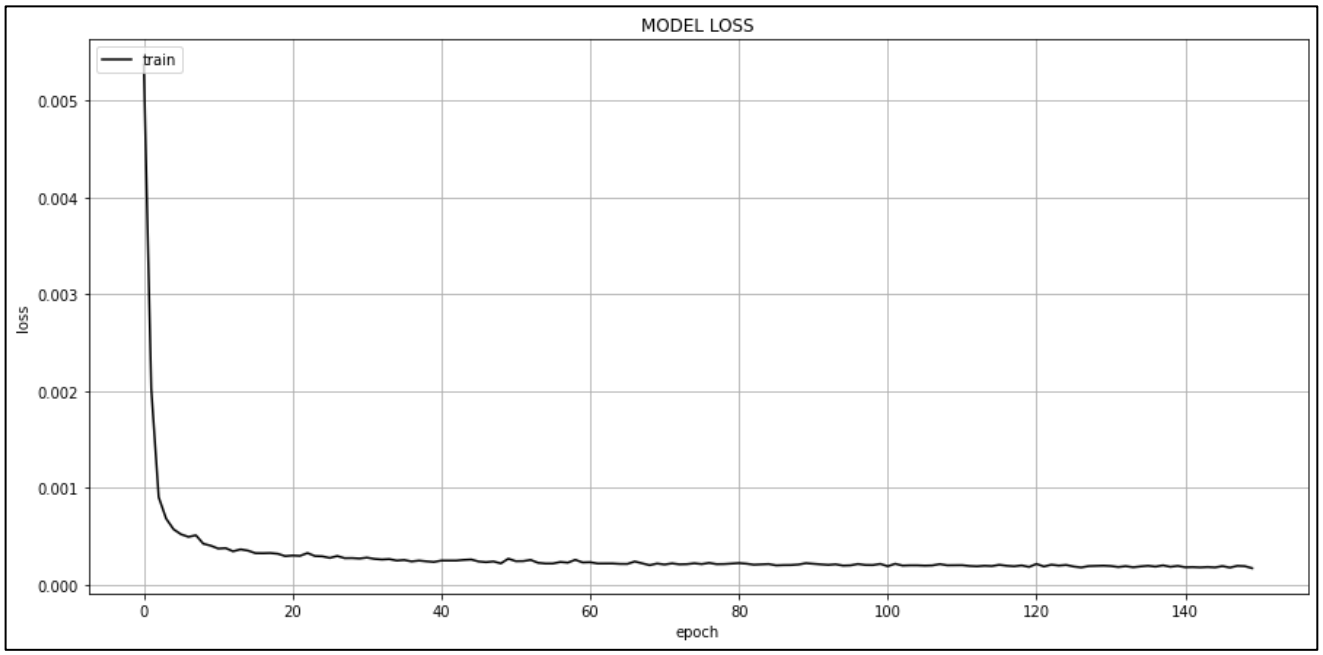

Рисунок 3.13 Графік функції втрат для прогнозу
кількості підтверджених випадків COVID-19 на 1 000 000 осіб щоденно

Якщо взяти до уваги вплив усіх перелічених критеріїв (рис. 3.14) у датасеті на значення критерію total_cases, тобто на загальну кількість випадків, то маємо результати, представлені на рисунках 3.15-3.16.

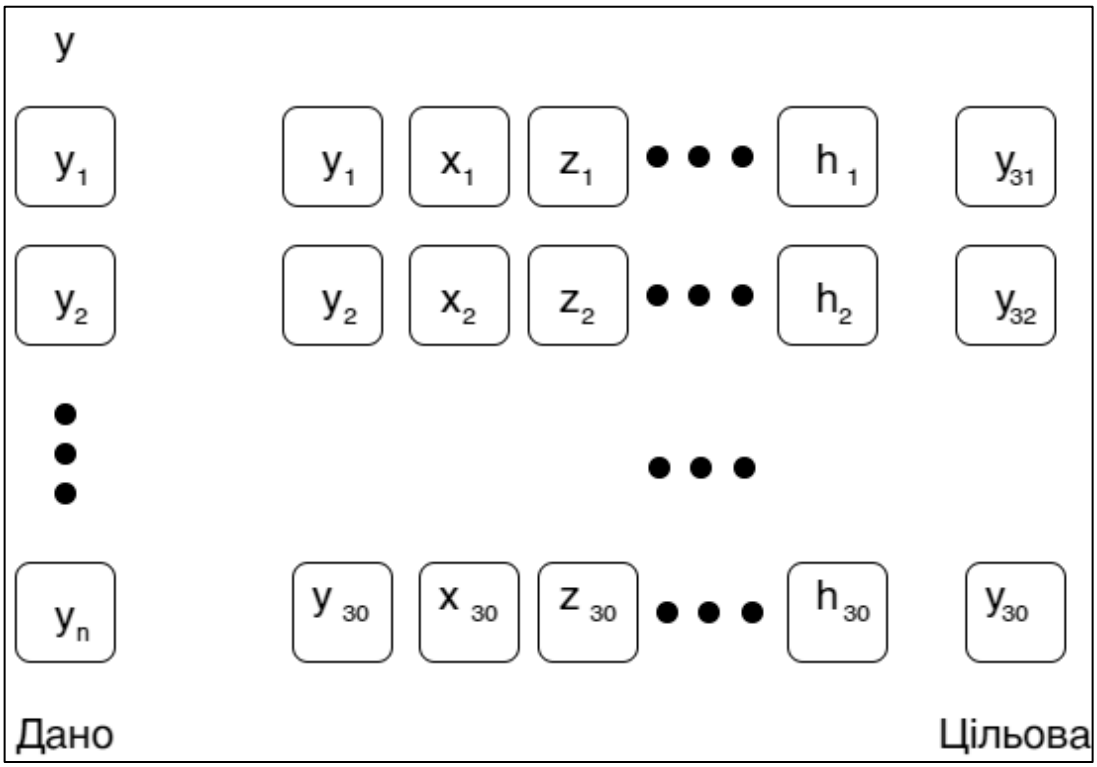

Рисунок 3.14 Схема формування набіру даних для прогнозування за багатьма критеріями

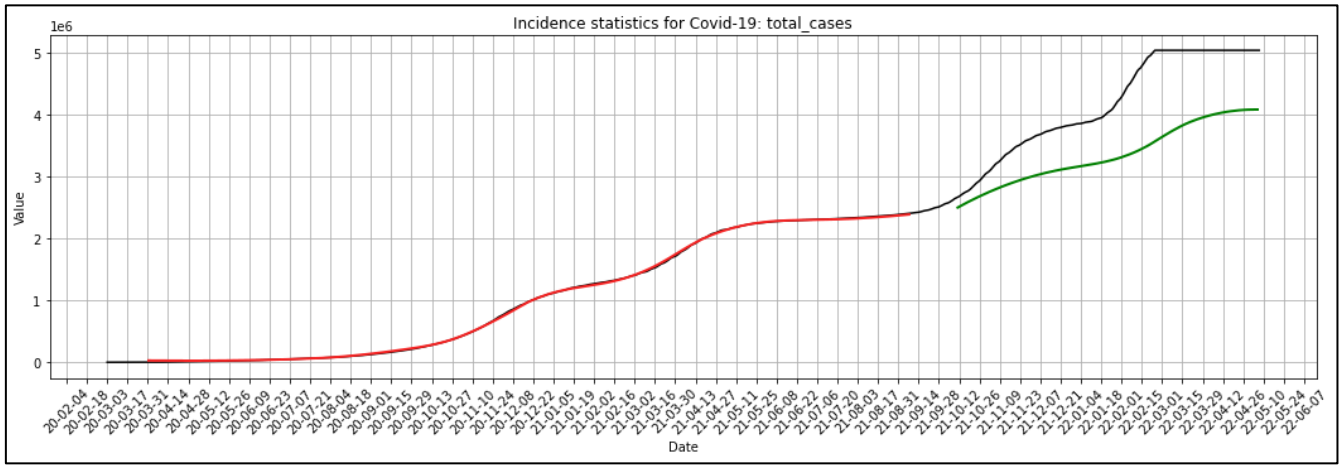

Рисунок 3.15 Прогнозування загальної кількості виявлених випадків

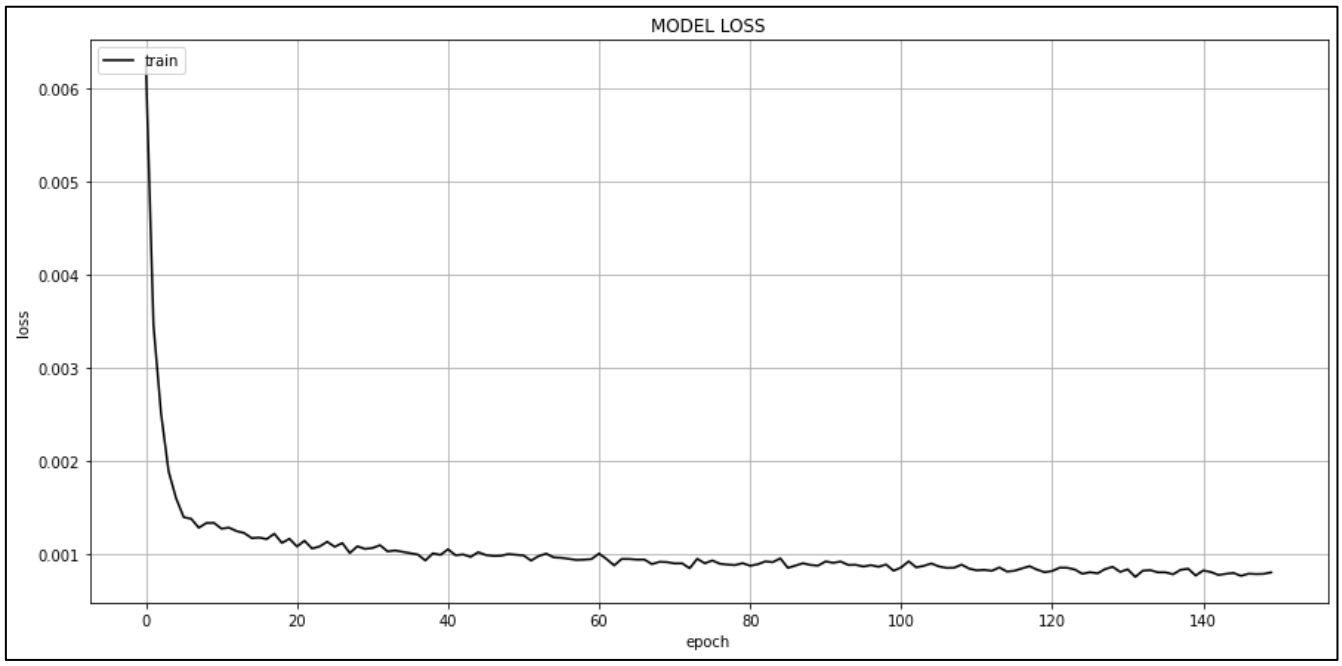

Рисунок 3.16 Графік функції втрат для прогнозу
загальної кількості виявлених випадків

Тобто, ми бачимо що прогноз для тестувальної вибірки має дещо інше значення. Це означає, що інші критерії мають свій вплив на даний показник і їх також варто враховувати при складанні прогнозів.

3.3. Оцінка адекватності експериментального дослідження

Отже, проаналізувавши отримані результати, можна сказати, що дана модель спрогнозувала спад розвитку епідемії, проте за прогнозом дещо повільніше, ніж це відбувалося в реальному житті. Для кожної функції втрат її графік прямує до нуля, що свідчить про достатньо високу точність отриманого прогнозу.

Для навчального набору даних прогноз є досить чітким, практично на 100% відповідає реальним даним. Для тестувального набору даних прогноз відрізняється від реальних значень. Це пов’язано з впливом інших чинників на перебіг хвороби, соціальних, політичних та інших обставин. Прогноз для тестувальної вибірки відрізняється в середньому на 30% від реальних значень. Крім того за прогнозом

епідемія мала піти на спад більш повільно, ніж це відбулося в реальності. Перш за все різкий спад захворюваності пов'язаний з тим, що в Україні спала кількість зроблених тестів на день та ускладнились умови підрахунку всіх зафіксованих випадків, тому дані статистики можуть бути не зовсім точними.

ВИСНОВКИ

У даній кваліфікаційній роботі було створено нейромережну технологію прогнозування розвитку пандемії COVID-19 на території України за допомогою LSTM мереж.

Тема роботи є дуже актуальною, адже прогнозування розвитку захворювань та їх поширення за допомогою нейронних мереж – це новий перспективний напрямок розвитку медичної сфери і штучного інтелекту.

Розроблена модель, що призначена для прогнозування коронавірусної хвороби, може використовуватись для передбачення можливих варіантів розвитку епідемій для вчасного проведення адекватних протиепідемічних заходів.

У першому розділі було проаналізовано статистику захворюваності в Україні, розглянуто основні методи вирішення задачі прогнозування на основі статистичних табличних даних.

У другому розділі було розглянуто основні методи прогнозування, описано архітектуру LSTM мереж та обрано оптимальний варіант для реалізації поставленої задачі.

У третьому розділі було описано особливості та переваги мови Python, розглянуто реалізацію обраної моделі нейронної мережі, наведено результати, що представлені у вигляді графіків, які візуалізують прогноз розвитку епідемії на території України.

Основні технології, що були використані у роботі - це мова програмування Python, нейромережна бібліотека Keras та інші програмні бібліотеки.

Для подальших досліджень та оптимізації рекомендується розширити список показників, що прогнозуються, наприклад, створити прогноз для кожної з областей України.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Statement on the second meeting of the International Health Regulations (2005) Emergency Committee regarding the outbreak of novel coronavirus (2019-nCoV) [Електронний ресурс] // Geneva, Switzerland. – 2020. – Режим доступу до ресурсу: https://www.who.int/news/item/30-01-2020-statement-on-the-second-meeting-of-the-international-health-regulations-(2005)-emergency-committee-regarding-the-outbreak-of-novel-coronavirus-(2019-ncov).

2. Smoking and COVID-19 [Електронний ресурс] // COVID-19: Scientific briefs. – 2020. – Режим доступу до ресурсу: https://www.who.int/publications/i/item/WHO-2019-nCoV-Sci_Brief-Smoking-2020.2.

3. Statement – Older people are at highest risk from COVID-19, but all must act to prevent community spread [Електронний ресурс]. – 2020. – Режим доступу до ресурсу: https://www.euro.who.int/en/health-topics/health-emergencies/coronavirus-covid-19/statements/statement-older-people-are-at-highest-risk-from-covid-19,-but-all-must-act-to-prevent-community-spread.

4. Three out of 10 people in the Americas are at increased risk of severe COVID-19 because of underlying conditions, PAHO Director says [Електронний ресурс]. – 2020. – Режим доступу до ресурсу: https://www.paho.org/en/news/21-7-2020-three-out-10-people-americas-are-increased-risk-severe-covid-19-because-underlying.

5. "Immunity passports" in the context of COVID-19 [Електронний ресурс]. – 2020. – Режим доступу до ресурсу: https://www.who.int/news-room/commentaries/detail/immunity-passports-in-the-context-of-covid-19.

6. Коронавірус в Україні [Електронний ресурс] – Режим доступу до ресурсу: https://covid19.com.ua/.

7. Мінфін. Коронавірус в Україні [Електронний ресурс] / Мінфін – Режим доступу до ресурсу: www.minfin.com.ua.

8. Ukraine: Coronavirus Pandemic Country Profile [Електронний ресурс] – Режим доступу до ресурсу: https://ourworldindata.org/coronavirus/country/ukraine.https://www.surgonco.ru/jour/article/view/505/407

9. Hasanov A.G., Shaybakov D.G., Zhernakov S.V., Men'shikov A.M., Badretdinova F.F., Sufiyarov I.F., Sagadatova J.R. Neural Networks in Forecasting Disease Dynamics. Creative surgery and oncology. 2020;10(3):198-204. (In Russ.) https://doi.org/10.24060/2076-3093-2020-10-3-198-204

10. М. А. Кондратьев. Методы прогнозирования и модели распространения заболеваний / М. А. Кондратьев. // КОМПЬЮТЕРНЫЕ ИССЛЕДОВАНИЯ И МОДЕЛИРОВАНИЕ. – 2013. – С. 863–882.

11. Forecasting COVID-19 cases: A comparative analysis between recurrent and convolutional neural networks / Khondoker NazmoonNabi, Md TokiTahmid, AbdurRafi та ін.]. // Results in Physics. – 2021.

12. COVID-19 in Iran: Forecasting Pandemic Using Deep Learning / Rahele Kafieh, Roya Arian, Narges Saeedizadeh. – 2021.

13. Forecasting of COVID-19 using deep layer Recurrent Neural Networks (RNNs) with Gated Recurrent Units (GRUs) and Long Short-Term Memory (LSTM) cells / K.E.ArunKumar, Dinesh V.Kalaga, Ch. Mohan SaiKumar. // Chaos, Solitons & Fractals Volume. – 2021.

14. Mohammad Ayoub Khan. Performance evaluation of regression models for COVID-19: A statistical and predictive perspective / Mohammad Ayoub Khan, Rijwan Khan, Fahad Algarni. // Ain Shams Engineering Journal. – 2022.

15. Светуньков И. С., Светуньков С. Г. Методы социально-экономического прогнозирования в 2 Т. Т.1 Теория и методология (учебник и практикум для вузов)

URL: https://www.urait.ru/viewer/metody-socialno-ekonomicheskogoprognozirovaniya-v-2-t-t-1-teoriya-i-metodologiya-467618#page/1 (дата звернення: 05.04.2021).

16. Коротков А.В. Маркетинговые исследования (Студми. Учебные материалы для студентов) URL: https://studme.org/63535/marketing/marketingovye_issledovaniya (дата звернення: 07.04.2021).

17. Robert M. Kunst Nonlinear time series URL: https://homepage.univie.ac.at/robert.kunst/nonlin1.pdf (дата звернення: 09.04.2021).

18. Чучуева Ирина Александровна Модель прогнозирования временных рядов по выборке максимального подобия: дис. … канд. тех. наук: 05.13.18 / Московский государственный технический университет им. Н.Э. Баумана. Москва, 2012.

19. Antonevych M. FORECASTING RETAIL TRADE REPORT DEPARTMENT STORES WITH LSTM NETWORKS [Електронний ресурс] / Antonevych M.. – 2021. – Режим доступу до ресурсу: https://www.kaggle.com/maricinnamon/retail-trade-report-department-stores-lstm.

ДОДАТОК А Лістинг програми навчання LSTM мережі

```
import numpy
import pandas
import matplotlib.pyplot as plt
from keras.models import Sequential
from keras.layers import Dense
from keras.layers import LSTM
from sklearn.preprocessing import MinMaxScaler
from sklearn.metrics import mean_squared_error
import matplotlib.dates as mdates
from matplotlib.dates import DateFormatter

for i in range (6):
    dataframe = pandas.read_excel('Covid-19-totalInfo_2022.xlsx',sheet_name=i )
    #print(dataframe.head())

    labelsDate=dataframe[dataframe.columns[0]]
    labelsDate= pandas.to_datetime(labelsDate).dt.date
    x_pos = numpy.arange(len(dataframe.index))

    dataframe.drop(dataframe.columns[0], axis=1, inplace=True)
    #print(dataframe.head())

    fig, ax = plt.subplots(figsize=(18, 5))
    plt.title('Incidence statistics for Covid-19: '+dataframe.columns[0])
    plt.xlabel('Date')
    ax.xaxis.set_major_locator(mdates.WeekdayLocator(interval=2))
    ax.xaxis.set_major_formatter(DateFormatter("%y-%m-%d"))
    ax.tick_params(axis='x', rotation=45)
    #plt.xticks(x_pos, labelsDate, rotation=90, fontsize=5)
    plt.ylabel('Value')
    plt.plot(labelsDate,dataframe, 'k-')
    plt.grid()
    plt.show()

    numpy.random.seed(7)

    dataset = dataframe.values
    dataset = dataset.astype('float32')
    #print(dataset[:5])

    scaler = MinMaxScaler(feature_range=(0, 1))
    dataset = scaler.fit_transform(dataset)
```

```
#print(dataset[:5])

train_size = int(len(dataset) * 0.70)
test_size = len(dataset) - train_size
train, test = dataset[0:train_size,:], dataset[train_size:len(dataset),:]

def create_dataset(dataset, look_back):
    dataX, dataY = [], []
    for i in range(len(dataset)-look_back-1):
        a = dataset[i:(i+look_back), 0]
        dataX.append(a)
        dataY.append(dataset[i + look_back, 0])
    return numpy.array(dataX), numpy.array(dataY)

look_back = 30
trainX, trainY = create_dataset(train, look_back)
testX, testY = create_dataset(test, look_back)
trainX = numpy.reshape(trainX, (trainX.shape[0], 1, trainX.shape[1]))
testX = numpy.reshape(testX, (testX.shape[0], 1, testX.shape[1]))

model = Sequential()
model.add(LSTM(4, input_shape=(1, look_back)))
model.add(Dense(1))
model.compile(loss='mean_squared_error', optimizer='adam')
history = model.fit(trainX, trainY, epochs=150, batch_size=1, verbose=0)

plt.figure(figsize=(15,7))
plt.plot(history.history['loss'], 'k')
plt.title('MODEL LOSS')
plt.ylabel('loss')
plt.xlabel('epoch')
plt.legend(['train'], loc='upper left')
plt.grid()
plt.show()

trainPredict = model.predict(trainX)
testPredict = model.predict(testX)

trainPredict = scaler.inverse_transform(trainPredict)
testPredict = scaler.inverse_transform(testPredict)

trainPredictPlot = numpy.empty_like(dataset)
trainPredictPlot[:, :] = numpy.nan
trainPredictPlot[look_back:len(trainPredict)+look_back, :] = trainPredict
testPredictPlot = numpy.empty_like(dataset)
testPredictPlot[:, :] = numpy.nan
```

```
testPredictPlot[len(trainPredict)+(look_back*2)+1:len(dataset)-1, :] = testPredict
fig, ax = plt.subplots(figsize=(18, 5))
ax.xaxis.set_major_locator(mdates.WeekdayLocator(interval=2))
ax.xaxis.set_major_formatter(DateFormatter("%y-%m-%d"))
ax.tick_params(axis='x', rotation=45)
plt.plot(labelsDate,scaler.inverse_transform(dataset), 'k')
plt.plot(labelsDate,trainPredictPlot, 'r')
plt.plot(labelsDate,testPredictPlot, 'g')
plt.title('Incidence statistics for Covid-19: '+dataframe.columns[0])
plt.ylabel('Value')
plt.xlabel('Date')
plt.grid()
plt.show()
```