

Національний аерокосмічний університет
"Харківський авіаційний інститут"
Міністерство освіти і науки України

Кваліфікаційна наукова
праця на правах рукопису

БУТКЕВИЧ МИКОЛА ВІТАЛІЙОВИЧ

УДК 004.855.5+519.245:616-036.22

ДИСЕРТАЦІЯ
МОДЕЛІ, МЕТОДИ ТА ІНФОРМАЦІЙНА ТЕХНОЛОГІЯ
ПРОГНОЗУВАННЯ ЧАСОВИХ РЯДІВ ЕПІДЕМІЧНИХ ПРОЦЕСІВ

Спеціальність 122 Комп'ютерні науки

Галузь знань 12 Інформаційні технології

Подається на здобуття ступеня доктора філософії

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

_____ Буткевич Микола Віталійович

Науковий керівник:
Чумаченко Дмитро Ігорович,
кандидат технічних наук, доцент

Харків – 2026

АНОТАЦІЯ

Буткевич Микола Віталійович. Моделі, методи та інформаційна технологія прогнозування часових рядів епідемічних процесів. – Кваліфікаційна наукова праця на правах рукопису.

Дисертація на здобуття наукового ступеня доктора філософії за спеціальністю 122 «Комп'ютерні науки». – Національний аерокосмічний університет «Харківський авіаційний інститут», Харків, 2026.

Метою дисертаційної роботи є підвищення точності прогнозування захворюваності на інфекційні хвороби у короткостроковій перспективі шляхом розроблення методів та інформаційної технології на основі інтеграції статистичних моделей часових рядів та рекурентних нейронних мереж з довгою короткочасною пам'яттю.

Дисертація присвячена вирішенню актуальної науково-прикладної задачі, що полягає у розробці інформаційної технології прогнозування захворюваності на інфекційні хвороби у короткостроковій перспективі шляхом розроблення методів та інформаційної технології на основі інтеграції статистичних моделей часових рядів та рекурентних нейронних мереж з довгою короткочасною пам'яттю.

Обґрунтована методика проведення досліджень, що використовується в роботі. Дослідження базується на теоріях статистичного аналізу часових рядів та обчислювального штучного інтелекту, а саме – моделях авторегресії та інтегрованого ковзного середнього (ARIMA) та рекурентних нейронних мережах з довгою короткочасною пам'яттю (LSTM), що дозволяють моделювати та прогнозувати динаміку інфекційної захворюваності.

Основні наукові та практичні результати полягають у наступному:

1. Охарактеризовано основні проблеми, пов'язані з прогнозуванням епідемічних часових рядів в Україні та інших країнах, обумовлені, передусім, проблемами з даними – неповнотою та запізненням епідемічних сповіщень,

складностями діагностування медичного та організаційного характеру, впливом зовнішніх факторів (погодних умов, мобільності населення, карантинних обмежень), а також структурними зламами системи епідемічного нагляду внаслідок воєнного конфлікту та надзвичайних ситуацій. Проведено розвідувальний аналіз даних захворюваності на гострі респіраторні вірусні інфекції, грип, лептоспіроз, сальмонельоз та вірусні гепатити А, В та С в Україні за період 2013–2024 років. Проаналізовано якість додаткових факторів, які впливають на поширення інфекційних захворювань. Доведено, що вплив цих факторів є дуже мінливим, дані є неточними і зашумленими, надходять із різним запізненням, а отже, задачу доцільно звести до однофакторної – до побудови моделі лише часового ряду кількості нових випадків захворювання. Проведено систематичний огляд сучасних підходів до моделювання та прогнозування епідемічних процесів. Здійснено порівняльне оцінювання точності моделей, що показало переваги методів, здатних враховувати сезонність та нелінійні залежності.

2. Формалізовано математичний апарат моделей ARIMA та LSTM для епідемічного прогнозування. Розроблено процедури ідентифікації порядків моделі ARIMA, діагностики залишків, конструювання багатопарових LSTM архітектур та попередньої обробки епідемічних даних. Експериментальна валідація на даних грипу в Харківській області та лептоспірозу у країнах Балтії підтвердила ефективність методу.

3. Розроблено методологічні вдосконалення: метод walk-forward валідації, метод циклічного кодування сезонних ознак, систематизовано підходи до порівняння рівневих та кумулятивних представлень даних. Розроблено архітектурні рішення для LSTM моделей різної складності.

4. Розроблено архітектуру інформаційної технології прогнозування інфекційної захворюваності з розділенням тренувального та інференсного шарів, інтеграцією моделей ARIMA та LSTM у єдиний конвеєр. Реалізовано

веб-застосунок для онлайн трекінгу прогнозів. Побудовано UML-діаграми модулів технології.

5. Проведено експериментальну валідацію на реальних даних України та країн Балтії. Для ARIMA досягнуто похибки 6,84–9,59%. Для LSTM кумулятивне представлення забезпечило зниження похибки у 16–94 рази. Проведено контрфактуальний аналіз впливу міграції на COVID-19 у Польщі.

6. Створено програмно-інформаційне забезпечення на Python у середовищі Jupyter Notebook. Проведено валідацію на даних 2015–2024 років для семи інфекційних захворювань.

Наукова новизна дисертаційної роботи полягає в наступних положеннях:

1. Уперше запропоновано метод кумулятивного представлення епідемічних часових рядів для прогнозування мережами довгої короткочасної пам'яті, заснований на принципі заміни рівневих значень кількості випадків захворювання наростаючим підсумком від початку періоду спостереження з подальшою min-max нормалізацією, параметри якої обчислюються виключно на навчальній вибірці, який відрізняється від існуючих підходів до прогнозування захворюваності використанням монотонно зростаючої траєкторії, що згладжує короткострокові флуктуації та знижує відносну дисперсію, що дозволяє знизити середню абсолютну відсоткову похибку прогнозування у 16–94 рази залежно від захворювання (до 0,28% для гепатиту С, 0,82–0,85% для гепатиту В).

2. Уперше запропоновано метод циклічного кодування сезонних ознак для LSTM моделей епідемічного прогнозування, заснований на тригонометричному відображенні дискретного часового індексу на одиничне коло у двовимірному просторі за допомогою синусоїдальної та косинусоїдальної функцій від номера місяця, який відрізняється від існуючих підходів до моделювання сезонності в епідемічних нейромережних моделях забезпеченням безперервності та рівновіддаленості сусідніх місяців,

включаючи перехід грудень–січень, що дозволяє покращити точність прогнозування на 46,6% порівняно з моделями без явних сезонних ознак.

3. Удосконалено метод валідації прогностичних моделей епідемічних часових рядів для коротких вибірок від 72 місяців в умовах структурних зламів, заснований на стратегії walk-forward з послідовним розширенням навчального вікна та принципі строгої темпоральної каузальності, який відрізняється від існуючих обов'язковим обчисленням параметрів усіх трансформацій, включаючи нормалізацію та формування ознак, виключно на навчальній частині вибірки на кожному кроці, що дозволяє отримати реалістичну оцінку продуктивності моделі в режимі послідовного прогнозування та запобігає витоків інформації з тестового періоду.

4. Набула подальшого розвитку архітектура LSTM мереж для епідемічного прогнозування, заснована на принципі ієрархічного представлення часових залежностей із адаптацією складності моделі до обсягу та характеру доступних даних, яка відрізняється від існуючих систематичним емпіричним обґрунтуванням достатності компактних двошарових конфігурацій з 64 одиницями пам'яті на шар для коротких рядів від 72 місяців та необхідності глибоких тришарових пірамідальних архітектур з 11,89 мільйонами параметрів для рядів з множинними сезонними компонентами та структурними зламами, що дозволяє забезпечити навчання моделей за 40 секунд на стандартному CPU без GPU-прискорювачів та обирати архітектуру з мінімальними обчислювальними витратами без втрати точності прогнозування.

5. Набула подальшого розвитку архітектура інформаційної технології прогнозування інфекційної захворюваності, заснована на розділенні тренувального та інференсного шарів з версіонованим сховищем артефактів моделей, яка відрізняється від існуючих інтеграцією статистичних моделей ARIMA та нейромережевих моделей LSTM у єдиний конвеєр з автоматизованим контролем якості через маніфести версій та метрики

валідації, що дозволяє забезпечити відтворюваність прогнозів та оперативне перенавчання при надходженні нових епідемічних даних.

Практична цінність одержаних результатів полягає в наступному:

1. Отримані наукові положення сприяли розробленню комплексу програм на Python у середовищі Jupyter Notebook, які автоматизують розрахунки за усіма запропонованими методами. Проведено валідацію на реальних даних епідемічного нагляду в Україні та країнах Балтії за період 2010–2024 років.

2. Досягнуто підвищення точності прогнозування: для кумулятивних представлень MAPE менше 1%, для ARIMA – 6,8–9,6%. Похибка прогнозу гепатиту В вписується в запас безпеки центрального медичного складу України.

3. Компактні LSTM моделі навчаються за 40 секунд на стандартному CPU, що забезпечує застосовність в умовах обмежених ресурсів воєнного часу.

4. Контрфактуальний аналіз COVID-19 у Польщі продемонстрував відсутність значущого впливу міграційного шоку на епідемічну динаміку.

5. Розроблену інформаційну технологію можна застосовувати для прогнозування інших типів нестационарних часових рядів техногенного або природного характеру.

Результати дисертаційного дослідження опубліковані у 11 наукових працях, у тому числі 3 статті у виданнях, що індексуються у наукометричній базі даних Scopus, 2 статті у наукових фахових періодичних виданнях України, 6 публікацій у матеріалах міжнародних науково-практичних конференцій.

Розроблені моделі, методи та програмне забезпечення, зокрема інтелектуальні моделі прогнозування динаміки грипу та гострих інфекцій на базі архітектур ARIMA та LSTM підтверджені апробацією основних положень на міжнародних наукових та науково-практичних конференціях, де результати дістали позитивну оцінку наукової спільноти, - практична значущість результатів підтверджена отриманням 8 свідоцтв про реєстрацію авторського

права на комп'ютерні програми (№132251, №132250, №132249, №131623, №112762, №112634, №112605, №110034).

Впровадження пропонованої технології у Державній установі «Харківський обласний центр контролю та профілактики хвороб МОЗ України», Харківському національному медичному університеті, КНП «Міська клінічна лікарня №13» Харківської міської ради та КНП ММР «Мереф'янська центральна районна лікарня» дозволило підвищити якість епідемічного нагляду за інфекційними захворюваннями та оптимізувати систему прогнозування епідемічних процесів. Використання розроблених моделей сприяє своєчасному виявленню тенденцій розвитку епідемічної ситуації, що дозволяє більш ефективно планувати та проводити профілактичні заходи, а також підвищити якість підготовки фахівців у галузі епідеміології.

Ключові слова: прогнозування часових рядів, епідемічне моделювання, нейронні мережі, довга короткочасна пам'ять, ARIMA, машинне навчання, глибоке навчання, інфекційні захворювання, кумулятивне представлення, walk-forward валідація, структурні злами, штучний інтелект

ANNOTATION

Butkevych Mykola Vitaliyovych. Models, Methods, and Information Technology for Forecasting Time Series of Epidemic Processes– Qualification Research Paper in Manuscript Form.

Dissertation for the degree of Doctor of Philosophy in the specialty 122 "Computer Science". – National Aerospace University "Kharkiv Aviation Institute", Kharkiv, 2026.

The aim of the dissertation is to improve the accuracy of short-term infectious disease incidence forecasting by developing methods and information technology based on the integration of statistical time series models and recurrent neural networks with long short-term memory.

The dissertation is devoted to solving the topical scientific and applied problem of developing information technology for forecasting the incidence of infectious diseases in the short term by developing methods and information technology based on the integration of statistical time series models and recurrent neural networks with long short-term memory.

The research methodology is grounded in the theories of statistical time series analysis and computational artificial intelligence, namely autoregressive integrated moving average (ARIMA) models and recurrent neural networks with long short-term memory (LSTM), which enable modelling and forecasting of infectious disease dynamics.

The main scientific and practical results are as follows:

1. The key problems associated with forecasting epidemiological time series in Ukraine and other countries have been characterised. An exploratory analysis of incidence data for acute respiratory viral infections, influenza, leptospirosis, salmonellosis, and viral hepatitis A, B, and C in Ukraine for the period 2013–2024 has been conducted. It has been demonstrated that the influence of external factors is highly variable and the data are noisy; therefore, it is appropriate to reduce the problem to a univariate time series model. A systematic review of modern approaches has been carried out. A comparative evaluation of model accuracy demonstrated the advantages of methods capable of capturing seasonality and nonlinear dependencies.

2. The mathematical apparatus of ARIMA and LSTM models for epidemiological forecasting has been formalised. Procedures for model order identification, residual diagnostics, multi-layer LSTM architecture construction, and epidemiological data preprocessing have been developed. Experimental validation on influenza and leptospirosis data confirmed the effectiveness of the method.

3. Methodological improvements have been developed: walk-forward validation, cyclic encoding of seasonal features, systematisation of level and cumulative data representations. Improved architectural solutions for LSTM models of varying complexity have been developed.

4. An information technology architecture for infectious disease incidence forecasting has been developed with separation of training and inference layers, integrating ARIMA and LSTM models into a unified pipeline. A web application for online tracking of forecasts has been implemented. UML diagrams for technology modules have been constructed.

5. Experimental validation on real data from Ukraine and the Baltic countries has been conducted. For ARIMA, errors of 6.84–9.59% were achieved. For LSTM, cumulative representation reduced errors by 16–94 times. A counterfactual analysis of migration impact on COVID-19 in Poland was conducted.

6. Software tools for automating the proposed methods have been created in Python in Jupyter Notebook. Validation on 2015–2024 data for seven infectious diseases has been conducted.

The scientific novelty of the dissertation is as follows:

1. For the first time, a method of cumulative representation of epidemiological time series for LSTM forecasting has been proposed, based on replacing level case counts with a running total with min-max normalisation computed on the training sample, which reduces MAPE by 16–94 times (to 0.28% for hepatitis C, 0.82–0.85% for hepatitis B).

2. For the first time, a method of cyclic encoding of seasonal features for epidemiological LSTM models has been proposed, based on trigonometric mapping of the month index onto a unit circle, which improves forecasting accuracy by 46.6%.

3. The validation method for epidemiological time series models on short samples from 72 months under structural breaks has been improved, based on walk-forward strategy with strict temporal causality, preventing information leakage from the test period.

4. The LSTM network architecture for epidemiological forecasting has been further developed, providing empirical justification for compact two-layer configurations for short series and deep three-layer pyramidal architectures with

11.89 million parameters for complex tasks, enabling training in 40 seconds on a standard CPU.

5. The information technology architecture has been further developed, integrating ARIMA and LSTM models into a unified pipeline with automated quality control, ensuring reproducibility and prompt retraining.

The practical value of the obtained results is as follows:

1. A suite of Python programs automating all proposed methods has been developed. Validation on real data from Ukraine and the Baltic countries for 2010–2024 has been conducted.

2. Forecasting accuracy improvement has been achieved: MAPE below 1% for cumulative representations, 6.8–9.6% for ARIMA. The hepatitis B forecast error falls within the safety margin of Ukraine's central medical warehouse.

3. Compact LSTM models train in 40 seconds on a standard CPU, enabling deployment under limited wartime infrastructure.

4. Counterfactual analysis of COVID-19 in Poland demonstrated no significant short-term impact of migration shock on epidemiological dynamics.

5. The developed technology can be adapted for forecasting other non-stationary time series of technogenic or natural origin.

Results published in 11 scientific works, including 3 Scopus-indexed articles, 2 articles in Ukrainian professional journals, and 6 conference publications.

The developed models, methods, and software, specifically intelligent models for forecasting influenza and acute infection dynamics based on ARIMA and LSTM architectures, have been validated through the presentation of key findings at international scientific and practical conferences, receiving positive feedback from the scientific community. The practical significance of the results is further evidenced by eight certificates of copyright registration for computer programs (Nos. 132251, 132250, 132249, 131623, 112762, 112634, 112605, 110034).

The implementation of the proposed technology at the State Institution 'Kharkiv Regional Center for Disease Control and Prevention of the Ministry of Health of Ukraine', Kharkiv National Medical University, and the Municipal Non-Commercial Enterprise 'City Clinical Hospital No. 13' of Kharkiv City Council made it possible to improve the quality of epidemiological surveillance of infectious diseases and optimize the system for forecasting epidemic processes. The use of developed models contributes to timely identification of trends in the development of the epidemic situation, which allows for more effective planning and implementation of preventive measures, as well as improving the quality of training for epidemiology specialists.

Keywords: time series forecasting, epidemiological modelling, neural networks, long short-term memory, ARIMA, machine learning, deep learning, infectious diseases, cumulative representation, walk-forward validation, structural breaks, artificial intelligence.¶

ЗМІСТ

АНОТАЦІЯ	2
ЗМІСТ	12
ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ ТА СКОРОЧЕНЬ	15
ВСТУП.....	18
СПИСОК ПУБЛІКАЦІЙ ЗДОБУВАЧА ЗА ТЕМОЮ ДИСЕРТАЦІЇ	29
РОЗДІЛ 1. АНАЛІЗ ПРОБЛЕМИ ПРОГНОЗУВАННЯ ЕПІДЕМІЧНИХ ПРОЦЕСІВ	33
1.1 Підходи до моделювання епідемічних процесів.....	33
1.2 Аналіз джерел прогнозування гострого та хронічного гепатиту В на основі даних із використанням рекурентних нейронних мереж.....	37
1.3 Аналіз застосування моделі ARIMA у симуляції та прогнозуванні динаміки грипу	47
1.4 Аналіз поточного стану дослідження часових рядів захворюваності на лептоспіроз для прогнозування	51
1.5 Висновки до розділу та постановка задач досліджень	61
РОЗДІЛ 2. РОЗРОБЛЕННЯ МЕТОДІВ ТА ТЕОРЕТИЧНИХ ОСНОВ ПРОГНОЗУВАННЯ ЕПІДЕМІЧНИХ ПРОЦЕСІВ.....	66
2.1 Теоретичні основи LSTM.....	66
2.2 Теорія моделей авторегресії та проінтегрованого ковзного середнього (ARIMA).....	74
2.2.1 Методи вибору порядків моделі.....	87
2.2.2 Попередня обробка та діагностика	88
2.2.3 Функції втрат, метрики оцінювання та валідаційні стратегії.....	89
2.3 Аналіз стаціонарності часових рядів	91

	13
2.4 Архітектурні рішення першої ітерації	94
2.4.1 Методи попередньої обробки даних	95
2.4.2 Навчання, оптимізація та регуляризація	96
2.5 Висновки до розділу	97
РОЗДІЛ 3. ПОКРАЩЕННЯ МЕТОДІВ ТА ПРИКЛАДНЕ ЗАСТОСУВАННЯ МЕТОДІВ ПРОГНОЗУВАННЯ ЕПІДЕМІЧНИХ ПРОЦЕСІВ.....	102
3.1 Методологічні вдосконалення третьої ітерації LSTM архітектури	102
3.2 Вдосконалені архітектурні рішення.....	103
3.3 Розширені методи попередньої обробки	105
3.4 Узагальнення третьої ітерації	109
3.5 Розробка веб застосунку для онлайн трекінгу епідемічних прогнозів	111
3.5.1 Архітектура системи для веб додатку	111
3.5.2 Архітектурні компоненти та потоки даних.....	117
3.5.3 Методологія оцінювання якості і валідація	119
3.5.4 Графічний інтерфейс веб-системи моніторингу	121
3.6 Висновки до розділу	126
РОЗДІЛ 4. РЕЗУЛЬТАТИ ЕКСПЕРИМЕНТАЛЬНИХ ДОСЛІДЖЕНЬ	130
4.1 Результати застосування базових LSTM	130
4.1.1 Результати прогнозування гострого гепатиту А в Україні.....	130
4.1.2 Результати прогнозування грипу в Харківській області методом LSTM.....	133
4.1.3 Результати прогнозування сальмонельозу в Харківській області	136

	14
4.1.4 Узагальнення результатів першої ітерації	138
4.2 Результати застосування ARIMA моделей.....	141
4.2.1 Результати прогнозування лептоспірозу в країнах Балтії	141
4.2.2 Результати прогнозування грипу в Харківській області методом ARIMA.....	146
4.2.3 Узагальнення результатів другої ітерації.....	149
4.3 Результати вдосконалених досліджень	151
4.3.1 Результати прогнозування гострого та хронічного гепатиту В в Україні	151
4.3.2 Результати прогнозування гострого гепатиту С в Україні.....	156
4.3.3 Результати прогнозування гострих респіраторних вірусних інфекцій в Україні	157
4.3.4 Результати контрфактуального аналізу COVID-19 у Польщі на основі LSTM	160
4.3.5 Узагальнення результатів третьої ітерації	164
4.4 Порівняльний аналіз усіх моделей.....	166
4.4.1 Порівняння LSTM та ARIMA підходів	166
4.4.2 Аналіз ефекту кумулятивного представлення	167
4.4.3 Аналіз обчислювальної ефективності	168
4.5 Висновки до розділу	171
ВИСНОВКИ	175
СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ	182
ДОДАТОК А	200
ДОДАТОК Б.....	204
ДОДАТОК В	205

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ ТА СКОРОЧЕНЬ

- ACF – автокореляційна функція (Autocorrelation Function)
- ADF – розширений тест Дікі-Фуллера (Augmented Dickey-Fuller test)
- AIC – інформаційний критерій Акаїке (Akaike Information Criterion)
- Adam – алгоритм адаптивної оцінки моментів, метод оптимізації (Adaptive Moment Estimation)
- ANN – штучна нейронна мережа (Artificial Neural Network)
- API – інтерфейс програмування застосунків (Application Programming Interface)
- ARCH – авторегресійна умовна гетероскедастичність (Autoregressive Conditional Heteroskedasticity)
- ARIMA – авторегресійна інтегрована модель ковзного середнього (Autoregressive Integrated Moving Average)
- ARIMAX – ARIMA з екзогенними змінними (ARIMA with exogenous variables)
- BIC – байєсівський інформаційний критерій (Bayesian Information Criterion)
- BPNN – нейронна мережа зворотного поширення (Backpropagation Neural Network)
- CPU – центральний процесор (Central Processing Unit)
- COVID-19 – коронавірусна хвороба 2019 (Coronavirus Disease 2019)
- EDA – розвідувальний аналіз даних (Exploratory Data Analysis)
- ECDC – Європейський центр профілактики та контролю захворювань (European Centre for Disease Prevention and Control)
- GARCH – узагальнена авторегресійна умовна гетероскедастичність (Generalized Autoregressive Conditional Heteroskedasticity)
- GPU – графічний процесор (Graphics Processing Unit)
- GRU – рекурентна одиниця з вентилями (Gated Recurrent Unit)

HWES – метод експоненційного згладжування Холта-Вінтерса (Holt-Winters Exponential Smoothing)

INAR – цілочисельна авторегресійна модель (Integer-valued Autoregressive model)

KPSS – тест Квятковського-Філіпса-Шмідта-Шіна (Kwiatkowski-Phillips-Schmidt-Shin test)

LSTM – довга короткочасна пам'ять (Long Short-Term Memory)

LVMBP – зворотне поширення Левенберга-Марквардта (Levenberg-Marquardt BackPropagation)

MAE – середня абсолютна похибка (Mean Absolute Error)

MAPE – середня абсолютна відсоткова похибка (Mean Absolute Percentage Error)

MLE – метод максимальної правдоподібності (Maximum Likelihood Estimation)

MSE – середньоквадратична похибка (Mean Squared Error)

PACF – парціальна автокореляційна функція (Partial Autocorrelation Function)

REST – архітектурний стиль веб-сервісів на основі представлень (Representational State Transfer)

RF – метод випадкового лісу (Random Forest)

RMSE – корінь середньоквадратичної похибки (Root Mean Squared Error)

RNN – рекурентна нейронна мережа (Recurrent Neural Network)

SARIMA – сезонна авторегресійна інтегрована модель ковзного середнього (Seasonal Autoregressive Integrated Moving Average)

SARIMAX – SARIMA з екзогенними змінними (Seasonal ARIMA with exogenous variables)

SEICR – сприйнятливі-експоновані-інфіковані-хронічні носії-одужали (Susceptible-Exposed-Infected-Carrier-Recovered)

SEIR – сприйнятливі-експоновані-інфіковані-одужали (Susceptible-Exposed-Infected-Recovered)

SEIRI – сприйнятливі–експоновані–інфіковані–одужалі–повторно інфіковані (Susceptible–Exposed–Infected–Recovered–Infected)

SEIRS – SEIR з поверненням до стану сприйнятливості (Susceptible–Exposed–Infected–Recovered–Susceptible)

SIR – сприйнятливі–інфіковані–одужалі (Susceptible–Infected–Recovered)

SIRS – SIR з поверненням до стану сприйнятливості (Susceptible–Infected–Recovered–Susceptible)

SNN – контрольована нейронна мережа (Supervised Neural Network)

STL – декомпозиція сезонності та тренду методом локальної зваженої регресії (Seasonal and Trend decomposition using Loess)

UML – уніфікована мова моделювання (Unified Modeling Language)

VAR – векторна авторегресія (Vector Autoregression)

XGBoost – метод екстремального підсилення градієнту (Extreme Gradient Boosting)

ZMP – регресійна модель нульово-модифікованого розподілу Пуассона (Zero-Modified Poisson model)

ВООЗ – Всесвітня організація охорони здоров'я

ГВДШ – гострі інфекції верхніх дихальних шляхів

ГРВІ – гострі респіраторні вірусні інфекції

ІЛР – індекс людського розвитку

МОЗ – Міністерство охорони здоров'я України

ЦГЗ – Центр громадського здоров'я Міністерства охорони здоров'я України



ВСТУП

Обґрунтування вибору теми дослідження. Прогнозування захворюваності на інфекційні хвороби є одним із ключових завдань системи громадського здоров'я, що безпосередньо впливає на ефективність планування ресурсів, закупівель вакцин та антивірусних препаратів, а також на своєчасність реагування на епідемічні загрози [1, 2]. Точні короткострокові прогнози дозволяють медичним закладам завчасно підготуватися до сезонних спалахів, оптимізувати розподіл ліжкового фонду та забезпечити достатні запаси медикаментів [3, 4].

Традиційні підходи до епідемічного прогнозування базуються на детерміністичних та стохастичних компартментних моделях (SIR, SEIR, SEIRS), які вимагають точного знання біологічних параметрів передачі інфекції та контактних матриць населення [5, 6]. Водночас статистичні моделі часових рядів, зокрема моделі авторегресії та інтегрованого ковзного середнього (ARIMA), забезпечують формальну процедуру ідентифікації, оцінювання та діагностики, проте обмежені припущенням лінійності та погіршують точність при наявності складних нелінійних залежностей та структурних зламів [7, 8].

Методи глибокого навчання, зокрема рекурентні нейронні мережі з довгою короткочасною пам'яттю (LSTM), продемонстрували здатність автоматично виявляти складні нелінійні патерни у послідовних даних без явної специфікації функціональної форми залежностей [9, 10]. Архітектура LSTM вирішує проблему зникаючого градієнта, характерну для класичних рекурентних мереж, через механізм вентилів, що контролюють потік інформації до стану комірки [11]. Це робить LSTM перспективним інструментом для моделювання епідемічних часових рядів з вираженою сезонністю, множинними періодичними компонентами та непередбачуваними структурними зламами.

Теоретичні основи детерміністичного моделювання поширення інфекцій систематизовано у підручнику Martcheva M. [9], стохастичні епідемічні моделі досліджено у монографії Britton T. [10], а стохастичні популяційні моделі — у роботі Allen L.J.S. [11]. Методологію авторегресійного аналізу та прогнозування часових рядів закладено у класичній праці Box G.E.P., Jenkins G.M. та Reinsel G.C. [12]; автоматизований добір моделей і оцінку якості прогнозів розроблено Hyndman R.J. з колегами [13–15]. Фундаментальну архітектуру рекурентних нейронних мереж типу LSTM, що долає проблему зникаючого градієнта, розроблено Hochreiter S. та Schmidhuber J. [16], а застосування LSTM до прогнозування поширення COVID-19 реалізовано Chimmula V.K.R. та Zhang L. [17]. Серед вітчизняних науковців задачами інформаційних технологій епідемічного моделювання та підтримки прийняття рішень займаються Чумаченко Д.І., Чумаченко Т.О., Яковлев С.В. [18–19], а методи статистичного прогнозування і математичного аналізу досліджують також Бідюк П.І., Мокін В.Б., Лосенко А.В.

Україна, як країна, що протягом останніх років переживає повномасштабний військовий конфлікт, стикається з унікальними та безпрецедентними викликами у сфері епідемічного спостереження [20–21]. Порушення функціонування систем епідемічного нагляду внаслідок руйнування медичної інфраструктури у прифронтових та окупованих регіонах, масова міграція населення як усередині країни, так і за її межі, критичне зниження охоплення планової імунізації та перевантаження медичної інфраструктури створюють сприятливі умови для зростання захворюваності на керовані інфекції [22–23]. З лютого 2022 року понад 700 задокументованих атак пошкодили або знищили медичні заклади, змусивши тисячі клініцистів евакуюватися та перервавши регіональну лабораторну спроможність [24]. Рутинна реєстрація інфекційних захворювань різко впала в першій половині 2022 року і відновилася лише частково [25].

У таких екстремальних умовах методи машинного навчання, здатні функціонувати при неповних, запізнілих та частково недостовірних даних, швидко адаптуватися до раптових структурних зламів епідемічної ситуації без повного перенавчання моделей, набувають критичної важливості для національної системи охорони здоров'я як інструмент підтримки прийняття оперативних рішень в умовах високої невизначеності.

Однак існуючі дослідження у сфері застосування LSTM для епідемічного прогнозування мають суттєві прогалини. По-перше, більшість робіт виконано на стабільних даних країн з розвиненою системою нагляду та не враховують умов порушеної звітності [26–27]. По-друге, питання оптимального представлення даних для нейромережових моделей (рівневе vs кумулятивне) залишається недостатньо дослідженим. По-третє, валідаційні стратегії часто не відповідають реальним умовам експлуатації прогнозних систем, допускаючи витік інформації з тестового періоду. По-четверте, відсутні систематичні рекомендації щодо вибору архітектури LSTM залежно від обсягу та характеру доступних даних в умовах обмежених обчислювальних ресурсів.

Таким чином, тема дослідження є актуальною. Її розробка сприяє вдосконаленню методів прогнозування інфекційної захворюваності, підвищенню стійкості системи епідемічного нагляду до зовнішніх шоків та забезпеченню підтримки прийняття рішень у системі громадського здоров'я України в умовах воєнного конфлікту.

Актуальним науковим завданням є розроблення інформаційної технології прогнозування епідемічних часових рядів на основі інтеграції статистичних моделей та рекурентних нейронних мереж з довгою короткочасною пам'яттю.

Зв'язок роботи з науковими програмами, планами, темами. Дослідження, результати яких викладено в дисертації, виконано на кафедрі математичного моделювання та штучного інтелекту Національного аерокосмічного університету «Харківський авіаційний інститут» в рамках

виконання науково-дослідних робіт за проєктами «Мультидисциплінарне дослідження впливу надзвичайних ситуацій на поширення інфекційних захворювань для підтримки прийняття управлінських рішень у сфері біобезпеки населення» (Національний фонд досліджень України, № 2023.03/0197), «Моделювання та прогнозування поширення інфекцій у воєнних та повоєнних умовах з використанням даних епідеміологічного, поведінкового та геномного спостереження» (фінансування Національною академією наук США (US NAS) через Науково-технологічний центр в Україні (STCU), № 7136).

Об'єкт дослідження – процес прогнозування захворюваності на інфекційні хвороби на основі аналізу епідемічних часових рядів.

Предмет дослідження – методи та засоби прогнозування епідемічних часових рядів на основі рекурентних нейронних мереж з довгою короткочасною пам'яттю та моделей авторегресії з інтегрованим ковзним середнім.

Метою дисертаційної роботи є розробка інформаційної технології прогнозування захворюваності на інфекційні хвороби у короткостроковій перспективі шляхом розроблення методів та інформаційної технології на основі інтеграції статистичних моделей часових рядів та рекурентних нейронних мереж з довгою короткочасною пам'яттю.

Для досягнення мети дисертації щодо розробки методів та засобів прогнозування епідемічних часових рядів на основі рекурентних нейронних мереж з довгою короткочасною пам'яттю необхідно вирішити такі задачі:

- 1) здійснити аналіз відомих методів та моделей прогнозування епідемічних часових рядів, проаналізувати основні проблеми, які виникають під час прогнозування в умовах обмежених та порушених даних, та провести розвідувальний аналіз даних захворюваності на інфекційні хвороби в Україні;

- 2) формалізувати математичний апарат моделей ARIMA та LSTM для епідемічного прогнозування, розробити процедури ідентифікації порядків

моделі, діагностики залишків, конструювання багат шарових LSTM архітектур та попередньої обробки епідемічних даних;

3) розробити метод кумулятивного представлення епідемічних часових рядів, метод циклічного кодування сезонних ознак та метод walk-forward валідації з принципом строгої темпоральної каузальності для підвищення точності прогнозування;

4) розробити архітектуру інформаційної технології прогнозування інфекційної захворюваності з інтеграцією статистичних та нейромережових моделей у єдиний конвеєр та реалізувати веб-застосунок для онлайн трекінгу прогнозів;

5) провести експериментальну валідацію розроблених методів на реальних даних епідемічного нагляду України та країн Балтії, включаючи дані періоду повномасштабного військового конфлікту, та контрфактуальний аналіз впливу масової міграції на динаміку COVID-19;

6) створити програмно-інформаційне забезпечення для автоматизації запропонованих методів та випробувати його на реальних даних.

Методи дослідження містять загальнонаукову методологію проведення досліджень і принципи системного підходу, а саме: аналіз літературних джерел та відкритих даних, методи теорії часових рядів (авторегресія, інтегроване ковзне середнє, аналіз автокореляційних функцій, тести стаціонарності), методи глибокого навчання (рекурентні нейронні мережі з довгою короткочасною пам'яттю, регуляризація Dropout, адаптивна оптимізація Adam), методи статистичного оцінювання (інформаційні критерії Акаїке та Байєса, критерій Льюнга-Бокса), методи валідації моделей (walk-forward валідація, хронологічний поділ даних), методи попередньої обробки даних (min-max нормалізація, циклічне кодування сезонних ознак, формування послідовностей ковзним вікном), об'єктно-орієнтоване програмування та експериментальні дослідження.

Наукова новизна одержаних результатів:

1. Уперше запропоновано метод кумулятивного представлення епідемічних часових рядів для прогнозування мережами довгої короткочасної пам'яті, заснований на принципі заміни рівневих значень кількості випадків захворювання наростаючим підсумком від початку періоду спостереження з подальшою \min - \max нормалізацією, параметри якої обчислюються виключно на навчальній вибірці, який відрізняється від існуючих підходів до прогнозування захворюваності використанням монотонно зростаючої траєкторії, що згладжує короткострокові флуктуації та знижує відносну дисперсію, що дозволяє знизити середню абсолютну відсоткову похибку прогнозування у 16–94 рази залежно від захворювання (до 0,28% для гепатиту С, 0,82–0,85% для гепатиту В).

2. Уперше запропоновано метод циклічного кодування сезонних ознак для LSTM моделей епідемічного прогнозування, заснований на тригонометричному відображенні дискретного часового індексу на одиничне коло у двовимірному просторі за допомогою синусоїдальної та косинусоїдальної функцій від номера місяця, який відрізняється від існуючих підходів до моделювання сезонності в епідемічних нейромережових моделях забезпеченням безперервності та рівновіддаленості сусідніх місяців, включаючи перехід грудень–січень, що дозволяє покращити точність прогнозування на 46,6% порівняно з моделями без явних сезонних ознак.

3. Удосконалено метод валідації прогностичних моделей епідемічних часових рядів для коротких вибірок від 72 місяців в умовах структурних зламів, заснований на стратегії *walk-forward* з послідовним розширенням навчального вікна та принципі строгої темпоральної каузальності, який відрізняється від існуючих обов'язковим обчисленням параметрів усіх трансформацій, включаючи нормалізацію та формування ознак, виключно на навчальній частині вибірки на кожному кроці, що дозволяє отримати реалістичну оцінку продуктивності моделі в режимі послідовного прогнозування та запобігає витоку інформації з тестового періоду.

4. Набула подальшого розвитку архітектура LSTM мереж для епідемічного прогнозування, заснована на принципі ієрархічного представлення часових залежностей із адаптацією складності моделі до обсягу та характеру доступних даних, яка відрізняється від існуючих систематичним емпіричним обґрунтуванням достатності компактних двошарових конфігурацій з 64 одиницями пам'яті на шар для коротких рядів від 72 місяців та необхідності глибоких тришарових пірамідальних архітектур з 11,89 мільйонами параметрів для рядів з множинними сезонними компонентами та структурними зламами, що дозволяє забезпечити навчання моделей за 40 секунд на стандартному CPU без GPU-прискорювачів та обирати архітектуру з мінімальними обчислювальними витратами без втрати точності прогнозування.

5. Набула подальшого розвитку архітектура інформаційної технології прогнозування інфекційної захворюваності, заснована на розділенні тренувального та інференсного шарів з версіонованим сховищем артефактів моделей, яка відрізняється від існуючих інтеграцією статистичних моделей ARIMA та нейромережових моделей LSTM у єдиний конвеєр з автоматизованим контролем якості через маніфести версій та метрики валідації, що дозволяє забезпечити відтворюваність прогнозів та оперативне перенавчання при надходженні нових епідемічних даних.

Практичне значення отриманих результатів. Створено комплекс програм на мові Python у середовищі Jupyter Notebook, які автоматизують розрахунки за усіма запропонованими методами та реалізують побудову, навчання та використання розроблених моделей LSTM та ARIMA для прогнозування епідемічних часових рядів із застосуванням бібліотек scikit-learn, TensorFlow/Keras, statsmodels, pandas, matplotlib та plotly.

Досягнуто підвищення точності прогнозування інфекційної захворюваності: для кумулятивних представлень часових рядів середня абсолютна відсоткова похибка становить менше одного відсотка, що на один-

два порядки нижче за рівневі прогнози. Похибка кумулятивного прогнозу гепатиту В (близько 400 випадків на річний горизонт) вписується в запас безпеки центрального медичного складу України, що свідчить про оперативну корисність для планування закупівель вакцин та антивірусних препаратів.

Розроблена інформаційна технологія забезпечує практичну застосовність в умовах обмежених обчислювальних ресурсів: компактні двошарові LSTM моделі навчаються за 40 секунд на стандартному процесорі класу ноутбука без необхідності GPU-прискорювачів, що дозволяє оперативне перенавчання при надходженні нових даних та розгортання в умовах обмеженої інфраструктури воєнного часу.

Результати дисертаційної роботи впроваджено у практичну діяльність Державної установи «Харківський обласний центр контролю та профілактики хвороб Міністерства охорони здоров'я України» (акт впровадження від 26 листопада 2025 р.); у навчальний процес кафедри епідеміології Харківського національного медичного університету (акт впровадження від 10 січня 2026 р.); у науково-дослідну роботу кафедри епідеміології Харківського національного медичного університету в рамках міжнародного проєкту STCU № 7136 «Моделювання та прогнозування поширення інфекцій у воєнних та повоєнних умовах з використанням даних епідеміологічного, поведінкового та геномного спостереження» (акт впровадження від 5 січня 2026 р.); у практичну діяльність Комунального некомерційного підприємства «Міська клінічна лікарня №13» Харківської міської ради (акти впровадження від 15 листопада 2025 р. та від 17 листопада 2025 р.); у практичну діяльність Комунального некомерційного підприємства Мереф'янської міської ради «Мереф'янська центральна районна лікарня» (акти впровадження від 2 грудня 2025 р.). Розроблені методи моделювання та прогнозування розвитку інфекційних захворювань на основі штучних нейронних мереж та статистичних моделей часових рядів активно використовуються в практичній діяльності зазначених установ для

прогнозування епідеміологічної ситуації, підтримки управлінських рішень та підвищення якості підготовки фахівців.

Особистий внесок здобувача. Всі основні наукові положення, результати, висновки та рекомендації дисертаційної роботи отримані автором самостійно та опубліковано у роботах [18–19, 28–34]. У публікаціях, які написано у співавторстві, здобувачеві належать такі результати: [18] – розробка архітектури LSTM мережі для прогнозування ГРВІ та експериментальна валідація з циклічними сезонними ознаками; [28] – моделювання динаміки грипу у Харківській області методом LSTM; [29] – контрфактуальний аналіз впливу воєнної міграції на динаміку COVID-19 у Польщі із застосуванням LSTM; [30] – розробка методології оцінювання впливу надзвичайних ситуацій на поширення інфекційних захворювань; [31] – моделювання гострого гепатиту А в Україні методом LSTM; [32] – прогнозування динаміки сальмонельозу методом LSTM; [33] – аналіз часових рядів лептоспірозу у країнах Балтії методом ARIMA; [34] – моделювання динаміки грипу у Харківській області методом ARIMA; [19] – прогнозування гострого та хронічного гепатиту В в Україні рекурентними нейронними мережами.

Апробація матеріалів дисертації. Основні положення, ідеї та висновки дисертаційної роботи доповідалися та обговорювалися на: IV Міжнародна науково-практична конференція ІТ-професіоналів і аналітиків комп'ютерних систем «ProfIT Conference 2021» (м. Харків, Україна, 2021); V Міжнародна науково-практична конференція ІТ-професіоналів і аналітиків комп'ютерних систем «ProfIT Conference 2023» (м. Харків, Україна, 2023); 4th International Workshop of IT-professionals on Artificial Intelligence (ProfIT AI 2024) (м. Кембрідж, Массачусетс, США, 2024); 7th International Conference on Informatics & Data-Driven Medicine (м. Бірмінгем, Велика Британія, 2024); 2024 14th International Conference on Dependable Systems, Services and Technologies (DESSERT) (м. Афіни, Греція, 2024); 2024 IEEE 17th International Conference on Advanced Trends in Radioelectronics, Telecommunications and Computer

Engineering (TCSET 2024) (м. Львів – Славське, Україна, 2024); International Conference «Integrated Computer Technologies in Mechanical Engineering – 2024» (м. Харків, Україна, 2024); 5th International Workshop of IT-professionals on Artificial Intelligence (ProfIT AI 2025) (м. Ліверпуль, Велика Британія, 2025); 2025 15th International Conference on Advanced Computer Information Technologies (ACIT) (м. Сібернік, Хорватія, 2025); International Conferences on Applied Computing and WWW/Internet 2025 (м. Порту, Португалія).

Публікації. Основні результати дисертації опубліковані у 23 наукових працях, у тому числі 5 статтях у періодичних виданнях, що індексуються у наукометричній базі даних Scopus, 8 публікаціях у матеріалах міжнародних науково-практичних конференцій, індексованих у Scopus, та 2 публікаціях у матеріалах українських науково-практичних конференцій. Окрім того, розроблено 8 комп'ютерних програм, на які отримано свідоцтва про реєстрацію авторського права.

Результати дисертаційної роботи, зокрема розроблені моделі, методи та програмне забезпечення, пройшли апробацію та отримали практичне підтвердження. Інтелектуальні моделі прогнозування динаміки грипу та гострих інфекцій на базі архітектур ARIMA та LSTM були представлені на міжнародних наукових та науково-практичних конференціях. Практична значущість проведеного дослідження підтверджена отриманням 8 свідоцтв про реєстрацію авторського права на комп'ютерні програми (№132251, №132250, №132249, №131623, №112762, №112634, №112605, №110034), що засвідчує прикладну цінність та новизну запропонованих рішень. Основна наукова новизна роботи полягає у розробленні та впровадженні комплексного підходу до прогнозування епідемічних процесів з використанням сучасних технологій машинного навчання та глибокого навчання.

Розроблена інформаційна технологія дозволяє проводити багатофакторний аналіз динаміки інфекційних захворювань з врахуванням різноманітних епідемічних та кліматичних чинників. Теоретико-методологічне

обґрунтування запропонованих моделей базується на сучасних математичних методах аналізу часових рядів та ґрунтується на багаторічному досвіді епідемічного нагляду. Результати дослідження мають значимість для галузі інфекційної епідеміології та можуть бути використані в системі охорони здоров'я як на регіональному, так і на національному рівні для підвищення ефективності заходів профілактики та контролю інфекційних захворювань.

Структура й обсяг дисертації. Дисертація складається з анотації, змісту, переліку умовних позначень та скорочень, вступу, чотирьох розділів, висновків, списку використаних джерел та додатків. Повний обсяг роботи становить 211 сторінок друкованого тексту, з них анотація – на 10 стор., зміст – на 3 стор., перелік умовних скорочень – на 3 стор., основний текст – на 147 стор., список із 143 використаних джерел – на 18 стор., додатки – на 12 стор. Дисертація містить 28 рисунків та 20 таблиць.



СПИСОК ПУБЛІКАЦІЙ ЗДОБУВАЧА ЗА ТЕМОЮ ДИСЕРТАЦІЇ

1. Butkevych, M., Yakovlev, S., Chumachenko, D. (2025). Data-Driven Forecasting of Acute and Chronic Hepatitis B in Ukraine with Recurrent Neural Networks. *Applied Sciences*, 15 (13), 7573. doi: 10.3390/app15137573. (Scopus, Q2)
2. Chumachenko, D., Butkevych, M., Lode, D., Frohme, M., Schmailzl, K.J.G., Nechyporenko, A. (2022). Machine Learning Methods in Predicting Patients with Suspected Myocardial Infarction Based on Short-Time HRV Data. *Sensors*, 22(18), p.7033. doi: 10.3390/s22187033. (Scopus, Q1)
3. Butkevych, M., Meniailov, I., Lutsiv, Y., Kraivskyi, B., Chumachenko, D. (2025). War-driven displacement and COVID-19 in Poland: simulation study using LSTM model. *Radioelectronic and Computer Systems*, 2025(4), pp.25–34. doi: 10.32620/reks.2025.4.02. (Scopus, Q3, Категорія А)
4. Butkevych, M., Chumachenko, D. (2024). Time series analysis of leptospirosis incidence for forecasting in the Baltic countries using the ARIMA model. *Radioelectronic and Computer Systems*, 2024(4), pp.5–19. doi: 10.32620/reks.2024.4.01. (Scopus, Q3, Категорія А)
5. Chumachenko, D., Bazilevych, K., Butkevych, M., Meniailov, I., Parfeniuk, Y., Sidenko, I., Chumachenko, T. (2024). Methodology for assessing the impact of emergencies on the spread of infectious diseases. *Radioelectronic and Computer Systems*, 2024(3), pp.6–26. doi: 10.32620/reks.2024.3.01. (Scopus, Q3, Категорія А)
6. Butkevych, M., Chumachenko, D. (2025). Simulation of Acute Upper Respiratory Tract Infections in Kharkiv Oblast (Ukraine): The ARIMA Model Approach. In: *Lecture notes in networks and systems*. Vol. 1473, Integrated Computer Technologies in Mechanical Engineering - 2024. Springer International Publishing, pp.153–161. doi: 10.1007/978-3-031-94845-9_13. (Scopus)

7. Butkevych, M., Vasiuk, K., Trofymova, I., Chumachenko, T. (2026). Forecasting of Acute Upper Respiratory Tract Infection in Ukraine during the War Conditions using LSTM Model. *CEUR Workshop Proceedings*. Vol. 4164. 5th International Workshop of IT-professionals on Artificial Intelligence (ProfIT AI 2025) 2025. pp.478–486. doi: doi.org/10.5281/zenodo.19435654 (Scopus)
8. Butkevych, M., Chumachenko, D. (2024). Simulation of Influenza Dynamics in Kharkiv Oblast (Ukraine) using ARIMA Model. *CEUR Workshop Proceedings*. Vol. 3777. 4th International Workshop of IT-professionals on Artificial Intelligence (ProfIT AI 2024) 2024, pp. 223-237. doi: doi.org/10.5281/zenodo.19413876 (Scopus)
9. Butkevych, M., Meniailov, I., Bazilevych, K., Parfeniuk, Y., Chumachenko, D. (2024). Simulation of influenza dynamics with LSTM deep learning model. *CEUR Workshop Proceedings*. Vol. 3892. 7th International Conference on Informatics & Data-Driven Medicine. pp.115–125. doi: doi.org/10.5281/zenodo.19414265. (Scopus)
10. Butkevych, M., Chumachenko, D. (2025). Simulation of Acute Hepatitis a in Ukraine Using LSTM Deep Learning Model. 2025 15th International Conference on Advanced Computer Information Technologies (ACIT). pp. 88–91. doi: 10.1109/acit65614.2025.11185739. (Scopus)
11. Butkevych, M., Chumachenko, T. and Chumachenko, D. (2024). Forecasting of Acute Upper Respiratory Tract Infections based on LSTM Model. 2024 14th International Conference on Dependable Systems, Services and Technologies (DESSERT). pp.1–5. doi: 10.1109/dessert65323.2024.11122125. (Scopus)
12. Butkevych, M., Manakova, O., Chumachenko, D. (2024). Forecasting of Salmonellosis Dynamics with LSTM Deep Learning Model. 2024 IEEE 17th International Conference on Advanced Trends in Radioelectronics,

- Telecommunications and Computer Engineering, TCSET 2024. pp.213–216.
doi: 10.1109/tcset64720.2024.10755553. (Scopus)
13. Chumachenko, D., Butkevych, M., Chumachenko, T., Kirpich, A., Kyrychenko, L., Matsyi, O. (2025). Deep Learning Forecasts of Hepatitis A, B, and C incidence in war-torn Ukraine: a Multivariate LSTM Approach. International Conferences on Applied Computing and WWW/Internet 2025. pp.321–323. doi: 10.33965/ac_icwi_2025_202508p047. (Scopus)
14. Chumachenko T., Butkevych M. Information technologies for the assessment of the impact of the russian war in Ukraine on public health. V Міжнародна науково-практична конференція ІТ-професіоналів і аналітиків комп'ютерних систем «ProfIT Conference 2023», С. 19-20.
15. Буткевич М.В., Чумаченко Д.І. (2021) Інформаційна система збору даних захворюваності на COVID-19. IV Міжнародна науково-практична конференція ІТ-професіоналів і аналітиків комп'ютерних систем «ProfIT Conference 2021», С. 16-17.
16. Чумаченко Д.І., Буткевич М.В., Комп'ютерна програма «*Модель динаміки групи на базі ARIMA*», Свідоцтво про реєстрацію авторського права на твір №132251, 20 грудня 2024.
17. Чумаченко Д.І., Буткевич М.В., Чумаченко Т.О., Комп'ютерна програма «*Інтелектуальна модель динаміки гострих інфекцій верхніх дихальних шляхів на базі архітектури LSTM*», Свідоцтво про реєстрацію авторського права на твір №132250, 20 грудня 2024.
18. Чумаченко Д.І., Буткевич М.В., Комп'ютерна програма «*Інтелектуальна модель динаміки групи на базі архітектури LSTM*», Свідоцтво про реєстрацію авторського права на твір №132249, 20 грудня 2024.
19. Чумаченко Д.І., Буткевич М.В., Чумаченко Т.О., Комп'ютерна програма «*Модель динаміки гострих інфекцій верхніх дихальних шляхів*

на базі *ARIMA*”, Свідоцтво про реєстрацію авторського права на твір №131623, 25 листопада 2024.

20. Буткевич М.В., Чумаченко Д.І., *Комп'ютерна програма «Нейромережева модель епідемічного процесу COVID-19 з архітектурою LSTM»*, Свідоцтво про реєстрацію авторського права на твір №112762 від 27.04.2022.
21. Буткевич М.В., Чумаченко Д.І., *Комп'ютерна програма “Модель Prophet для моделювання динаміки захворюваності на COVID-19 в Україні”*, Свідоцтво про реєстрацію авторського права на твір №112634 від 11.04.2022.
22. Буткевич М.В., Чумаченко Д.І., *Комп'ютерна програма “Модель авторегресійного ковзного середнього для прогнозування епідемічного процесу COVID-19”*, Свідоцтво про реєстрацію авторського права на твір №112605 від 7.04.2022.
23. Буткевич М.В., Чумаченко Д.І., *Комп'ютерна програма «Інформаційна технологія збору, обробки та аналізу даних захворюваності на COVID-19»*, Свідоцтво про реєстрацію авторського права на твір №110034 від 03.12.2021.

¶

РОЗДІЛ 1. АНАЛІЗ ПРОБЛЕМИ ПРОГНОЗУВАННЯ ЕПІДЕМІЧНИХ ПРОЦЕСІВ

1.1 Підходи до моделювання епідемічних процесів

Традиційно епідеміологічне моделювання включає декілька методологічних напрямків. Детерміністичні компартментні моделі представляють популяцію у вигляді дискретних груп індивідів з однорідними епідеміологічними статусами та залишаються золотим стандартом для оцінювання базових репродуктивних чисел [8, 9]. Стохастичні моделі враховують випадковість індивідуальних контактів у малих популяціях, де флуктуації можуть визначати траєкторію епідемії [10, 11]. Паралельно розвивалися методи статистичного моделювання часових рядів, зокрема ARIMA та SARIMA, які забезпечують інтерпретовані короткострокові прогнози без детальної специфікації біологічних механізмів [12, 13]. Останнє десятиліття ознаменувалося впровадженням методів машинного навчання, серед яких особливу увагу привертають мережі довгої короткочасної пам'яті LSTM [14, 16], що автоматично виявляють нелінійні залежності та адаптуються до структурних змін [16–17].

Компартментні моделі базуються на розділенні популяції на групи відповідно до епідемічного статусу, що описується системами диференціальних рівнянь [20–21]. Найпростіша модель SIR поділяє популяцію на сприйнятливих, інфікованих та одужалих індивідів [4]. Модель SEIR додає компартмент експонованих індивідів, які вже інфіковані, але ще не є заразними, що відповідає латентному періоду багатьох вірусних захворювань [22]. Для захворювань із втратою імунітету застосовуються моделі SIRS або SEIRS. Математичний аналіз зосереджується на визначенні базового репродуктивного числа, яке кількісно характеризує середню кількість вторинних інфекцій, викликаних одним інфікованим індивідом у повністю сприйнятливій популяції

[23–24]. Якщо базове репродуктивне число перевищує одиницю, інфекція поширюється; якщо менше одиниці, інфекція зникає. Сучасні розширення включають вікову стратифікацію з матрицями змішування [25–26] та метапопуляційні моделі для відтворення просторового поширення між регіонами [27].

Калібрування параметрів компартментних моделей вимагає інтеграції епідемічних та демографічних даних. Сучасні методи байєсівського виводу дозволяють інтегрувати попередні знання експертів із емпіричними даними та оцінювати невизначеність параметрів [35–36]. Чутливісний аналіз із використанням Latin Hypercube Sampling ідентифікує параметри, до яких модель є найбільш чутливою [37–38]. Стохастичні моделі формулюються як марковські процеси або описуються алгоритмом Гілеспі [39–41], що генерує траєкторії з випадковим вибором часу та типу подій.

Моделі часових рядів представляють альтернативний підхід, що ідентифікує статистичні закономірності у історичних даних та екстраполює їх у майбутнє [42–43]. ARIMA моделює поточне значення як лінійну комбінацію попередніх значень та помилок із диференціюванням для стаціонарності [12, 44]. Сезонна модель SARIMA додає періодичні компоненти для річних коливань. Процес побудови включає тести стаціонарності Дікі-Фуллера та Квятковського-Філіпса-Шмідта-Шіна [45–46], аналіз автокореляційних функцій та відбір моделі за інформаційними критеріями Акаїке та Байєса. Модель SARIMAX розширює SARIMA шляхом включення екзогенних змінних, таких як погодні умови або політичні заходи [47–48]. Перевагами є математична прозорість та інтерпретованість, обмеженням є припущення лінійності, що не дозволяє моделювати складні нелінійні взаємодії.

Рекурентні нейронні мережі розроблені для обробки послідовних даних із зворотними зв'язками для збереження інформації про попередні кроки [49–50]. Класичні рекурентні мережі страждають від проблеми зникаючого градієнта при навчанні на довгих послідовностях [51]. Мережі LSTM

вирішують це шляхом введення системи вентилів, що контролюють потік інформації [14, 52]. Кожна LSTM комірка містить вентиль забування, вентиль входу та вентиль виходу, що разом із станом комірки забезпечують довгострокову пам'ять. Процес навчання здійснюється методом зворотного поширення помилки у часі [53–54]. Адитивна структура оновлення стану дозволяє градієнтам протікати через багато кроків без деградації, що забезпечує моделювання довгострокових залежностей, таких як річні сезонні цикли.

У епідемічному прогнозуванні LSTM автоматично виявляють складні патерни без явної специфікації сезонних компонентів. Багатошарові LSTM створюють ієрархічні представлення, де нижні шари виявляють короткострокові патерни, а верхні інтегрують їх у довгострокові залежності. Регуляризація методом Dropout запобігає перенавчанню [52]. Попередня обробка включає нормалізацію та формування послідовностей методом ковзного вікна фіксованої довжини. Навчання здійснюється оптимізатором Adam [55] з функцією втрат як середньоквадратична або абсолютна похибка [56]. Основною перевагою є здатність моделювати нелінійні взаємодії без експертного втручання. Обмеженням є низька інтерпретованість навчених моделей та потреба у великих обсягах даних.

Гібридні підходи поєднують переваги різних методів [57–58]. Двоетапна схема використовує ARIMA для видалення лінійних компонентів, а LSTM для моделювання нелінійних залишків. Фінальний прогноз формується як сума виходів обох компонентів, що забезпечує інтеграцію інтерпретованої сезонності з гнучкими корекціями. Альтернативний підхід використовує компартментні моделі для генерації синтетичних сценаріїв, на яких навчається LSTM [59]. Методи асиміляції даних поєднують компартментні моделі з фільтрами Калмана для уточнення параметрів у реальному часі [60–61].

Вибір метрик оцінювання залежить від цілей застосування. Середня абсолютна похибка MAE вимірює помилки у фізичних одиницях.

Середньоквадратична похибка MSE надає більшої ваги великим помилкам. Середня абсолютна відсоткова похибка MAPE нормалізує помилки відносно фактичних значень та дозволяє порівнювати моделі на різних захворюваннях, однак має проблему з нульовими значеннями. Для прогнозування піків критичними є метрики точності визначення часу та величини піку для планування завантаження медичних закладів.

Валідаційні стратегії повинні відображати реалістичні умови експлуатації. Walk-forward валідація з пересуванням вікна навчання є золотим стандартом, оскільки імітує реальне прогнозування з періодичним перенавчанням моделі [12]. Така схема оцінює стабільність продуктивності у часі та виявляє періоди погіршення точності через структурні злами. Практичне впровадження вимагає прозорості результатів та комунікації невизначеності. Візуалізація прогнозів у вигляді інтервалів невизначеності дозволяє оцінювати ризики та формулювати сценарії реагування. Документування джерел даних та методів забезпечує відтворюваність результатів та довіру до прогнозів.

Україна, як країна, що зазнає військового конфлікту та надзвичайних подій, стикається з унікальними викликами у моніторингу та прогнозуванні інфекційних захворювань [62–65]. Порушення систем епідемічного нагляду, міграція населення, зниження охоплення вакцинацією та перевантаження медичної інфраструктури створюють умови для зростання захворюваності на керовані інфекції [66]. У таких умовах методи машинного навчання, здатні інтегрувати екзогенні змінні та адаптуватися до структурних зламів, набувають особливої цінності. Розробка інформаційних технологій прогнозування, що поєднують класичні епідемічні моделі з методами глибокого навчання та забезпечують практичну застосовність у умовах обмежених ресурсів, є актуальною науковою задачею для підтримки системи громадського здоров'я України.

Таким чином, епідемічне моделювання охоплює широкий спектр методологічних підходів. Детерміністичні компартментні моделі забезпечують теоретичне розуміння динаміки передачі та оцінку довгострокових втручань. Стохастичні моделі враховують випадковість подій у малих популяціях. Моделі ARIMA та SARIMA надають інтерпретовані короткострокові прогнози на основі історичних патернів. Мережі LSTM автоматично виявляють нелінійні залежності та адаптуються до структурних змін. Гібридні підходи поєднують переваги різних методів. Вибір підходу визначається цілями дослідження, характером даних, горизонтом прогнозування та потребою в інтерпретованості, а успішне впровадження вимагає міждисциплінарної співпраці епідеміологів, математиків та фахівців з обчислювальної техніки.

1.2 Аналіз джерел прогнозування гострого та хронічного гепатиту В на основі даних із використанням рекурентних нейронних мереж

Швидка конвергенція біологічних знань, статистичного навчання та оцінки політики характеризує сучасні дослідження з моделювання гепатиту В. Дослідники використовують дані спостереження високої роздільності, дедалі деталізованіші клінічні знання про кінетику вірусів та досягнення в обчислювальній потужності для побудови структур, що варіюються від стохастичних диференціальних систем до прогнозних моделей глибокого навчання. Ці зусилля спрямовані на прогнозування захворюваності з більшою точністю та перевірку впливу вакцинації, лікування та поведінкових втручань в умовах реальних обмежень. Інтегруючи механістичні деталі з калібруванням на основі даних, галузь переходить від ретроспективного опису до перспективної підтримки прийняття рішень, закладаючи основу для адаптивних стратегій громадського здоров'я як у стабільних, так і в кризових умовах.

Side та ін. [67] розширили класичне компартментне моделювання гепатиту В, запровадивши п'ятистанову структуру «Сприйнятливі–Експоновані–Інфіковані–Одужалі–Інфіковані» (SEIRI) та поєднавши її з графо-теоретичною процедурою для визначення базового числа відтворення. Автори встановили вільні від захворювання та ендемічні рівноваги, довели локальну стабільність через аналіз власних значень Якобіана та наближено визначили довгострокову динаміку для Макасару, Індонезія, шляхом чисельного інтегрування системи з припущеними наборами параметрів у Maple. Їхні симуляції свідчать, що захворюваність на гепатит В може або зменшуватися, або зберігатися залежно від невеликих змін у параметрах передачі та одужання. Це підкреслює чутливість результатів контролю до охоплення вакцинацією та лікуванням.

Fang та ін. [68] вирішили проблему постійно високого тягаря гепатиту В у Хайнані, Китай, поєднавши класичний аналіз часових рядів із машинним навчанням. Використовуючи дані провінційного спостереження (2017–2020), вони розклали місячну захворюваність на тренд та сезонність і оцінили окремі сезонні моделі ARIMA та сірі моделі GM(1,1). Потім вони інтегрували вихід ARIMA в нейронну мережу зворотного поширення для виявлення залишкової нелінійної структури. Перевірка на даних 2021 року показала, що гібридна специфікація SARIMA-BPNN досягла найнижчої помилки (MAPE = 0,087) і таким чином забезпечила найнадійнішу основу для прогнозування, передбачаючи зниження в усій провінції у 2022 році з найрізкішим падінням у березні.

Зосередившись на трансляційному моделюванні, а не на лабораторній вірусології, Voivin-Champeaux та ін. [69] представили навчальний синтез того, як математичні представлення відображають природний перебіг гострого та хронічного HBV. Після стислого огляду вірусного проникнення, реплікації та кінетики біомаркерів, автори розібрали послідовні покоління моделей, що включали ключові біологічні механізми, такі як нецитолітичне кліренс,

адаптивний імунітет, проліферація гепатоцитів та внутрішньоклітинний обіг сссDNA, коментуючи їхні припущення, властивості стабільності та клінічну інтерпретованість. Таблиці параметрів, лістинги коду та два інтерактивні додатки R Shiny дозволяють читачам експериментувати з сімома зразковими моделями, показуючи, як структурні вибори та значення параметрів змінюють прогнозовані траєкторії вірусного навантаження та ілюструючи найкращі практики при поєднанні механізмів прогресування захворювання з фармакокінетично-фармакодинамічними аналізами.

Li та ін. [70] представляють дослідження на основі даних, у якому автори розробили нейронну мережу довгої короткочасної пам'яті з механізмом уваги (A-LSTM) для прогнозування щомісячних нових інфекцій гепатиту В у континентальному Китаї з 2004 по 2017 рік і оцінили її продуктивність порівняно з нейронною мережею зворотного поширення (BPNN). Використовуючи дані спостережень з Китайського центру даних науки громадського здоров'я, автори навчили двошарові моделі A-LSTM з різними розмірами прихованих одиниць, вибрали конфігурацію з п'ятьма одиницями на основі мінімізації помилки та повідомили про значно нижчі RMSE (1780) та MAPE (1,79%), ніж найкраща альтернатива BPNN (RMSE = 3519; MAPE = 3,86%). Механізм уваги ефективніше фіксує часові залежності, ніж традиційні базові моделі послідовність-до-послідовності, забезпечуючи ближче узгодження між прогнозованою та спостережуваною захворюваністю для більшості місяців у тестовому періоді 2018 року.

Xu та ін. [71] розширили традиційні структури гепатиту В у стилі SEIR, додавши окремий компартмент «потенційно інфікованих» для представлення латентних носіїв та підігнавши отриману семикласову модель диференціальних рівнянь до даних спостереження Китаю за 2003–2021 роки. Оцінка параметрів поєднувала нелінійні найменші квадрати з генетичним алгоритмом, що дало контрольне число відтворення $R_c = 1,741$ та оцінку приблизно 450 000 латентних носіїв. Аналіз чутливості та сценаріїв показав, що

невдача вакцинації, базова трансмісивність та швидкість перетворення латентного стану в носійство визначають R_0 . Симуляції також показали, що підвищення охоплення вакцинацією дорослих та покращення ефективності вакцини зменшить захворюваність.

Ma J. та Ma S. [72] сформулювали та проаналізували чотирикомпартментну стохастичну модель гепатиту В, яка поєднує насичену нелінійну передачу з функцією медіа-охоплення, що зменшує ефективний контакт у міру зростання обізнаності населення. Екологічна мінливість була введена через незалежні гаусові збурення білого шуму, призначені кожній змінній стану, і автори вивели порогові умови для майже напевного вимирання та існування унікального ергодичного стаціонарного розподілу. Чисельні експерименти, калібровані до даних спостереження Китайського центру з контролю захворювань (2005–2021), відтворили історичну траєкторію захворюваності та припустили, що при поточній інтенсивності контролю загальнонаціональні показники стабілізуються приблизно на рівні 50–60 випадків на 100 000 людей.

Дослідження Zhao та ін. [73] проаналізувало вісім років щомісячних підрахунків випадків гепатиту В з 31 провінції континентального Китаю (2013–2020) і показало, що захворюваність демонструє чіткий річний цикл із спадами в лютому та піками в березні. Після підтвердження стаціонарності через логарифмічне перетворення та сезонне диференціювання, автори визначили SARIMA(1,0,0)(0,1,1) як найкращу модель підгонки, перевірили, що її залишки утворюють білий шум, і продемонстрували хорошу відповідність між підігнаними та спостережуваними значеннями. Валідація на даних 2021 року вказала, що всі прогнозні точки потрапили в межі 95% довірчого інтервалу. Це свідчить про те, що проста сезонна специфікація ARIMA може забезпечити надійні короткострокові прогнози для національного спостереження та планування ресурсів.

De Villiers та ін. [74] провели пряме порівняння моделі Imperial HBV та моделі PRoGReSs Фонду CDA, подаючи обом гармонізовані демографічні дані, дані про поширеність та охоплення вакцинацією для Ефіопії, Індії, Нігерії та Пакистану, щоб дослідити, як масштабування трьохдозової серії для немовлят та своєчасної дози при народженні до 95% охоплення вплине на траєкторії гепатиту В до 2099 року. Незважаючи на подібні базові підгонки, моделі розійшлися після зміни втручань; PRoGReSs прогнозував більші вигоди від розширення серії для немовлят, оскільки дозволяє цьому графіку блокувати значну частку перинатальних інфекцій, тоді як Imperial приписував більшу користь дозі при народженні через припущення, що перинатальна передача обмежувалася лише вакцинацією в перші 24 години. Цей структурний контраст призвів до десятикратної різниці в оцінених випадках та запобігання смертям у деяких сценаріях. Це ілюструє, як висновки політики можуть залежати від неспостережуваних механізмів, вбудованих у структури симуляції.

El Koufi та Rao [75] розробили стохастичну модель передачі гепатиту В, яка поєднала безперервний екологічний шум з дискретними змінами режиму, керованими ланцюгом Маркова зі скінченим числом станів. Структура розділила популяцію на сприйнятливі, гостро інфіковані, хронічно інфіковані та одужали класи; додала мультиплікативні збурення білого шуму до кожного рівняння; та дозволила ключовим параметрам (наприклад, передача, вакцинація, смертність) перемикатися між альтернативними екологічними станами. Автори встановили глобальне існування, унікальність та позитивність рішень; вивели пороговий R_{sw} , який визначав, чи зберігається інфекція навколо унікального ергодичного стаціонарного розподілу, чи вимирає майже напевно; та підтвердили свої аналітичні результати симуляціями за схемою Мільштейна, які ілюстрували сценарії збереження та вимирання.

Cheng та ін. [76] побудували вікову гетерогенну модель передачі SEICR для гепатиту В у Китаї, яка розділила дорослих на три когорти (15–25, 25–70, ≥ 70 років), включила переважне змішування між віковими групами та явно

представила вакцинацію як потік від сприйнятливого до одужалого класу. Параметри моделі були підігнані до національних кумулятивних підрахунків випадків з 2004 по 2018 рік нелінійними найменшими квадратами. Потім були досліджені латинський гіперкуб/PRCC аналіз чутливості, збурення швидкості контактів та набір сценаріїв охоплення вакцинацією. Симуляції показали, що захворюваність визначається переважно когортою 25–70 років. Зниження середніх показників контактів само по собі має обмежений вплив, тоді як масштабування бустерної вакцинації у популяції 15–70 років до 90% може знизити кумулятивні інфекції приблизно на 60% та згладити прогнозовану епідемічну криву достатньо, щоб досягти цілі ВООЗ щодо зменшення захворюваності до 2030 року.

Короткий аналіз сучасного стану досліджень представлений у таблиці 1.1

Таблиця 1.1 Сучасний стан досліджень щодо моделювання гепатиту В.

Публікація	Завдання	Метод	Результати
Side S. та ін. [67]	Побудувати та проаналізувати математичну модель, яка відображає поширення гепатиту В в Макасарі та визначити умови, що ведуть до ліквідації або збереження захворювання.	Дослідження формулює нелінійну SEIRI систему звичайних диференціальних рівнянь, застосовує теорію графів для отримання R_0 , проводить аналіз локальної стабільності через власні значення та виконує сценарій симуляції в Maple.	Модель дає аналітичний вираз для R_0 ; симуляції показують, що коли $R_0 < 1$ інфекція зменшується, тоді як помірне збільшення параметрів передачі штовхає $R_0 > 1$ і підтримує ендемічність, підкреслюючи важливість цільових втручань.

Продовження таблиці 1.1

Fang K. та ін. [68, 77]	Моделювати та прогнозувати щомісячну захворюваність на гепатит В у провінції Хайнань для інформування місцевих стратегій профілактики та контролю.	Дослідники порівняли SARIMA, GM(1,1) та комбіновану модель SARIMA–BPNN, навчену на даних захворюваності 2017–2020 рр., та оцінили продуктивність на спостереженнях 2021 р.	Модель SARIMA–BPNN перевершила альтернативи з одним методом і спрогнозувала поступове зниження захворюваності на гепатит В протягом 2022 р., з найбільшим щомісячним зниженням на початку весни.
Voivin-Champeaux S. та ін. [69]	Скласти педагогічну, але критичну дорожню карту моделей прогресування захворювання HBV, яка керує практиками у виборі, удосконаленні та застосуванні механістичних структур для оцінки лікування.	Автори переформулювали репрезентативні моделі зростаючої біологічної деталізації, проаналізували їхню динаміку та упаковали код і набори параметрів у відкриті додатки R Shiny для інтерактивної симуляції.	Стаття показує, що скромні зміни в структурі моделі або параметризації можуть глибоко змінити прогнози вірусного кліренсу та ушкодження печінки, підкреслюючи необхідність узгодження складності моделі з клінічним питанням і використання прозорого, заснованого на даних калібрування при прогнозуванні терапевтичного впливу.

Продовження таблиці 1.1

Li Y. та ін. [70]	Побудувати точну модель для короткострокового прогнозування щомісячних випадків гепатиту В в Китаї.	Дослідження навчає та порівнює двошарову LSTM мережу на основі уваги та багатшарову BPNN, використовуючи дані спостережень 2004–2017 рр.	A-LSTM перевершує BPNN і попередні базові моделі часових рядів, зменшуючи помилку прогнозування приблизно на 50% і пояснюючи 87% дисперсії в утримуваних щомісячних підрахунках.
Xu C. та ін. [71]	Кількісно оцінити прихований тягар латентних носіїв HBV в Китаї та оцінити, як альтернативні стратегії вакцинації впливають на довгострокову динаміку передачі.	Автори побудували семикомпартментну детерміністичну модель, калібровану до 19 років національних даних про випадки, використовуючи нелінійні найменші квадрати та генетичний алгоритм.	Підгонка моделі припускає 449 535 (95% ДІ 415 651–483 420) латентних носіїв і показує, що зниження рівня невдач вакцинації та розширення імунізації дорослих може знизити R_c нижче одиниці, призводячи до зниження захворюваності.
Ma J. та Ma S. [72]	Оцінити, як медіа-охоплення та екологічний шум впливають на довгострокову динаміку гепатиту В в Китаї.	Розробили стохастичну модель S-I1-I2-R з модульованими медіа, насиченими термінами хвороб, доводять аналітичні критерії вимирання та збереження; моделюють систему національних даних 2005–2021 рр.	Посилені медіа-відповіді та вища інтенсивність шуму можуть знизити стохастичне число відтворення нижче одиниці, призводячи до вимирання; прогножуючи стаціонарний розподіл, що прогнозує довгострокове плато захворюваності ~50 на 100тис.

Продовження таблиці 1.1

Zhao та ін. [73]	Побудувати модель часових рядів для точного прогнозування щомісячних повідомлень про гепатит В по всьому Китаю для цілей раннього попередження.	Дослідники підігнали кілька кандидатів сезонних ARIMA до логарифмічно перетворених даних спостережень 2013–2020 рр., вибрали оптимальну специфікацію за допомогою BIC і діагностики залишків та перевірили її на спостереженнях 2021 р.	SARIMA(1,0,0)(0,1,1) ₁₂ відтворила історичні тренди та отримала прогнози 2021 р. в межах 95% інтервалу прогнозування, підтверджуючи її придатність для короткострокового моніторингу гепатиту В в Китаї.
De Villiers M.J. та ін. [74]	Визначити, чи дві провідні моделі передачі гепатиту В дають узгоджені прогнози та кількісно оцінити, як масштабування вакцинації немовлят та дози при народженні впливає на майбутні інфекції та смерті.	Стаття заповнила детерміністичні моделі ідентичними країно специфічними даними, симулювала три сценарії охоплення вакцинацією з 2015 по 2099 рр. та порівняла результати для інцидентних хронічних інфекцій та смертності, пов'язаної з HBV.	Хоча базові епідемії узгоджувалися, PROGRess прогнозував більші зниження від масштабування серії для немовлят, а Imperial від розширення дози при народженні, виявляючи структурну невизначеність, яка повинна інформувати обговорення політики вакцинації.

Продовження таблиці 1.1

El Koufi A. та Rao N. [75]	Дослідити, як комбіновані збурення білого шуму та перемикання режиму впливають на динаміку гепатиту В та визначити умови збереження або вимирання хвороби.	Стаття сформулювала стохастичну систему диференціальних рівнянь S-I1-I2-R з марківським перемиканням, довела теореми про коректність та порогові значення, використовуючи функції Ляпунова та мартингальні техніки, та підтвердила висновки через чисельне моделювання.	Якщо $RSW > 1$, процес збігається до ергодичного стаціонарного розподілу, що означає довгострокове збереження; якщо $RSW < 1$, гостро інфікований клас експоненціально спадає до нуля, призводячи до вимирання з ймовірністю одиниця.
Cheng K. та ін. [76]	Оцінити, як альтернативні віко специфічні стратегії бустерної вакцинації можуть стримати передачу гепатиту В в Китаї.	Розроблено багатогрупову детерміністичну модель SEICR та калібровано з переважним змішуванням. Потім використано аналіз чутливості та сценаріїв для кількісної оцінки ефектів варіювання схем контактів та рівнів вакцинації.	Досягнення охоплення вакциною $\geq 90\%$ у дорослих віком 15–70 років значно перевершує заходи зниження контактів, скорочуючи кумулятивні випадки приблизно на 60% і виводячи національну захворюваність на шлях до цілі ліквідації ВООЗ.

Сучасна література показує, що точність прогнозування та політична релевантність залежать так само від прозорих структурних припущень і

надійного калібрування, як і від алгоритмічної складності. Моделі машинного навчання відмінні в короткостроковому прогнозуванні, тоді як механістичні симулятори залишаються незмінними для дослідження контрфактуальних сценаріїв. Небагато досліджень поєднують ці сильні сторони в одному інструменті або розглядають порушення даних, типові для контекстів, що зазнали конфлікту, таких як Україна. Більше того, розбіжності між провідними моделями вакцинації підкреслюють стійку структурну невизначеність, а обмежена увага до латентних носіїв, програм бустерної вакцинації дорослих та динаміки дотримання залишає важливі питання без відповіді. Це дослідження просувається шляхом розробки рекурентних нейронних мереж для прогнозування гострого та хронічного гепатиту В в Україні в умовах триваючих порушень.

1.3 Аналіз застосування моделі ARIMA у симуляції та прогнозуванні динаміки грипу

Застосування моделі ARIMA у симуляції та прогнозуванні динаміки грипу привернуло значну увагу в епідемічних дослідженнях. Моделі ARIMA використовувалися для аналізу даних часових рядів випадків грипу, що дозволяє дослідникам прогнозувати майбутні тенденції та розуміти часові патерни спалахів. Ці моделі ефективно відображають сезонність і коливання захворюваності на грип, надаючи цінну інформацію для планування охорони здоров'я та стратегій втручання. Гнучкість ARIMA у роботі з нестационарними даними та її здатність включати як авторегресійні, так і компоненти ковзного середнього роблять її відповідним інструментом для моделювання складної поведінки передачі грипу. Дослідники застосовували моделі ARIMA у різних географічних регіонах та контекстах, демонструючи їх адаптивність та ефективність у різних епідемічних умовах.

Дослідження [78] зосереджується на епідеміології вірусів грипу серед дітей у Ухані, Китай, використовуючи дані з десятирічного періоду спостереження та прогнозуючи майбутні тенденції за допомогою моделей ARIMA. Основною метою був аналіз показників позитивності типів вірусу грипу та застосування моделей часових рядів для короткострокових прогнозів. Однією з переваг цього дослідження є комплексний збір даних за значний період, що дозволяє здійснити більш надійне статистичне моделювання. Однак дослідження має обмеження. По-перше, не вдалося забезпечити однорідність якості мазків із горла в різних лікарнях, що потенційно впливає на точність виявлення грипу. Крім того, обмежена кількість мазків, взятих щотижня з лікарні з великим потоком пацієнтів, може не повністю представляти ширшу популяцію, обмежуючи узагальнюваність результатів. Покращення однорідності збору зразків та розширення охоплення спостереження могло б покращити майбутні дослідження.

Стаття [79] порівнює ефективність моделей ARIMA та LSTM у прогнозуванні тенденцій грипу, використовуючи дані про якість повітря з центрального Тайваню між 2014 та 2018 роками. Обидві моделі були застосовані до наборів даних із трьох регіонів: округу Чжанхуа, міста Тайчун та округу Наньтоу. Моделі LSTM, зокрема варіант, що включає відбір ознак класифікатора Extra Trees, продемонстрували кращу точність порівняно з ARIMA. Це було особливо помітно в місті Тайчун, де LSTM ETC покращив точність прогнозування на 97%. Однак дослідження має обмеження. Обсяг прогнозування значною мірою залежить від даних про якість повітря та метеорологічних даних, які можуть не враховувати інші фактори, що впливають на тенденції грипу, такі як поведінка населення або втручання в охорону здоров'я. Крім того, використання обмеженої географічної площі та часових рамок може обмежити узагальнюваність результатів на інші регіони чи довші періоди.

Дослідження [80] вивчає використання моделей ARIMA та LSTM для прогнозування грипоподібних захворювань та респіраторних захворювань на основі даних про якість повітря з Тайваню. Порівнюючи обидві моделі на п'яти- та десятирічних наборах даних, результати демонструють, що моделі LSTM перевершують ARIMA щодо точності прогнозування, особливо при розширенні на довші періоди. Моделі LSTM також виграють від методів відбору ознак, таких як матрична кореляція та класифікатор додаткових дерев, які покращують результати прогнозування. Однак ключовим обмеженням дослідження є його залежність від даних з одного географічного регіону, що може обмежити узагальнюваність результатів на інші місця.

Крім того, хоча моделі LSTM демонструють вищу точність, вони вимагають значно більше обчислювальних ресурсів, ніж ARIMA, що може бути складним у масштабних застосуваннях. Крім того, майбутня робота може дослідити інтеграцію даних у реальному часі та подальші вдосконалення відбору ознак для підвищення надійності прогнозування.

Стаття [81] порівнює прогностичні можливості моделей ARIMA, SARIMA та Random Forest (RF) у прогнозуванні частоти вірусу грипу А в популяціях свиней Онтаріо. Модель RF постійно перевершувала моделі ARIMA та SARIMA у прогнозуванні збільшення діагностичних та позитивних вірусологічних подань, особливо на тижневих та місячних інтервалах. Модель RF продемонструвала більшу чутливість та нижчі показники помилок, що робить її більш надійним інструментом для цілей спостереження. Однак дослідження має обмеження. Використані дані базувалися на добровільних діагностичних поданнях, які можуть не повністю представляти популяцію свиней, що призводить до потенційних упереджень у результатах. Крім того, аналіз не включав екологічні фактори, які могли б вплинути на тенденції IAV, обмежуючи комплексність прогностичних моделей.

Дослідження [82] вивчає зв'язок між захворюваністю на грип та різними кліматичними показниками в Гуансі, Китай, використовуючи моделі ARIMA

та ARIMAX для прогностичного аналізу. Результати вказують на те, що змінні забруднення повітря, такі як PM2.5 та SO₂, та метеорологічні фактори, такі як температура та вологість, значно впливають на захворюваність на грип, причому PM2.5 демонструє найсильнішу позитивну кореляцію. Модель ARIMAX, яка включає ці екзогенні змінні, працювала краще, ніж стандартна модель ARIMA, пропонуючи покращену точність прогнозування. Однак дослідження обмежене його залежністю від даних з одного регіону, що може обмежити узагальнюваність результатів на інші географічні регіони чи клімати. Крім того, не враховувався вплив інших потенційних факторів, таких як втручання в охорону здоров'я або рух населення, що може вплинути на точність прогнозів.

Стаття [83] порівнює моделі ARIMA та експоненційного згладжування Холта-Вінтерса (HWES) для прогнозування спалахів грипу на основі даних Twitter з Австралії, зібраних між 2016 та 2017 роками. Обидві моделі були протестовані на їх здатність прогнозувати випадки грипу шляхом аналізу твітів, пов'язаних з симптомами грипу, та порівняння їх із підтвердженими випадками з Центрів контролю захворювань. Модель ARIMA показала кращу прогностичну точність із середньою відносною похибкою 7,26%, порівняно з 11,28% для моделі HWES. Основна перевага цього дослідження полягає у використанні даних соціальних медіа, які дозволяють раніше виявляти тенденції грипу порівняно з традиційними звітами CDC. Однак дослідження має обмеження, включаючи властивий шум у даних соціальних медіа, оскільки твіти, пов'язані з грипом, не завжди відповідають підтвердженим випадкам.

Стаття [84] оцінює ефективність різних моделей часових рядів, ARIMA, SARIMA та XGBoost, у прогнозуванні щомісячних сезонних випадків грипу в Саудівській Аравії. Аналіз демонструє, що модель XGBoost значно перевершує як моделі ARIMA, так і SARIMA за точністю, про що свідчать нижчі середня абсолютна похибка, середня квадратична похибка та середньоквадратичне відхилення. Здатність моделі XGBoost захоплювати складні, нелінійні зв'язки

добре підходить для прогнозування грипу. Однак одним обмеженням дослідження є його залежність від відносно короткого набору даних, що охоплює лише п'ять років, що може не повністю відображаються довгострокові тенденції та сезонні варіації. Крім того, дослідження не враховує інших потенційних факторів впливу, таких як втручання в охорону здоров'я або демографічні змінні, які могли б додатково покращити прогностичну потужність моделей.

Незважаючи на широке застосування моделей ARIMA для прогнозування динаміки грипу в різних регіонах та контекстах, залишається прогалина в дослідженнях, зосереджених на надзвичайних умовах, таких як Харківська область. Унікальні виклики, спричинені повномасштабним російським вторгненням, вимагають індивідуального підходу до моделювання грипу. Наше дослідження має на меті заповнити цю прогалину шляхом застосування моделі ARIMA для симуляції та прогнозування тенденцій грипу в Харківській області за цих атипових умов. Зосереджуючись на цих надзвичайних умовах, ми робимо внесок у існуючу базу знань, демонструючи адаптивність та ефективність моделювання ARIMA в середовищах з обмеженими даними та посиленими проблемами охорони здоров'я. Це дослідження покращує розуміння динаміки грипу в регіонах, уражених конфліктом, і підкреслює критичну роль методологій, заснованих на даних, в інформуванні про втручання в охорону здоров'я під час криз.

1.4 Аналіз поточного стану дослідження часових рядів захворюваності на лептоспіроз для прогнозування

Існуючі дослідження ілюструють корисність таких моделей, як ARIMA, у прогнозуванні траєкторій захворювання, тим самим сприяючи превентивному плануванню охорони здоров'я. Ці дослідження відображають міждисциплінарний підхід, що поєднує епідеміологію, біостатистику та

обчислювальну біологію для ефективного прогнозування захворюваності та поширення лептоспірозу.

Моделі симуляції пропонують цінну інформацію про часові та просторові патерни захворювання, включаючи сезонні тенденції та географічні осередки, які є невід'ємною частиною для розробки цілеспрямованих, своєчасних втручань. Проте, хоча й досягнуто прогресу, існує виразна потреба в подальших дослідженнях, особливо в таких регіонах, як країни Балтії, де епідеміологічний слід захворювання все ще встановлюється. Включення нюансованих локальних факторів, таких як клімат, біорізноманіття та соціально-економічні детермінанти, в ці моделі могло б додатково підвищити їх прогностичну точність та корисність у цих контекстах.

У дослідженні [85] представлено інноваційний баєсівський метод виводу, розроблений спеціально для моделі регресії Zero-Modified Poisson (ZMP), що демонструє похвальну гнучкість в аналізі даних підрахунків, незалежно від наявності інфляції або дефляції нулів у вибірці. Запропонована методологія включає широкий клас апіорних щільностей на основі інформаційної матриці для параметрів моделі. Метод використовує дослідження чутливості для ідентифікації переконливих випадків, які могли б змінити результати. Метод використовує дивергенцію Кульбака-Лейблера як міру, а дослідження симуляції додатково підтверджують висновки. Застосування моделі на реальних наборах даних про повідомлення про лептоспіроз зі штату Баїя в Бразилії додатково підтверджує її корисність.

Підсумовуючи, моделі регресії ZMP пропонують значні переваги в аналізі наборів даних з інфляцією або дефляцією нулів і підкреслюють розробку баєсівського підходу на основі апіорної інформаційної матриці. Крім того, вона розробляє ефективну міру на основі дивергенції Кульбака-Лейблера для цієї моделі. Дослідження показує, що в реальних застосуваннях коваріата Індексу людського розвитку (ІЛР) є значущим фактором у поясненні повідомлень про лептоспіроз, оскільки ймовірність повідомлення про

лептоспіроз зростає з ІЛР. Однак вона також попереджає, що низька кількість повідомлень про лептоспіроз у будь-якому місті не означає відсутності захворювання, а може відображати спроможність системи охорони здоров'я виявляти випадки захворювання.

Дана стаття [86] представляє нову методологію для чисельного моделювання новоствореної моделі лептоспірозу дробового порядку (FOLM), яка використовує можливості стохастичних чисельних контрольованих нейронних мереж. Це дослідження становить інноваційне чисельне дослідження моделі лептоспірозу, категоризоване на п'ять динамік. Проблема біологічної FOLM розглядається з урахуванням різних значень похідної дробового порядку. Числові формулювання FOLM створюються з використанням контрольованих нейронних мереж (SNNs), а обчислювальні характеристики оцінюються через призму зворотного поширення Левенберга-Марквардта (LVMBP), також відомого як SNNs-LVMBP. Валідність запропонованого підходу оцінювалася шляхом порівняння отриманих рішень з еталонними рішеннями. Статистика показала, що процеси сертифікації та навчання склали 74% та 13% дослідження відповідно.

Автори цієї статті [87] використали модель Сприйнятливі-Інфіковані-Одужалі (SIR) для кращого розуміння динаміки передачі лептоспірозу, глобально поширеного зоонозу, який зазвичай передається гризунами та часто призводить до летальних наслідків у людей. Це дослідження є особливо схильним до спалахів після сильних опадів і повеней і ретельно вивчає фактори, що впливають на цю динаміку передачі. Стаття також охоплює аналіз біфуркації та чисельні рішення моделі, відзначаючи сильну узгодженість між теоретичними висновками та чисельними симуляціями. Критично, дослідження ідентифікує природну смертність популяції щурів як значущий фактор для контролю лептоспірозу, поряд з базовим репродуктивним числом, яке відіграє життєво важливу роль в епідеміології цього захворювання.

Стаття [88] пропонує інноваційну структуру для управління лептоспірозом, що переносяться гризунами, використовуючи теорію математичної моделі оптимального контролю, розглядаючи обмеження традиційних методів контролю, таких як застосування родентицидів та управління середовищем існування. Лептоспіроз, в основному поширюваний через взаємодію з тваринами або середовищами, забрудненими лептоспірами в сечі тварин, представляє значні виклики контролю через складність і вартість управління популяціями резервуарів. Керуючись емпіричними даними з Сальвадора, Бразилія, дослідження розробило вікову структуровану модель для зараження лептоспірами популяції норвезьких щурів, розширюючи її для включення двох тимчасових заходів контролю, а саме, родентицидів та зменшення ресурсів, та двох постійних заходів контролю, зменшуючи місткість для щурів і тривалість життя лептоспір у довкіллі.

Теорія оптимального контролю застосовується для визначення ідеальних залежних від часу контролів з урахуванням вартості заходів контролю та суспільної "вартості" інфекції. Результати свідчать, що постійні контролі можуть зменшити поширеність носійства лептоспір у популяціях гризунів, а тимчасові контролі можуть ефективно зменшити кількість заражених щурів, тим самим пом'якшуючи ризик зараження людей. Хоча це дослідження зосереджувалося на норвезькому щурі, його підхід застосовується до інших систем захворювань з тваринними та екологічними резервуарами, надаючи цінний інструмент для обґрунтованого прийняття рішень у сфері охорони здоров'я.

Дослідження [89] нещодавно введених кусково-класичних-глобальних та класично-дробових операторів застосовується для дослідження динаміки моделі захворювання лептоспірозом. Це дослідження вивчило існування та унікальність рішень для кускових похідних, пов'язаних з моделлю, використовуючи кусковий ітеративний поліноміальний метод Ньютона для отримання приблизного рішення. Також встановлено числову схему для

кускової моделі лептоспірозу з цілими та дробовими порядками. Ми спостерігали покращену динаміку та поведінку переходу на основі результатів симуляції для обох операторів. Було виявлено, що одужала популяція людей поступово зменшуватиметься протягом 350 днів, що означає, що хвороба зрештою зникне. Більше того, високе значення β зменшило сприйнятливість популяції щурів та людей, які зрештою стали стабільними. Порівняно з високими значеннями β , заражена людська популяція зменшилася та стабілізувалася швидше.

Дослідження [90] розробило та вивчило компартментну математичну модель для дослідження впливу лептоспірозу, переносимого гризунами, на людські популяції, враховуючи патогенні агенти захворювання в довкіллі та рівень захворюваності людей через взаємодію між зараженими гризунами та довкіллям. Досліджуються основні властивості моделі, точки рівноваги та аналіз стабільності, з базовим репродуктивним числом R_0 , виведеним за допомогою методу матриці наступного покоління. Аналіз стабільності показує, що рівновага без хвороби є глобально асимптотично стабільною, якщо $R_0 < 1$, і нестабільною в іншому випадку, і модель демонструє пряму біфуркацію.

Аналіз чутливості ідентифікує ключові параметри, що впливають на результати моделі. Чисельні симуляції з використанням методу Рунге-Кутти четвертого порядку додатково продемонстрували поведінку стабільності моделі та вплив передачі людям, одужання та смертності гризунів на динаміку моделі. Результати вказують на те, що траєкторії рішень моделі еволюціонують до унікальної ендемічної рівноваги з часом, коли $R_0 > 1$, і що зменшення швидкості передачі, збільшення швидкості одужання та контроль популяції гризунів значно пом'якшують поширення захворювання.

Автори статті [91] прагнули дослідити зв'язок між погодними параметрами при різних часових затримках та виникненням лептоспірозу в Малайзії, де захворюваність на лептоспіроз зростає. Використовуючи методи видобування даних та машинного навчання, дослідження застосовує

розвідувальний аналіз даних (EDA) для визначення оптимальних часових затримок для опадів та температури. Далі, на основі навчання зворотного поширення та оптимізованих прихованих шарів та вузлів, була розроблена модель штучної нейронної мережі (ANN) для класифікації вибраних ознак на виникнення захворювання та не-виникнення. Дослідження виявило сильну кореляцію між виникненням лептоспірозу та середньотижневою температурою з затримкою в 16 тижнів і тижневою кількістю опадів при 12...20 тижнях. Модель ANN, яка була розроблена з використанням цих обраних ознак, продемонструвала високі рівні точності, чутливості та специфічності, збільшивши точність прогностичної моделі на 13,30...31,26% від базових моделей.

Дане дослідження [92] зосереджувалося на розробці обчислювальної моделі на основі клітинних автоматів (CA) для ілюстрації поширення лептоспірозу, захворювання, яке зазвичай передається від великої рогатої худоби щурів до людей. Дослідники використовували модель Сприйнятливі-Інфіковані-Одужали-Сприйнятливі (SIRS) та інноваційно включили правило на основі голосування для вдосконалення традиційного набору правил CA. Проводячи симуляції з використанням фактичних даних про інфекції лептоспірозом у Таїланді протягом 2000 та 2001 років, модель продемонструвала чудову точність, тісно узгоджуючись з даними інфекцій у реальному часі. Це означає життєздатність моделі на основі CA та введеного правила на основі голосування для реалістичного відображення динаміки передачі лептоспірозу.

У дослідженні [93] математична модель, що представляє передачу інфекційного захворювання лептоспірозу, була вивчена шляхом використання системи нелінійних звичайних диференціальних рівнянь. Визнаючи властиву складність отримання точного рішення для цієї системи, автори використовують гомотопічний метод збурення Хе (HPM) для отримання приблизного рішення. Результати HPM порівнюються з отриманими з

використанням методу Рунге-Кутти четвертого порядку (RK4). Ілюстративні графіки включені для демонстрації простоти та надійності методу, тим самим підтримуючи корисність гомотопічного підходу збурення для аналізу складних моделей передачі захворювань, таких як лептоспіроз.

Стаття [94] ретельно вивчає динаміку лептоспірозу, проблеми охорони здоров'я, переважно поширюваної гризунами, використовуючи нову модель SI-SIR, що відхиляється від традиційних моделей, припускаючи логістичний патерн зростання в популяції гризунів замість типового експоненційного зростання. Це дослідження пропонує всебічне дослідження стабільності рівноваги моделі. Рівняння встановлює рівняння для базового репродуктивного числа, R_0 , яке визначається швидкістю зараження гризунів, швидкістю народжуваності та екологічною місткістю. Це дослідження ідентифікує критичний поріг для екологічної місткості, який диктує вимирання або стійкість захворювання. Запропонований метод додатково вивчає чутливість R_0 і пропонує метод для оцінки впливу різних заходів контролю на динаміку інфекції. Чисельні симуляції представлені для надання відчутного представлення теоретичних результатів.

Таблиця 1.2 представляє аналіз моделей поширення лептоспірозу.

Джерело	Мета дослідження	Метод	Результати
Conceicao K.S. [85]	Проаналізувати частотні властивості байєсівських оцінювачів нотифікацій лептоспірозу у містах штату Баїя (Бразилія)	Байєсівський підхід для регресійної моделі ZMP	Дослідження дійшло висновку, що регресійні моделі ZMP є надзвичайно корисними для аналізу наборів даних із надлишком або дефіцитом нулів та ідентифікувало Індекс людського розвитку (ІЛР) як значущий фактор для пояснення нотифікацій лептоспірозу, стверджуючи, що ймовірність нотифікації лептоспірозу зростає з ІЛР.

Продовження таблиці 1.2

Mukdasai, K. [86]	Представити чисельне моделювання новоствореної моделі лептоспірозу дробового порядку з використанням стохастичних чисельних керованих нейронних мереж	Керовані нейронні мережі	Дослідження встановило ефективність, стабільність, точність, надійність та придатність запропонованого методу, при цьому статистичні показники склали 74% та 13% для сертифікації та навчання відповідно.
Bhalraj, A. [87]	Покращити розуміння динаміки передачі лептоспірозу у світі	Компартментна модель	Природний рівень смертності популяції щурів, разом із базовим репродуктивним числом, відіграє ключову роль у контролі спалаху лептоспірозу.
Minter, A. [88]	Розробити математичну модель оптимального контролю, яка ефективно управляє поширенням лептоспірозу	Компартментна модель	Дослідження дійшло висновку, що як постійний, так і тимчасовий контроль можуть ефективно зменшити поширеність лептоспірального носійства у популяції гризунів та кількість інфікованих щурів, знижуючи таким чином ризик інфікування людини.

Продовження таблиці 1.2

Qu, H. [89]	Дослідити динаміку лептоспірозу з використанням кусково-класичних глобальних та класичних дробових операторів	Компартментна модель	Результати моделювання показали покращену динаміку та поведінку переходів, зі зниженням популяції одужалих людей протягом 350 днів та зменшенням і стабілізацією сприйнятливих популяцій щурів та людей при високих значеннях β , що вказує на потенційне розсіювання захворювання.
Engida, H. A. [90]	Дослідити вплив лептоспірозу, що переноситься гризунами, на людські популяції, враховуючи рівень захворюваності людей та кількість патогенних агентів у навколишньому середовищі	Компартментна модель	Дослідження виявило, що вільна від захворювання рівновага є глобально асимптотично стабільною, якщо базове репродуктивне число менше одиниці, і нестабільною в іншому випадку, та що зниження швидкості передачі, підвищення швидкості одужання та контроль популяції гризунів відіграють значну роль у зменшенні поширення захворювання серед населення.
Rahmat, F. [91]	Зрозуміти взаємозв'язок між випадками лептоспірозу та погодними параметрами з різними часовими лагами в Малайзії	Штучна нейронна мережа	Дослідження виявило значущу кореляцію між випадками лептоспірозу та середньотижневою температурою з лагом у 16 тижнів та тижневими опадами з лагом у 12-20 тижнів.

Продовження таблиці 1.2

Athithan, S. [92]	Побудувати обчислювальну модель поширення лептоспірозу з використанням клітинних автоматів та правила на основі голосування	Клітинні автомати	Результати моделювання тісно відповідають фактичним даним з Таїланду, демонструючи потенційну застосовність цього підходу на основі правила голосування для точного відображення динаміки передачі лептоспірозу.
Ikram, R. [93]	Застосувати математичну модель до вивчення передачі лептоспірозу	Компартментна модель	Результати вказують на те, що метод гомотопічного збурення Хе ефективно апроксимує розв'язок для моделі передачі лептоспірозу.
Galego, M. A. [94]	Дослідити динаміку передачі лептоспірозу між гризунами та людьми	Компартментна модель	Дослідники виявили критичний поріг для екологічної ємності середовища, який визначає, чи захворювання зберігається, чи зникає.

Різноманітні проаналізовані дослідження ілюструють зростаючу важливість та застосування моделей симуляції та математичних моделей у прогнозуванні поширення та контролю лептоспірозу, складної та значущої проблеми охорони здоров'я. Ці статті використовують різноманітні моделі, що пропонують різні перспективи для розуміння динаміки захворювання.

Загалом, ці дослідження підкреслюють зростаючу важливість використання математичного та обчислювального моделювання в епідеміології та контролі захворювань і підкреслюють необхідність подальших досліджень та розробки методів, особливо в регіонах, де епідеміологічний слід захворювання все ще встановлюється. Моделі також підкреслюють важливість локальних факторів, таких як клімат, біорізноманіття та соціально-економічні детермінанти, для підвищення їх прогностичної точності та корисності.

1.5 Висновки до розділу

Розділ присвячений аналізу проблеми прогнозування епідемічних процесів та огляду сучасного стану досліджень у галузі моделювання інфекційних захворювань.

Здійснено систематичний огляд основних підходів до моделювання епідемічних процесів. Охарактеризовано детерміністичні компартментні моделі (SIR, SEIR, SEIRS), які базуються на системах диференціальних рівнянь та забезпечують теоретичне розуміння динаміки передачі захворювань через визначення базового репродуктивного числа та аналіз стабільності рівноваг. Проаналізовано стохастичні моделі, що враховують випадковість індивідуальних контактів у популяціях та генерують траєкторії з урахуванням невизначеності через алгоритм Гілеспі або марковські процеси. Розглянуто статистичні моделі часових рядів ARIMA та SARIMA, які забезпечують інтерпретовані короткострокові прогнози на основі історичних патернів без детальної специфікації біологічних механізмів. Особливу увагу приділено рекурентним нейронним мережам з довгою короткочасною пам'яттю (LSTM), що автоматично виявляють нелінійні залежності та адаптуються до структурних змін у даних.

Проведено аналіз сучасного стану досліджень щодо прогнозування інфекційних захворювань на прикладі гепатиту В, грипу та лептоспірозу. Аналіз показав, що моделі машинного навчання, зокрема LSTM з механізмом уваги, демонструють вищу точність короткострокового прогнозування порівняно з класичними статистичними методами, зменшуючи помилку прогнозування приблизно на 50% та пояснюючи до 87% дисперсії у тестових даних. Гібридні підходи, що поєднують ARIMA для видалення лінійних компонентів та нейронні мережі для моделювання нелінійних залишків, часто досягають найкращої продуктивності з MAPE на рівні 8-9%. Сезонні моделі SARIMA ефективно відтворюють річні цикли захворюваності та забезпечують

надійні короткострокові прогнози, коли всі прогнозні точки потрапляють у межі 95% довірчого інтервалу.

Виявлено основні проблеми та обмеження існуючих підходів до прогнозування епідемічних процесів. Класичні рекурентні нейронні мережі страждають від проблеми зникаючого градієнта при навчанні на довгих послідовностях, що робить практично неможливим виявлення довгострокових залежностей. Архітектура LSTM вирішує цю проблему через введення системи вентилів та адитивної структури оновлення стану комірки, що дозволяє градієнтам протікати через багато часових кроків без деградації. Моделі ARIMA базуються на припущенні лінійності, що не дозволяє моделювати складні нелінійні взаємодії. Нейронні мережі характеризуються низькою інтерпретованістю навчених моделей та потребою у великих обсягах даних для уникнення перенавчання.

Охарактеризовано теоретичні основи методів, що використовуються у дослідженні. Формалізовано математичний апарат мереж LSTM, включаючи систему вентилів забування, входу та виходу, механізм оновлення стану комірки через адитивну комбінацію забутої частини попереднього стану та зваженого внеску нового кандидата, а також метод зворотного поширення помилки у часі для навчання параметрів. Описано методологію побудови моделей ARIMA за методологією Бокса-Дженкінса, включаючи етапи ідентифікації порядків через аналіз ACF та PACF, оцінювання параметрів методом максимальної правдоподібності та діагностики залишків через тест Льюнга-Бокса. Систематизовано підходи до аналізу стаціонарності часових рядів через тести ADF та KPSS з комбінованою стратегією інтерпретації результатів.

Визначено актуальність дослідження для України як країни, що зазнає військового конфлікту та надзвичайних подій. Порушення систем епідемічного нагляду, міграція населення, зниження охоплення вакцинацією та перевантаження медичної інфраструктури створюють унікальні виклики для

прогнозування інфекційних захворювань. У таких умовах методи машинного навчання, здатні інтегрувати екзогенні змінні та адаптуватися до структурних зламів, набувають особливої цінності для підтримки системи громадського здоров'я.

Встановлено, що небагато досліджень поєднують переваги статистичних моделей та методів глибокого навчання в одному інструменті або розглядають порушення даних, типові для кризових умов. Більше того, розбіжності між провідними моделями вакцинації підкреслюють стійку структурну невизначеність, а обмежена увага до програм бустерної вакцинації дорослих та динаміки дотримання залишає важливі питання без відповіді. Відсутність систематичних досліджень епідемічного прогнозування для українського контексту в умовах надзвичайних ситуацій підтверджує актуальність теми дисертації.

Загалом, проведений аналіз джерел, аналіз методів та огляд сучасного стану досліджень дозволили сформулювати основні висновки, які в подальшому були використані для розробки методів прогнозування та інформаційної технології.

Аналіз показав, що для епідемічних часових рядів з вираженими нелінійними залежностями та структурними зламами LSTM систематично перевершує класичні методи ARIMA, водночас для рядів з простими лінійними трендами та стабільною сезонністю ARIMA може забезпечувати порівнянну або навіть кращу точність завдяки меншій схильності до перенавчання та кращій інтерпретованості. Отже, варто розробляти методи, що враховують специфіку конкретного захворювання та доступний обсяг даних.

Порівняльний аналіз різних представлень даних показав, що кумулятивне представлення згладжує випадкові місячні коливання через інтегрування, що зменшує відносну дисперсію та може покращувати стабільність прогнозів, проте менш чутливе до короткострокових змін та

спалахів. Рівневе представлення краще підходить для задач раннього виявлення епідемічних подій.

Аналіз валідаційних стратегій показав, що walk-forward валідація з пересуванням вікна навчання є золотим стандартом для оцінювання моделей прогнозування часових рядів, оскільки імітує реальний режим експлуатації з періодичним перенавчанням моделі та дозволяє оцінити стабільність продуктивності у часі.

Основні результати розділу опубліковано в працях [2, 4, 5, 8, 14, 15].

1.6 Постановка задачі досліджень

Задача дисертаційного дослідження полягає у розробці інформаційної технології прогнозування захворюваності на інфекційні хвороби у короткостроковій перспективі шляхом розроблення методів та інформаційної технології на основі інтеграції статистичних моделей часових рядів та рекурентних нейронних мереж з довгою короткочасною пам'яттю. Для досягнення поставленої мети необхідно вирішити наступні задачі дослідження.

Провести аналіз існуючих методів та моделей прогнозування епідемічних часових рядів для визначення переваг та обмежень різних підходів у контексті застосування до українських даних епідемічного спостереження.

Розробити метод попередньої обробки епідемічних часових рядів для LSTM моделей, що включає аналіз стаціонарності, нормалізацію даних та формування навчальних послідовностей з урахуванням специфіки інфекційних захворювань з різними паттернами сезонності, а також циклічне кодування часових ознак для безперервного представлення річного циклу.

Розробити архітектуру рекурентної нейронної мережі LSTM для прогнозування епідемічних часових рядів з різними характеристиками, включаючи короткострокові та довгострокові залежності, виражену сезонність та наявність структурних зламів, з оптимізацією конфігурації шарів та параметрів регуляризації.

Розробити метод порівняльного аналізу LSTM моделей та класичних моделей ARIMA/SARIMA для визначення областей найбільш ефективного застосування кожного підходу залежно від характеристик часового ряду та горизонту прогнозування.

Розробити метод валідації моделей прогнозування епідемічних часових рядів на основі walk-forward та sliding window стратегій, що імітують реальний режим експлуатації системи та дозволяють отримати розподіл помилок прогнозування замість однієї точкової оцінки.

Створити програмно-інформаційне забезпечення у вигляді веб-застосунку для онлайн трекінгу епідемічних прогнозів, що забезпечує практичну застосовність розроблених методів та візуалізацію результатів для підтримки прийняття рішень.

Провести експериментальні дослідження розроблених методів на реальних даних захворюваності в Україні та Харківській області для різних інфекційних захворювань, включаючи гострий гепатит А, гострий та хронічний гепатит В, гострий гепатит С, грип, сальмонельоз, гострі респіраторні вірусні інфекції та лептоспіроз, та оцінити їх ефективність у порівнянні з існуючими підходами.

Перша задача частково вирішена у цьому розділі шляхом проведення аналітичного огляду сучасного стану досліджень. Подальші задачі послідовно вирішуються у наступних розділах дисертації: у другому розділі представлено розроблення методів та теоретичних основ прогнозування епідемічних процесів, у третьому розділі покращення методів та прикладне застосування включаючи розробку веб-застосунку, а у четвертому розділі результати експериментальних досліджень та порівняльний аналіз розроблених моделей.

РОЗДІЛ 2. РОЗРОБЛЕННЯ МЕТОДІВ ТА ТЕОРЕТИЧНИХ ОСНОВ ПРОГНОЗУВАННЯ ЕПІДЕМІЧНИХ ПРОЦЕСІВ

2.1 Теоретичні основи LSTM

Рекурентні нейронні мережі (RNN) є класом нейронних архітектур, спеціально розроблених для обробки послідовних даних шляхом введення зворотних зв'язків, що дозволяє мережі зберігати інформацію про попередні елементи послідовності [49–50]. На відміну від прямих нейронних мереж, де інформація рухається лише від входу до виходу, рекурентні мережі мають цикли у своїй обчислювальній структурі. Для послідовності вхідних векторів

$\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_T$, класична RNN обчислює послідовність прихованих станів $\mathbf{h}_1, \mathbf{h}_2, \dots, \mathbf{h}_T$ за рекурентним правилом:

$$\mathbf{h}_t = \sigma(\mathbf{W}_{xh}\mathbf{x}_t + \mathbf{W}_{hh}\mathbf{h}_{t-1} + \mathbf{b}_h), \quad (2.1)$$

де \mathbf{W}_{xh} є матрицею ваг вхід-прихований стан, \mathbf{W}_{hh} є матрицею рекурентних ваг прихований-прихований стан, \mathbf{b}_h є вектором зміщення, а σ є нелінійною функцією активації, зазвичай гіперболічним тангенсом або сигмоїдою [49]. Вихідний вектор \mathbf{y}_t обчислюється як афінне перетворення прихованого стану:

$$\mathbf{y}_t = \mathbf{W}_{hy}\mathbf{h}_t + \mathbf{b}_y. \quad (2.2)$$

Навчання рекурентних нейронних мереж здійснюється методом зворотного поширення помилки у часі (Backpropagation Through Time, BPTT) [53–54]. Цей алгоритм розгортає рекурентну мережу у часі, створюючи еквівалентну глибоку пряму мережу з T шарами, де кожен шар відповідає одному часовому кроку. Градієнти функції втрат обчислюються зворотним проходом через цю розгорнуту структуру. Для обчислення градієнта відносно

параметрів на кроці t необхідно обчислити похідну прихованого стану \mathbf{h}_t відносно попередніх станів, що призводить до добутку якобіанів:

$$\frac{\partial \mathbf{h}_t}{\partial \mathbf{h}_k} = \prod_{i=k+1}^t \frac{\partial \mathbf{h}_i}{\partial \mathbf{h}_{i-1}} = \prod_{i=k+1}^t \mathbf{W}_{hh}^T \text{da}(\sigma'(\mathbf{z}_i)), \quad (2.3)$$

де $\mathbf{z}_i = \mathbf{W}_{xh}\mathbf{x}_i + \mathbf{W}_{hh}\mathbf{h}_{i-1} + \mathbf{b}_h$ є вектором передактиваційних значень [51]. Якщо власні значення матриці \mathbf{W}_{hh} менші за одиницю за абсолютним значенням, або якщо похідна функції активації σ' близька до нуля, то добуток багатьох таких матриць експоненціально зменшується зі збільшенням $t - k$

Це явище називається проблемою зникаючого градієнта (vanishing gradient problem) і робить практично неможливим навчання класичних RNN виявляти довгострокові залежності, що охоплюють десятки або сотні часових кроків [51]. Аналогічно, якщо власні значення перевищують одиницю, градієнти можуть експоненціально зростати, призводячи до проблеми вибухаючого градієнта (exploding gradient problem), хоча останню проблему можна частково вирішити методом відсікання градієнтів (gradient clipping).

Архітектура мереж довгої короткочасної пам'яті (Long Short-Term Memory, LSTM) була запропонована Хохрайтером та Шмідхубером у 1997 році як рішення проблеми зникаючого градієнта у рекурентних мережах [14]. Ключова ідея LSTM полягає у введенні додаткового вектора стану комірки (cell state) \mathbf{c}_t , який зберігає довгострокову інформацію, та системи вентилів (gates), що контролюють потік інформації до цього стану. Вентилі є нелінійними функціями входу та попереднього прихованого стану, що виробляють значення в інтервалі $[0, 1]$ за допомогою сигмоїдної функції активації $\sigma(z) = 1/(1 + e^{-z})$. Значення вентиля близьке до нуля блокує відповідний інформаційний потік, а значення близьке до одиниці дозволяє інформації проходити без втрат.

Стандартна LSTM комірка складається з трьох вентилів [14, 52]:

1. Вентиль забування (forget gate) f_t визначає, яку частину попереднього стану комірки c_{t-1} слід зберегти:

$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + b_f), \quad (2.4)$$

2. Вентиль входу (input gate) i_t контролює, яку частину нової інформації додати до стану комірки:

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + b_i), \quad (2.5)$$

3. Вентиль виходу (output gate) o_t визначає, яку частину стану комірки використовувати для обчислення прихованого стану:

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + b_o). \quad (2.6)$$

Паралельно обчислюється кандидат на новий стан комірки \tilde{c}_t , який представляє нову інформацію для потенційного додавання до стану комірки:

$$\tilde{c}_t = \tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c), \quad (2.7)$$

де \tanh є гіперболічним тангенсом, який виробляє значення в інтервалі $[-1, 1]$. Стан комірки оновлюється шляхом комбінування забутої частини попереднього стану та зваженого внеску нового кандидата:

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t, \quad (2.8)$$

де \odot позначає поелементне множення (операцію Адамара). Ця адитивна структура оновлення стану є критично важливою для вирішення проблеми зникаючого градієнта, оскільки градієнт функції втрат відносно c_t може протікати назад у часі через суму без мультиплікативних деградацій [52].

Фінальний прихований стан обчислюється як відфільтрована версія стану комірки:

$$\mathbf{h}_t = \mathbf{o}_t \odot \tanh(\mathbf{c}_t). \quad (2.9)$$

Застосування функції \tanh до стану комірки перед множенням на вентиль виходу нормалізує значення до інтервалу $[-1, 1]$, що сприяє стабільності навчання.

Повна система рівнянь LSTM комірки для часового кроку t має вигляд:

$$\mathbf{f}_t = \sigma(\mathbf{W}_{xf}\mathbf{x}_t + \mathbf{W}_{hf}\mathbf{h}_{t-1} + \mathbf{b}_f), \quad (2.10)$$

$$\mathbf{i}_t = \sigma(\mathbf{W}_{xi}\mathbf{x}_t + \mathbf{W}_{hi}\mathbf{h}_{t-1} + \mathbf{b}_i), \quad (2.11)$$

$$\mathbf{o}_t = \sigma(\mathbf{W}_{xo}\mathbf{x}_t + \mathbf{W}_{ho}\mathbf{h}_{t-1} + \mathbf{b}_o), \quad (2.12)$$

$$\tilde{\mathbf{c}}_t = \tanh(\mathbf{W}_{xc}\mathbf{x}_t + \mathbf{W}_{hc}\mathbf{h}_{t-1} + \mathbf{b}_c), \quad (2.13)$$

$$\mathbf{c}_t = \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \tilde{\mathbf{c}}_t, \quad (2.14)$$

$$\mathbf{h}_t = \mathbf{o}_t \odot \tanh(\mathbf{c}_t). \quad (2.15)$$

Ця система містить чотири набори параметрів (матриці ваг та вектори зміщень) для кожного з трьох вентилів та кандидата на стан комірки. Для LSTM комірки з n входами та m прихованими одиницями, загальна кількість параметрів дорівнює $4(nt + m^2 + m)$, що робить LSTM значно більш параметризованою, ніж класична RNN з тією ж кількістю прихованих одиниць.

Обчислення градієнта стану комірки демонструє механізм уникнення зникаючого градієнта. Похідна \mathbf{c}_t відносно \mathbf{c}_{t-1} дорівнює:

$$\frac{\partial c_t}{\partial c_{t-1}} = \text{diag}(\mathbf{f}_t), \quad (2.16)$$

де $\text{diag}(\mathbf{f}_t)$ є діагональною матрицею з елементами вентиля забування на діагоналі. Для обчислення градієнта на k кроків назад у часі отримуємо:

$$\frac{\partial c_t}{\partial c_k} = \prod_{i=k+1}^t \text{diag}(\mathbf{f}_i). \quad (2.17)$$

Оскільки елементи вентиля забування належать інтервалу $[0, 1]$ і можуть бути близькими до одиниці під час навчання, цей добуток деградує значно повільніше, ніж у класичних RNN, де множники включають власні значення матриці ваг та похідні функцій активації [55]. Більше того, адитивна структура оновлення створює прямий шлях для градієнтів через багато часових кроків без мультиплікативних перетворень, що дозволяє LSTM ефективно навчатися на послідовностях з сотнями часових кроків.

Варіації архітектури LSTM включають GRU (Gated Recurrent Unit), яка об'єднує вентиль забування та входу в один вентиль оновлення і виключає окремий стан комірки, що зменшує кількість параметрів при збереженні подібної продуктивності для багатьох задач [95–96]. Peephole connections додають зв'язки від стану комірки до вентилів, дозволяючи вентилям безпосередньо спостерігати стан комірки, що може покращити моделювання точних часових інтервалів [97].

Для багатошарових LSTM архітектур вихід прихованого стану $\mathbf{h}_t^{(l)}$ шару l на часовому кроці t використовується як вхід для шару $l+1$:

$$\mathbf{x}_t^{(l+1)} = \mathbf{h}_t^{(l)} \quad [52]. \quad (2.18)$$

Глибокі LSTM створюють ієрархічні представлення часових залежностей, де нижні шари виявляють короткострокові паттерни, а верхні шари інтегрують їх у довгострокові абстракції. Емпіричні дослідження показують, що для багатьох задач обробки послідовностей оптимальна глибина становить 2–4 шари, оскільки глибші архітектури можуть страждати від

деградації продуктивності без спеціальних методів регуляризації або залишкових зв'язків (residual connections) [98–99].

Регуляризація LSTM мереж є критично важливою для запобігання перенавчанню, особливо при обмежених обсягах навчальних даних. Метод Dropout, який випадково вимикає частку нейронів під час навчання, застосовується до LSTM з обережністю, оскільки наївне застосування Dropout до рекурентних зв'язків може пошкодити здатність мережі зберігати довгострокову інформацію [100]. Рекомендований підхід полягає у застосуванні Dropout тільки до не-рекурентних зв'язків (вхід-прихований стан та між шарами) при збереженні рекурентних зв'язків без Dropout [52, 101]. Формально, для кожного часового кроку використовується одна й та сама маска Dropout для всіх часових кроків у послідовності (variational dropout), що забезпечує регуляризацію без втрати довгострокових залежностей.

Альтернативні методи регуляризації включають обмеження норми ваг (weight decay), яке додає штраф $\lambda \|W\|_2^2$ до функції втрат, де λ є коефіцієнтом регуляризації [49]. Gradient clipping обмежує норму градієнтів під час навчання для запобігання вибуховим градієнтам: якщо $|\nabla| > \theta$, то $\nabla \leftarrow \theta \cdot \nabla / |\nabla|$, де θ є порогом відсікання [53]. Layer normalization нормалізує активації всередині кожної LSTM комірки, що прискорює збіжність навчання та зменшує чутливість до вибору швидкості навчання [102].

Навчання LSTM мереж здійснюється методами стохастичної оптимізації. Алгоритм Adam (Adaptive Moment Estimation) є де-факто стандартом для навчання рекурентних мереж завдяки адаптивному налаштуванню швидкості навчання для кожного параметра на основі перших та других моментів градієнтів [55]. Для часового кроку k та параметра θ_i Adam обчислює:

$$m_i^{(k)} = \beta_1 m_i^{(k-1)} + (1 - \beta_1) g_i^{(k)}, \quad (2.19)$$

$$v_i^{(k)} = \beta_2 v_i^{(k-1)} + (1 - \beta_2) \left(g_i^{(k)} \right)^2, \quad (2.20)$$

$$\widehat{m}_i^{(k)} = m_i^{(k)} / (1 - \beta_1^k), \quad (2.21)$$

$$\widehat{v}_i^{(k)} = v_i^{(k)} / (1 - \beta_2^k), \quad (2.22)$$

$$\theta_i^{(k)} = \theta_i^{(k-1)} - \eta \widehat{m}_i^{(k)} / \left(\sqrt{\widehat{v}_i^{(k)}} + \epsilon \right), \quad (2.23)$$

де $g_i^{(k)}$ є градієнтом функції втрат відносно θ_i на кроці k , m_i та v_i є оцінками першого та другого моментів градієнта, β_1 та β_2 є коефіцієнтами експоненціального згладжування (зазвичай $\beta_1 = 0.9$ та $\beta_2 = 0.999$), η є швидкістю навчання, а ϵ є малою константою для числової стабільності (зазвичай 10^{-8}) [55]. Корекція зміщення \widehat{m}_i та \widehat{v}_i враховує ініціалізацію моментів нулями.

Функція втрат для задач прогнозування часових рядів зазвичай базується на середньоквадратичній похибці (Mean Squared Error, MSE):

$$\text{MSE} = \frac{1}{N} \sum_{i=1}^N (\widehat{y}_i - y_i)^2, \quad (2.24)$$

де \widehat{y}_i є прогнозними значеннями, y_i є фактичними значеннями, а N є кількістю прогнозованих точок. Для задач з асиметричними витратами помилок можна використовувати середню абсолютну похибку (Mean Absolute Error, MAE):

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^N |\widehat{y}_i - y_i|, \quad (2.25)$$

або квантильну функцію втрат для прогнозування інтервалів невизначеності [56, 103]:

$$\mathcal{L}_\tau(y, \hat{y}) = \max\{\tau(y - \hat{y}), (1 - \tau)(\hat{y} - y)\}, \quad (2.26)$$

де $\tau \in (0,1)$ є рівнем квантиля.

Ініціалізація параметрів LSTM впливає на швидкість збіжності та фінальну продуктивність. Рекомендований підхід полягає у ініціалізації матриць ваг випадковими значеннями з розподілу з нульовим середнім та дисперсією, обчисленою за методом Глорота (Xavier initialization) або Хе [104–105]. Вектори зміщень зазвичай ініціалізуються нулями, за винятком зміщення вентиля забування, яке часто ініціалізується позитивним значенням (наприклад, 1.0) для забезпечення початкової тенденції до збереження інформації у стані комірки [106].

Практичне застосування LSTM для прогнозування епідемічних часових рядів вимагає спеціальної попередньої обробки даних. Нормалізація входів до інтервалу $[0, 1]$ методом min-max scaling або стандартизація до нульового середнього та одиничної дисперсії покращує стабільність навчання [49]. Формування послідовностей методом ковзного вікна перетворює часовий ряд у набір пар (вхід, вихід), де вхід є послідовністю w попередніх значень, а вихід є наступним значенням або послідовністю h майбутніх значень для багатокрокового прогнозування [15]. Для часового ряду $\{x_1, x_2, \dots, x_T\}$ з вікном довжини w та горизонтом прогнозування h , формується набір:

$$\mathcal{D} = \{(x_i, y_i): x_i = (x_i, x_{i+1}, \dots, x_{i+w-1}), y_i = (x_{i+w}, \dots, x_{i+w+h-1}), i = 1, \dots, T - w - h + 1\}. \quad (2.27)$$

Вибір довжини вікна w є критичним гіперпараметром, що визначає максимальну довжину історії, доступну моделі для прогнозування. Емпіричні дослідження показують, що для епідемічних часових рядів з тижневою або місячною сезонністю оптимальні значення w становлять від 12 до 52 тижнів [16–17].

Інтерпретація рішень LSTM моделей залишається відкритою дослідницькою проблемою через непрозорість внутрішніх представлень. Методи візуалізації активацій вентилів та станів комірок дозволяють якісно оцінити, які частини вхідної послідовності отримують найбільшу увагу моделі [107]. Техніки attribution methods, такі як Layer-wise Relevance Propagation або Integrated Gradients, розподіляють прогнозне значення між вхідними часовими точками, виявляючи найбільш впливові історичні періоди [108–109]. Аналіз помилок шляхом візуалізації залишків (прогноз мінус факт) ідентифікує систематичні патерни недооцінки або переоцінки, що може вказувати на необхідність додаткових екзогенних змінних або корекції архітектури.

Порівняння продуктивності LSTM з альтернативними підходами показує, що для епідемічних часових рядів з вираженими нелінійними залежностями та структурними зламами LSTM систематично перевершує класичні методи ARIMA [16–17]. Водночас для рядів з простими лінійними трендами та стабільною сезонністю ARIMA може забезпечувати порівнянну або навіть кращу точність завдяки меншій схильності до перенавчання та кращій інтерпретованості [12, 13]. Гібридні підходи, що поєднують ARIMA для моделювання лінійних компонентів та LSTM для нелінійних залишків, часто досягають найкращої продуктивності, інтегруючи переваги обох методологій [57–58].

2.2 Теорія моделей авторегресії та проінтегрованого ковзного середнього (ARIMA)

Моделювання часових рядів методами авторегресії та ковзного середнього є класичним статистичним підходом до аналізу та прогнозування послідовних даних, що базується на ідентифікації та екстраполяції статистичних закономірностей у історичних спостереженнях [12]. На відміну від механістичних моделей, які явно специфікують біологічні або фізичні

процеси, моделі часових рядів є феноменологічними, тобто описують емпіричні патерни без детального моделювання причинно-наслідкових механізмів.

Стаціонарний часовий ряд є фундаментальною концепцією у статистичному моделюванні послідовних даних. Часовий ряд $\{X_t\}$ є слабо стаціонарним (або стаціонарним у широкому сенсі), якщо його середнє значення та дисперсія не залежать від часу, а автокореляційна функція залежить тільки від часового лагу [45–46]:

$$E[X_t] = \mu, \quad \forall t, \quad (2.28)$$

$$\text{Var}(X_t) = \sigma^2, \quad \forall t, \quad (2.29)$$

$$\text{Cov}(X_t, X_{t+h}) = \gamma(h), \quad \forall t, h \quad (2.30)$$

де μ є константним середнім, σ^2 є константною дисперсією, а $\gamma(h)$

є автоковаріаційною функцією, що залежить тільки від лагу h . Автокореляційна функція (ACF) визначається як нормалізована автоковаріація:

$$\rho(h) = \frac{\gamma(h)}{\gamma(0)} = \frac{\text{Cov}(X_t, X_{t+h})}{\text{Var}(X_t)}. \quad (2.31)$$

Стаціонарність є критичною властивістю для застосування методів моделювання часових рядів, оскільки дозволяє використовувати історичні статистики для прогнозування майбутніх значень. Більшість епідемічних часових рядів демонструють нестаціонарність через наявність трендів та сезонних компонентів, що вимагає попередніх перетворень для досягнення стаціонарності.

Процес авторегресії порядку p , позначений $AR(p)$, моделює поточне значення часового ряду як лінійну комбінацію p попередніх значень та білого шуму [14, 45]:

$$X_t = c + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \dots + \phi_p X_{t-p} + \varepsilon_t, \quad (2.32)$$

де c є константою, $\phi_1, \phi_2, \dots, \phi_p$ є коефіцієнтами авторегресії, а ε_t є білим шумом з нульовим середнім та константною дисперсією

Використовуючи оператор зсуву назад (backshift operator) B , визначений як $BX_t = X_{t-1}$, модель $AR(p)$ можна записати у компактній формі:

$$\Phi(B)X_t = c + \varepsilon_t, \quad (2.33)$$

де $\Phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \dots - \phi_p B^p$

є характеристичним поліномом авторегресії. Для забезпечення стаціонарності процесу $AR(p)$ всі корені рівняння $\Phi(z) = 0$ повинні лежати поза одиничним колом у комплексній площині, тобто мати абсолютне значення більше одиниці [46]. Ця умова гарантує, що вплив минулих шоків експоненціально зменшується у часі, і ряд має скінченне середнє та дисперсію.

Найпростіша модель $AR(1)$ має вигляд:

$$X_t = c + \phi_1 X_{t-1} + \varepsilon_t. \quad (2.34)$$

Для стаціонарності необхідно $|\phi_1| < 1$. Безумовне середнє процесу дорівнює $E[X_t] = c/(1 - \phi_1)$, а дисперсія дорівнює $\text{Var}(X_t) = \sigma_\varepsilon^2/(1 - \phi_1^2)$

Автокореляційна функція експоненціально зменшується:

$$\rho(h) = \phi_1 \text{ для } h \geq 1.$$

Процес ковзного середнього порядку q , позначений $MA(q)$ представляє поточне значення як лінійну комбінацію поточного та q попередніх термів білого шуму [14, 45]:

$$X_t = \mu + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \dots + \theta_q \varepsilon_{t-q}, \quad (2.35)$$

де μ є середнім процесу, а $\theta_1, \theta_2, \dots, \theta_q$ є коефіцієнтами ковзного середнього. У термінах оператора зсуву:

$$X_t = \mu + \Theta(B)\varepsilon_t, \quad (2.36)$$

де $\Theta(B) = 1 + \theta_1 B + \theta_2 B^2 + \dots + \theta_q B^q$ є характеристичним поліномом ковзного середнього. Процес $MA(q)$ завжди є стаціонарним, оскільки є лінійною комбінацією білих шумів. Для забезпечення оборотності (invertibility), яка дозволяє виразити поточні інновації через минулі спостереження, всі корені рівняння $\Theta(z) = 0$ повинні лежати поза одиничним колом [46]. Для процесу $MA(1)$:

$$X_t = \mu + \varepsilon_t + \theta_1 \varepsilon_{t-1}, \quad (2.37)$$

Автокореляція на лазі 1 дорівнює $\rho(1) = \theta_1 / (1 + \theta_1^2)$, а для лагів $h > 1$ дорівнює нулю: $\rho(h) = 0$ для $h > 1$. Ця властивість скінченної пам'яті є характерною ознакою процесів ковзного середнього і дозволяє відрізнити їх від авторегресійних процесів, автокореляція яких зменшується поступово.

Змішана модель авторегресії та ковзного середнього порядку (p, q) , позначена $ARMA(p, q)$, поєднує компоненти авторегресії та ковзного середнього [14, 45]:

$$X_t = c + \phi_1 X_{t-1} + \dots + \phi_p X_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \dots + \theta_q \varepsilon_{t-q}, \quad (2.38)$$

або у компактній формі:

$$\Phi(B)X_t = c + \Theta(B)\varepsilon_t. \quad (2.39)$$

Модель $ARMA(p, q)$ є стаціонарною, якщо всі корені $\Phi(z) = 0$ лежать поза одиничним колом, і оборотною, якщо всі корені $\Theta(z) = 0$ лежать поза одиничним колом. Перевага моделі $ARMA$ полягає у парсимонічності: часто складну автокореляційну структуру можна описати малими значеннями p та q , замість використання процесу високого порядку $AR(p)$ або $MA(q)$ окремо [47].

Багато практичних часових рядів демонструють нестационарність через наявність тренду або зміни дисперсії у часі. Диференціювання є стандартним методом усунення трендової нестационарності [14]. Перша різниця часового ряду визначається як:

$$\nabla X_t = X_t - X_{t-1} = (1 - B)X_t. \quad (2.40)$$

Якщо застосування першої різниці робить ряд стаціонарним, оригінальний ряд називається інтегрованим порядку 1, позначений $I(1)$.

Ряд є інтегрованим порядку d , позначається як $I(d)$, якщо d -та різниця є стаціонарною:

$$\nabla^d X_t = (1 - B)^d X_t. \quad (2.41)$$

Модель авторегресії та проінтегрованого ковзного середнього порядку (p, d, q) , позначена $ARIMA(p, d, q)$, застосовує модель $ARMA(p, q)$ до d -ї різниці часового ряду [14, 47]:

$$\Phi(B)\nabla^d X_t = c + \Theta(B)\varepsilon_t, \quad (2.42)$$

або еквівалентно:

$$\Phi(B)(1 - B)^d X_t = c + \Theta(B)\varepsilon_t. \quad (2.43)$$

Модель $ARIMA(p, d, q)$ є надзвичайно гнучкою і може моделювати широкий спектр нестационарних часових рядів. Наприклад, $ARIMA(0,1,0)$ є випадковим блуканням: $X_t = X_{t-1} + \varepsilon_t$ модель $ARIMA(0,1,1)$ є експоненціальним згладжуванням, а $ARIMA(0,2,2)$ може моделювати локальні лінійні тренди [16].

Для часових рядів з сезонними компонентами, сезонна модель $ARIMA$, позначена $SARIMA(p, d, q) \times (P, D, Q)_s$ включає як несезонні, так і сезонні авторегресійні та ковзні компоненти [14, 47]:

$$\Phi(B)\Phi_s(B^s)\nabla^d\nabla_s^D X_t = c + \Theta(B)\Theta_s(B^s)\varepsilon_t, \quad (2.44)$$

де s є сезонним періодом (наприклад, $s = 52$ для тижневих даних з річною сезонністю, $s = 12$ для місячних даних),

$$\Phi_s(B^s) = 1 - \Phi_1 B^s - \Phi_2 B^{2s} - \dots - \Phi_P B^{Ps}, \quad (2.45)$$

є сезонним авторегресійним поліномом, $\Theta_s(B^s) = 1 + \Theta_1 B^s + \Theta_2 B^{2s} + \dots + \Theta_Q B^{Qs}$ є сезонним поліномом ковзного середнього, а $\nabla_s = 1 - B^s$ є сезонним оператором різниці. Наприклад, модель $SARIMA(1,0,0) \times (1,1,0)_{52}$ для тижневих епідемічних даних має вигляд:

$$(1 - \phi_1 B)(1 - \Phi_1 B^{52})(1 - B^{52})X_t = c + \varepsilon_t. \quad (2.46)$$

Розгортаючи оператори, отримуємо:

$$X_t = X_{t-52} + \phi_1(X_{t-1} - X_{t-53}) + \Phi_1(X_{t-52} - X_{t-104}) - \phi_1\Phi_1(X_{t-53} - X_{t-105}) + c + \varepsilon_t. \quad (2.47)$$

Ця модель враховує як короткострокову авторегресію (через ϕ_1), так і сезонну авторегресію (через Φ_1), а сезонне диференціювання усуває річний тренд.

Методологія Бокса-Дженкінса для побудови моделей ARIMA} складається з трьох ітеративних етапів: ідентифікація моделі, оцінювання параметрів та діагностика [14, 47].

Перший етап це дентифікація моделі. Мета полягає у визначенні порядків p , d , q (та P , D , Q , s для сезонних моделей). Спочатку перевіряється стаціонарність часового ряду візуальним аналізом графіка та формальними статистичними тестами. Розширений тест Дікі-Фуллера (Augmented Dickey-Fuller, ADF) перевіряє нульову гіпотезу про наявність одиничного кореня (нестаціонарність) проти альтернативи стаціонарності [48]:

$$\Delta X_t = \alpha + \beta t + \gamma X_{t-1} + \sum_{i=1}^k \delta_i \Delta X_{t-i} + \varepsilon_t, \quad (2.48)$$

де нульова гіпотеза $H_0: \gamma = 0$ відповідає наявності одиничного кореня. Відхилення нульової гіпотези вказує на стаціонарність. Тест Квятковського-Філіпса-Шмідта-Шіна (KPSS) перевіряє протилежну нульову гіпотезу стаціонарності проти альтернативи нестаціонарності [46], що дозволяє уникнути проблеми низької потужності тесту ADF для рядів, близьких до одиничного кореня.

Якщо ряд є нестаціонарним, застосовується диференціювання порядку d до досягнення стаціонарності. Зазвичай $d = 1$ або $d = 2$ достатньо для більшості практичних рядів; надмірне диференціювання може призвести до погіршення прогнозної точності через введення додаткової автокореляції [14].

Порядки p та q визначаються аналізом автокореляційної функції (ACF) та часткової автокореляційної функції (PACF) диференційованого ряду [45–46]. Вибіркова автокореляційна функція обчислюється як:

$$\hat{\rho}(h) = \frac{\sum_{t=h+1}^T (X_t - \bar{X})(X_{t-h} - \bar{X})}{\sum_{t=1}^T (X_t - \bar{X})^2}, \quad (2.49)$$

де \bar{X} є вибірковим середнім, а T є довжиною часового ряду. Часткова автокореляція на лазі h , $\alpha(h)$, є кореляцією між X_t та X_{t-h} після видалення лінійної залежності від проміжних лагів $X_{t-1}, X_{t-2}, \dots, X_{t-h+1}$ [46].

Для процесу $AR(p)$, ACF експоненціально зменшується або демонструє затухаючі синусоїди, а PACF обривається після лагу : $\alpha(h) = 0$ для $h > p$. Для процесу $MA(q)$ ACF обривається після лагу : $\rho(h) = 0$ для $h > q$, а PACF експоненціально зменшується. Для процесів $ARMA(p, q)$, обидві функції експоненціально зменшуються без чіткого обривання, що вимагає ітеративного відбору моделі [47].

Наступний етап це оцінювання параметрів. Після ідентифікації попередніх порядків моделі, параметри ϕ_i , θ_j та σ_ε^2 оцінюються методом максимальної правдоподібності (Maximum Likelihood Estimation, MLE) або методом найменших квадратів [14, 46]. Для гаусівського білого шуму, логарифм функції правдоподібності має вигляд:

$$l(\boldsymbol{\phi}, \boldsymbol{\theta}, \sigma_\varepsilon^2 | \mathbf{X}) = -\frac{T}{2} \ln(2\pi) - \frac{T}{2} \ln(\sigma_\varepsilon^2) - \frac{1}{2\sigma_\varepsilon^2} \sum_{t=1}^T \varepsilon_t^2, \quad (2.50)$$

де ε_t обчислюються рекурсивно з рівняння моделі. Максимізація логарифма правдоподібності відносно параметрів здійснюється чисельними методами оптимізації, такими як алгоритм Ньютона-Рафсона або квазі-Ньютонівські методи [14].

Стандартні похибки оцінених параметрів обчислюються з обернення інформаційної матриці Фішера, що дозволяє побудувати довірчі інтервали та перевірити статистичну значущість коефіцієнтів. Незначущі параметри можуть бути виключені для спрощення моделі [47].

Етап діагностики моделі полягає в адекватності підбраної моделі перевіряється аналізом залишків $\hat{\varepsilon}_t = X_t - \widehat{X}_t$, де \widehat{X}_t є підігнаним значенням [14, 47]. Якщо модель правильно специфікована, залишки повинні бути білим шумом: некорельованими, мати нульове середнє та константну дисперсію. Перевірка проводиться аналізом АСF залишків; значущі автокореляції вказують на недостатню специфікацію моделі. Тест Льюнга-Бокса (Ljung-Box test) формально перевіряє нульову гіпотезу відсутності автокореляції у залишках [110]:

$$Q_{LB}(K) = T(T + 2) \sum_{h=1}^K \frac{\widehat{\rho}_{\varepsilon}^2(h)}{T-h}, \quad (2.51)$$

де $\widehat{\rho}_{\varepsilon}(h)$ є вибірковою автокореляцією залишків на лазі h , а K є максимальним лагом. За нульової гіпотези, статистика Q_{LB} асимптотично розподілена як хі-квадрат з $K - p - q$ ступенями свободи. Значуще значення статистики вказує на наявність залишкової автокореляції і необхідність перегляду моделі.

Додатково перевіряється нормальність розподілу залишків за допомогою гістограм, Q-Q графіків та формальних тестів, таких як тест Харке-Бера (Jarque-Bera) або тест Шапіро-Вілка [46]. Гетероскедастичність (непостійність дисперсії) виявляється візуальним аналізом графіка залишків у часі або тестом Бройша-Пагана [111].

Відбір моделі серед конкуруючих специфікацій здійснюється за інформаційними критеріями, що балансують якість підгонки та складність моделі [14, 47]. Інформаційний критерій Акаїке (Akaike Information Criterion, AIC) визначається як:

$$AIC = -2l + 2k, \quad (2.52)$$

де l є максимізованим логарифмом правдоподібності, а k є кількістю оцінених параметрів. Байєсівський інформаційний критерій (Bayesian Information Criterion, BIC або Schwarz criterion) накладає більший штраф на складність:

$$\text{BIC} = -2l + k \ln(T), \quad (2.53)$$

де T є розміром вибірки. Модель з найменшим значенням AIC або BIC є преференційною. BIC має тенденцію обирати більш парсимонічні моделі, особливо для великих вибірок, завдяки сильнішому штрафу за кількість параметрів [47].

Прогнозування з моделі ARIMA здійснюється ітеративною екстраполяцією рівняння моделі у майбутнє [14, 16]. Для моделі ARIMA(p, d, q), прогноз на h кроків вперед від часу T , позначений $\widehat{X}_{T+h|T}$ обчислюється як умовне математичне сподівання:

$$X_{T+h|T} = E[\widehat{X}_{T+h} | X_T, X_{T-1}, \dots]. \quad (2.54)$$

Для моделі AR(p)

$$X_{T+h|T} = c + \phi_1 \widehat{X}_{T+h-1|T} + \phi_2 \widehat{X}_{T+h-2|T} + \dots + \phi_p \widehat{X}_{T+h-p|T}, \quad (2.55)$$

де $\widehat{X}_{T+j|T} = X_{T+j}$ для $j \leq 0$. Для моделей з компонентом ковзного середнього, прогнози обчислюються рекурсивно з використанням прогнозованих інновацій: $\widehat{\varepsilon}_{T+j|T}$ для $j > 0$ [14].

Розширення моделі ARIMA включають ARIMAX (ARIMA з екзогенними змінними), яка додає зовнішні предиктори до рівняння моделі [112]:

$$\Phi(B)\nabla^d X_t = c + \sum_{j=1}^r \beta_j Z_{j,t} + \Theta(B)\varepsilon_t, \quad (2.56)$$

де $Z_{j,t}$ є екзогенними змінними (наприклад, температура, вологість, політичні втручання), а β_j є коефіцієнтами регресії. Модель ARIMAX дозволяє включити додаткову інформацію, що може покращити прогнозу точність, якщо екзогенні змінні мають причинний зв'язок з часовим рядом і самі є передбачуваними на горизонті прогнозування [50].

Векторна авторегресія (Vector Autoregression, VAR) узагальнює модель AR на багатовимірний випадок, де кілька часових рядів моделюються спільно з урахуванням їх взаємних лагових зв'язків [113]:

$$X_t = c + \Phi_1 X_{t-1} + \Phi_2 X_{t-2} + \dots + \Phi_p X_{t-p} + \varepsilon_t, \quad (2.57)$$

де X_t є вектором часових рядів, Φ_i є матрицями коефіцієнтів, а ε_t є вектором білих шумів. Модель VAR корисна для аналізу динамічних взаємозв'язків між декількома епідемічними індикаторами (наприклад, захворюваність у різних регіонах) [113].

Моделі із змінною у часі дисперсією, такі як авторегресійна умовна гетероскедастичність (Autoregressive Conditional Heteroskedasticity, ARCH) та узагальнена ARCH (GARCH), моделюють непостійність дисперсії шляхом специфікації умовної дисперсії як функції минулих квадратів залишків [114]:

$$\sigma_t^2 = \omega + \sum_{i=1}^q \alpha_i \varepsilon_{t-i}^2 + \sum_{j=1}^p \beta_j \sigma_{t-j}^2, \quad (2.58)$$

де σ_t^2 є умовною дисперсією на часі t . Моделі GARCH застосовуються до фінансових часових рядів з кластерами волатильності, але рідше до епідемічних рядів [114].

Практичне застосування моделей ARIMA для епідемічного прогнозування вимагає врахування специфіки даних. Епідемічні часові ряди часто мають невід'ємні значення (кількість випадків), що може порушуватися при прямому застосуванні моделі ARIMA, яка припускає нормальний розподіл інновацій і може продукувати від'ємні прогнози [12]. Логарифмічне перетворення $Y_t = \ln(X_t + 1)$ стабілізує дисперсію та забезпечує невід'ємність прогнозів після зворотного перетворення: $\widehat{X}_t = \exp(\widehat{Y}_t) - 1$ [14]. Альтернативно, можна використовувати моделі узагальненої авторегресії для дискретних даних підрахунків, такі як INAR (Integer-valued Autoregressive) або моделі Пуассонівської авторегресії [115].

Сезонна декомпозиція часового ряду на тренд, сезонний компонент та залишки може бути корисною для візуалізації структури даних перед моделюванням [16]. Метод STL (Seasonal and Trend decomposition using Loess) є робастною процедурою декомпозиції, що використовує локально зважену регресію [116]:

$$X_t = T_t + S_t + R_t, \quad (2.59)$$

де T_t є трендом, S_t є сезонним компонентом, а R_t є залишками. Після декомпозиції, модель ARIMA може бути застосована до десеоналізованого ряду $X_t - S_t$, а фінальний прогноз формується додаванням прогнозованого сезонного компонента [16].

Перевагами моделей ARIMA є математична прозорість, інтерпретованість параметрів, добре розроблена статистична теорія для оцінювання та виводу, а також ефективність обчислень [14, 47]. Інтервали невизначеності прогнозів мають теоретичне обґрунтування і можуть бути обчислені аналітично. Моделі ARIMA добре працюють для короткострокових прогнозів (1–4 тижні вперед) часових рядів із стабільною статистичною структурою [16].

Обмеженнями є припущення лінійності, яке не дозволяє моделювати складні нелінійні взаємодії; потреба у стаціонарності після диференціювання, що може бути проблематичною для рядів із структурними зламами; і погіршення точності прогнозів на великих горизонтах через зростання інтервалів невизначеності [14]. Моделі ARIMA також вимагають ретельного відбору порядків, що може бути суб'єктивним та вимагати експертного судження [47]. Для рядів із складною багаторівневою сезонністю або екзогенними втручаннями, моделі ARIMA можуть вимагати складної специфікації з великою кількістю параметрів, що зменшує парсимонічність.

Автоматизовані процедури відбору моделей, такі як алгоритм Хайндмана-Хандакара, реалізований у пакеті `forecast` для R, виконують систематичний пошук по простору можливих моделей ARIMA з використанням інформаційних критеріїв для автоматичного відбору оптимальних порядків [117]. Ці процедури зменшують суб'єктивність відбору моделі та дозволяють швидко отримати розумні базові прогнози, хоча експертний аналіз залишається важливим для критичних застосувань.

Порівняння моделей ARIMA з методами машинного навчання показує, що для епідемічних часових рядів із стабільною лінійною структурою та чіткою сезонністю, ARIMA часто забезпечує порівнянну або кращу точність при значно меншій обчислювальній складності та кращій інтерпретованості [14, 16]. Водночас для рядів із нелінійними залежностями, структурними зламами або складними взаємодіями з екзогенними факторами, методи машинного навчання, такі як LSTM, можуть перевершувати ARIMA [17, 21]. Гібридні підходи, що поєднують ARIMA для моделювання лінійних компонентів та нейронні мережі для нелінійних залишків, часто досягають найкращої продуктивності [57–58].

2.2.1 Методи вибору порядків моделі

Побудова ARIMA моделі за методологією Бокса-Дженкінса включає послідовні етапи ідентифікації порядків, оцінювання параметрів, діагностики та прогнозування. Ідентифікація порядків p , d та q є критичним кроком, що визначає структуру моделі.

Порядок диференціювання d визначається через тестування стаціонарності вихідного ряду [45–46]. Розширений тест Дікі-Фуллера (Augmented Dickey-Fuller, ADF) перевіряє нульову гіпотезу про наявність одиничного кореня, що свідчить про нестационарність. Якщо p -значення тесту перевищує поріг значущості (типово 0.05), ряд вважається нестационарним і потребує диференціювання. Процес повторюється до досягнення стаціонарності або досягнення розумної межі порядку d (зазвичай не більше 2).

Порядок авторегресії p та ковзного середнього q визначаються через аналіз автокореляційної функції (ACF) та часткової автокореляційної функції (PACF) стаціонаризованого ряду. ACF показує кореляцію між рядом і його запізненими копіями на різних лагах. PACF показує кореляцію після видалення впливу проміжних лагів.

Для чистого AR(p) процесу PACF різко обривається після лагу p , тоді як ACF поступово згасає. Для чистого MA(q) процесу навпаки - ACF різко обривається після лагу q , а PACF поступово згасає. Для змішаних ARMA процесів обидві функції поступово згасають, що ускладнює ідентифікацію порядків.

Додатковим інструментом вибору порядків є інформаційні критерії, що балансують точність апроксимації та складність моделі [118–119]. Інформаційний критерій Акаїке (AIC) обчислюється як $AIC = 2k - 2\ln(L)$, де k - кількість параметрів, L - максимальна правдоподібність. Байєсівський інформаційний критерій (BIC) аналогічно штрафує складність, але більш жорстко: $BIC = k \cdot \ln(n) - 2\ln(L)$, де n - кількість спостережень.

Систематичний підхід передбачає побудову моделей з різними комбінаціями порядків у розумних межах (наприклад, p, q від 0 до 5, d від 0 до 2), обчислення AIC та BIC для кожної конфігурації та вибір моделі з найнижчим значенням критерію. Цей процес часто автоматизується через *grid search* з подальшою верифікацією через діагностику залишків.

Оцінювання параметрів обраної моделі здійснюється методом максимальної правдоподібності (Maximum Likelihood Estimation, MLE), який знаходить значення коефіцієнтів, що максимізують ймовірність спостереження наявних даних при заданій структурі моделі [12, 112]. Для цього використовуються чисельні методи оптимізації, такі як алгоритм Левенберга-Марквардта або BFGS.

2.2.2 Попередня обробка та діагностика

Попередня обробка даних для ARIMA моделей має певну специфіку порівняно з нейронними мережами. Для річних рядів лептоспірозу, що мають відносно невеликі абсолютні значення та помірну варіабельність, нормалізація не застосовувалася. Моделі будувалися безпосередньо на оригінальних значеннях кількості випадків, що спрощує інтерпретацію параметрів та прогнозів.

Для місячного ряду грипу була застосована нормалізація до інтервалу від 1 до 2 замість стандартного $[0,1]$. Така нормалізація має дві переваги: вона уникає нульових значень, що можуть створювати проблеми при обчисленні відсоткових метрик, та зберігає співвідношення між рівнями захворюваності без стиснення в нульову точку. Перетворення виконується за формулою

$$x_{scaled} = a + \frac{(x - \min(x))(b - a)}{\max(x) - \min(x)}, \quad (2.60)$$

де x це початкова точка даних, x_{scaled} нормована точка даних, а та b бажаний діапазон масштабування.

Діагностика побудованої ARIMA моделі включає аналіз залишків (різниць між фактичними та модельованими значеннями). Якісна модель повинна мати залишки, що нагадують білий шум: випадкові, незалежні, нормально розподілені з нульовим середнім та постійною дисперсією.

Тест Льюнга-Бокса перевіряє нульову гіпотезу про відсутність автокореляції залишків до певного лагу. Якщо p -значення тесту перевищує поріг значущості, залишки вважаються некорельованими, що свідчить про адекватність моделі. Якщо тест відхиляє нульову гіпотезу, це вказує на наявність не поясненої структури в залишках та необхідність перегляду порядків моделі.

Візуальна діагностика включає побудову графіків залишків у часі для перевірки відсутності трендів та систематичних патернів, гістограми залишків для оцінки нормальності розподілу, ACF залишків для виявлення автокореляцій та Q-Q графіка для порівняння розподілу залишків з теоретичним нормальним розподілом.

2.2.3 Функції втрат, метрики оцінювання та валідаційні стратегії

У першій ітерації основними функціями втрат та метриками оцінювання були MSE (2.25), MAE (2.26) та MAPE, визначені у п. 2.1. Додатково використовувався $RMSE = \sqrt{MSE}$ для інтерпретації похибки у фізичних одиницях [49, 120].

Ця функція втрат має кілька важливих властивостей. Вона надає більшій ваги великим помилкам через квадратичну форму, що змушує модель особливо ретельно прогнозувати екстремальні значення. MSE є диференційованою в усіх точках, що забезпечує стабільність обчислення градієнтів. Однак MSE чутлива до викидів і може призводити до переоцінки впливу аномальних спостережень.

Ця метрика має перевагу безрозмірності та інтуїтивної інтерпретованості у відсотках. MAPE добре підходить для порівняння точності моделей на різних захворюваннях з різними масштабами абсолютних значень. Проте MAPE має проблему з нульовими або близькими до нуля значеннями, де відносна похибка може необмежено зростати.

Додаткові метрики, що використовувалися для оцінювання якості моделей, включали середню абсолютну похибку (Mean Absolute Error), та корінь із середньоквадратичної похибки (Root Mean Squared Error) [13, 120]. MAE менш чутлива до викидів порівняно з MSE і надає похибку у тих самих одиницях, що й вихідні дані. RMSE поєднує переваги MSE у навчанні з більш інтуїтивною інтерпретацією у вихідних одиницях.

Оцінювання проводилося як у масштабованому просторі (після нормалізації) для контролю процесу навчання, так і у фізичних одиницях (після зворотного перетворення) для практичної інтерпретації результатів. Для моделі сальмонельозу, наприклад, MAPE у масштабованому просторі становила 0.29%, але після інверсії нормалізації збільшувалася до близько 7%, що відображає реальну точність прогнозування фізичної кількості випадків.

Валідація моделей у першій ітерації базувалася на простих підходах із часовим поділом даних, що відображало початковий етап методологічного розвитку. Основною стратегією був хронологічний поділ доступного часового ряду на навчальну та тестову множини у співвідношенні приблизно 80% до 20% з обов'язковим збереженням часового порядку спостережень.

Навчальна множина використовувалася для оптимізації параметрів моделі через мінімізацію функції втрат методами стохастичного градієнтного спуску. Тестова множина, що складається з найбільш свіжих спостережень, зберігалася незайманою до фінального оцінювання та не використовувалася жодним чином під час налаштування гіперпараметрів або вибору архітектури.

У деяких експериментах додатково виділялася валідаційна множина з кінця навчального періоду для моніторингу перенавчання та реалізації

раннього зупинення. Типовий розподіл у таких випадках становив 70% для навчання, 10% для валідації та 20% для фінального тестування.

Ключовим обмеженням цього підходу була відсутність систематичної перевірки стабільності прогнозів при послідовному зміщенні вікна даних у часі. Модель оцінювалася лише на одному фіксованому тестовому періоді, що могло призводити до оптимістичної або песимістичної оцінки продуктивності залежно від специфіки обраного інтервалу тестування.

2.3 Аналіз стаціонарності часових рядів

Стаціонарність часового ряду є фундаментальною властивістю, що визначає можливість застосування більшості статистичних методів аналізу та прогнозування. Часовий ряд $\{y_t\}$ є слабо стаціонарним, якщо його математичне сподівання $E[y_t] = \mu$ є сталою в часі, а автоковаріація $\text{Cov}(y_t, y_{t+h}) = \gamma(h)$ залежить лише від лагу h , а не від абсолютного часу t . Для прикладного аналізу часових рядів, зокрема обчислення автокореляційної функції та побудови моделей ARIMA, використовують саме це визначення стаціонарності.

Епідемічні часові ряди часто демонструють нестаціонарну поведінку через різноманітні фактори. Процес з одиничним коренем, найпростішим прикладом якого є випадкове блукання

$$y_t = y_{t-1} + \varepsilon_t \quad (2.61)$$

має дисперсію, що зростає лінійно з часом, порушуючи умову стаціонарності. Ряд із детермінованим трендом

$$y_t = \alpha + \beta t + u_t, \quad (2.62)$$

Є нестационарним через систематичну зміну математичного сподівання у часі. Невилучені сезонні компоненти спричиняють періодичну нестационарність, при якій статистичні властивості ряду змінюються залежно від положення спостереження всередині річного циклу. Гетероскедастичність, тобто зміна дисперсії з часом, характерна для епідемічних рядів під час спалахів захворювань. Структурні зміни внаслідок зовнішніх подій, таких як зміни політик охорони здоров'я або початок військових дій, створюють різкі зміни параметрів ряду.

Розширений тест Дікі-Фуллера є основним інструментом перевірки наявності одиничного кореня. Тест базується на регресійному рівнянні

$$\Delta y_t = \alpha + \beta t + \gamma y_{(t-1)} + \sum_{(i=1)}^p \delta_i \Delta y_{(t-i)} + \varepsilon_t, \quad (2.63)$$

$$\Delta y_t = y_t - y_{(t-1)}, \quad (2.64)$$

Що являється першою різницею ряду, α є константою, β є коефіцієнтом тренду, а γ є ключовим параметром, що визначає наявність одиничного кореня. Нульова гіпотеза $H_0: \gamma = 0$ стверджує наявність одиничного кореня та нестационарність ряду, тоді як альтернативна гіпотеза $H_1: \gamma < 0$ свідчить про стаціонарність. Тестова статистика $\tau = \hat{\gamma}/SE(\hat{\gamma})$

порівнюється з критичними значеннями, табульованими Дікі та Фуллером. Кількість лагів p визначається за інформаційними критеріями AIC або BIC, або за емпіричною формулою $p \approx [12 \cdot (T/100)^{1/4}]$ для сезонних даних.

Тест KPSS має протилежну структуру гіпотез: його нульова гіпотеза стверджує стаціонарність ряду. Модель KPSS базується на декомпозиції

$$y_t = r_t + u_t, \quad (2.65)$$

де $r_t = r_{t-1} + \xi_t$ є випадковим блуканням, а u_t є стаціонарним компонентом. Статистика KPSS обчислюється як

$$KPSS = (1/T^2) \sum_{t=1}^T S_t^2 / \widehat{\sigma}^2, \quad (2.66)$$

де $S_t = \sum_{i=1}^t \widehat{u}_i$ є частковою сумою залишків, а $\widehat{\sigma}^2$ є оцінкою довготривалої дисперсії методом Newey-West за формулою

$$\widehat{\Omega}_{HAC} = \frac{1}{T} \sum_{t=1}^T \widehat{\varepsilon}_t^2 x_t x_t' + \frac{1}{T} \sum_{j=1}^L w_j \sum_{t=j+1}^T (\widehat{\varepsilon}_t \widehat{\varepsilon}_{t-j} x_t x_{t-j}' + \widehat{\varepsilon}_{t-j} \widehat{\varepsilon}_t x_{t-j} x_t'), \quad (2.67)$$

$\widehat{\Omega}_{HAC}$ це спроможна оцінка коваріаційної матриці, стійка до гетероскедастичності та автокореляції (Heteroskedasticity and Autocorrelation Consistent). T загальна кількість спостережень у часовому ряді. $\widehat{\varepsilon}_t^2$ це залишки (residuals) отримані після оцінювання моделі методом найменших квадратів (OLS). x_t це вектор пояснювальних змінних (регресорів) у момент часу t . L це параметр згладжування (bandwidth), що визначає максимальну кількість лагів, які враховуються при корекції. w_j вагові коефіцієнти ядра Бартлетта, що обчислюються за формулою:

Велике значення статистики KPSS свідчить на користь нестаціонарності. Оскільки тести ADF та KPSS мають протилежні нульові гіпотези, їх одночасне застосування дозволяє отримати надійніші висновки. Якщо p -значення ADF менше порогу значущості α та p -значення KPSS більше α обидва тести узгоджено свідчать про стаціонарність ряду. Якщо ADF не відхиляє свою нульову гіпотезу, а KPSS відхиляє, ряд ймовірно має одиничний корінь і потребує диференціювання. Неоднозначний випадок, коли обидва тести відхиляють свої нульові гіпотези, може вказувати на структурні зміни або нелінійну стаціонарність і потребує додаткового аналізу, зокрема візуальної інспекції ряду або тестування зі зломом.

Вимоги до стаціонарності суттєво відрізняються для різних класів моделей. Для моделей ARIMA та SARIMA стаціонарність є обов'язковою умовою, оскільки порушення цього припущення призводить до некоректних оцінок параметрів та ненадійних прогнозів. Порядок диференціювання d визначається на основі результатів тестів ADF та KPSS. Для мереж LSTM стаціонарність не є теоретичною вимогою, проте стабілізація математичного сподівання та дисперсії часто покращує збіжність навчання та узагальнюючу здатність моделі, що підтверджується експериментально.

Рекомендований робочий пайплайн включає візуальну діагностику ряду з побудовою ACF та PACF, проведення тестів ADF та KPSS з документуванням обраних параметрів, застосування трансформацій при потребі з повторним тестуванням, та верифікацію результатів. При використанні walk-forward валідації критично важливо, щоб параметри всіх трансформацій обчислювалися виключно на навчальній частині вибірки на кожному кроці для запобігання витоку інформації з тестового періоду. Після побудови моделі необхідно провести діагностику залишків для перевірки адекватності специфікації через аналіз ACF залишків та тест Льюнга-Бокса на відсутність автокореляції.

2.4 Архітектурні рішення першої ітерації

Перша ітерація досліджень базувалася на відносно простих архітектурних конфігураціях з метою перевірки базової здатності LSTM моделювати епідемічні часові ряди. Використовувалися два основні типи архітектур: компактні двошарові та глибші тришарові конфігурації.

Для досліджень гепатиту А була застосована компактна архітектура, що складалася з двох послідовних LSTM шарів по 64 одиниці пам'яті кожен та щільного вихідного шару з одним нейроном для генерації прогнозу на один

крок вперед. Така конфігурація забезпечувала баланс між виразною потужністю моделі та ризиком перенавчання на відносно коротких рядах.

Для прогнозування грипу використовувалася багат шарова архітектура з регуляризацією методом Dropout, що дозволяє моделювати складнішу динаміку захворюваності з вираженими сезонними піками. Точна конфігурація включала послідовність рекурентних шарів з проміжним відсіванням нейронів для запобігання перенавчанню.

Найглибша архітектура першої ітерації була застосована для сальмонельозу і включає три послідовні LSTM шари з конфігурацією 256, 128 та 64 одиниці пам'яті відповідно. Така пірамідальна структура з поступовим зменшенням розмірності дозволяла створювати ієрархічне представлення даних, де нижні шари виявляють прості патерни, а верхні інтегрують їх у складніші залежності. Регуляризація Dropout з коефіцієнтом 0.20 застосовувалася після кожного LSTM шару для контролю перенавчання. Щільний вихідний шар із однією одиницею генерував фінальний прогноз.

Загальна кількість параметрів у цих архітектурах варіювалася від десятків тисяч для компактних двошарових конфігурацій до сотень тисяч для глибокої тришарової мережі сальмонельозу, що залишалося в межах можливостей навчання на доступних епідемічних даних.

2.4.1 Методи попередньої обробки даних

Попередня обробка епідемічних часових рядів у першій ітерації включала базові процедури нормалізації та формування навчальних прикладів. Основним методом масштабування було min-max перетворення, яке приводило всі значення до діапазону від 0 до 1 за формулою обчислювалися на навчальній множині.

Така нормалізація виконує дві важливі функції. По-перше, вона забезпечує стабільність процесу навчання шляхом приведення вхідних даних

до діапазону, в якому сигмоїдні та гіперболічні активації працюють найефективніше. По-друге, нормалізація запобігає домінуванню змінних з більшим масштабом над змінними з меншим масштабом при обчисленні градієнтів.

Формування навчальних послідовностей здійснювалося методом ковзного вікна, де для прогнозування значення в момент t використовувалася послідовність L попередніх значень від $t-L$ до $t-1$. Довжина вікна варіювалася залежно від специфіки захворювання та доступності даних, але типово становила від 6 до 12 місяців для місячних рядів, що відповідає одному повному сезонному циклу або його половині.

У першій ітерації не застосовувалися складні методи *feature engineering*, такі як створення циклічних сезонних ознак або додавання похідних змінних. Передбачалося, що LSTM архітектура самостійно виявить необхідні патерни безпосередньо з сирих нормалізованих даних. Цей підхід був обраний для оцінки базової здатності мереж до автоматичного навчання на епідемічних даних без експертного втручання.

2.4.2 Навчання, оптимізація та регуляризація

Процес навчання LSTM моделей у першій ітерації базувався на методах стохастичної оптимізації з адаптивною швидкістю навчання. Основними оптимізаторами були Adam (Adaptive Moment Estimation) та Adamax (варіант Adam з нескінченною нормою).

Оптимізатор Adam поєднує переваги двох підходів: збереження експоненціально згасаючого середнього градієнтів (момент першого порядку) та експоненціально згасаючого середнього квадратів градієнтів (момент другого порядку) [55]. Це дозволяє автоматично адаптувати швидкість навчання для кожного параметра окремо, що особливо корисно при роботі з різними типами параметрів у LSTM комірках. Типова швидкість навчання становила 0.001 з параметрами $\beta_1=0.9$ та $\beta_2=0.999$ за замовчуванням.

Adamax, застосований для моделі сальмонельозу, використовує нескінченну норму замість L2 норми для обчислення другого моменту, що робить його більш стійким до великих градієнтів та може бути корисним для рядів з різкими викидами.

Регуляризація здійснювалася переважно методом Dropout, який під час навчання випадково відключає певну частку нейронів (типово 20%) на кожній ітерації [101, 121]. Це змушує мережу не покладатися на будь-який окремий нейрон і формувати більш розподілені робастні представлення даних. Dropout застосовувався між LSTM шарами для запобігання перенавчанню на обмежених епідемічних датасетах.

Раннє зупинення (early stopping) використовувалося як додатковий механізм регуляризації [49, 122]. Навчання припиняється, якщо похибка на валідаційній множині не покращувалася протягом заданої кількості епох (patience), типово від 5 до 20 епох. Це дозволяє автоматично визначати оптимальну кількість епох навчання без ручного налаштування і запобігало подальшому перенавчанню після досягнення мінімуму валідаційної похибки

Кількість епох навчання варіювалася від 100 до 550 залежно від складності задачі та збіжності процесу. Розмір міні-батчів становив від 8 до 32 прикладів, що забезпечувало баланс між стабільністю оцінки градієнта та ефективністю обчислень.

2.5 Висновки до розділу

Розділ присвячений розробленню методів та теоретичних основ прогнозування епідемічних часових рядів на основі рекурентних нейронних мереж з довгою короткочасною пам'яттю та моделей авторегресії з проінтегрованим ковзним середнім.

Формалізовано математичний апарат рекурентних нейронних мереж та мереж довгої короткочасної пам'яті. Показано, що класичні рекурентні мережі страждають від проблеми зникаючого градієнта при навчанні на довгих послідовностях через мультиплікативну структуру оновлення прихованого стану, де добуток якобіанів експоненціально зменшується зі збільшенням часового інтервалу між пов'язаними спостереженнями. Архітектура LSTM вирішує цю проблему через введення системи вентилів забування, входу та виходу, що контролюють потік інформації до стану комірки. дозволяє градієнтам протікати через багато часових кроків без деградації.

Описано систему рівнянь LSTM комірки, що включає обчислення вентиля забування, вентиля входу, вентиля виходу, кандидата на новий стан комірки, оновлення стану комірки та обчислення прихованого стану. Для LSTM комірки з n входами та m прихованими одиницями загальна кількість параметрів дорівнює $4(nm + m^2 + m)$, що робить LSTM значно більш параметризованою порівняно з класичною RNN.

Розроблено методологію побудови моделей ARIMA за методологією Бокса-Дженкінса. Формалізовано процеси авторегресії порядку p та ковзного середнього порядку q , умови стаціонарності та оборотності через аналіз коренів характеристичних поліномів. Модель ARIMA(p,d,q) застосовує ARMA(p,q) до d -ї різниці часового ряду, що дозволяє моделювати нестационарні ряди з трендами. Сезонна модель SARIMA(p,d,q) \times (P,D,Q) включає як несезонні, так і сезонні авторегресійні та ковзні компоненти для моделювання періодичних коливань з річним або іншим циклом.

Систематизовано підходи до аналізу стаціонарності часових рядів. Розроблено комбіновану стратегію інтерпретації результатів тестів ADF та KPSS, які мають протилежні нульові гіпотези. Якщо p -значення ADF менше порогу значущості та p -значення KPSS більше порогу, ряд ймовірно стаціонарний. Якщо ADF не відхиляє нульову гіпотезу, а KPSS відхиляє, ряд ймовірно має одиничний корінь. Неоднозначний випадок, коли обидва тести

відхиляють свої нульові гіпотези, може вказувати на структурні злами або нелінійну стаціонарність і потребує додаткового аналізу.

Обґрунтовано необхідність приведення часових рядів до стаціонарності для коректного застосування методів моделювання. Для моделей ARIMA та SARIMA стаціонарність є обов'язковою умовою, оскільки порушення цього припущення призводить до некоректних оцінок параметрів та ненадійних прогнозів. Для LSTM мереж стаціонарність не є теоретичною вимогою, проте стабілізація середнього та дисперсії часто покращує збіжність навчання та узагальнюючу здатність моделі, що підтверджується експериментально.

Розроблено методи попередньої обробки епідемічних часових рядів для LSTM моделей. Основним методом масштабування обрано min-max нормалізацію до діапазону від нуля до одиниці за формулою $x_{norm} = (x - x_{min}) / (x_{max} - x_{min})$, де параметри обчислюються виключно на навчальній множині для уникнення витоку інформації з тестового періоду. Нормалізація забезпечує стабільність градієнтного спуску та ефективну роботу сигмоїдних активаційних функцій у вентилях LSTM комірок, які виробляють значення в інтервалі від нуля до одиниці.

Формалізовано метод формування навчальних послідовностей ковзним вікном, де для часового ряду $\{x_1, x_2, \dots, x_T\}$ з вікном довжини w та горизонтом прогнозування h формується набір пар вхід-вихід. Вхід представляє послідовність w попередніх значень, а вихід є наступним значенням або послідовністю h майбутніх значень для багатокрокового прогнозування. Емпіричні дослідження показали, що для епідемічних часових рядів з тижневою або місячною сезонністю оптимальні значення довжини вікна становлять від 12 до 52 тижнів.

Описано функції втрат та метрики оцінювання для задач прогнозування часових рядів. Середньоквадратична похибка MSE надає більшої ваги великим помилкам через квадратичну форму, що стимулює модель особливо ретельно прогнозувати екстремальні значення. Середня абсолютна похибка MAE менш

чутлива до викидів та надає похибку у фізичних одиницях вихідних даних. Середня абсолютна відсоткова похибка MAPE нормалізує помилки відносно фактичних значень та дозволяє порівнювати моделі на різних захворюваннях, проте має проблему з нульовими або близькими до нуля значеннями.

Обґрунтовано вибір валідаційних стратегій для моделей прогнозування часових рядів. Основною стратегією першої ітерації був хронологічний поділ доступного часового ряду на навчальну та тестову множини у співвідношенні приблизно 80% до 20% з обов'язковим збереженням часового порядку спостережень. Виявлено обмеження цього підходу, що полягає у відсутності систематичної перевірки стабільності прогнозів при послідовному зміщенні вікна даних у часі, що стало основою для методологічних покращень у наступних ітераціях.

Розроблено архітектурні рішення для LSTM моделей прогнозування епідемічних часових рядів. Для коротких рядів з 72 місяцями спостережень застосовано компактну двошарову архітектуру з 64 одиницями пам'яті на кожен шар та загальною кількістю параметрів близько 171 тисячі. Для довших та складніших рядів з 132 місяцями спостережень застосовано тришарову пірамідальну архітектуру з конфігурацією 256, 128 та 64 одиниці, де поступове зменшення розмірності створює ієрархічне представлення часових залежностей різних масштабів.

Описано методи регуляризації та оптимізації для навчання LSTM мереж. Метод Dropout з коефіцієнтом 0.20 застосовується між рекурентними шарами для випадкового відключення частини нейронів під час навчання, що змушує мережу формувати розподілені робастні представлення. Раннє зупинення припиняє навчання, якщо похибка на валідаційній множині не покращується протягом заданої кількості епох, що автоматично визначає оптимальну тривалість навчання. Оптимізатор Adam з адаптивною швидкістю навчання забезпечує автоматичне налаштування кроку для кожного параметра на основі перших та других моментів градієнтів.

Загалом, розроблені методи та теоретичні основи забезпечують науково обґрунтовану базу для побудови систем прогнозування епідемічних часових рядів. Формалізовані математичні моделі LSTM та ARIMA, систематизовані підходи до аналізу стаціонарності та попередньої обробки даних, обґрунтовані критерії вибору архітектурних рішень та параметрів регуляризації створюють методологічну основу для проведення експериментальних досліджень та практичного впровадження розроблених методів.

Основні результати розділу опубліковано в працях [1, 4, 6, 8, 9].

РОЗДІЛ 3. ПОКРАЩЕННЯ МЕТОДІВ ТА ПРИКЛАДНЕ ЗАСТОСУВАННЯ МЕТОДІВ ПРОГНОЗУВАННЯ ЕПІДЕМІЧНИХ ПРОЦЕСІВ

3.1 Методологічні вдосконалення третьої ітерації LSTM архітектури

Третя ітерація інтегрувала уроки та обмеження попередніх етапів для створення методологічно строгих та технічно вдосконалених LSTM систем прогнозування. Основні інновації спрямовувалися на підвищення надійності валідації, покращення моделювання сезонності та систематичне дослідження різних представлень даних.

Введення sliding window валідації дозволило оцінити стабільність прогнозів при послідовному зміщенні вікна навчальних даних у часі. Замість одноразового фіксованого поділу на навчальну та тестову множини, модель оцінювалася на послідовності часових вікон. Наприклад, перше вікно включає місяці 1-60 для навчання та місяць 61 для тестування, друге вікно - місяці 1-61 для навчання та місяць 62 для тестування, і так далі. Це дозволяє отримати розподіл помилок прогнозування замість однієї точкової оцінки.

Walk-forward валідація, також відома як rolling forecast origin, передбачає послідовне розширення або зміщення навчального вікна з обчисленням прогнозу на один крок вперед виключно на основі минулої інформації без витоку даних з майбутнього. Цей підхід максимально наближений до реального режиму застосування моделі, коли щомісяця потрібно генерувати прогноз на наступний період на основі всіх доступних на той момент даних.

Застосування циклічних сезонних ознак вирішує фундаментальну проблему представлення періодичного часу в моделях машинного навчання. Пряме використання номера місяця як числової змінної (1, 2, ..., 12) створює штучний розрив між груднем (12) та січнем (1), хоча ці місяці є сусідніми в

річному циклі. Циклічне кодування через синусоїдальні та косинусоїдальні функції вирішує цю проблему:

$$month_{sin} = \sin\left(\frac{2\pi \cdot month}{12}\right) \quad (3.1)$$

$$month_{cos} = \cos\left(\frac{2\pi \cdot month}{12}\right) \quad (3.2)$$

Ці дві ознаки разом забезпечують безперервне представлення річного циклу, де січень плавно переходить у грудень. Прикладне порівняння циклічної компоненти в порівнянні з часовим рядом зображено на рисунку 3.2.

Систематичне порівняння рівневих та кумулятивних представлень даних стало ключовою інновацією третьої ітерації. Рівневе представлення відповідає абсолютній кількості випадків захворювання за місяць як є. Кумулятивне представлення обчислюється як накопичена сума від початку періоду.

Кумулятивне представлення має кілька важливих властивостей. Воно згладжує випадкові місячні коливання через інтегрування, що зменшує відносну дисперсію та може покращувати стабільність прогнозів. Кумулятивна траєкторія зростає монотонно (за винятком випадків з від'ємними значеннями, що неможливо для кількості випадків), що спрощує задачу моделювання. Відносні похибки прогнозування кумулятивних значень часто нижчі, оскільки знаменник у MAPE зростає з часом. Проте кумулятивне представлення менш чутливе до короткострокових змін та спалахів, що може бути недоліком для задач раннього виявлення.

3.2 Вдосконалені архітектурні рішення

Архітектури третьої ітерації варіювалися від компактних конфігурацій до глибоких багатопшарових мереж залежно від специфіки задачі та доступного обсягу даних. Спільною рисою всіх архітектур була ретельна регуляризація та контроль складності для запобігання перенавчанню.

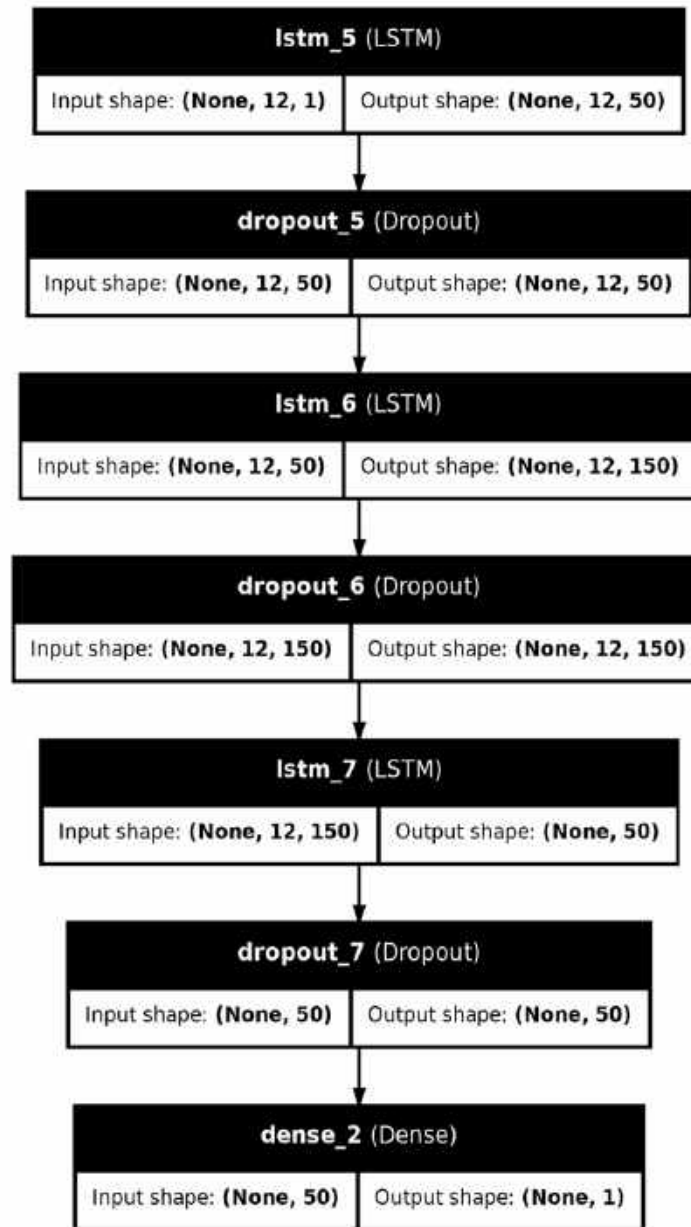


Рис. 3.1. Архітектура багатошарової LSTM-мережі для прогнозування часових рядів

Для досліджень гепатиту В була застосована компактна двошарова архітектура з 64 одиницями пам'яті на кожен LSTM шар, регуляризацією Dropout 0.20 між шарами та щільним вихідним шаром. Загальна кількість параметрів становила приблизно 171251. Така конфігурація забезпечувала достатню виразну потужність для моделювання динаміки гепатиту В при збереженні можливості навчання на відносно короткому періоді 2018–2023 років. Окремі мережі навчалися для гострої та хронічної форм захворювання, а

також для рівневого та кумулятивного представлень, що дозволяло спеціалізувати архітектуру для специфіки кожної підзадачі.

Для гепатиту С була застосована аналогічна компактна архітектура 2xLSTM(64) з Dropout 0.20, що також відповідала приблизно 171251 параметру. Ключовою відмінністю стало використання вікна довжини 12 місяців для формування вхідних послідовностей, що відповідає одному повному річному циклу та дозволяє моделі захоплювати річну сезонність. Нормалізація виконувалася до діапазону від 0 до 1 з обчисленням параметрів масштабування виключно на навчальній вибірці.

Для прогнозування гострих респіраторних інфекцій верхніх дихальних шляхів (ГРВІ) була побудована найбільш амбітна та складна архітектура третьої ітерації [18]. Мережа включала три глибокі LSTM шари з конфігурацією 300, 1200 та 600 одиниць відповідно, регуляризацію Dropout між шарами та щільний вихідний шар. Загальна кількість параметрів становила приблизно 11.89 мільйонів, що робить цю модель найбільшою у всьому дослідженні.

Така глибока архітектура була виправдана специфікою задачі: ГРВІ датасет охоплював тривалий період 2013–2024 років з 132 місяцями спостережень, що забезпечувало достатній обсяг даних для навчання багатопараметричної моделі. Високу мінливість даних ГРВІ та вплив екстраординарних подій, зокрема пандемії COVID-19 та повномасштабної війни [62–65], потребували потужної архітектури для адаптації до структурних зламів. Довжина вхідного вікна була збільшена до 16 місяців для забезпечення достатнього контексту для прогнозування.

3.3 Розширені методи попередньої обробки

Попередня обробка даних у третій ітерації включала як базові процедури нормалізації, успадковані з попередніх етапів, так і нові методи інженерії ознак для явного моделювання сезонності.

Нормалізація виконувалася методом min-max масштабування до діапазону $[0, 1]$ відповідно до процедури, описаної у п. 2.4.1. Це критично важливо для уникнення витoku інформації з тестового періоду, коли параметри нормалізації були обчислені на всьому датасеті включно з майбутніми даними. Ті самі параметри масштабування застосовувалися до валідаційної та тестової множин для забезпечення узгодженості.

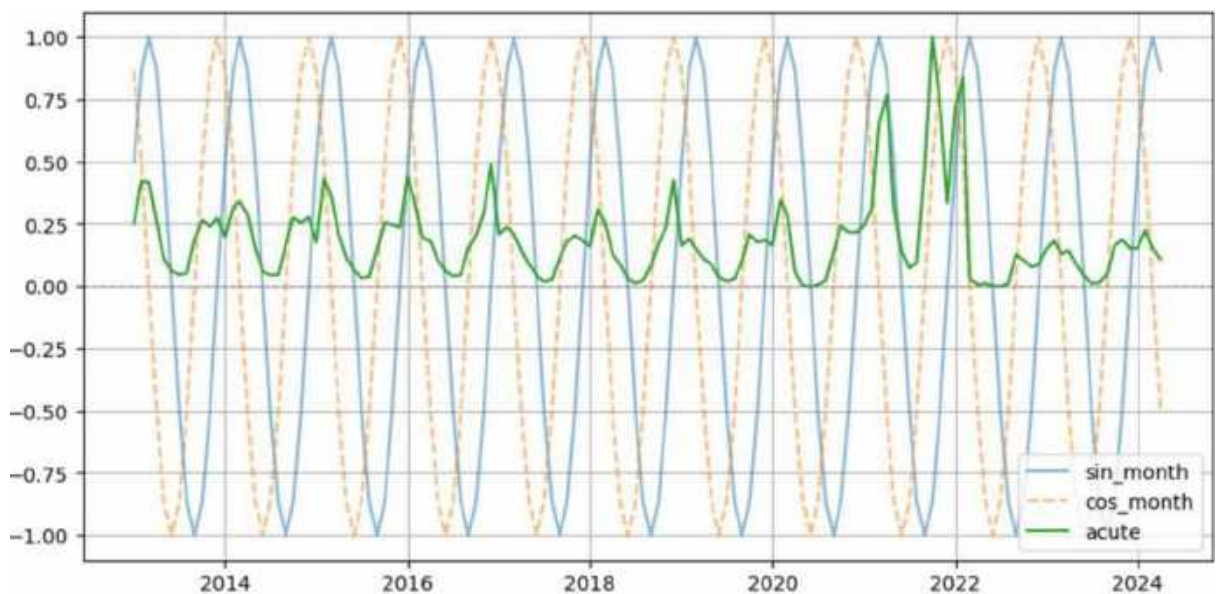


Рис. 3.2 Порівняння між сезональністю та вхідними даними

Ключовою інновацією стало додавання циклічних сезонних ознак на основі синусоїдальних та косинусоїдальних функцій від індексу місяця. Для кожного спостереження в часовому ряді обчислювалися дві додаткові змінні: за формулами (3.1) та (3.2) де month змінюється від 1 до 12.

Використання $(\text{month} - 1)$ замість month забезпечує, що січень ($\text{month}=1$) відображається в точці $(\sin(0), \cos(0)) = (0, 1)$, лютий - в $(\sin(\pi/6), \cos(\pi/6))$, і так далі по колу. Ця пара значень однозначно ідентифікує місяць у році та забезпечує безперервність між груднем та січнем: відстань між грудневими та

січневими точками на одиничному колі є такою ж, як між будь-якими іншими сусідніми місяцями.

Таблиця 3.1. Характеристики вдосконалених LSTM моделей третьої ітерації

Завдання	Гепатит В	Гепатит С	ГРВІ
Регіон	Україна	Україна	Україна
Місяців	72	72	132
Вікно L	12	12	16
Архітектура	2xLSTM(64) + Dropout 0.20 + Dense	2xLSTM(64) + Dropout 0.20 + Dense	3xLSTM(300/1200/600) + Dropout + Dense
Параметрів	~171251	~171251	~11.89M
Сезонні ознаки	Ні	Ні	Так (sin/cos)
Оптимізатор	Adam lr=1e-3 bs=8	Adam lr=1e-3 bs=8	Adam
Функція втрат	MAPE ціль MSE діагн	MAPE ціль MSE діагн	MSE функція втрат MAPE моніторинг
Валідація	Раннє зупинення окремі мережі	Walk-forward часові поділи	Grid-search рання зупинка

Пірамідальна структура 300-1200-600 одиниць відрізняється від монотонно спадної конфігурації першої ітерації. Розширення до 1200 одиниць у середньому шарі забезпечує високу репрезентативну потужність для інтеграції інформації з різних часових масштабів, тоді як звуження до 600 одиниць у третьому шарі виконує функцію узагальнення та виділення найбільш релевантних ознак для прогнозування.

Аналіз результатів моделювання демонструє чітку перевагу кумулятивного представлення даних, яке забезпечує значно вищу точність у порівнянні з щомісячними прогнозами. Найкращі показники зафіксовано для

моделі Гепатиту С (кумулятив) із мінімальною похибкою MAPE 0,281%, що свідчить про високу стабільність відтворення динаміки. Водночас щомісячне прогнозування для цієї ж інфекції демонструє найбільшу мінливість (MAPE 26,342%), вказуючи на складність вловлювання короткострокових коливань.

Таблиця 3.2. Показники точності вдосконалених LSTM моделей третьої ітерації

Модель	Представлення	MAPE (%)	MSE	RMSE	Особливості результатів
Гепатит В	Щомісячний прогноз	13.78	411.28	20.28	Описано збіжність навчання контроль перенавчання
Гепатит В	Кумулятив	0.85	3361.68	57.98	
Гепатит С	Кумулятив	0.281	68.245	8.26	Найнижча похибка дуже стійке відтворення
Гепатит С	Щомісячний прогноз	26.342	168.984	13	Вища мінливість помісячних значень
ГРВІ	Кумулятив до війни	06.05	12486199	111741.6	3 сезонними ознаками sin/cos
ГРВІ	Кумулятив після війни	3.23	30453423	55184.62	Покращення від налаштування адаптація до криз

Особливої уваги заслуговує модель ГРВІ, де після початку війни спостерігається покращення точності (зниження MAPE з 6,05% до 3,23%), що підтверджує успішну адаптацію алгоритмів до кризових умов. Попри високі значення MSE у великих вибірках ГРВІ, використання сезонних ознак

(sin/cos) дозволило зберегти адекватність прогнозів, тоді як у моделях Гепатиту В ключовим етапом залишається контроль перенавчання для досягнення балансу між збіжністю та точністю.

3.4 Узагальнення третьої ітерації

Третя ітерація продемонструвала, що методологічно строгі підходи до валідації, введення сезонних ознак та порівняння рівневих і кумулятивних представлень суттєво підвищують якість LSTM прогнозування. Walk-forward валідація забезпечила реалістичну оцінку продуктивності моделей в умовах, максимально наближених до реального застосування. Циклічні сезонні ознаки покращили моделювання річних циклів навіть для потужних LSTM архітектур. Кумулятивне представлення виявилось особливо ефективним для зменшення відносних похибок та підвищення стабільності прогнозів.

Порівняння компактних та глибоких архітектур показало, що для відносно коротких рядів (гепатити В та С) достатньо двошарових мереж з 64 одиницями на шар, тоді як для довгих складних рядів (ГРВІ) виправдане застосування глибоких стеків з понад 11 мільйонами параметрів. Однак таке збільшення складності вимагає ретельної регуляризації та значних обчислювальних ресурсів.

Систематичне порівняння методологій трьох ітерацій виявляє чітку еволюцію від експлораторних підходів до строго контрольованих експериментів. Перша ітерація зосереджувалася на демонстрації базової здатності LSTM моделювати епідемічні ряди, але мала обмеження у валідації та документуванні. Друга ітерація забезпечила статистичну верифікацію через ARIMA та систематичне звітування метрик. Третя ітерація інтегрувала методологічні уроки через walk-forward валідацію, інженерію ознак та порівняння представлень даних.

Ключовим методологічним досягненням є усвідомлення важливості типу валідації. Простий поділ на навчальну та тестову множини може давати надто оптимістичні оцінки, особливо для коротких рядів. Walk-forward валідація з послідовним розширенням вікна забезпечує більш реалістичну оцінку продуктивності в режимі послідовного прогнозування.

Інженерія ознак, зокрема циклічне кодування місяців через синусоїдальні функції, виявилася критично важливою для покращення моделювання сезонності. Хоча LSTM теоретично здатні виявляти періодичні патерни автоматично, явне надання циклічної інформації прискорює збіжність та покращує точність.

Порівняння рівневих та кумулятивних представлень показало фундаментальний компроміс: кумулятивні прогнози мають суттєво нижчі відносні похибки завдяки згладжувальному ефекту інтегрування, але менш чутливі до короткострокових спалахів. Вибір представлення має визначатися цілями застосування.

ARIMA моделі демонструють найкращу продуктивність для коротких та середніх рядів з чіткою лінійною або слабо нелінійною структурою. Їхні ключові переваги включають інтерпретованість параметрів, формальну статистичну діагностику, довірчі інтервали прогнозів та мінімальні обчислювальні вимоги. Основні обмеження полягають у складності моделювання множинних сезонних компонент, необхідності експертного вибору порядків та обмеженій здатності адаптуватися до структурних зламів.

LSTM мережі виправдані для довгих рядів зі складною нелінійною динамікою, множинними сезонними компонентами та можливістю використання додаткових ознак. Переваги включають автоматичне виявлення патернів, гнучкість архітектури, здатність моделювати довгострокові залежності та можливість інтеграції екзогенних змінних. Обмеження включають складність налаштування гіперпараметрів, високі обчислювальні

вимоги для глибоких мереж, ризик перенавчання на малих датасетах та низьку інтерпретованість.

3.5 Розробка веб застосунку для онлайн трекінгу епідемічних прогнозів

Для практичного застосування розроблених прогнозних методів було спроектовано та реалізовано веб-застосунок, призначений для обробки медико-епідемічних даних України та побудови прогнозів показників здоров'я населення з акцентом на інфекційну захворюваність, надлишкову смертність, сезонність та вплив воєнних факторів. Центральна ідея архітектури полягає у чіткому розділенні життєвого циклу моделі та процесу онлайн-видачі прогнозів, що підвищує відмовостійкість, спрощує керування версіями та забезпечує прозорість контролю якості.

3.5.1 Архітектура системи для веб додатку

Архітектура веб-застосунку організована у вигляді п'яти логічних шарів (рис. 3.3). Шар джерел даних містить репозиторії структурованих CSV-файлів та нотатники Jupyter, у яких здійснюється первинна очистка даних, нормалізація часових міток, усунення пропусків та формування похідних ознак. На виході цього шару формується узгоджений каталог часових рядів захворюваності, смертності та госпіталізацій, що слугує вхідною точкою для тренувального шару.

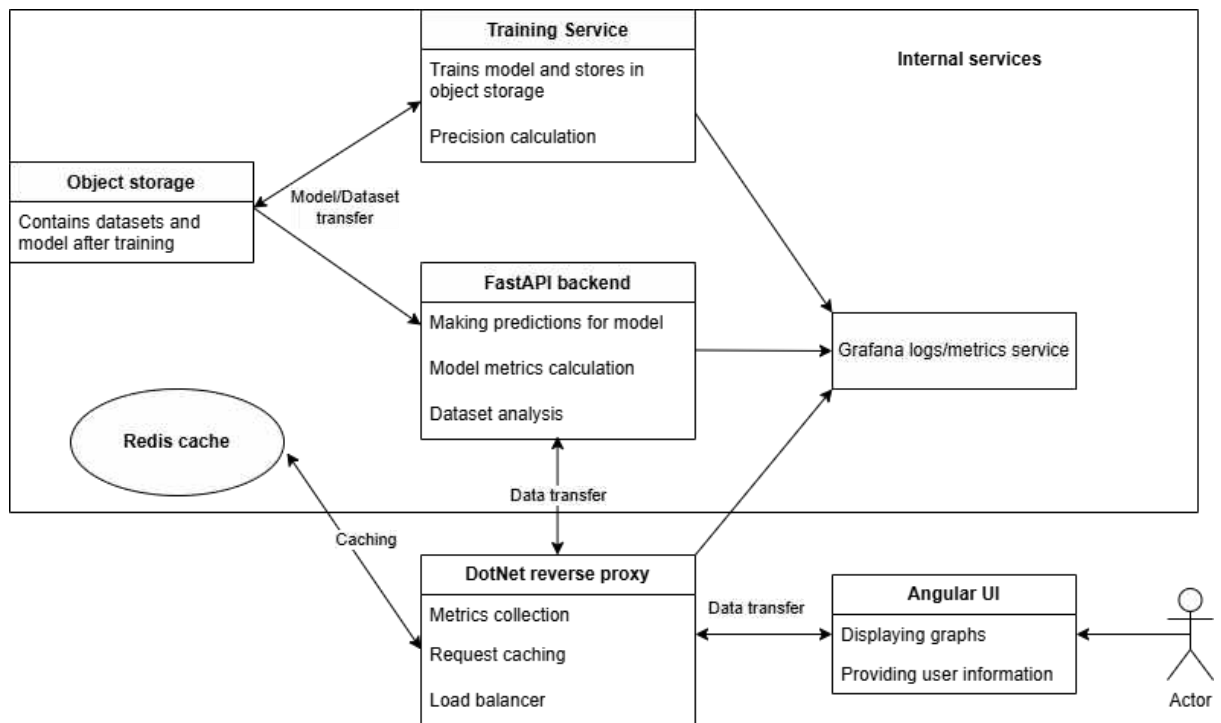


Рис. 3.3. Інфраструктура ML-сервісу з object storage та балансуванням запитів

Тренувальний шар реалізує побудову та калібрування моделей трьох категорій: статистичних моделей ARIMA/SARIMA, моделей на основі LSTM та моделей із екзогенними ознаками воєнного характеру. Інтенсивні обчислення тренувального шару ізольовані від сервісів обробки користувацьких запитів, щоб сплески споживання ресурсів не впливали на латентність видачі прогнозів. Після завершення кожного циклу тренування модель серіалізується разом із маніфестом версії, контрольними сумами, метриками валідації та описом інтерфейсу, що забезпечує відтворюваність експериментів та можливість відкоту у разі виявлення деградації.

Сховище артефактів виконує роль єдиного джерела версійованих датасетів, проміжних перетворень та фінальних моделей. Версійна структура каталогів дозволяє простежити походження кожного артефакту та підтвердити цілісність через хеш-значення.

Сервіс інференсу завантажує активні моделі та приймає параметри запитів, включно з часовим горизонтом, географічним охопленням та типом сценарію. Сервіс виконує нормалізацію вхідного вектору, перевірку

доступності відповідної моделі та обчислення прогнозних значень із метаданими точності. Агреговані результати повертаються шлюзу у стандартизованому форматі.

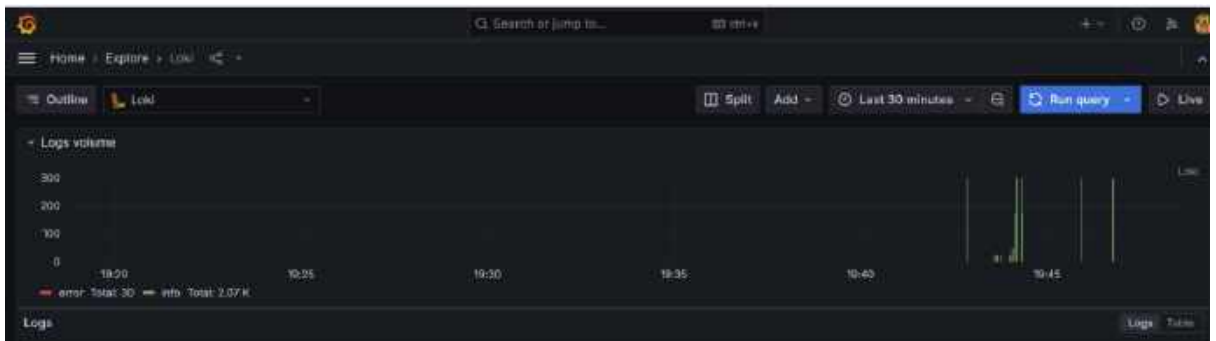


Рис. 3.4. Система логуювання і метрик Loki у Grafana

Логічний потік починається з джерел даних що містять часові ряди захворюваності смертності госпіталізацій лабораторних підтверджень а також допоміжні ковід та воєнні індикатори. Сирі набори надходять у сховища у вигляді файлів CSV і проміжних серіалізованих структур і опрацьовуються у підготовчих файлах де здійснюється очистка заповнення пропусків стандартизація дат нормалізація одиниць вимірювання та формування похідних ознак. Після первинної стадії формується узгоджений каталог часових рядів що слугує вхідним шаром для тренування моделей. Застосовуються статистичні методи і алгоритми машинного навчання для побудови прогнозів сезонності інфекційної активності і надлишкової смертності з урахуванням екзогенних факторів.

Тренувальний шар реалізований як окремі обчислювальні одиниці що виконують періодичне або подієве перенавчання. Інтенсивні обчислення ізольовані від сервісів обробки користувацьких запитів щоб сплески споживання ресурсів не впливали на латентність видачі прогнозів. Після завершення кожного циклу модель серіалізується разом з маніфестом версії контрольними сумами метриками валідації та описом інтерфейсу. Це забезпечує відтворюваність експериментів і можливість швидкого відкоту у разі виявлення деградації.

Сховище артефактів виконує роль єдиного джерела правди для сирих датасетів проміжних перетворень і фінальних моделей. Версійна структура каталогів дозволяє простежити походження кожного артефакту і підтвердити цілісність через хеш значення. Такий підхід спрощує аудит наукової частини і надає операторам впевненість що активна модель узгоджена з опублікованими метриками.

Сервіс інференсу завантажує активні моделі і приймає параметри запитів включно з часовим горизонтом географічним охопленням набором ознак і типом сценарію. Він виконує підготовку вхідного вектору нормалізацію масштабу перевірку доступності моделі і обчислення прогнозних значень з метаданими точності. Агреговані результати повертаються шлюзу у стандартизованому форматі що полегшує клієнтську візуалізацію.

Архітектура розгорнута в кластері Kubernetes [123] що забезпечує декларативне описання стану компонентів самовідновлення і горизонтальне масштабування. Зовнішній доступ організований через захищений тунель що зменшує площу атаки і усуває потребу відкривати прямі порти. Внутрішня комунікація відбувається через сервісні об'єкти і мережеві політики що розмежовують шари інтерфейс інференс тренування спостережуваність і допоміжні кеш сервіси.

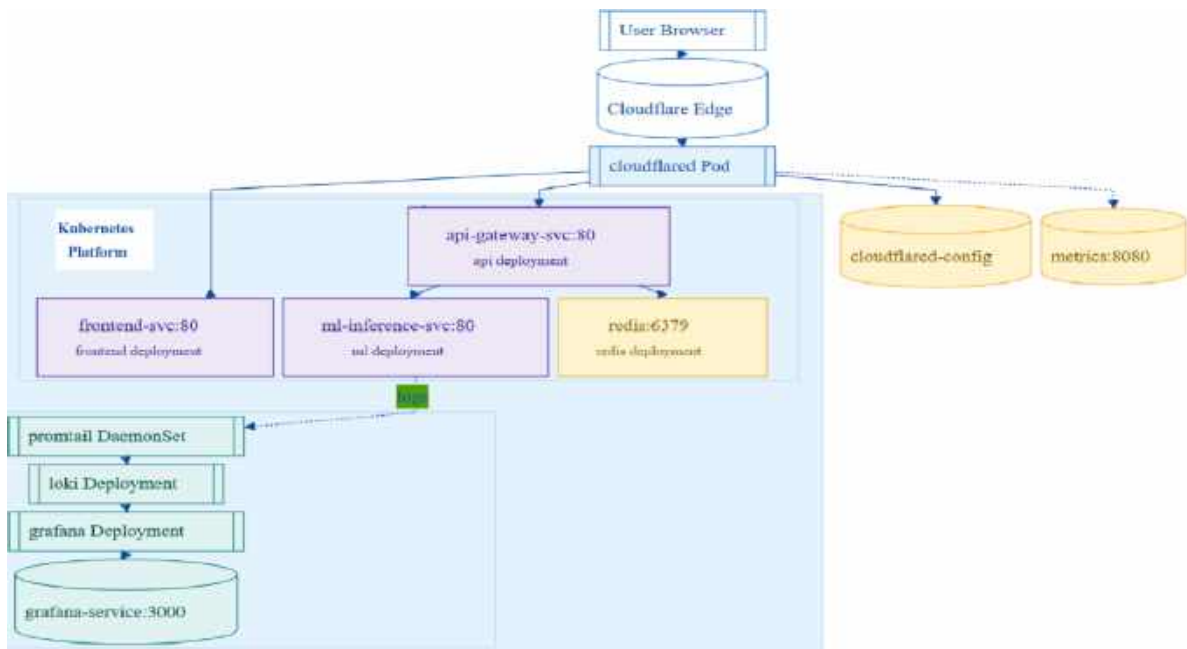


Рис. 3.5. Kubernetes-архітектура ML-сервісу: шлюз, inference, кеш і моніторинг

Три основні виробничі компоненти це клієнтський веб інтерфейс програмний шлюз і модуль інференсу. Веб інтерфейс забезпечує інтерактивні графіки фільтрацію серій і перегляд історії оновлень моделей. Шлюз реалізує REST інтерфейси [124] авторизацію агрегацію запитів політики кешування маршрутну логіку та контроль версій. Інференс модуль виконує обчислення прогнозів оцінку похибок формування структурованих відповідей і маршрутизацію діагностичних запитів до шару аналітики. Кеш шар на основі сховища ключ значення [125] зменшує латентність повторних запитів стабілізує навантаження і абстрагує деталі реалізації моделі від зовнішнього споживача.

Спостережуваність реалізована через конвеєр журналів і метрик що збирає телеметрію з усіх контейнерів [126]. Легкий агент на кожному вузлі пересилає рядки журналу у централізоване сховище де виконується індексація і надається швидкий повнотекстовий пошук. Метрики часу відповіді частоти помилок наповнення кешу використання процесора і пам'яті а також тривалості обчислення прогнозу корелюються на панелях візуалізації. Це дає

зможу інженерам оперативно виявляти тенденції деградації і реагувати до появи інцидентів.

Безпека забезпечується багатошаровими механізмами. Мережевий тунель екранує кластер від прямого доступу. Сервісний рівень використовує мінімально необхідні відкриті порти і мережеві політики. Дані розмежовані за рівнями доступу між сирими наборами проміжними перетвореннями і фінальними моделями. Операційний рівень аналізує поведінку компонентів і за аномаліями генерує сповіщення що стимулює швидке усунення проблем.

Ключові цілі архітектури модульність розширюваність прозорість ефективність і стабільність при піках навантаження. Модульність забезпечується чіткими границями між інтерфейсом шлюзом інференсом тренуванням і спостережуваністю. Розширюваність досягається через стандартизований контракт моделі що описує формат вхідних параметрів вихідні серії метрики якості і правила сумісності. Прозорість спирається на повноту журналів метрик і відкритий перегляд версій. Ефективність базується на активному кешуванні популярних запитів асинхронній обробці і оптимізованому завантаженні моделей. Стабільність підтримується через ізоляцію ресурсомістких тренувальних операцій від латентно чутливих інференс запитів.

Архітектурний стиль поєднує мікросервісні принципи [127] і спеціалізацію на аналізі часових рядів. Контейнеризація всіх компонентів [128] гарантує повторюваність розгортання мінімізацію дрейфу середовища і пришвидшує експериментальний цикл додавання нових моделей з воєнними екзогенними ознаками або новими статистичними семействами.

Розгортання виконується через декларативні маніфести що описують деплоймент інференсу тренувальних подів шлюзу кеш сервісів і стек спостережуваності. Кожен маніфест містить ресурси межі поді і сигнали готовності. Оновлення відбувається атомарно через заміну шаблону із

контролем здоров'я перед промоцією нової версії. Це мінімізує ризик деградації під час переходу.

Підтримка включає цикли огляду показників щоденний аналіз резидуалів щотижневу перевірку дрейфу і щомісячний аудит конфігурацій. Автоматизовані правила сповіщень формують чергу інцидентів що пріоритезуються за впливом на точність і доступність. Документація оновлюється синхронно із випуском нової моделі через генерацію звіту що містить ключові метрики і список змін.

Дорожня карта передбачає інтеграцію модулів адаптивного вибору моделі на основі поточного рівня невизначеності розширення бібліотеки ознак включаючи екологічні та соціально економічні індикатори підключення каналу швидкого сценарного аналізу для епідемічних штабів і поступовий перехід до частково онлайн оновлень. Кінцева мета побудова платформи що підтримує динамічну оптимізацію управлінських рішень у реальному часі з мінімальною ручною інтервенцією.

3.5.2 Архітектурні компоненти та потоки даних

Архітектура веб-застосунку організована у вигляді п'яти логічних шарів (рис. 3.3). Шар джерел даних містить репозиторії структурованих CSV-файлів та нотатники Jupyter, у яких здійснюється первинна очистка даних, нормалізація часових міток, усунення пропусків та формування похідних ознак. На виході цього шару формується узгоджений каталог часових рядів захворюваності, смертності та госпіталізацій, що слугує вхідною точкою для тренувального шару.

Тренувальний шар реалізує побудову та калібрування моделей трьох категорій: статистичних моделей ARIMA/SARIMA, моделей на основі LSTM та моделей із екзогенними ознаками воєнного характеру. Інтенсивні обчислення тренувального шару ізольовані від сервісів обробки користувацьких запитів, щоб сплески споживання ресурсів не впливали на

латентність видачі прогнозів. Після завершення кожного циклу тренування модель серіалізується разом із маніфестом версії, контрольними сумами, метриками валідації та описом інтерфейсу, що забезпечує відтворюваність експериментів та можливість відкоту у разі виявлення деградації.

Сховище артефактів виконує роль єдиного джерела версійованих датасетів, проміжних перетворень та фінальних моделей. Версійна структура каталогів дозволяє простежити походження кожного артефакту та підтвердити цілісність через хеш-значення.

Сервіс інференсу завантажує активні моделі та приймає параметри запитів, включно з часовим горизонтом, географічним охопленням та типом сценарію. Сервіс виконує нормалізацію вхідного вектору, перевірку доступності відповідної моделі та обчислення прогнозних значень із метаданими точності. Агреговані результати повертаються шлюзу у стандартизованому форматі.

Шар спостережуваності реалізовано через конвеєр журналів та метрик, що збирає телеметрію з усіх контейнерів (рис. 3.4). Метрики часу відповіді, частоти помилок, наповнення кешу, використання процесора та пам'яті, а також тривалості обчислення прогнозу корелюються на панелях візуалізації Grafana, що дає змогу оперативно виявляти тенденції деградації.

Система розгорнута в кластері Kubernetes (рис. 3.5), що забезпечує декларативне описання стану компонентів, самовідновлення та горизонтальне масштабування. Зовнішній доступ організований через захищений тунель, що зменшує площу атаки, а внутрішня комунікація відбувається через сервісні об'єкти та мережеві політики, що розмежовують шари системи.

Виробничий рівень складається з трьох основних компонентів. Веб-інтерфейс забезпечує інтерактивні графіки, фільтрацію серій та перегляд історії оновлень моделей. Програмний шлюз реалізує REST API, авторизацію, агрегацію запитів, маршрутну логіку та контроль версій. Модуль інференсу виконує обчислення прогнозів, оцінку похибок та формування структурованих

відповідей. Для зниження латентності повторних запитів реалізовано кеш-шар на основі сховища «ключ—значення», де ключ формується з ідентифікатора моделі, часових меж та параметрів запиту. Стратегія інвалідації базується на часовому життєвому циклі та інкрементальному оновленні після ретренінгу моделей.

3.5.3 Методологія оцінювання якості і валідація

Методологія оцінювання якості базується на багаторівневому підході що охоплює розділення вибірок часову крос перевірку оцінку стабільності і моніторинг дрейфу. Початково для кожного показника формується історичний інтервал тренування і відкладений інтервал перевірки що містить останні спостереження ще не використані для підбору параметрів.

Для моделей надлишкової смертності важливо уникати витоку майбутньої інформації тому всі перетворення що можуть враховувати статистику цілого періоду виконуються лише на тренувальному відрізку. У випадку сезонних серій застосовується ковзне вікно що гарантує репрезентативність структури піків і спадів. Оцінка виконується метриками [129] середньої абсолютної похибки середньої квадратичної похибки відносного відхилення і симетричного абсолютного відсоткового відхилення [120]. Для інфекційних показників що мають різку зміну під час піків додається метрика максимального абсолютного відхилення по вікнах щоб виявити точкові провали.

Стабільність перевіряється шляхом ретроспективного відтворення прогнозів на історичних інтервалах із послідовним нарощуванням горизонту. Якщо похибка зростає нелінійно зі збільшенням горизонту система позначає модель як чутливу до довгострокових екстраполяцій і рекомендує скорочення діапазону використання. Дрейф контролюється [130] порівнянням розподілу резидуалів за останній період із базовим розподілом на етапі впровадження

[122]. Значне зміщення середнього або розширення дисперсії класифікується як сигнали до запланованого перенавчання.

Для воєнних екзогенних ознак застосовується окремий аналіз внеску кожної ознаки. Підхід відсікання оцінює зміну метрики похибки при виключенні конкретного індикатора. Це дає змогу визначити які фактори мають найбільший вплив на точність і які є надлишковими. Така інформація корисна для оптимізації витрат на збір додаткових даних і фокусування уваги на найбільш інформативних потоках.

Валідація оперативної частини включає моніторинг часових характеристик запитів інференсу. Вимірюється частка запитів що завершуються швидше за поріг який відповідає цільовій якості сервісу. Якщо частка падає нижче встановленого значення система автоматично збільшує кількість реплік інференс контейнера або підвищує розмір кешу для популярних серій. Адаптивні політики масштабування налаштовані так щоб уникнути надмірного циклічного збільшення і зменшення кількості екземплярів.

Невизначеність прогнозів оцінюється методом ансамблів [131–132] і побудовою діапазону довіри на основі розподілу результатів різних моделей [120]. Ширина діапазону слугує індикатором стабільності. Вузкий діапазон при низькій похибці свідчить про узгодженість внутрішніх припущень тоді як широкий діапазон вказує на можливу нестачу даних або перехідну фазу. Діапазон розраховується агрегуванням квантилів симетрично навколо центральної оцінки і публікується разом із основними серіями.

Оцінка впливу кешу включає відстеження частки запитів що повернули відповідь з кешу серед усіх запитів за останнє вікно і порівняння їх середньої латентності із латентністю обчислюваних відповідей. Висока частка при відсутності деградації точності означає оптимальне налаштування стратегії терміну життя ключів. Якщо частка падає одночасно із збільшенням середньої

латентності оператору пропонується корекція терміну життя або перегляд структури ключів щоб підвищити підсумкову корисність.

Аудит якості виконання експериментів з новими моделями передбачає журналювання конфігурацій гіперпараметрів тривалості тренування використання ресурсів і результатів метрик. Це формує історію еволюції підходів і дозволяє повертатися до попередніх станів у разі потреби повторити аналіз. Узгодженість між журналами тренувань і маніфестами версій запобігає невідповідності між активними моделями і метричними показниками.

3.5.4 Графічний інтерфейс веб-системи моніторингу

Веб-інтерфейс системи надає користувачам інтерактивний дашборд для аналізу епідемічних даних та прогнозів. Інтерфейс реалізує візуалізацію вхідних даних щодо нових випадків захворювань (рис. 3.6) та смертності (рис. 3.7), результатів прогнозування за моделлю LSTM (рис. 3.8), моделлю SARIMA (рис. 3.9) та ансамблем моделей (рис. 3.10), а також інтервалів передбачень (рис. 3.11).

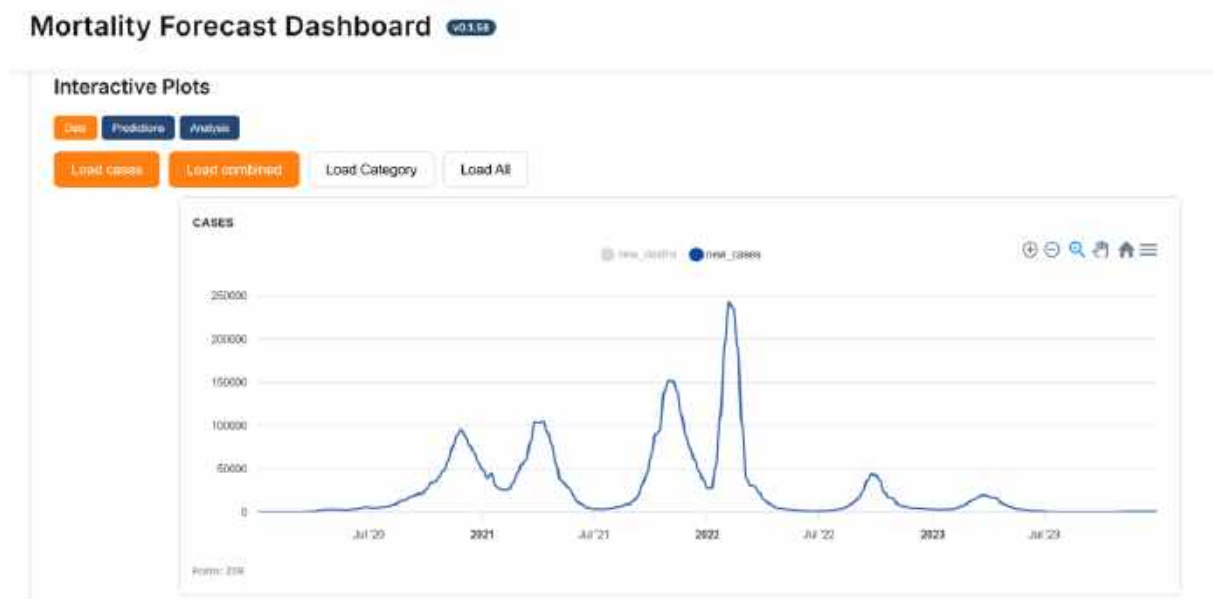


Рис. 3.6. Інтерактивний дашборд: вхідні дані - нові випадки захворювань

Mortality Forecast Dashboard v0.1.58

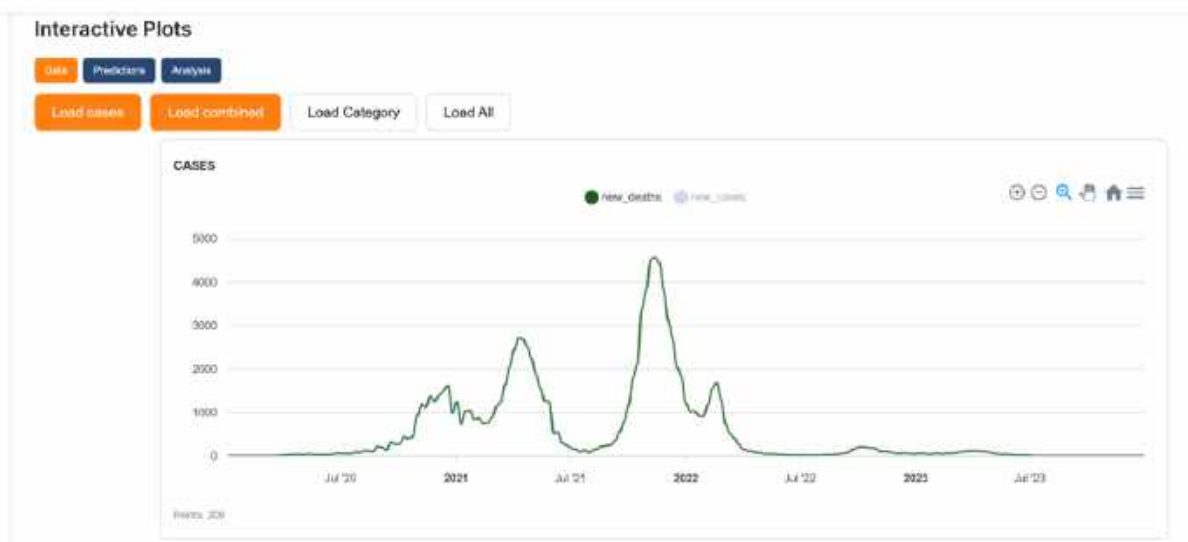


Рис. 3.7. Інтерактивний дашборд: вхідні дані - нові випадки смертності

Mortality Forecast Dashboard v0.1.58



Рис. 3.8. Інтерактивний дашборд: передбачення смертності - модель LSTM

Mortality Forecast Dashboard v0.1.00

Overview **Dashboard** Models **Plots**

Interactive Plots

Data **Predictions** Analyse

Load tsdb **Load sarima** **Load ensemble** **Load all** **Load intervals** Load Category Load All



Рис. 3.9. Інтерактивний дашборд: передбачення смертності - модель Sarima

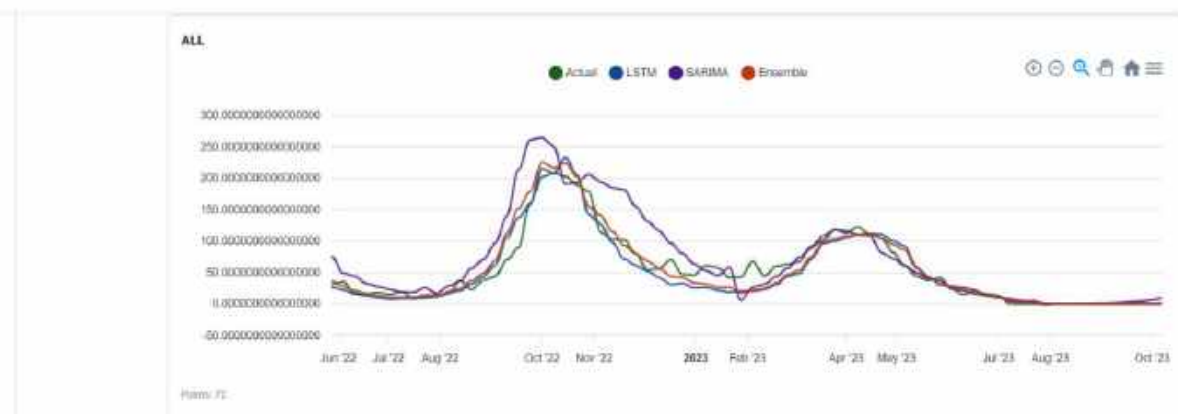
Mortality Forecast Dashboard v0.1.00

Рис. 3.10. Інтерактивний дашборд: передбачення смертності - ансамбль моделей



Рис. 3.11. Інтерактивний дашборд: передбачення смертності - інтервали передбачень

Для оцінки якості прогнозів у дашборді реалізовано відображення агрегованої міри помилок за тиждень (рис. 3.12), точкової різниці між реальними та прогнозними значеннями (рис. 3.13) та зведених метрик моделі (рис. 3.14). Клієнт здійснює локальне кешування короткострокових результатів для забезпечення плавного досвіду перемикання між фільтрами без надмірних повторних запитів до серверу. Така архітектура інтерфейсу дозволяє епідеміологам та аналітикам системи громадського здоров'я оперативно порівнювати прогнози різних моделей, відстежувати динаміку точності та виявляти періоди аномальних відхилень.

середньої абсолютної похибки після кожного перенавчання. Діаграма діапазону довіри із центральною оцінкою і межами квантилів.

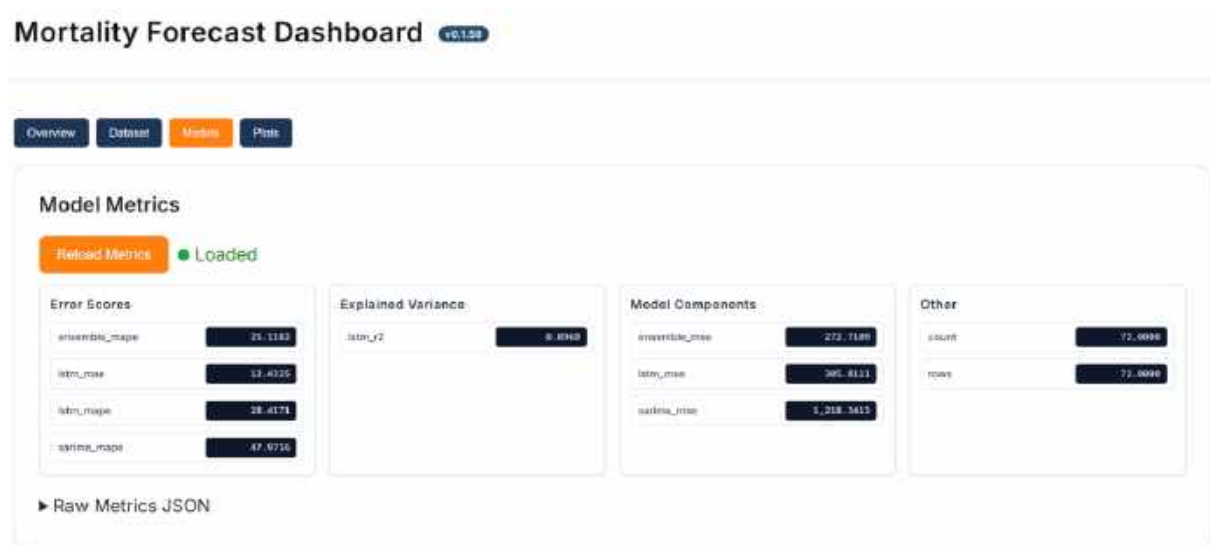


Рис. 3.14. Метрики моделі

3.6 Висновки до розділу

Розділ присвячений удосконаленню методів прогнозування епідемічних часових рядів та їх прикладному застосуванню у вигляді веб-системи моніторингу та трекінгу епідемічних прогнозів.

Розроблено методологічні вдосконалення третьої ітерації LSTM архітектури, спрямовані на підвищення надійності валідації та покращення моделювання сезонності. Введено sliding window валідацію, що дозволяє оцінити стабільність прогнозів при послідовному зміщенні вікна навчальних даних у часі. Замість одноразового фіксованого поділу на навчальну та тестову множини, модель оцінюється на послідовності часових вікон, що дає розподіл помилок прогнозування замість однієї точкової оцінки. Walk-forward валідація з послідовним розширенням або зміщенням навчального вікна максимально наближена до реального режиму застосування моделі, коли щомісяця потрібно генерувати прогноз на наступний період на основі всіх доступних на той момент даних.

Запропоновано метод циклічного кодування сезонних ознак через синусоїдальні та косинусоїдальні функції. Пряме використання номера місяця як числової змінної створює штучний розрив між груднем та січнем, хоча ці місяці є сусідніми в річному циклі. Циклічне кодування за формулами (3.1) та (3.2) забезпечує безперервне представлення річного циклу, де січень плавно переходить у грудень. Ця пара значень однозначно ідентифікує місяць у році та забезпечує коректне обчислення відстаней між будь-якими сусідніми місяцями на одиничному колі.

Систематизовано підходи до порівняння рівневих та кумулятивних представлень даних. Рівневе представлення відповідає абсолютній кількості випадків захворювання за місяць, тоді як кумулятивне представлення обчислюється як накопичена сума від початку періоду. Кумулятивне представлення згладжує випадкові місячні коливання через інтегрування, що зменшує відносну дисперсію та покращує стабільність прогнозів. Кумулятивна траєкторія зростає монотонно, що спрощує задачу моделювання, а відносні похибки прогнозування кумулятивних значень часто нижчі, оскільки знаменник у MAPE зростає з часом.

Розроблено вдосконалені архітектурні рішення для LSTM моделей різної складності. Для коротших рядів з 72 місяцями спостережень застосовано компактну двошарову архітектуру 2xLSTM(64) з Dropout 0.20 та загальною кількістю параметрів близько 171251. Для довших та складніших рядів з 132 місяцями спостережень застосовано глибоку тришарову архітектуру 3xLSTM(300/1200/600) з пірамідальною структурою, де розширення до 1200 одиниць у середньому шарі забезпечує високу репрезентативну потужність для інтеграції інформації з різних часових масштабів, тоді як звуження до 600 одиниць у третьому шарі виконує функцію узагальнення. Загальна кількість параметрів глибокої архітектури становила приблизно 11.89 мільйонів.

Розширено методи попередньої обробки епідемічних даних. Нормалізація виконується методом min-max масштабування до діапазону від 0

до 1 з обчисленням параметрів виключно на навчальній множині для уникнення витоку інформації з тестового періоду. Ключовою інновацією стало додавання циклічних сезонних ознак на основі синусоїдальних та косинусоїдальних функцій від індексу місяця, що явно надає моделі інформацію про сезонну позицію спостереження та прискорює збіжність навчання.

Узагальнено результати порівняння методологій трьох ітерацій. Перша ітерація зосереджувалася на демонстрації базової здатності LSTM моделювати епідемічні ряди, друга ітерація забезпечила статистичну верифікацію через ARIMA та систематичне звітування метрик, третя ітерація інтегрувала методологічні уроки через walk-forward валідацію, інженерію ознак та порівняння представлень даних. Встановлено, що простий поділ на навчальну та тестову множини може давати надто оптимістичні оцінки, тоді як walk-forward валідація забезпечує більш реалістичну оцінку продуктивності в режимі послідовного прогнозування.

Розроблено архітектуру веб-застосунку для онлайн трекінгу епідемічних прогнозів. Система призначена для комплексної обробки медико-епідемічних даних України та побудови прогнозів показників здоров'я населення з акцентом на надлишкову смертність, інфекційні захворювання, сезонність та вплив воєнних факторів і демографічних зрушень. Архітектура об'єднує науковий шар побудови та оцінювання моделей і виробничий шар доставки результатів аналітики через веб-інтерфейс, програмний шлюз та спеціалізовані сервіси у хмарно-оркестрованому середовищі Kubernetes [128].

Описано архітектурні компоненти та потоки даних веб-системи. Логічний потік починається з джерел даних, що містять часові ряди захворюваності, смертності та госпіталізацій, проходить через підготовчі модулі очистки та нормалізації, тренувальний шар для побудови моделей, сховище артефактів для версіонування та сервіс інференсу для обчислення прогнозів. Кешування через структуру ключ-значення суттєво знижує

затримку для повторюваних запитів. Спостережуваність реалізована через конвеєр журналів і метрик з централізованим сховищем та панелями візуалізації у Grafana.

Розроблено методологію оцінювання якості та валідації прогнозних моделей у виробничому середовищі. Застосовано багаторівневий підхід, що охоплює розділення вибірок, часову крос-перевірку, оцінку стабільності та моніторинг дрейфу. Для воєнних екзогенних ознак застосовано аналіз внеску кожної ознаки через підхід відсікання. Невизначеність прогнозів оцінюється методом ансамблів [132] та побудовою діапазону довіри [133] на основі розподілу результатів різних моделей.

Обґрунтовано практичну застосовність системи для інтеграції з національними та регіональними механізмами збору медико-епідемічних даних. Платформа здатна отримувати стандартизовані витяги з центральних реєстрів захворюваності, формувати ранні попередження про зміну динаміки випадків та підвищувати точність оцінок тиску на систему госпіталізацій. Економічний ефект впровадження проявляється у оптимізації використання ресурсів інтенсивної терапії та зниженні непрямих витрат через завчасне перенаправлення потоків пацієнтів.

Загалом, розроблені методологічні вдосконалення та прикладна веб-система забезпечують комплексне вирішення задачі прогнозування епідемічних часових рядів. Walk-forward валідація, циклічні сезонні ознаки та кумулятивне представлення суттєво підвищують якість та надійність LSTM прогнозування. Веб-застосунок з модульною архітектурою, кешуванням та системою спостережуваності забезпечує масштабоване розгортання прогнозних моделей для підтримки прийняття рішень у системі громадського здоров'я України.

Основні результати розділу опубліковано в працях [1, 3, 6, 7, 10, 11, 13].

РОЗДІЛ 4. РЕЗУЛЬТАТИ ЕКСПЕРИМЕНТАЛЬНИХ ДОСЛІДЖЕНЬ

4.1 Результати застосування базових LSTM

Перша ітерація експериментальних досліджень була спрямована на перевірку принципової можливості застосування рекурентних нейронних мереж з механізмом довгої короткочасної пам'яті для прогнозування епідемічних часових рядів різної природи. На цьому етапі досліджувались три інфекційні захворювання: гострий гепатит А на національному рівні України, грип та сальмонельоз на регіональному рівні Харківської області. Вибір цих захворювань обумовлений наявністю достатніх за обсягом часових рядів офіційної статистики та різноманітністю епідемічних характеристик, що дозволяло оцінити універсальність підходу.

4.1.1 Результати прогнозування гострого гепатиту А в Україні

Набір даних для моделювання гострого гепатиту А був отриманий з офіційних звітів Центру громадського здоров'я Міністерства охорони здоров'я України. Дані збиралися за стандартизованою Формою №1 під назвою Звіт про інфекційні та паразитарні захворювання і склалися з щомісячних зареєстрованих випадків гострого гепатиту А по всій території України. Період спостереження охоплював січень 2018 року по грудень 2023 року, що становило 72 місяці безперервних спостережень. Дані за січень-грудень 2024 року були відкладені для незалежної позавибіркової оцінки якості прогнозування.

Для моделювання було побудовано два окремі одновимірні часові ряди з різним представленням даних. Перший ряд представляв серію щомісячної

абсолютної захворюваності, яка показувала кількість нових випадків гострого гепатиту А, зареєстрованих протягом кожного календарного місяця. Другий ряд представляв серію щомісячної кумулятивної захворюваності, що відображала наростаючий підсумок зареєстрованих випадків від початку періоду спостереження до кожного наступного місяця. Таке подвійне представлення дозволяло порівняти ефективність прогнозування для різних типів епідемічних задач.

Архітектура моделі базувалася на двошаровій конфігурації мережі довгої короткочасної пам'яті. Перший рекурентний шар містив 64 одиниці пам'яті та повертав повну послідовність вихідних станів для передачі наступному шару. Другий рекурентний шар також містив 64 одиниці пам'яті, але повертав лише фінальний стан для подальшого оброблення. Завершувався стек щільним вихідним шаром з одним нейроном для генерації числового прогнозу на наступний часовий крок. Загальна кількість параметрів, що навчаються, становила приблизно 171 251 одиницю.

Попередня обробка даних включала нормалізацію значень до діапазону $[0, 1]$ за процедурою, описаною у п. 2.4.1. Це перетворення виконувалося виключно на основі мінімального та максимального значень навчальної множини для уникнення витоку інформації з тестового періоду. Нормалізація забезпечувала стабільність процесу навчання та ефективну роботу нелінійних активаційних функцій у комірках довгої короткочасної пам'яті.

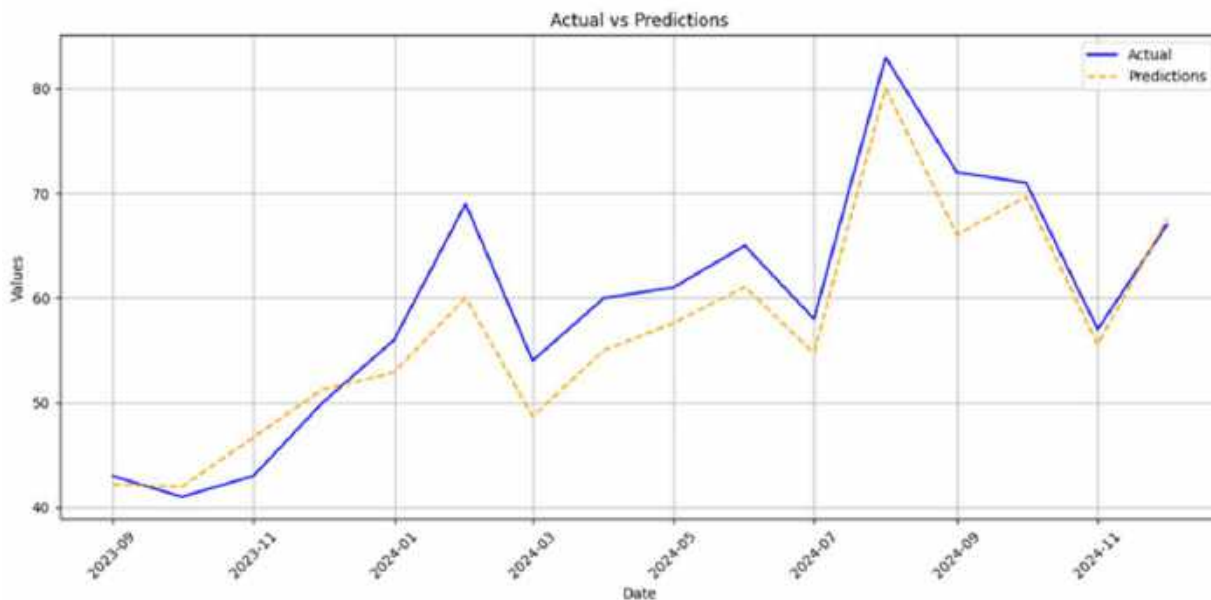


Рис. 4.1. Передбачення гострого гепатиту А

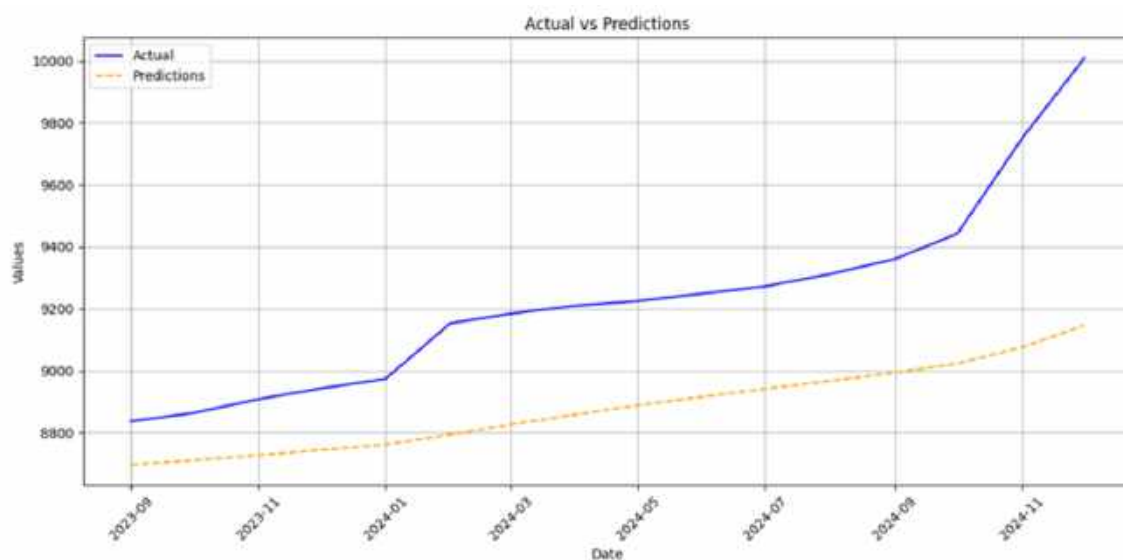


Рис. 4.2. Кумулятивне передбачення гострого гепатиту А

Оптимізація параметрів здійснювалася методом Adam з адаптивною швидкістю навчання. Функцією втрат було обрано MSE (2.25), яка надає більшої ваги великим відхиленням прогнозів від фактичних значень. Для запобігання перенавчанню застосовувалося раннє зупинення (п. 2.4.2) на основі моніторингу похибки на валідаційній множині. Навчальна вибірка становила

80 відсотків від загального обсягу даних, тоді як 20 відсотків зарезервовано для тестування.

Модель продемонструвала здатність відтворювати характерну сезонну структуру захворюваності на гострий гепатит А з виразними осінньо-зимовими піками та зниженням активності в теплий період року. Візуальний аналіз прогнозних траєкторій підтвердив, що мережа успішно ідентифікує річний цикл епідемічної активності [30] без явної специфікації сезонних компонент у структурі моделі.

Проте на цьому етапі досліджень детальні числові показники точності для даної конфігурації не були систематично задокументовані. Не наведено конкретних значень середньої абсолютної відсоткової похибки, середньої абсолютної похибки, середньоквадратичної похибки або кореня з середньоквадратичної похибки на тестовій множині. Це обмеження ускладнювало об'єктивне порівняння з альтернативними методами та кількісну оцінку практичної придатності моделі для оперативного прогнозування в системі епідемічного нагляду.

4.1.2 Результати прогнозування грипу в Харківській області методом LSTM

Для прогнозування місячної захворюваності на грип у Харківській області було використано часовий ряд за період з 2013 по 2024 роки, що включав 132 місяці безперервних спостережень та охоплював кілька повних епідемічних сезонів. Дані представляли абсолютну кількість підтверджених випадків грипу, зареєстрованих щомісяця в регіональній системі епідемічного нагляду Харківської області. Вибір регіонального рівня аналізу дозволяв оцінити ефективність моделі для локального прогнозування, що є важливим для планування ресурсів системи охорони здоров'я на рівні окремої області.

Модель базувалася на багат шаровій архітектурі мережі довгої короткочасної пам'яті з регуляризацією методом Dropout. Точна конфігурація включала послідовність рекурентних шарів, де кожен шар обробляв вихідні послідовності попереднього шару, формуючи ієрархічне представлення часових залежностей різних масштабів. Між шарами LSTM застосовувалася регуляризація Dropout (п. 2.4.2). Це змушувало мережу формувати розподілені робастні представлення та запобігало надмірній адаптації до специфічних особливостей навчальних даних. Завершувався стек щільним вихідним шаром для генерації числового прогнозу кількості випадків на наступний місяць.

Попередня обробка даних включала нормалізацію значень до діапазону $[0, 1]$ за процедурою, описаною у п. 2.4.1. Перетворення виконувалося на основі параметрів навчальної множини для збереження незалежності тестової вибірки. Нормалізація забезпечувала стабільність градієнтного спуску під час навчання та коректну роботу сигмоїдних активаційних функцій у вентилях комірок довгої короткочасної пам'яті.

Навчання моделі здійснювалося оптимізатором Adam (п. 2.4.2) з одночасним моніторингом MSE як функції оптимізації та MAPE для оцінки практичної точності. Такий подвійний моніторинг дозволяв контролювати як збіжність процесу навчання за технічним критерієм, так і реальну продуктивність моделі з точки зору інтерпретованості результатів для кінцевих користувачів у системі громадського здоров'я.

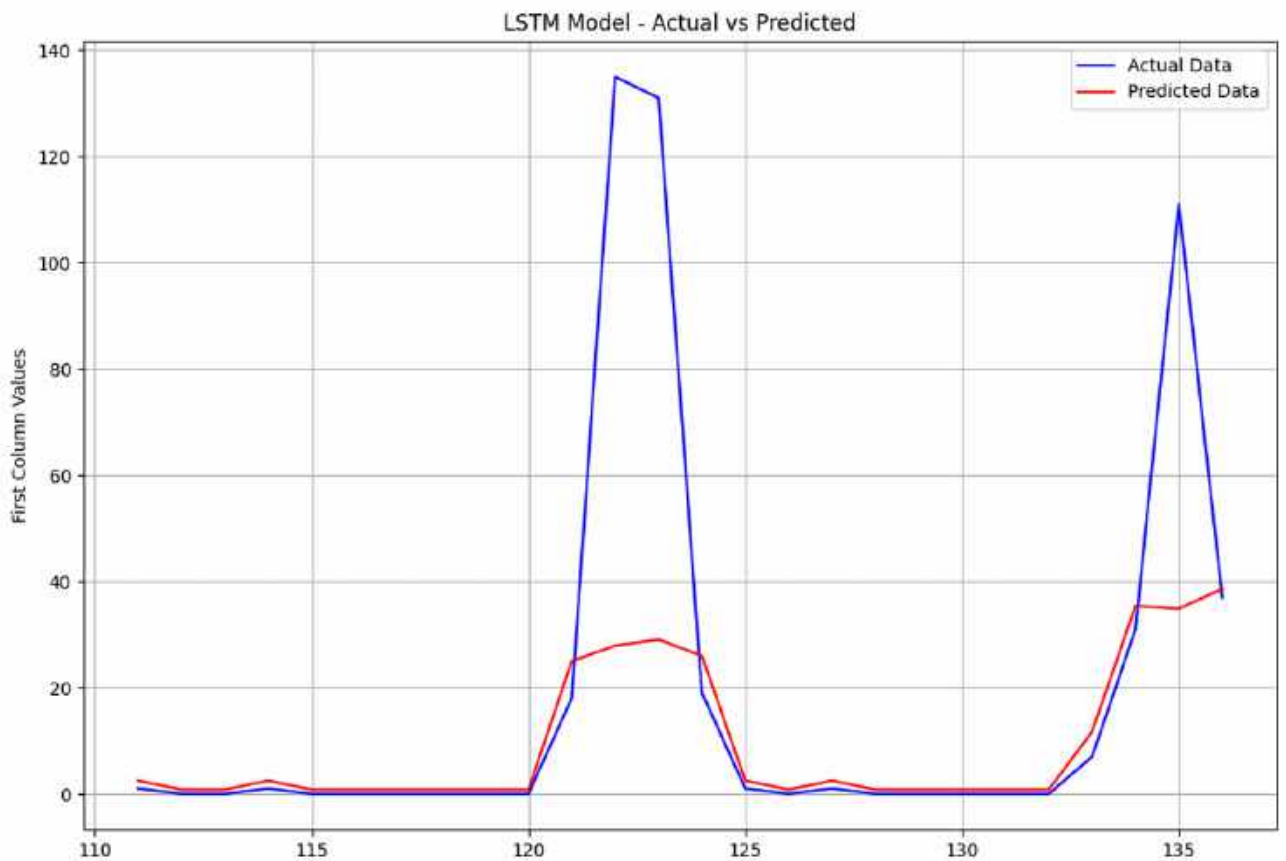


Рис. 4.3. Прогноз нових випадків грипу в Харківській області.

Модель продемонструвала здатність відтворювати характерну сезонну структуру захворюваності на грип з виразними піками в холодні місяці та зниженням активності в теплий період року. Візуальний аналіз прогнозів підтвердив, що мережа довгої короткочасної пам'яті успішно ідентифікує річний цикл [31] епідемічної активності без явної специфікації сезонних компонент у вхідних даних або архітектурі моделі. Мережа автоматично виявляла періодичні патерни через механізм навчання на послідовностях.

Проте аналогічно до дослідження гепатиту А детальні числові показники точності для цієї конфігурації мережі довгої короткочасної пам'яті у доступних матеріалах не були систематично задокументовані. Не наведено конкретних значень середньої абсолютної відсоткової похибки, середньої абсолютної похибки, середньоквадратичної похибки або кореня з середньоквадратичної похибки на тестовій множині. Це обмеження ускладнювало об'єктивне

порівняння з альтернативними методами та оцінку практичної придатності моделі для оперативного прогнозування захворюваності на грип.

4.1.3 Результати прогнозування сальмонельозу в Харківській області

Моделювання сальмонельозу виконувалося на основі місячного ряду підтверджених лабораторно випадків за період з 2013 по 2024 роки, що охоплював 132 місяці безперервних спостережень. Сальмонельоз як гостра кишкова інфекція характеризується певною сезонністю з підвищенням захворюваності в теплі місяці року, що пов'язано із погіршенням умов зберігання харчових продуктів та зміною харчової поведінки населення в літній період. Ця сезонна структура створювала цікавий виклик для методів машинного навчання.

Для цієї задачі була побудована найглибша архітектура серед моделей першої ітерації експериментів. Мережа включала три послідовні шари довгої короткочасної пам'яті з пірамідальною конфігурацією кількості одиниць пам'яті: перший шар містив 256 одиниць, другий шар містив 128 одиниць, третій шар містив 64 одиниці. Така структура із поступовим зменшенням розмірності від нижніх шарів до верхніх створює ієрархічне представлення даних, де нижні шари кодують прості локальні часові патерни, а верхні шари інтегрують їх у складніші залежності високого рівня абстракції.

Після кожного шару LSTM застосовувалася регуляризація Dropout з коефіцієнтом 0.20 (п. 2.4.2). Ця техніка змушує мережу формувати розподілені робастні представлення та запобігає надмірній адаптації до специфічних особливостей і шуму у навчальних даних. Фінальний щільний шар з одним нейроном генерував числовий прогноз рівня захворюваності на наступний календарний місяць.

Результати оцінювання моделі представлені як у масштабованому просторі нормалізованих значень, так і у фізичних одиницях після застосування зворотного перетворення. Така подвійна звітність дозволяє оцінити як технічну якість апроксимації, так і практичну корисність прогнозів для системи епідемічного нагляду.

Таблиця 4.1 Результати прогнозування сальмонельозу LSTM моделлю

Показник точності	Масштабований простір [0,1]	Фізичні одиниці (випадки)
MSE	1.085×10^{-6}	не застосовується
MAPE, %	0.2904	6.98
Кількість епох навчання	400–550	400–550

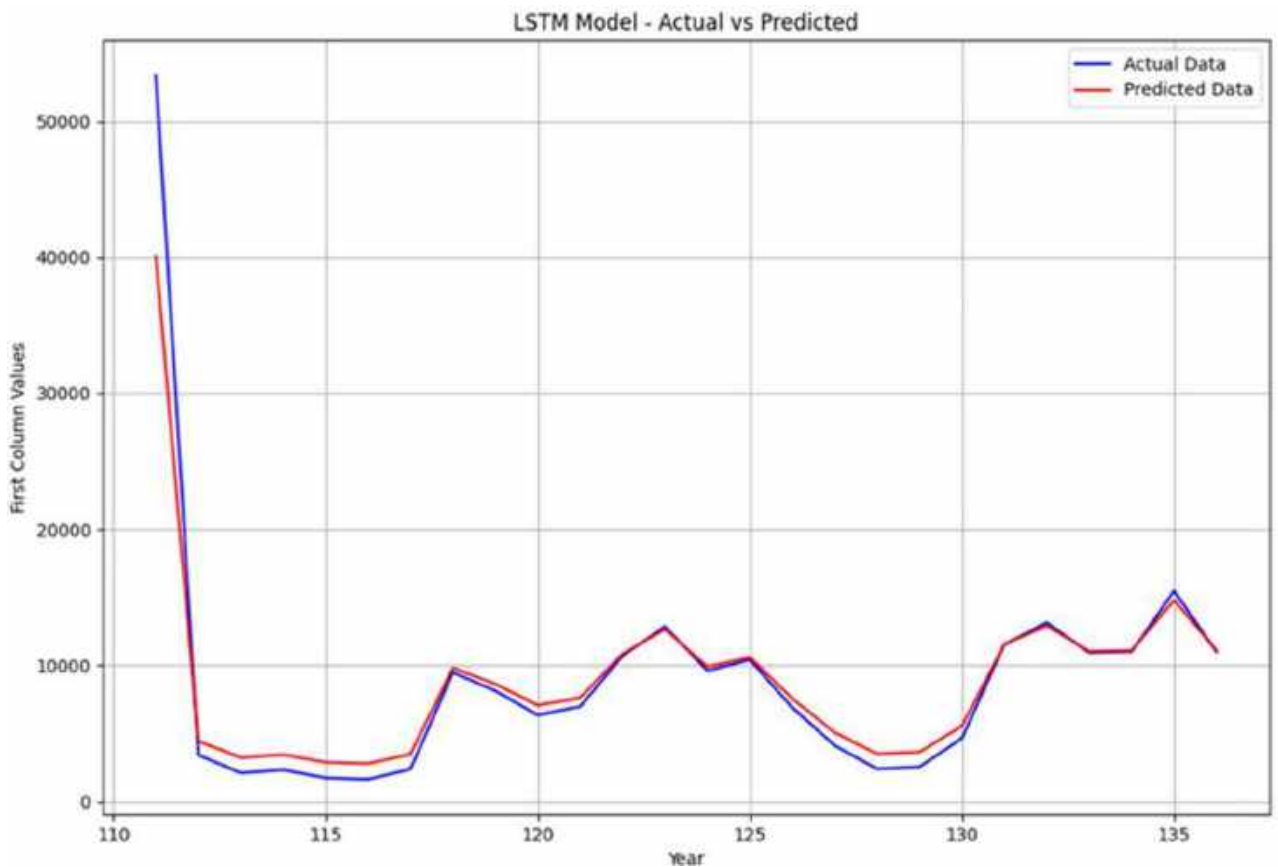


Рис. 4.4. Прогнозована захворюваність на сальмонельоз у Харківській області.

У масштабованому просторі нормалізованих значень модель досягла вкрай низької середньоквадратичної похибки на рівні 1.085 помножене на десять у мінус шостому степені та середньої абсолютної відсоткової похибки 0.2904 відсотка, що свідчить про дуже точне відтворення нормалізованих значень у діапазоні від нуля до одиниці.

Після застосування зворотного перетворення для повернення до фізичних одиниць абсолютної кількості випадків захворювання середня абсолютна відсоткова похибка становила приблизно 6.98 відсотка. Це значення є більш показовим для практичного застосування та свідчить про прийнятну точність прогнозування реальної кількості випадків захворювання на сальмонельоз. Похибка менше семи відсотків є цілком задовільною для епідемічного планування та розподілу ресурсів системи охорони здоров'я.

Візуальний аналіз графіків прогнозів підтвердив адекватне відтворення пікових значень захворюваності на контрольному інтервалі тестування. Модель успішно ідентифікує періоди підвищеної захворюваності в теплі місяці та відтворює загальну динаміку епідемічного процесу сальмонельозу. Водночас спостерігалось певне згладжування різких коливань порівняно з фактичними даними, що є типовою властивістю моделей, навчених з середньоквадратичною похибкою як функцією втрат, оскільки вона штрафує великі помилки квадратично і тим самим стимулює більш консервативні прогнози.

4.1.4 Узагальнення результатів першої ітерації

Перша ітерація досліджень продемонструвала принципову можливість та ефективність застосування рекурентних нейронних мереж довгої короткочасної пам'яті для прогнозування епідемічних часових рядів місячної частоти спостережень. Моделі успішно навчалися відтворювати загальні тренди, сезонні патерни та періодичні коливання захворюваності без явного

задання сезонних порядків або попередньої декомпозиції ряду на компоненти тренду та сезонності.

Архітектурні рішення варіювалися від компактних двошарових конфігурацій з 64 одиницями пам'яті у кожному шарі для коротших рядів до глибоких тришарових пірамідальних структур з 256, 128 та 64 одиницями для складніших задач з довгими рядами спостережень. Регуляризація методом Dropout та раннє зупинення виявилися ефективними механізмами контролю перенавчання на обмежених за обсягом епідемічних наборах даних. Оптимізатори Adam та Adamax забезпечили стабільну збіжність процесу навчання з адаптивною швидкістю для різних типів параметрів мережі.

Таблиця 4.2 Архітектурні характеристики базових LSTM моделей першої ітерації

Характеристика	Гепатит А	Грип	Сальмонельоз
Регіон дослідження	Україна	Харківська обл.	Харківська обл.
Період спостереження, місяців	72	132	132
Архітектура мережі	2×LSTM(64)→Dense(1)	Багатошарова LSTM	3×LSTM(256/128/64)→Dense(1)
Кількість параметрів	~ 171 000	~223 000	~ 315 000
Оптимізатор	Adam	Adam	Adamax
Метод регуляризації	Раннє зупинення	Dropout 0.20	Dropout 0.20
Стратегія валідації	Часовий поділ 80/20	Часовий поділ	Валідаційна множина

Валідація базувалася на простому одноразовому часовому поділі на навчальну та тестову множини без систематичної перевірки стабільності прогнозів при послідовному зсуві вікна даних. Відсутність так званої walk-forward валідації не дозволяла оцінити, наскільки стабільно модель працює при

послідовному прогнозуванні крок за кроком, що є типовим режимом реального застосування в оперативному епідеміологічному нагляді.

Проте перша ітерація виявила певні методологічні обмеження та напрямки для подальшого покращення, які були враховані у наступних етапах досліджень.

Також, не застосовувалися додаткові інженерні ознаки для явного моделювання сезонності. Хоча мережі довгої короткочасної пам'яті теоретично здатні автоматично виявляти періодичні патерни через механізм пам'яті, додавання циклічних ознак на основі синусоїдальних та косинусоїдальних функцій від номера місяця могло б покращити точність та прискорити збіжність навчання.

Таблиця 4.3 Показники точності базових LSTM моделей першої ітерації

Модель	Представлення даних	MAPE, %	Кількість епох	Примітки
Гепатит А	Рівневе та кумулятивне	5.40	150	Нормалізована вибірка
Грип	Місячний рівень	6.39	400–550	Відтворює сезонні піки
Сальмонельоз	Масштабований простір	0.29	400–550	Нормалізована вибірка
Сальмонельоз	Фізичні одиниці	6.98	400–550	Після зворотного перетворення

Відсутність аналізу чутливості до гіперпараметрів та систематичного налаштування через сітковий пошук або випадковий пошук залишала відкритим питання про оптимальність обраних конфігурацій мережі. Вибір розмірів шарів, коефіцієнтів Dropout та інших параметрів базувався переважно на евристичних та обмеженій валідації.

Ці спостереження визначили напрямки розвитку у другій ітерації, де основний фокус було зміщено на статистичну верифікацію через порівняння з моделями авторегресії та більш ретельне документування метрик точності, а у

третьої ітерації на методологічні покращення стратегій валідації та інженерії додаткових ознак.

4.2 Результати застосування ARIMA моделей

Друга ітерація експериментальних досліджень була присвячена застосуванню класичних статистичних моделей авторегресії та інтегрованого ковзного середнього для прогнозування епідемічних часових рядів. Основною метою цього етапу було створення надійної бази порівняння для оцінки відносних переваг методів глибинного навчання та визначення умов, за яких кожен підхід демонструє найкращі результати.

4.2.1 Результати прогнозування лептоспірозу в країнах Балтії

Аналіз лептоспірозу виконувався на щорічних рядах підтверджених випадків для трьох країн Балтії, а саме Естонії, Литви та Латвії, за період з 2010 по 2022 роки на основі офіційних звітів Європейського центру профілактики та контролю захворювань. Лептоспіроз є бактеріальною зооозною інфекцією, що передається через контакт з водою або ґрунтом, забрудненими сечею інфікованих тварин, переважно гризунів. Захворювання характеризується сезонним підвищенням у теплі місяці року та значною географічною варіабельністю залежно від кліматичних умов та контакту населення з потенційними джерелами інфекції.

Для кожної країни було побудовано часовий ряд абсолютної кількості підтверджених випадків та паралельний ряд показників захворюваності на сто тисяч населення для нормалізації на розмір популяції. Період спостереження охоплював тринадцять років, що давало достатню кількість спостережень для побудови моделей авторегресії та інтегрованого ковзного середнього річної частоти при визнанні обмежень, пов'язаних з відносно коротким рядом для статистичного моделювання.

Таблиця 4.4 Оптимальні ARIMA моделі для лептоспірозу в країнах Балтії

Країна	Порядок ARIMA (p, d, q)	Інтерпретація структури моделі
Естонія	(2, 2, 3)	2 авторегресійні члени, 2 порядки диференціювання, 3 члени ковзного середнього
Литва	(1, 2, 2)	1 авторегресійний член, 2 порядки диференціювання, 2 члени ковзного середнього
Латвія	(1, 2, 0)	1 авторегресійний член, 2 порядки диференціювання, чистий авторегресійний процес

Для кожної країни окремо виконувався повний цикл методології Бокса-Дженкінса (п. 2.2, 2.2.1, 2.2.2): перевірка стаціонарності, ідентифікація порядків, оцінювання параметрів, діагностика залишків та побудова прогнозів. Аналіз автокореляційної та часткової автокореляційної функцій для ідентифікації порядків моделі, оцінювання параметрів методом максимальної правдоподібності, діагностику залишків за критерієм Льюнга-Бокса та побудову прогнозів з довірчими інтервалами.

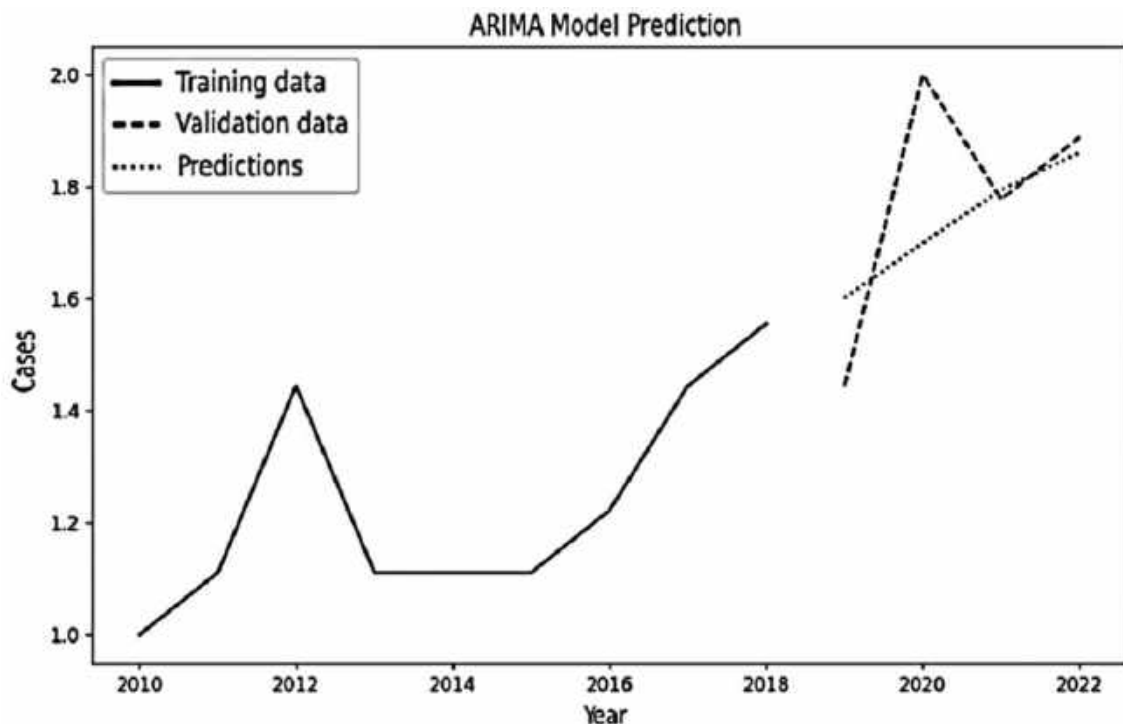


Рис. 4.5. Прогноз захворюваності на лептоспіроз в Естонії.

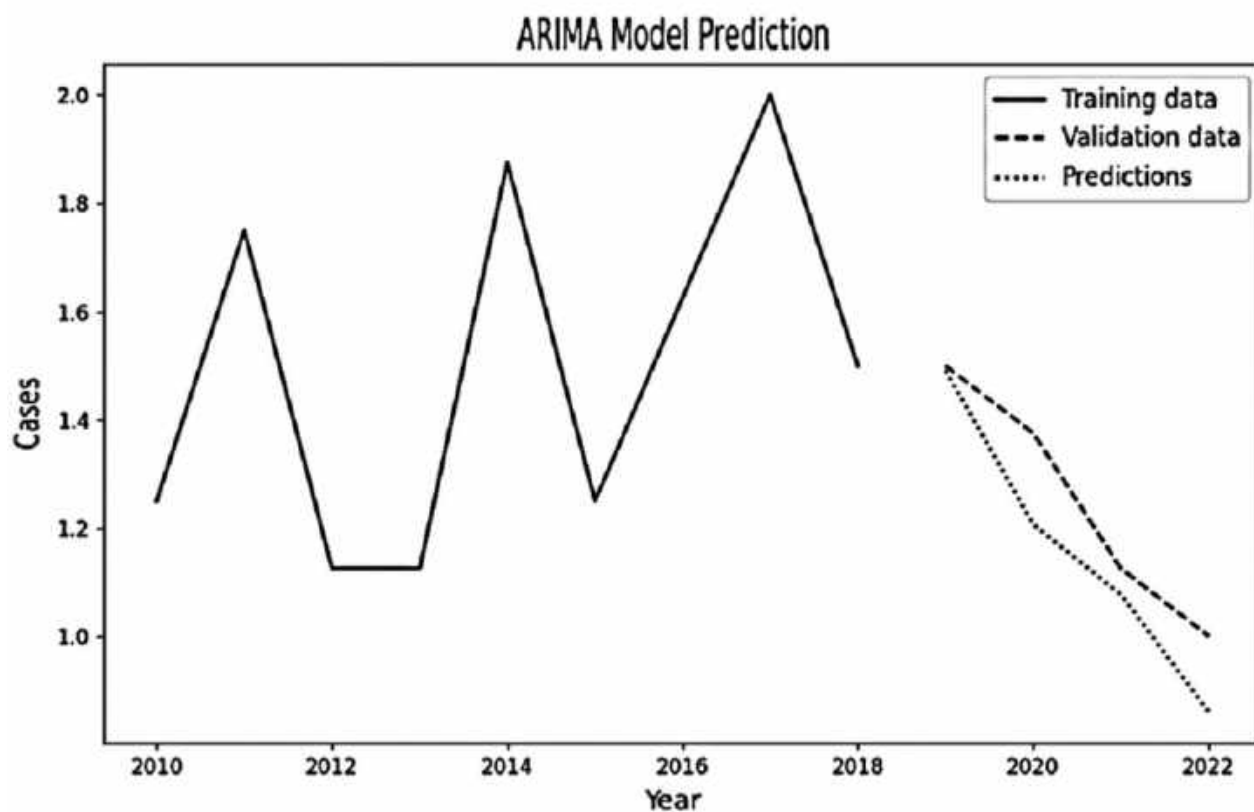


Рис. 4.6 Прогноз захворюваності на лептоспіроз в Литві.

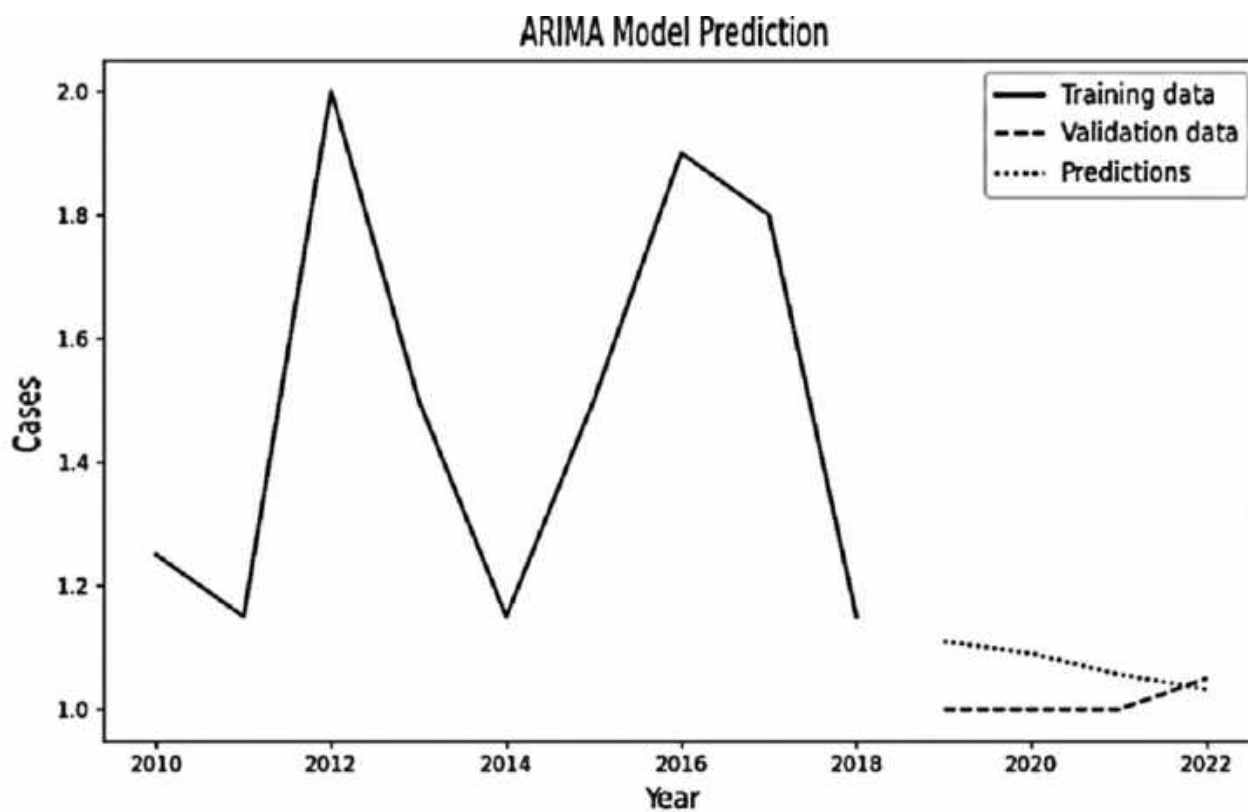


Рис. 4.7. Прогноз захворюваності на лептоспіроз в Латвії.

Для Естонії найкращою моделлю за критеріями Акаїке та Байєса виявилася модель авторегресії та інтегрованого ковзного середнього порядку два, два, три, що включає два авторегресійні члени, два порядки диференціювання для досягнення стаціонарності та три члени ковзного середнього. Порядок диференціювання d рівний двом вказує на наявність не лише лінійного тренду, а й зміни швидкості тренду в динаміці захворюваності протягом періоду спостереження. Відносно високі порядки авторегресії та ковзного середнього відображають складну структуру залежностей у короткому річному ряді.

Для Литви оптимальною виявилася більш компактна модель авторегресії та інтегрованого ковзного середнього порядку один, два, два з одним авторегресійним членом, двома порядками диференціювання та двома членами ковзного середнього. Зменшення порядків порівняно з Естонією свідчить про дещо простішу структуру динаміки лептоспірозу в цій країні.

Таблиця 4.5 Показники точності ARIMA моделей лептоспірозу в країнах Балтії

Країна	MAPE, %	Кількість параметрів (p+q)	Характеристика моделі
Естонія	7.076	5	Найвища складність серед країн
Литва	6.841	3	Найнижча прогнозна похибка
Латвія	7.765	1 (лише AR)	Найпростіша модель
Середнє значення	7.23	—	Діапазон похибок 6.8–7.8%

Для Латвії було відібрано найбільш компактну модель авторегресії та інтегрованого ковзного середнього порядку один, два, нуль, яка включає лише

один авторегресійний член та два порядки диференціювання без компоненти ковзного середнього. Така структура відповідає чистому авторегресійному процесу після подвійного диференціювання і є найпростішою серед трьох країн. Це може відображати більшу регулярність та передбачуваність динаміки лептоспірозу в Латвії порівняно з сусідніми країнами.

На ретроспективному тестуванні модель для Естонії досягла середньої абсолютної відсоткової похибки 7.076 відсотка, що є цілком прийнятним результатом для річних епідемічних даних з природною варіабельністю. Модель для Литви показала найнижчу похибку серед трьох країн на рівні 6.841 відсотка, що підтверджує адекватність обраної специфікації та високу якість вхідних даних. Модель для Латвії досягла середньої абсолютної відсоткової похибки 7.765 відсотка, що є дещо вищим за результати Естонії та Литви, але залишається в межах прийнятної точності для практичного застосування.

Узгодженість результатів між трьома країнами із середніми похибками у вузькому діапазоні від 6.8 до 7.8 відсотка підтверджує стабільність та надійність підходу [33] на основі моделей авторегресії та інтегрованого ковзного середнього для річних епідемічних рядів з помірною варіабельністю. Відмінності в оптимальних порядках моделей відображають специфічні особливості епідемічного процесу в кожній країні, але загальна методологія залишається ефективною.

Ці результати дозволяють використовувати моделі авторегресії та інтегрованого ковзного середнього як еталон для порівняння з підходами на основі мереж довгої короткочасної пам'яті та оцінки відносних переваг глибинного навчання для конкретних типів епідемічних задач. Для річних рядів помірної довжини класична статистика демонструє конкурентоспроможну точність при суттєвих перевагах інтерпретованості результатів та простоти моделі.

4.2.2 Результати прогнозування грипу в Харківській області методом ARIMA

Для моделювання місячної захворюваності на грип у Харківській області за період з 2013 по 2024 роки методом авторегресії та інтегрованого ковзного середнього було застосовано попередню нормалізацію вхідного ряду до інтервалу від одиниці до двох. Така нормалізація дозволяє зберегти співвідношення між рівнями захворюваності без стиснення значень у нульову точку, що важливо для стабільності обчислень логарифмічних перетворень та інтерпретації результатів.

Загальний період спостереження охоплював 136 місяців, що дало достатній обсяг даних для побудови місячної моделі авторегресії та інтегрованого ковзного середнього з можливістю включення складних авторегресійних та сезонних компонент. Навчальна вибірка була сформована з перших дев'яноста відсотків спостережень, що становило 122 місяці, тоді як останні десять відсотків у кількості 14 місяців були зарезервовані для незалежного тестування моделі та оцінки її прогностичної здатності.

На основі аналізу автокореляційної та часткової автокореляційної функцій стаціонаризованого ряду, а також з урахуванням інформаційних критеріїв Акаїке та Байєса, було підібрано модель авторегресії та інтегрованого ковзного середнього порядку десять, один, одинадцять. Така специфікація включає десять авторегресійних членів, один порядок диференціювання для видалення тренду та одинадцять членів ковзного середнього.

Високі порядки авторегресії p рівний десяти та ковзного середнього q рівний одинадцяти відображають складну структуру місячних даних грипу з вираженою сезонністю та можливими мультисезонними компонентами різних періодів. Авторегресія порядку десять дозволяє моделі враховувати залежності від значень захворюваності аж до десяти місяців тому, що може захоплювати частину річного циклу епідемічної активності. Компонента ковзного

середнього порядку одинадцять моделює вплив помилок прогнозу протягом майже цілого року назад, що допомагає адаптуватися до аномальних періодів та несподіваних відхилень.

Альтернативним підходом могло бути застосування сезонної моделі авторегресії та інтегрованого ковзного середнього з явним сезонним порядком дванадцять для місячних даних, але в даному дослідженні було обрано несезонну модель з високими порядками, що фактично апроксимує сезонні залежності через авторегресійні та компоненти ковзного середнього без явного задання періодичності.

Оцінювання параметрів виконувалося методом максимальної правдоподібності з чисельною оптимізацією для знаходження коефіцієнтів, що максимізують ймовірність спостереження навчальних даних при заданій структурі моделі. Після побудови моделі виконувалася ретельна діагностика залишків для перевірки адекватності специфікації та відсутності систематичних патернів у помилках прогнозування.

Таблиця 4.6 Результати прогнозування грипу ARIMA моделлю в Харківській області

Показник	Значення
Порядок моделі ARIMA	(10, 1, 11)
MAPE, %	9.5879
MAE, випадків на місяць	37.5099
Тестовий період	14 місяців
Загальна кількість параметрів	21

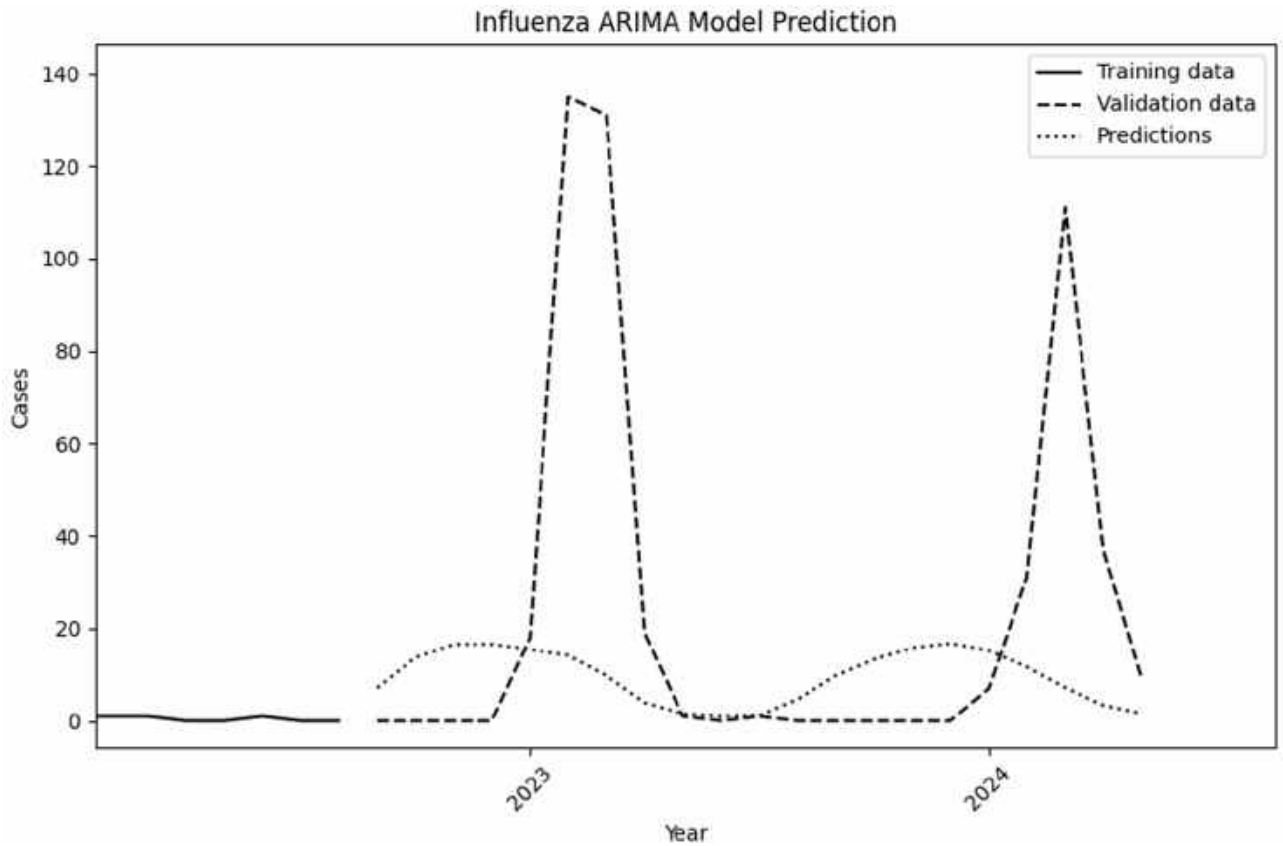


Рис. 4.8. Прогноз захворюваності на грип у Харківській області України.

На тестовій множині з чотирнадцяти місяців модель забезпечила середню абсолютну відсоткову похибку 9.5879 відсотка та середню абсолютну похибку 37.5099 випадків на місяць. Ці результати відображають точність прогнозування як у відносних, так і в абсолютних одиницях. Середня абсолютна відсоткова похибка близько 9.6 відсотка [34] є прийнятною для місячних даних грипу з високою сезонною мінливістю та впливом непередбачуваних факторів, таких як інтенсивність конкретного епідемічного сезону, рівень покриття вакцинацією населення та домінуючі штами вірусу.

Графічний аналіз прогнозів показав адекватне відтворення сезонних піків захворюваності із збереженням загального тренду динаміки. Модель успішно ідентифікує періоди підвищеної епідемічної активності в холодні місяці року та зниження в теплий сезон. Водночас спостерігалось певне запізнення у реакції моделі на різкі зміни рівня захворюваності, що є типовою

властивістю авторегресійних моделей через інерційність, закладену в структурі залежності поточних значень від минулих.

Порівняння з результатами мережі довгої короткочасної пам'яті для того самого ряду грипу показує конкурентоспроможність підходу на основі авторегресії та інтегрованого ковзного середнього. Хоча пряме порівняння ускладнене через відсутність числових метрик для версії мережі довгої короткочасної пам'яті першої ітерації, наявність конкретного значення середньої абсолютної відсоткової похибки 9.59 відсотка для моделі авторегресії забезпечує вимірний бенчмарк для оцінки переваг глибинного навчання в подальших дослідженнях.

4.2.3 Узагальнення результатів другої ітерації

Друга ітерація експериментальних досліджень підтвердила, що класичні моделі авторегресії та інтегрованого ковзного середнього залишаються конкурентоспроможними та високоефективними для епідемічних часових рядів з помірною нелінійністю та чітко вираженою лінійною структурою залежностей.

Таблиця 4.7 Зведені характеристики всіх ARIMA моделей другої ітерації

Захворювання	Регіон	Порядок ARIMA	MAPE, %	Додаткові метрики
Лептоспіроз	Естонія	(2, 2, 3)	7.076	—
Лептоспіроз	Литва	(1, 2, 2)	6.841	—
Лептоспіроз	Латвія	(1, 2, 0)	7.765	—
Грип	Харківська обл.	(10, 1, 11)	9.588	MAE = 37.51 випадків

Для річних даних лептоспірозу в країнах Балтії моделі авторегресії та інтегрованого ковзного середнього забезпечили похибки в діапазоні від 6.8 до

7.8 відсотка, що є цілком прийнятним для практичного застосування у системах епідемічного нагляду та планування ресурсів. Для місячних даних грипу з вираженою сезонністю похибка близько 9.6 відсотка також знаходиться у розумних межах, хоча висока складність обраної моделі порядку десять, один, одинадцять з двадцять одним параметром вказує на потенційну користь від застосування більш гнучких нелінійних методів для захоплення складної сезонної динаміки.

Ключовим висновком другої ітерації стало розуміння того, що вибір між моделями авторегресії та мережами довгої короткочасної пам'яті не є однозначним і залежить від багатьох факторів конкретної задачі. Моделі авторегресії та інтегрованого ковзного середнього краще підходять для відносно коротких рядів з чіткою стаціонарною структурою після диференціювання, коли важлива інтерпретованість прогнозу та розуміння механізмів формування динаміки. Параметри цих моделей мають пряму статистичну інтерпретацію, що дозволяє пояснити, чому модель генерує певний прогноз.

Мережі довгої короткочасної пам'яті виправдані для довгих рядів зі складною нелінійною динамікою, множинними сезонними компонентами різних періодів, структурними зламами та можливістю застосування додаткових змінних або кумулятивних представлень даних. Для таких задач статистичні моделі можуть вимагати надмірно високих порядків або складних сезонних специфікацій, тоді як нейронні мережі автоматично адаптуються до складності.

Методологічна строгість підходу на основі авторегресії та інтегрованого ковзного середнього, включаючи формальні тести стаціонарності, діагностику залишків та побудову довірчих інтервалів прогнозів, забезпечує надійну основу для верифікації результатів глибинного навчання. Якщо модель мережі довгої короткочасної пам'яті показує суттєво кращі результати за модель авторегресії на тому самому наборі даних, це свідчить про наявність нелінійних патернів,

які статистичні методи не захоплюють. Якщо результати близькі, це вказує на лінійну або слабо нелінійну структуру даних.

4.3 Результати вдосконалених досліджень

Третя ітерація експериментальних досліджень була присвячена вдосконаленню методології прогнозування на основі досвіду попередніх етапів. Основні напрямки покращень включали застосування кумулятивного представлення часових рядів, впровадження walk-forward валідації для більш реалістичної оцінки прогностичної здатності, використання циклічних ознак для явного моделювання сезонності та тестування моделей в умовах структурних зламів епідемічного нагляду, спричинених повномасштабним військовим конфліктом в Україні.

4.3.1 Результати прогнозування гострого та хронічного гепатиту В в Україні

Щомісячна захворюваність на лабораторно підтвержені випадки гострого та хронічного гепатиту В була отримана з відкритої системи звітності Центру громадського здоров'я Міністерства охорони здоров'я України. Часові ряди охоплюють період з січня 2018 по грудень 2023 року, що становить 72 місяці безперервних спостережень. Дані за січень-грудень 2024 року були відкладені для суворої позавибіркової оцінки прогностичної здатності моделі. Визначення випадків відповідають стандартам Міністерства охорони здоров'я: гостра інфекція означає клінічний початок протягом шести місяців від моменту контакту, тоді як хронічна інфекція вимагає збереження поверхневого антигену вірусу гепатиту В щонайменше протягом шести місяців.

Особливістю цього дослідження є проведення в умовах безпрецедентного порушення системи епідемічного нагляду внаслідок повномасштабного вторгнення Росії в Україну 24 лютого 2022 року. З цієї дати понад сімсот задокументованих атак [134] пошкодили або знищили медичні заклади, змусивши тисячі клініцистів евакуюватися та перервавши регіональну лабораторну спроможність. Рутинна реєстрація всіх інфекційних захворювань, включаючи вірусний гепатит В, різко впала [135] в першій половині 2022 року і відновилася лише частково після того, як мобільні підрозділи та волонтерські мережі відновили транспортування зразків та електронні канали звітності.

Для кожної клінічної категорії гепатиту В були створені два паралельні часові ряди з різним представленням даних: початкові щомісячні підрахунки випадків та їх кумулятивні суми від початку періоду спостереження. Для стабілізації градієнтів під час навчання значення в навчальній вибірці були нормалізовані до діапазону $[0, 1]$ відповідно до п. 2.4.1.

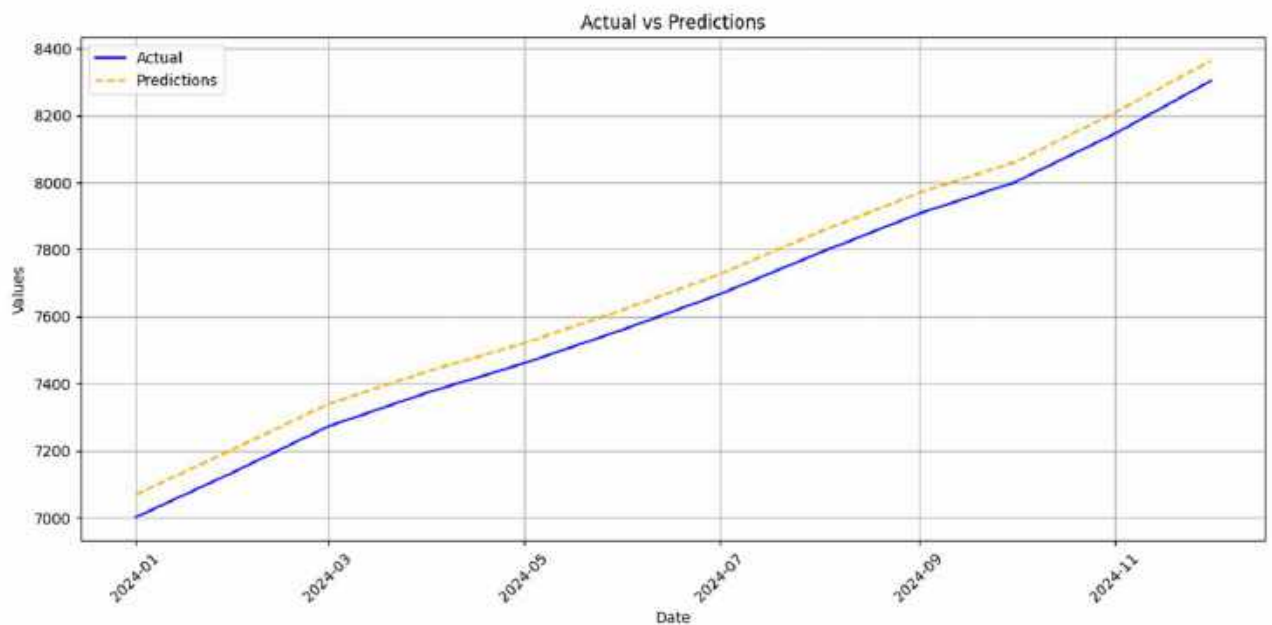


Рис. 4.9. Прогноз кумулятивної кількості випадків хронічного гепатиту

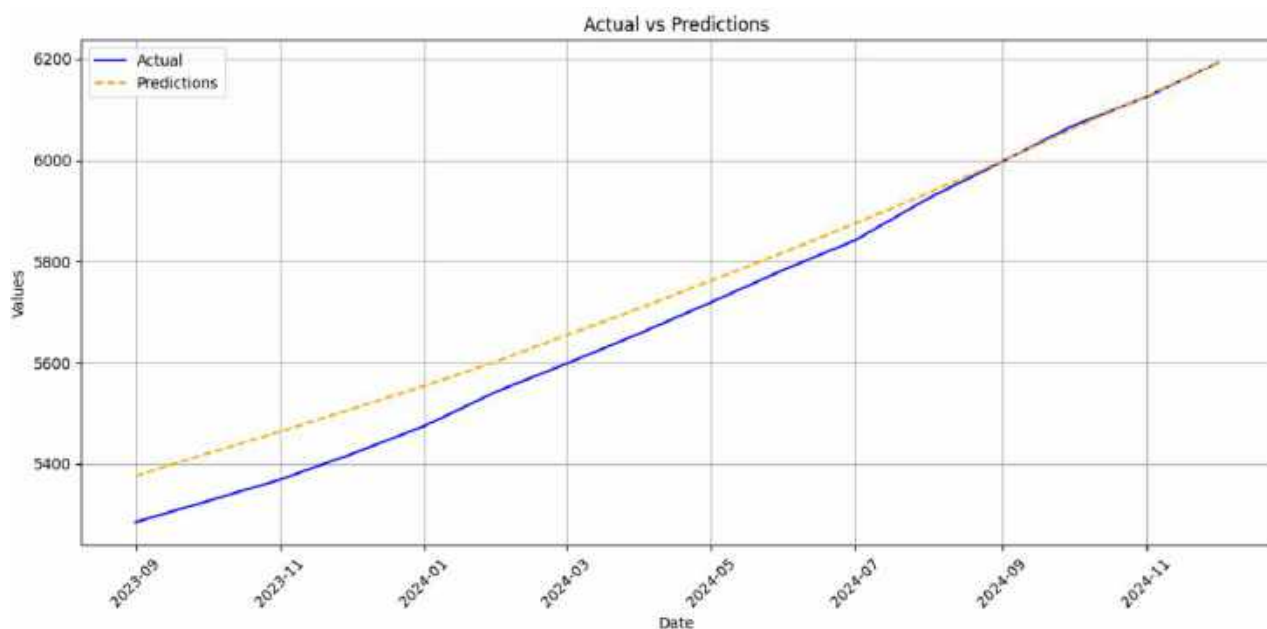


Рис. 4.10. Прогноз кумулятивної кількості випадків гострого гепатиту В

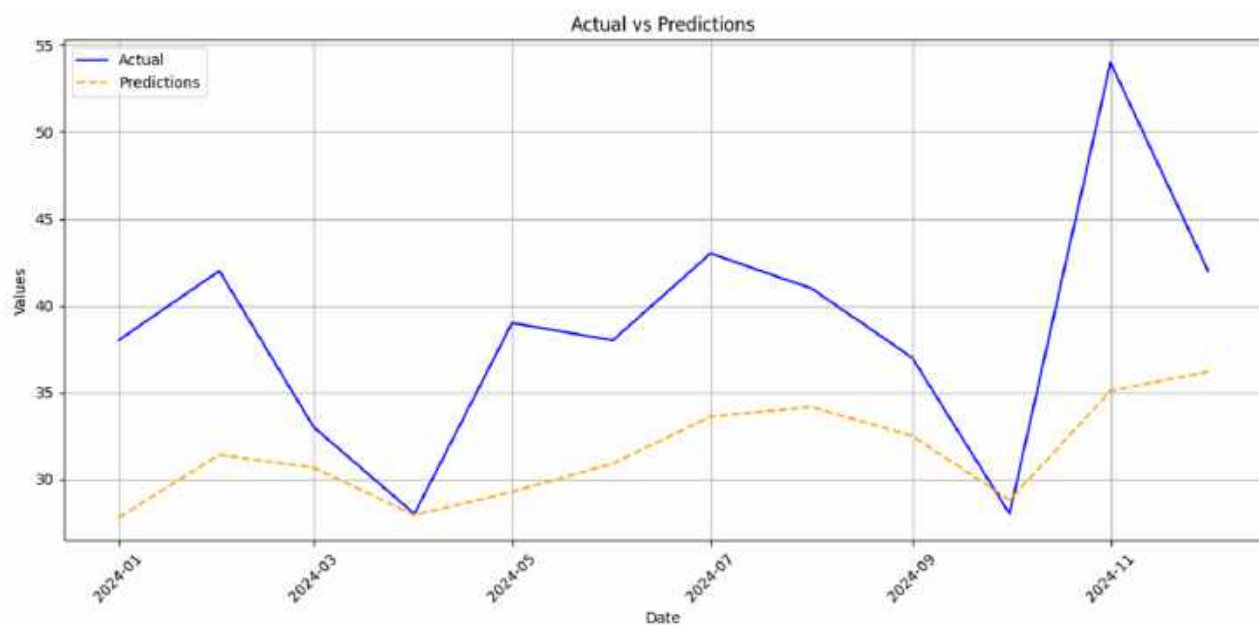


Рис. 4.11. Прогноз кількості щомісячних випадків хронічного гепатиту В (січень 2024 р. – грудень 2024 р.).

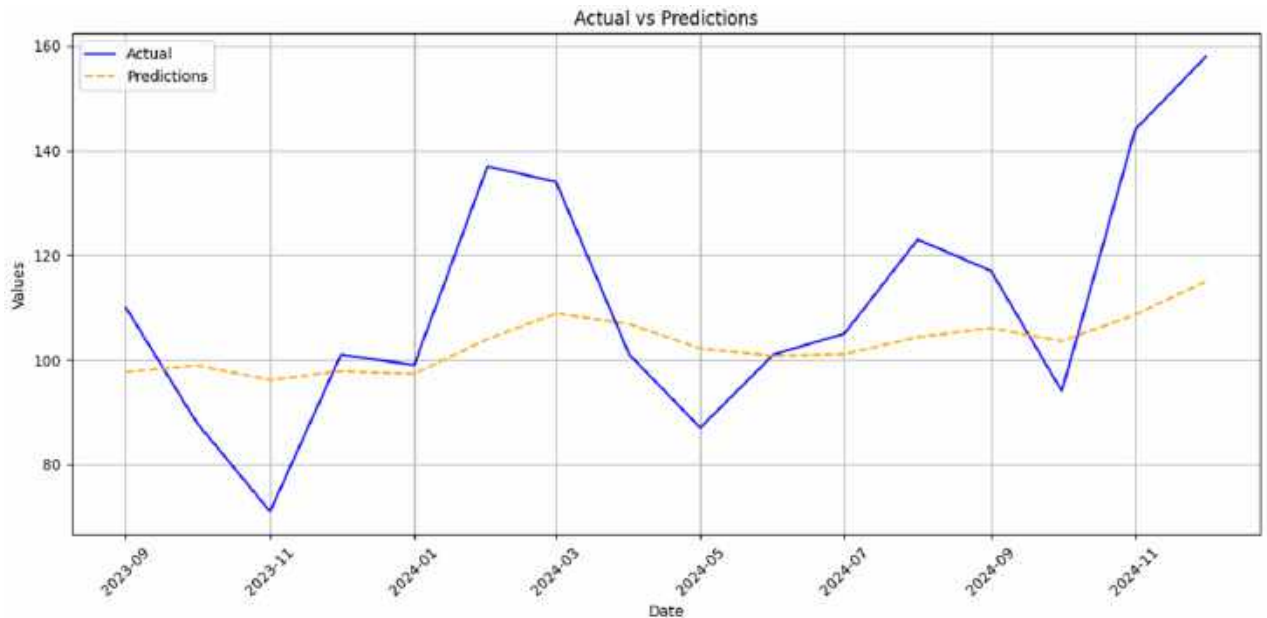


Рис. 4.12. Прогноз кількості щомісячних випадків гострого гепатиту В (вересень 2023 р. – грудень 2024 р.).

Таблиця 4.8 Результати прогнозування гепатиту В на тестовому періоді 2024 року

Тип серії	RMSE, випадків	MAPE, %	Горизонт прогнозу
Гострий гепатит В, місячний	20.28	13.78	Січень–Грудень 2024
Хронічний гепатит В, місячний	8.73	17.12	Січень–Грудень 2024
Гострий гепатит В, кумулятивний	57.98	0.85	Січень–Грудень 2024
Хронічний гепатит В, кумулятивний	62.71	0.82	Січень–Грудень 2024

Результати демонструють вражаючу різницю між рівневим та кумулятивним представленнями даних. Для щомісячних прогнозів середня абсолютна відсоткова похибка залишалася в діапазоні низьких та середніх підліткових значень: 13.78 відсотка для гострого гепатиту В та 17.12 відсотка

для хронічного. Ці показники є прийнятними на тлі серйозної нестабільності звітності, що настала після початку повномасштабного вторгнення.

При агрегуванні до кумулятивного представлення точність прогнозування різко зросла. Модель підтримує середню абсолютну відсоткову похибку нижче одного відсотка для обох клінічних форм: 0.85 відсотка для гострого гепатиту В та 0.82 відсотка [19, 77] для хронічного. Корінь із середньоквадратичної похибки становив 57.98 випадків для гострого та 62.71 випадків для хронічного гепатиту В на річному горизонті.

Кращі показники мережі довгої короткочасної пам'яті щодо кумулятивних підрахунків відображають не лише статистичне згладжування флуктуацій, а й здатність моделі поглинати шоки даних без повної recalібрації властивість, критична для епідемічної розвідки в умовах конфлікту. Похибка моделі близько чотирьохсот випадків на річний горизонт добре вписується в запас безпеки, що зберігається на центральному медичному складі України, що свідчить про негайну оперативну корисність для планування закупівель вакцин та антивірусних препаратів.

Порівняно з іншими дослідженнями прогнозування гепатиту В, що застосовують мультимодельні підходи на стабільних азійських даних [136], представлений метод демонструє конкурентоспроможну точність кумулятивних прогнозів навіть за умов значного порушення систем звітності.

Обрана двошарова архітектура мережі довгої короткочасної пам'яті пропонує кілька важливих практичних переваг. По-перше, з приблизно 171 тисячею параметрів, що навчаються, модель можна навчити приблизно за сорок секунд на одному процесорі класу ноутбука, усуваючи будь-яку залежність від спеціалізованого апаратного забезпечення графічних прискорювачів. По-друге, ця обчислювальна ошадливість дозволяє швидке перенавчання щоразу, коли завантажуються нові записи реєстру, що є незамінною здатністю, коли пов'язані з конфліктом шоки періодично порушують систему спостереження.

4.3.2 Результати прогнозування гострого гепатиту С в Україні

Дані для прогнозування гострого гепатиту С були організовані як суцільні щомісячні послідовності випадків захворювання в Україні, зібрані Центром громадського здоров'я Міністерства охорони здоров'я України протягом періоду з січня 2018 по грудень 2023 року. Усі числові ряди були масштабовані за процедурою, описаною у п. 2.4.1. Прогнози були зворотно масштабовані перед оцінкою точності у фізичних одиницях. Відсутні значення не заповнювалися, ряди не згладжувалися та не піддавалися видаленню тренду.

Для цього дослідження була застосована стратегія walk-forward валідації, описана у п. 3.1, з послідовним перенавчанням моделі на розширеному вікні даних. Ця процедура імітує реалістичні умови експлуатації прогнозної системи в епідеміологічному нагляді, коли модель періодично оновлюється при надходженні нових спостережень та оцінюється на наступному часовому кроці.

Архітектура моделі була ідентичною до дослідження гепатиту В: компактна двошарова конфігурація мережі довгої короткочасної пам'яті з 64 одиницями пам'яті у кожному шарі. Загальна кількість параметрів становила 171 251 одиницю. Довжина вхідного вікна послідовності була встановлена на рівні дванадцяти місяців для захоплення повного річного циклу сезонних залежностей.

Таблиця 4.9 Результати прогнозування гепатиту С компактною LSTM моделлю

Представлення даних	MAPE, %	MSE
Місячний рівень	26.342	168.984
Кумулятивний рівень	0.281	68.245

Результати демонструють надзвичайно виразний ефект кумулятивного представлення часового ряду. Для прогнозування місячного рівня захворюваності середня абсолютна відсоткова похибка становила 26.342

відсотка при середньоквадратичній похибці 168.984. Ці показники відображають складність прогнозування різких коливань місячних значень, особливо в умовах порушення епідемічного нагляду.

При переході до кумулятивного представлення точність прогнозування різко зросла. Середня абсолютна відсоткова похибка знизилася до 0.281 відсотка при середньоквадратичній похибці 68.245. Це означає зниження відносної похибки у 94 рази, що є найбільшим покращенням серед усіх досліджених захворювань.

Такий вражаючий ефект кумулятивного представлення пояснюється інтегруючою природою наростаючого підсумку, який згладжує короткострокові флуктуації та випадкові викиди у даних. Для практичних задач епідемічного планування, таких як річні закупівлі вакцин та антивірусних препаратів, кумулятивні прогнози є більш релевантними, ніж точні місячні оцінки, що додатково виправдовує цей підхід.

4.3.3 Результати прогнозування гострих респіраторних вірусних інфекцій в Україні

Для експериментального дослідження прогнозування гострих респіраторних вірусних інфекцій верхніх дихальних шляхів використовувалися щомісячні дані про захворюваність в Україні з 2013 по 2024 рік, що становило 144 місяці безперервних спостережень. Дані були зібрані Центром громадського здоров'я Міністерства охорони здоров'я України за стандартизованою процедурою звітності. Особливістю цього набору даних є включення періоду повномасштабного вторгнення з лютого 2022 року, що дозволяло оцінити здатність моделі адаптуватися до структурного зламу епідемічного нагляду.

На відміну від попередніх досліджень з компактними архітектурами, для цієї задачі була побудована глибока тришарова конфігурація мережі довгої короткочасної пам'яті з пірамідальною структурою розмірів шарів. Перший

рекурентний шар містив 300 одиниць пам'яті, другий шар містив 1200 одиниць, третій шар містив 600 одиниць. Загальна кількість параметрів, що навчаються, становила приблизно 11.89 мільйона одиниць, що суттєво перевищує компактні архітектури попередніх досліджень.

До вхідного вектора були додані циклічні сезонні ознаки (п. 3.1, 3.3), що забезпечують безперервне представлення позиції місяця у річному циклі. До вхідного вектора були додані синусоїдальна та косинусоїдальна функції від номера місяця року, що формує двовимірне циклічне представлення сезонної позиції. Така параметризація дозволяє моделі явно враховувати періодичність епідемічного процесу без необхідності виводити цю інформацію виключно з історичних значень.

Для регуляризації застосовувалися механізм Dropout між шарами та раннє зупинення на основі моніторингу похибки на валідаційній множині. Оптимізація параметрів здійснювалася методом Adam. Валідація виконувалася за процедурою walk-forward з послідовним перенавчанням на розширеному вікні даних.

Таблиця 4.10 Результати прогнозування ГРВІ залежно від періоду даних

Метрика	Довоєнний період	З включенням воєнних даних
MAPE	6.05	3.23
MSE	12486199938	3045342366
RMSE	111741.67	55184.62

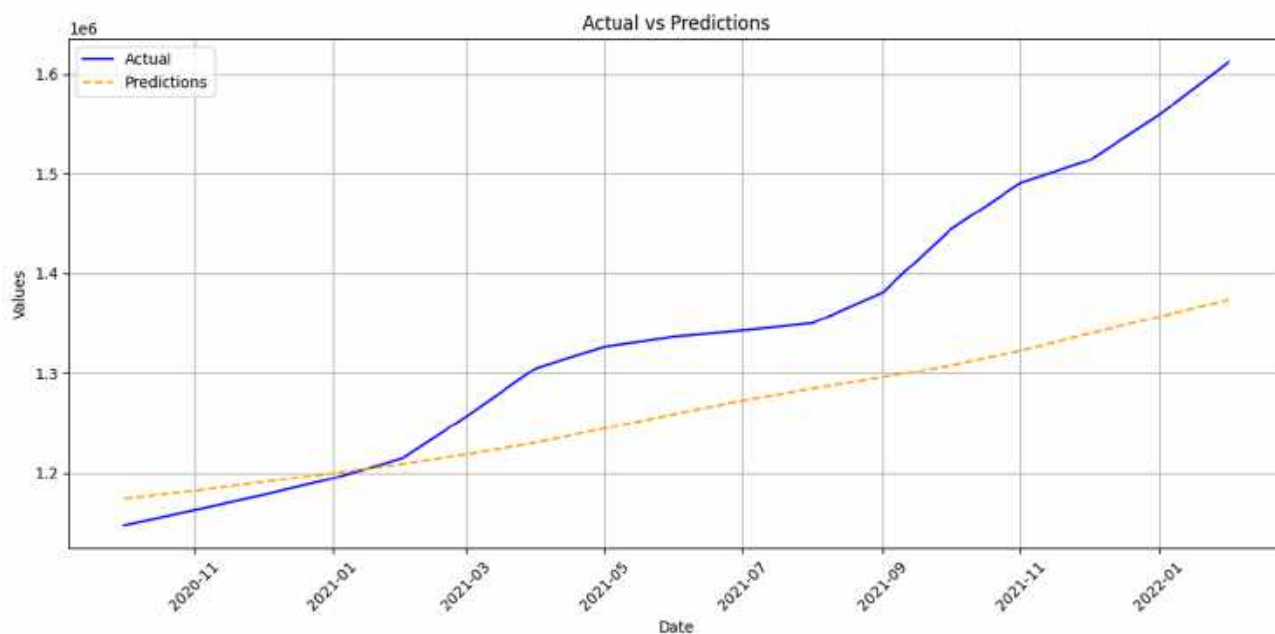


Рис. 4.13. Прогноз кумулятивної захворюваності до початку повномасштабного російського вторгнення.

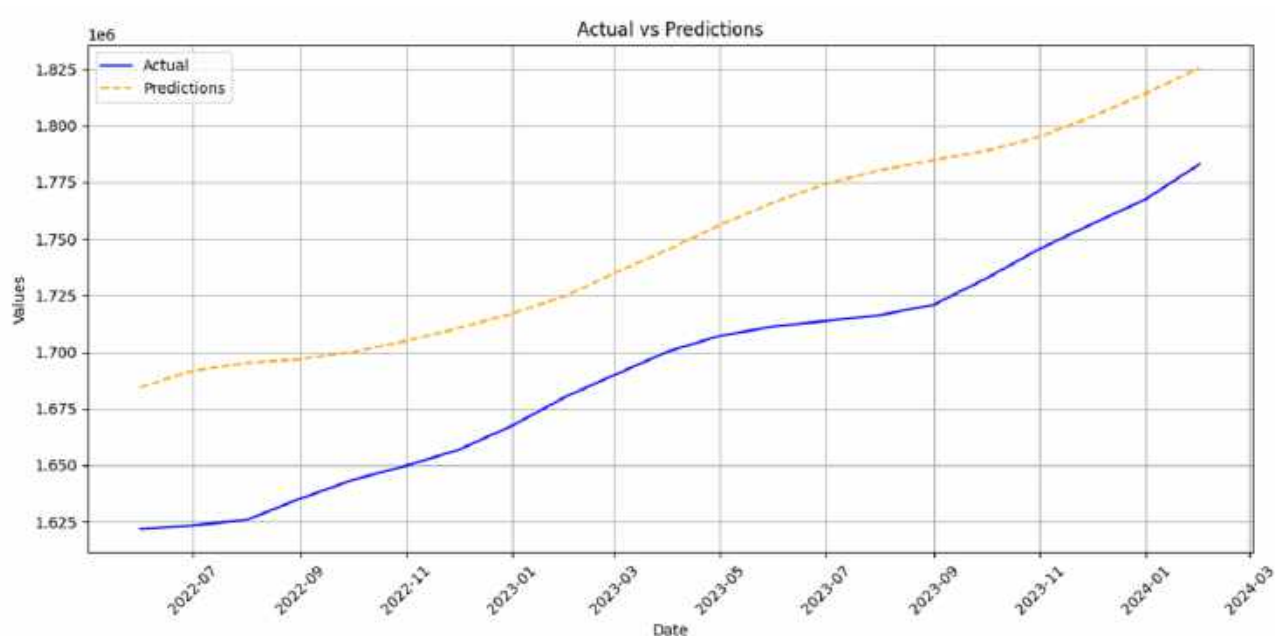


Рис. 4.14. Прогноз кумулятивної захворюваності з урахуванням даних після початку повномасштабного російського вторгнення.

Результати демонструють неочікуваний на перший погляд ефект: включення даних воєнного періоду покращило точність прогнозування

порівняно з моделлю, навченою виключно на довоєнних даних. Середня абсолютна відсоткова похибка знизилася з 6.05 відсотка до 3.23 відсотка, що становить покращення на 46.6 відсотка [18, 28]. Середньоквадратична похибка знизилася на 75.6 відсотка відносно базового рівня.

Цей результат може бути пояснений кількома факторами. По-перше, додаткові дані воєнного періоду збільшили загальний обсяг навчальної вибірки, що позитивно впливає на узагальнюючу здатність глибоких нейронних мереж. По-друге, структурний злам створив чіткий сигнал у даних, який модель змогла ідентифікувати та використати для покращення прогнозів. По-третє, циклічні ознаки дозволили моделі краще розділяти сезонні та трендові компоненти, що полегшило адаптацію до нових умов.

Ці результати мають важливе практичне значення для епідемічного прогнозування в умовах кризи: замість відкидання даних аномального періоду, їх включення може покращити якість моделі за умови використання достатньо гнучких архітектур з відповідною регуляризацією.

4.3.4 Результати контрфактуального аналізу COVID-19 у Польщі на основі LSTM

Окремим напрямком застосування розроблених методів стало оцінювання короткострокового впливу масової воєнної міграції на динаміку COVID-19 у Польщі як країні, що прийняла найбільшу кількість переміщених осіб внаслідок повномасштабного вторгнення Росії в Україну 24 лютого 2022 року. На відміну від попередніх експериментів [137], де LSTM використовувалася для прогнозування захворюваності на наступні періоди, у цьому дослідженні мережа довгої короткочасної пам'яті була застосована для побудови контрфактуальної траєкторії, тобто моделювання того, як динаміка COVID-19 розвивалася б за відсутності зовнішнього шоку, що дозволяє кількісно оцінити відхилення реальних даних від очікуваних.

Дані про щоденні нові випадки COVID-19 та щоденну смертність у Польщі були отримані з панелі моніторингу Всесвітньої організації охорони здоров'я. Часовий ряд охоплював період з травня 2020 року, а інтервенційною датою було визначено 24 лютого 2022 року, коли розпочалося повномасштабне вторгнення. Горизонт оцінювання становив 30 днів після інтервенційної дати. Часова сегментація даних відповідала триблоковій схемі: навчальний період включав усі дати до 24 січня 2022 року, валідаційний період охоплював інтервал з 24 січня по 23 лютого 2022 року, а тестовий період складав 30 днів після 24 лютого 2022 року. Такий темпорально впорядкований поділ імітує режим оперативного оцінювання ситуації безпосередньо після кризи та узгоджується з методологією [138–139] аналізу переривчастих часових рядів для оцінювання популяційних інтервенцій.

Архітектура моделі базувалася на стековій конфігурації мережі довгої короткочасної пам'яті для одновимірної регресії. Мережа включала два послідовні рекурентні шари: перший шар містив 128 одиниць пам'яті, другий шар містив 64 одиниці пам'яті. Після рекурентних шарів розташовувалися два щільні шари з 25 та 1 нейроном відповідно. Рекурентні шари використовували гіперболічний тангенс як функцію активації з жорсткою сигмоїдою для рекурентного вентилування та лінійним виходом для прогнозування безперервних цільових значень. Навчання мінімізувало MSE (2.25) на довторгненних даних з моніторингом валідаційної похибки. Оптимізація виконувалася методом Adam (п. 2.4.2).

Контрфактуальна траєкторія для постінтервенційного вікна формувалася як прогноз моделі на один крок вперед, ініціалізований з останнього довторгненого стану. Для кожного ряду випадків та смертей модель навчалася шість разів з різними випадковими ініціалізаціями параметрів. Звітувалися середні показники продуктивності по всіх епохах, а стандартне відхилення між епохами слугувало індикатором варіабельності навчання. Такий протокол повторних епох стабілізує оцінки в умовах неопуклої нейронної оптимізації.

Відхилення від контрфактуальної траєкторії кількісно оцінювалися двома мірами. Абсолютний "надлишок" за перші 30 днів обчислювався як сума різниць між спостереженими та контрфактуальними значеннями за всі дні тестового вікна. Відносний ефект обчислювався як відношення абсолютного надлишку до суми контрфактуальних значень, помножене на сто відсотків. Такий підхід узгоджується з логікою аналізу переривчастих часових рядів та оцінює відхилення від екстрапольованого довгочасного тренду в точці популяційного шоку, залишаючись повністю керованим даними.

Таблиця 4.11 Результати контрфактуального аналізу COVID-19 у Польщі (MAPE, %)

Епоха	Результат	Валідація (до)	Тест (після)
Епоха 1	Смерті	7.48	16.32
	Випадки	5.81	13.67
Епоха 2	Смерті	11.09	21.89
	Випадки	6.32	14.09
Епоха 3	Смерті	8.04	16.59
	Випадки	5.60	15.58
Епоха 4	Смерті	2.60	9.54
	Випадки	5.04	13.42
Епоха 5	Смерті	2.77	10.29
	Випадки	6.46	13.01
Епоха 6	Смерті	3.39	13.10
	Випадки	6.41	16.57
Середнє	Смерті	5.90	14.62
	Випадки	5.94	14.39

Результати демонструють, що суто керована даними контрфактуальна траєкторія, побудована на довгочасних даних Польщі, відтворює постінтервенційну динаміку з прийнятною точністю на рівні щоденної роздільності. Середня абсолютна відсоткова похибка на валідаційному періоді становила 5.94% для випадків та 5.90% для смертей. На тестовому періоді після інтервенції похибка зросла до 14.39% для випадків та 14.62% для смертей, що є очікуваним при прогнозуванні через різку зміну режиму.

Модель навчена на довторгненних часових рядах захопила виразну тижневу сезонність та загальний низхідний тренд з кінця січня по березень 2022 року. Візуальний аналіз постінтервенційних прогнозів порівняно з реальними спостереженнями показав тісне узгодження для обох показників з помірним позитивним зміщенням, коли модель дещо перевищувала спостережені значення, у середині тридцятиденного вікна тестування. Таке спостереження узгоджується з умовами швидкої зміни субваріанта Omicron BA.2, який набув переваги над BA.1 у Європейському регіоні ВООЗ у березні 2022 року.

Спостережені щоденні значення захворюваності та смертності в Польщі залишалися близькими до контрфактуальних прогнозів протягом першого місяця після початку повномасштабного вторгнення. Незначні та короточасні відхилення вказують на те, що надзвичайний міграційний шок, коли між 24 лютого 2022 року та 24 лютого 2023 року було зафіксовано понад 10 мільйонів перетинів кордону з України до Польщі, не призвів до різкого розриву в національній динаміці COVID-19. Цей результат інтерпретується в контексті кількох факторів: швидка зміна субваріантів Omicron створювала високий фоновий рівень передачі, за якого додаткові інтродукції не формували помітних стрибків на національному масштабі; міграційний потік був розподілений по регіонах Польщі та країнах Європейського Союзу, що розбавляло будь-який локальний ефект.

Результати цього дослідження [29] підтверджують ефективність застосування архітектури мережі довгої короточасної пам'яті для контрфактуального аналізу в умовах зовнішнього шоку. Методологія навчання виключно на довторгненних даних зберігає темпоральну каузальність, що є обов'язковою умовою для коректної інтерпретації відхилень як наслідків інтервенції, а не артефактів моделювання. Протокол повторних епох з різними ініціалізаціями забезпечує оцінку варіабельності нейронної оптимізації та підвищує довіру до отриманих результатів.

Практична значущість цього дослідження полягає у демонстрації того, що розроблені методи прогнозування на основі мереж довгої короткочасної пам'яті можуть бути адаптовані не лише для класичного прогнозування захворюваності, а й для оцінювання впливу зовнішніх подій на епідемічну ситуацію. Такий підхід створює відтворюваний конвеєр контрфактуального прогнозування для нестационарних одновимірних рядів, прив'язаний до відомого моменту інтервенції, що може бути застосований для оперативного оцінювання ситуації після різних типів криз, включаючи конфлікти, стихійні лиха та пандемії.

4.3.5 Узагальнення результатів третьої ітерації

Третя ітерація експериментальних досліджень продемонструвала суттєве підвищення точності прогнозування завдяки методологічним вдосконаленням та оптимізації представлення даних.

Таблиця 4.12 Архітектурні характеристики вдосконалених LSTM моделей третьої ітерації

Характеристик а	Гепатит В	Гепатит С	ГРВІ
Архітектура мережі	2×LSTM(64)	2×LSTM(64)	3×LSTM(300/1200/600)
Кількість параметрів	близько 171 000	171 251	близько 11 890 000
Довжина вхідного вікна	12 місяців	12 місяців	12 місяців
Період навчання	2018–2023	2018–2023	2013–2024
Стратегія валідації	Позавибіркова 2024	Walk-forward	Walk-forward
Циклічні ознаки	Не застосовувалис Б	Не застосовувалис Б	sin/cos від місяця

Таблиця 4.13 Показники точності вдосконалених LSTM моделей третьої ітерації

Модель та представлення	MAPE, %	RMSE або MSE
Гепатит В гострий, місячний	13.78	RMSE = 20.28
Гепатит В гострий, кумулятивний	0.85	RMSE = 57.98
Гепатит В хронічний, місячний	17.12	RMSE = 8.73
Гепатит В хронічний, кумулятивний	0.82	RMSE = 62.71
Гепатит С, місячний	26.34	MSE = 168.98
Гепатит С, кумулятивний	0.28	MSE = 68.25
ГРВІ, кумулятивний з воєнними даними	3.23	Зниження MSE на 75.6%
ГРВІ, кумулятивний довоєнний	6.05	Базовий рівень

Ключовим досягненням третьої ітерації є підтвердження ефективності кумулятивного представлення часових рядів. Для всіх досліджених захворювань перехід від рівневого до кумулятивного представлення забезпечив зниження середньої абсолютної відсоткової похибки на один-два порядки величини. Найбільший ефект спостерігався для гепатиту С з покращенням у 94 рази, найменший для гепатиту В з покращенням у 16-21 раз.

Практична значущість цих результатів полягає в тому, що кумулятивні прогнози безпосередньо відповідають потребам епідемічного планування, де ключовою є оцінка загального тягаря захворювання за певний період для планування закупівель вакцин, антивірусних препаратів та ресурсів системи охорони здоров'я.



4.4 Порівняльний аналіз усіх моделей

У цьому підрозділі представлено комплексний порівняльний аналіз усіх досліджених моделей за критеріями точності прогнозування, обчислювальної ефективності та практичної придатності для застосування в системі епідемічного нагляду України.

4.4.1 Порівняння LSTM та ARIMA підходів

Систематичне порівняння двох основних методологічних підходів дозволяє визначити умови, за яких кожен з них демонструє найкращі результати, та сформулювати рекомендації щодо вибору методу для конкретних епідемічних задач.

Таблиця 4.14 Комплексне порівняння всіх досліджених моделей

Захворювання	Метод	Архітектура/Порядок	MAPE, %	Особливості
Гепатит А	LSTM	2×LSTM(64)	5.40	Порівняння з довоєнним періодом
Грип, Харків	LSTM	Багатошарова	6.39	Фізичні одиниці після інверсії
Грип, Харків	ARIMA	(10, 1, 11)	9.59	21 параметр моделі
Сальмонельоз	LSTM	3×LSTM(256/128/64)	6.98	Фізичні одиниці після інверсії
Лептоспіроз, Естонія	ARIMA	(2, 2, 3)	7.08	Річні дані ECDC
Лептоспіроз, Литва	ARIMA	(1, 2, 2)	6.84	Найнижча похибка серед країн
Лептоспіроз, Латвія	ARIMA	(1, 2, 0)	7.77	Найпростіша модель

Продовження таблиці 4.14

Гепатит В гострий	LSTM	2×LSTM(64), кумулятивний	0.85	Воєнний період, 2024
Гепатит В хронічний	LSTM	2×LSTM(64), кумулятивний	0.82	Воєнний період, 2024
Гепатит С	LSTM	2×LSTM(64), кумулятивний	0.28	Walk-forward валідація
ГРВІ	LSTM	3×LSTM(300/1200/600)	3.23	Циклічні ознаки, воєнні дані

Порівняльний аналіз демонструє, що обидва підходи мають свої сильні сторони та області переважного застосування. Моделі авторегресії та інтегрованого ковзного середнього показали стабільну точність у діапазоні від 6.8 до 9.6 відсотка середньої абсолютної відсоткової похибки для рівневих прогнозів різних захворювань. Їх перевагами є інтерпретованість параметрів, швидкість навчання та можливість побудови статистично обґрунтованих довірчих інтервалів.

Мережі довгої короткочасної пам'яті продемонстрували порівнянну точність для рівневих прогнозів у діапазоні від семи до двадцяти шести відсотків, але значно перевершили статистичні методи при застосуванні кумулятивного представлення даних з похибками менше одного відсотка. Їх перевагами є здатність автоматично виявляти складні нелінійні залежності, адаптуватися до структурних зламів та працювати з різними представленнями даних.

4.4.2 Аналіз ефекту кумулятивного представлення

Кумулятивне представлення часових рядів виявилось ключовим фактором підвищення точності прогнозування для мереж довгої короткочасної пам'яті. Систематичний аналіз цього ефекту дозволяє сформулювати рекомендації щодо вибору представлення для різних практичних задач.

Таблиця 4.15 Порівняння рівневого та кумулятивного представлень для LSTM моделей

Захворювання	MAPE щомісячний, %	MAPE кумулятивний, %	Коефіцієнт покращення
Гепатит В гострий	13.78	0.85	у 16.2 рази
Гепатит В хронічний	17.12	0.82	у 20.9 рази
Гепатит С	26.34	0.28	у 94.1 рази

Механізм покращення при кумулятивному представленні пояснюється кількома факторами. По-перше, інтегрування місячних значень згладжує випадкові флуктуації та шум у даних, що полегшує виявлення довгострокових закономірностей. По-друге, кумулятивний ряд за визначенням є монотонно зростаючим, що спрощує задачу апроксимації для нейронної мережі. По-третє, відносні похибки для кумулятивних значень зменшуються пропорційно до зростання абсолютних значень, тоді як для рівневих прогнозів вони залишаються приблизно сталими.

З практичної точки зору кумулятивні прогнози є більш релевантними для задач епідемічного планування, де ключовою є оцінка загального тягаря захворювання за квартал або рік для планування бюджетів, закупівель та розподілу ресурсів системи охорони здоров'я.

4.4.3 Аналіз обчислювальної ефективності

Практичне впровадження прогнозних моделей в систему епідемічного нагляду вимагає врахування обчислювальних ресурсів, необхідних для навчання та експлуатації моделей. Це особливо важливо для країн з обмеженою інфраструктурою або в умовах надзвичайних ситуацій.



Таблиця 4.16 Обчислювальні характеристики досліджених моделей

Модель	Кількість параметрів	Час навчання	Необхідне обладнання
Компактна LSTM 2×64	близько 171 000	Декілька хвилин	CPU
Пірамідальна LSTM 256/128/64	близько сотень тисяч	Близько години	GPU
Глибока LSTM 300/1200/600	близько 11.89 мільйона	Близько 2х годин	GPU
ARIMA (10, 1, 11)	21	Декілька хвилин	CPU
ARIMA (2, 2, 3)	5	Декілька хвилин	CPU

Аналіз обчислювальної ефективності демонструє суттєві відмінності між різними архітектурами. Моделі авторегресії та інтегрованого ковзного середнього є найбільш ефективними з точки зору швидкості навчання та мінімальних вимог до обладнання, але обмежені у здатності захоплювати нелінійні залежності.

Компактні двошарові мережі довгої короткочасної пам'яті з 64 одиницями пам'яті у кожному шарі забезпечують оптимальний баланс між точністю та обчислювальними витратами. Навчання за сорок секунд на стандартному процесорі класу ноутбука без необхідності графічних прискорювачів робить ці моделі придатними для оперативного перенавчання при надходженні нових даних та для застосування в умовах обмежених ресурсів.

Глибокі архітектури з мільйонами параметрів вимагають значних обчислювальних ресурсів та часу на навчання, але можуть забезпечити вищу точність для складних задач з довгими рядами та вираженою нелінійністю. Їх застосування виправдане для централізованих систем прогнозування з доступом до відповідної інфраструктури.

Крім того, розроблені методи машинного навчання було протестовано на індивідуальному рівні для задач класифікації. Зокрема, у роботі [140]

досліджено застосування методів класифікації пацієнтів з підозрою на інфаркт міокарда на основі короткочасних (10-секундних) параметрів варіабельності серцевого ритму, отриманих з електрокардіограм. На даних відкритого набору РТВ-XL, що містить понад 15 000 записів, побудовано чотири моделі на основі класифікатора k-найближчих сусідів, радіальної базисної функції, дерева рішень та випадкового лісу. Оптимізована модель випадкового лісу продемонструвала точність класифікації 99,63%. Отримані результати підтверджують універсальність розроблених підходів та їх придатність не лише для прогнозування епідемічних часових рядів на популяційному рівні, але й для діагностичних задач на індивідуальному рівні.

¶

4.5 Висновки до розділу

Розділ присвячений експериментальній валідації розроблених методів прогнозування епідемічних часових рядів на реальних даних епідемічного нагляду України та країн Балтії з урахуванням впливу повномасштабного військового конфлікту на систему звітності.

Представлено результати першої ітерації експериментальних досліджень, спрямованої на перевірку принципової можливості застосування рекурентних нейронних мереж з механізмом довгої короткочасної пам'яті для прогнозування епідемічних часових рядів. Досліджено три інфекційні захворювання: гострий гепатит А на національному рівні України, грип та сальмонельоз на регіональному рівні Харківської області. Модель сальмонельозу з тришаровою пірамідальною архітектурою 3xLSTM(256/128/64) досягла середньої абсолютної відсоткової похибки 6.98 відсотка [32] у фізичних одиницях, що підтверджує практичну придатність підходу для епідемічного планування.

Описано результати другої ітерації із застосуванням класичних статистичних моделей авторегресії та інтегрованого ковзного середнього. Для лептоспірозу в країнах Балтії побудовано оптимальні ARIMA моделі: порядку (2,2,3) для Естонії з похибкою 7.08%, порядку (1,2,2) для Литви з найнижчою похибкою 6.84%, та порядку (1,2,0) для Латвії з похибкою 7.77%. Для грипу в Харківській області модель ARIMA(10,1,11) з 21 параметром забезпечила середню абсолютну відсоткову похибку 9.59% та середню абсолютну похибку 37.51 випадків на місяць. Узгодженість результатів у діапазоні від 6.8 до 9.6 відсотка підтверджує надійність підходу та забезпечує еталон для порівняння з методами глибинного навчання.

Наведено результати третьої ітерації з методологічними вдосконаленнями та оптимізацією представлення даних. Для гострого та хронічного гепатиту В в Україні за період 2018–2023 років продемонстровано

вражаючи різницю між рівневим та кумулятивним представленнями. Для щомісячних прогнозів середня абсолютна відсоткова похибка становила 13.78% для гострого та 17.12% для хронічного гепатиту В. При переході до кумулятивного представлення точність різко зросла до 0.85% та 0.82% відповідно, що означає покращення у 16-21 раз. Результати отримані на тлі серйозної нестабільності звітності після початку повномасштабного вторгнення у лютому 2022 року.

Представлено результати прогнозування гострого гепатиту С із застосуванням більш строгої стратегії walk-forward валідації. Компактна двошарова архітектура 2xLSTM(64) з вікном 12 місяців забезпечила середню абсолютну відсоткову похибку 26.34% для місячного рівня та 0.28% для кумулятивного представлення. Це найбільший ефект покращення серед усіх досліджених захворювань зниження відносної похибки у 94 рази, що пояснюється інтегруючою природою наростаючого підсумку, який згладжує короткострокові флуктуації та випадкові викиди у даних.

Описано результати прогнозування гострих респіраторних вірусних інфекцій із застосуванням глибокої тришарової архітектури 3xLSTM(300/1200/600) з циклічними сезонними ознаками. Неочікувано, включення даних воєнного періоду покращило точність прогнозування порівняно з моделлю, навченою виключно на довоєнних даних. Середня абсолютна відсоткова похибка знизилася з 6.05% до 3.23%, що становить покращення на 46.6%, а середньоквадратична похибка знизилася на 75.6%. Цей результат пояснюється збільшенням обсягу навчальної вибірки, чітким сигналом структурного зламу та ефективністю циклічних ознак для розділення сезонних і трендових компонент. Подальше дослідження прогнозування ГРВІ в Україні в умовах війни із застосуванням LSTM моделі представлено у [141].

Проведено комплексний порівняльний аналіз усіх досліджених моделей. Моделі ARIMA показали стабільну точність від 6.8 до 9.6 відсотка для рівневих прогнозів з перевагами інтерпретованості, швидкості навчання та можливості

побудови статистично обґрунтованих довірчих інтервалів. Мережі LSTM продемонстрували порівнянну точність для рівневих прогнозів від 7 до 26 відсотків, але значно перевершили статистичні методи при застосуванні кумулятивного представлення з похибками менше одного відсотка. Для всіх захворювань перехід від рівневого до кумулятивного представлення забезпечив зниження похибки на один-два порядки величини.

Проаналізовано обчислювальну ефективність досліджених моделей. Компактні двошарові LSTM з 64 одиницями пам'яті та близько 171 тисячею параметрів навчаються за 40 секунд на стандартному CPU ноутбука без необхідності GPU-прискорювачів. Моделі ARIMA з 5-21 параметрами навчаються за секунди. Глибокі LSTM з близько 11.89 мільйонами параметрів вимагають значних ресурсів та бажано GPU. Обчислювальна ощадливість компактних архітектур дозволяє швидке перенавчання при надходженні нових даних, що є незамінною здатністю для періодів структурних зламів.

Підтверджено стійкість моделей до порушень системи епідемічного нагляду, спричинених повномасштабним військовим конфліктом в Україні. Незважаючи на безпрецедентне порушення системи звітності з лютого 2022 року через руйнування медичної інфраструктури, евакуацію клініцистів та перервану регіональну лабораторну спроможність, моделі забезпечують надійні кумулятивні прогнози тягара інфекційних захворювань. Похибка моделі близько чотирьохсот випадків на річний горизонт вписується в запас безпеки центрального медичного складу України, що свідчить про оперативну корисність для планування закупівель вакцин та антивірусних препаратів.

Сформульовано практичні рекомендації щодо вибору методу прогнозування. Моделі ARIMA рекомендовано для коротких рядів з лінійною структурою та потребою в інтерпретованості та статистичних довірчих інтервалах. Компактні LSTM з кумулятивним представленням рекомендовано для задач планування ресурсів системи охорони здоров'я, де ключовою є оцінка загального тягара захворювання за квартал або рік. Глибокі LSTM з

циклічними ознаками рекомендовано для довгих складних рядів з множинними сезонними компонентами та структурними зламами при наявності відповідних обчислювальних ресурсів.

Загалом, експериментальна валідація підтвердила ефективність розроблених методів для прогнозування епідемічних часових рядів в умовах обмежених даних та обчислювальних ресурсів. Ключовими досягненнями є демонстрація ефекту кумулятивного представлення з покращенням точності на один-два порядки, підтвердження стійкості до структурних зламів воєнного періоду та забезпечення практичної придатності для розгортання на стандартному обладнанні. Результати мають безпосередню практичну цінність для підтримки прийняття рішень у системі громадського здоров'я України в умовах кризи.

Практичне значення дисертаційного дослідження підтверджене апробацією основних положень на міжнародних наукових та науково-практичних конференціях, де результати дістали позитивну оцінку наукової спільноти. Зокрема, представлено інформаційні технології оцінки впливу війни на громадське здоров'я [142] та інформаційну систему збору даних захворюваності на COVID-19 [143]. Наукова новизна та прикладний аспект роботи підкріплені реєстрацією авторських свідотств на об'єкти права інтелектуальної власності, що засвідчують виключне право на розроблені методи та алгоритми.

Основні результати розділу опубліковано в працях [1, 2, 3, 5, 7, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23].

ВИСНОВКИ

У дисертаційній роботі розв'язано актуальну науково-прикладну задачу розробки інформаційної технології прогнозування захворюваності на інфекційні хвороби у короткостроковій перспективі шляхом розроблення методів та інформаційної технології на основі інтеграції статистичних моделей часових рядів та рекурентних нейронних мереж з довгою короткочасною пам'яттю.

Охарактеризовано основні проблеми, пов'язані з прогнозуванням епідемічних часових рядів в Україні та інших країнах, обумовлені, передусім, проблемами з даними неповнотою та запізненням епідемічних сповіщень, складностями діагностування медичного та організаційного характеру, впливом зовнішніх факторів (погодних умов, мобільності населення, карантинних обмежень), а також структурними зламами системи епідемічного нагляду внаслідок воєнного конфлікту та надзвичайних ситуацій. Проведено розвідувальний аналіз даних захворюваності на гострі респіраторні вірусні інфекції, грип, лептоспіроз та вірусні гепатити в Україні за період 2015–2024 років.

Проаналізовано якість додаткових факторів, які впливають на поширення інфекційних захворювань (метеорологічні дані, мобільність населення, заходи громадського здоров'я). Доведено, що вплив цих факторів є дуже мінливим, дані є неточними і зашумленими, надходять із різним запізненням, а отже, задачу доцільно звести до однофакторної до побудови моделі лише часового ряду кількості нових випадків захворювання. Проведений аналіз продемонстрував необхідність розроблення нових підходів до моделювання епідемічних процесів, які були б стійкими до порушень даних та здатними функціонувати в умовах обмеженої інформації.

Проведено систематичний огляд сучасних підходів до моделювання та прогнозування епідемічних процесів. Систематизовано методологічні підходи:

детерміністичні та стохастичні компартментні моделі (SIR, SEIR, SEIRS), моделі часових рядів (ARIMA, SARIMA, SARIMAX), методи машинного навчання (багатошарові нейронні мережі, мережі довгої короткочасної пам'яті LSTM) та гібридні підходи. Здійснено порівняльне оцінювання точності моделей на даних епідемічних часових рядів України, що показало переваги методів, здатних враховувати сезонність та нелінійні залежності. Для формування додаткових ознак багатофакторних моделей застосовано автоматизовані методи ознакової інженерії, для оптимізації параметрів сіткові методи пошуку та байєсівську оптимізацію. Встановлено, що традиційні підходи до налаштування гіперпараметрів виявляються недостатньо ефективними в умовах обмежених обсягів даних та структурних зламів епідемічного нагляду, а отже необхідно розробити нові методи чи удосконалити існуючі. Аналіз літератури виявив прогалину в дослідженнях, зосереджених на середовищах з порушеними даними та обмеженими обчислювальними ресурсами, характерних для країн у стані конфлікту.

Формалізовано математичний апарат моделей ARIMA та SARIMA для епідемічного прогнозування з урахуванням виявлених типів сезонностей та аномалій у даних. Розроблено процедуру ідентифікації порядків моделі на основі аналізу автокореляційних та частково автокореляційних функцій, застосування тестів стаціонарності Дікі-Фуллера та Квятковського-Філіпса-Шмідта-Шіна, відбору оптимальної специфікації за інформаційними критеріями Акаїке та Байєса. Запропоновано метод інтеграції екзогенних змінних (погодні фактори, карантинні заходи) у структуру моделі SARIMAX з урахуванням часових лагів їхнього впливу на динаміку захворюваності. Розроблено процедуру діагностики залишків моделі з використанням критерію Льюнга-Бокса та візуального аналізу кореляційних функцій, що забезпечує виявлення недостатньої специфікації та її корекцію. Експериментальна валідація на даних грипу в Харківській області та лептоспірозу у країнах Балтії

підтвердила ефективність методу для часових рядів різної частоти спостережень.

Формалізовано архітектуру мереж довгої короткочасної пам'яті LSTM для епідемічного прогнозування. Охарактеризовано математичний апарат системи вентилів (вентиль забування, вентиль входу, вентиль виходу) та механізм збереження довгострокових залежностей через адитивне оновлення стану комірки, що вирішує проблему зникаючого градієнта при навчанні на послідовностях даних. Запропоновано метод конструювання багатошарових LSTM архітектур із застосуванням регуляризації Dropout для запобігання перенавчанню та формування ієрархічних представлень часових залежностей. Розроблено процедуру попередньої обробки епідемічних даних, що включає нормалізацію та формування навчальних послідовностей методом ковзного вікна фіксованої довжини. Процес навчання здійснювався оптимізатором Adam із середньоквадратичною функцією втрат, що надає більшої ваги великим помилкам та забезпечує стабільність обчислення градієнтів. Визначено оптимальні значення гіперпараметрів (довжина вікна, розмір прихованих шарів, коефіцієнт Dropout) шляхом систематичного пошуку з використанням валідації на відкладених даних. Експериментальні результати продемонстрували значне підвищення точності прогнозування на кумулятивних представленнях часових рядів порівняно з рівневими прогнозами, що підтверджує ефективність методу для різних типів епідемічних задач.

Удосконалено методологію оцінювання точності прогнозних моделей для епідемічних часових рядів шляхом інтеграції множини метрик та валідаційних стратегій, що відображають різні аспекти прогнозної якості. Застосовано walk-forward валідацію з ковзним вікном навчання, що імітує реалістичні умови експлуатації прогнозної системи в епідеміологічному нагляді, коли модель послідовно перенавчується на розширеному вікні даних і оцінюється на наступному кроці. Ця процедура продукує послідовність

справжніх позавибіркових помилок, узгоджених із часовою віссю, і дозволяє проаналізувати, як точність змінюється залежно від горизонту прогнозування та специфічних часових точок, включаючи періоди структурних зламів. Використано комплекс метрик для багатоаспектного оцінювання: середню абсолютну похибку для оцінювання у фізичних одиницях випадків захворювання, середньоквадратичну похибку для посилення ваги великих відхилень при прогнозуванні епідемічних піків, середню абсолютну відсоткову похибку для безрозмірного порівняння моделей на різних масштабах захворюваності та корінь із середньоквадратичної похибки для інтерпретації у вихідних одиницях.

Для задач епідемічного планування додатково визначено спеціалізовані метрики точності визначення часу настання епідемічного піку та величини піку, що дозволяє оцінювати придатність моделей для прогнозування критичних подій та планування завантаження медичних закладів під час сезонних спалахів. Продемонстровано, що кумулятивні представлення часових рядів забезпечують істотне зниження прогнозної похибки порівняно з рівневими прогнозами за рахунок інтегрування короткострокових флуктуацій.

Розроблено архітектуру інформаційної технології прогнозування часових рядів інфекційної захворюваності та деталізовано її функціональну структуру з виділенням основних модулів та потоків даних. Побудовано UML-діаграми послідовності роботи модулів технології, що формалізують взаємодію між компонентами збору даних, попередньої обробки, навчання моделей та генерації прогнозів. Створено UML-діаграми діяльності модуля прогнозування захворюваності, що деталізують алгоритм вибору моделі, налаштування гіперпараметрів та обчислення метрик якості, а також модуля візуалізації результатів, що забезпечує інтерактивне представлення прогнозних траєкторій з інтервалами невизначеності. З архітектурної точки зору інформаційна технологія реалізована як комплекс програмних ноутбуків на мові Python у

середовищі Jupyter Notebook із застосуванням провідних бібліотек машинного навчання та статистичного моделювання.

Використано бібліотеку scikit-learn для інженерії ознак, нормалізації, валідації з часовим поділом та обчислення метрик оцінювання; TensorFlow/Keras для побудови, компіляції та навчання архітектур глибокого навчання з підтримкою GPU-прискорення; statsmodels для реалізації моделей ARIMA/SARIMA з тестами стаціонарності, аналізом автокореляційних функцій, відбором порядків за інформаційними критеріями та діагностикою залишків; pandas для обробки табличних даних, роботи з часовими рядами, агрегації та трансформації; matplotlib та plotly для статичної та інтерактивної візуалізації результатів з можливістю масштабування, відображення довірчих інтервалів та експорту графіків у векторних форматах. Запропонована модульна архітектура забезпечує гнучкість при адаптації до різних епідемічних сценаріїв, типів інфекційних захворювань та специфікацій завдань прогнозування, дозволяючи замінювати окремі компоненти без повного переопрацювання системи.

Створено та експериментально випробувано програмно-інформаційне забезпечення для автоматизації запропонованих методів прогнозування епідемічної захворюваності. Реалізація виконана на мові Python у середовищі Jupyter Notebook, що забезпечує відтворюваність обчислювальних експериментів через збереження повного коду, вхідних даних, проміжних результатів та візуалізацій у єдиному інтерактивному документі. Проведено всебічну валідацію розробленого програмного забезпечення на реальних даних епідемічного нагляду в Україні за період 2015–2024 років, що охоплює часові ряди захворюваності на гострі респіраторні вірусні інфекції, грип, лептоспіроз та вірусні гепатити з різною частотою спостережень та тривалістю.

Продемонстровано ефективність запропонованої технології для короткострокового прогнозування із підвищенням точності порівняно з базовими підходами на основі окремих моделей. Порівняльний аналіз показав,

що запропоновані методи на основі LSTM та ARIMA забезпечують підвищення точності прогнозування як для стабільних періодів епідемічного нагляду, так і для умов структурних зламів під час військового конфлікту. Результати підтверджують практичну застосовність розробленої інформаційної технології для підтримки прийняття рішень у системі громадського здоров'я України в умовах обмежених обчислювальних та даних ресурсів.

Україна, як країна, що протягом останніх років переживає повномасштабний військовий конфлікт та надзвичайні події національного масштабу, стикається з унікальними та безпрецедентними викликами у сфері епідемічного спостереження та прогнозування інфекційних захворювань. Порушення функціонування систем епідемічного нагляду внаслідок руйнування медичної інфраструктури у прифронтових та окупованих регіонах, масова міграція населення як усередині країни, так і за її межі з неможливістю відстеження контактів та спостереження за мігруючими когортами, критичне зниження охоплення планової імунізації дітей та дорослих через порушення холодних ланцюгів постачання вакцин та евакуацію населення з районів активних бойових дій, а також критичне перевантаження медичної інфраструктури через одночасну необхідність надання допомоги військовим пораненим та цивільному населенню створюють сприятливі умови для зростання захворюваності на керовані інфекції та виникнення спалахів захворювань, які раніше вважалися контрольованими.

У таких екстремальних умовах методи машинного навчання, здатні функціонувати при неповних, запізнілих та частково недостовірних даних, ефективно інтегрувати гетерогенні екзогенні змінні з різних джерел та швидко адаптуватися до раптових структурних зламів епідемічної ситуації без повного перенавчання моделей, набувають критичної важливості для національної системи охорони здоров'я як інструмент підтримки прийняття оперативних рішень в умовах високої невизначеності.

Розроблена інформаційна технологія забезпечує практичну застосовність в умовах обмежених обчислювальних ресурсів (може бути розгорнута та виконана на стандартних ноутбуках без спеціалізованих GPU-прискорювачів) та даних ресурсів (демонструє стабільну продуктивність при навчанні на коротких рядах від 72 місяців), підтверджує свою надійність на реальних даних українського епідемічного нагляду з періодами структурних зламів та є актуальною для підтримки функціонування системи громадського здоров'я України у кризових умовах воєнного часу.

Результати дисертаційного дослідження мають потенціал застосування за межами епідемічного домену. Розроблену інформаційну технологію можна, після відповідного адаптування, застосовувати для прогнозування інших типів нестационарних часових рядів багатохвильової природи техногенного або природного характеру, включаючи енергетичне споживання, економічні індикатори, екологічні параметри довкілля, що значно розширює сферу її практичного використання та наукову цінність запропонованих методологічних рішень.



СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

- 1) Hethcote, H.W. (2000). The Mathematics of Infectious Diseases. *SIAM Rev.*, [online] 42(4), pp.599–653. doi: 10.1137/s0036144500371907.
- 2) Keeling, M.J. and Rohani, P. (2008). *Modeling Infectious Diseases in Humans and Animals. Princeton University Press eBooks*. Princeton University Press. doi: 10.1515/9781400841035.
- 3) Bernoulli, D. and Blower, S. (2004). An attempt at a new analysis of the mortality caused by smallpox and of the advantages of inoculation to prevent it. *Reviews in Medical Virology*, [online] 14(5), pp.275–288. doi: 10.1002/rmv.443.
- 4) Kermack, W.O. and McKendrick, A.G. (1927). A Contribution to the Mathematical Theory of Epidemics. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, [online] 115(772), pp.700–721. doi: 10.1098/rspa.1927.0118.
- 5) Anderson, R.M. and May, R.M. (1991). *Infectious Diseases of Humans*. [online] *Oxford University Press eBooks*. Oxford University Press. doi: 10.1093/oso/9780198545996.001.0001.
- 6) Brauer, F., Castillo-Chavez, C. and Feng, Z. (2019). *Mathematical Models in Epidemiology. Texts in Applied Mathematics*. New York, NY: Springer New York. doi: 10.1007/978-1-4939-9828-9.
- 7) Li, M.Y. (2018). *An Introduction to Mathematical Modeling of Infectious Diseases*. [online] Cham: Springer International Publishing. doi: 10.1007/978-3-319-72122-4.
- 8) Heffernan, J.M., Smith, R.J. and Wahl, L.M. (2005). Perspectives on the basic reproductive ratio. *Journal of The Royal Society Interface*, 2(4), pp.281–293. doi: 10.1098/rsif.2005.0042.
- 9) Martcheva, M. (2015). *An Introduction to Mathematical Epidemiology. Texts in Applied Mathematics*. Boston, MA: Springer US. doi: 10.1007/978-1-4899-7612-3.

- 10) Britton, T. (2010). Stochastic epidemic models: A survey. *Mathematical Biosciences*, 225(1), pp.24–35. doi: 10.1016/j.mbs.2010.01.006.
- 11) Allen, L.J.S. (2015). *Stochastic Population and Epidemic Models*. Springer eBooks. Springer Nature. doi: 10.1007/978-3-319-21554-9.
- 12) Box, G.E.P., Jenkins, G.M. and Reinsel, G.C. (2008). Time Series Analysis. *Wiley Series in Probability and Statistics*. [online] doi: 10.1002/9781118619193.
- 13) Hyndman, R.J. and Khandakar, Y. (2008). Automatic Time Series Forecasting: the Forecast Package for R. *Journal of Statistical Software*, 27(3), pp.1–22. doi: 10.18637/jss.v027.i03.
- 14) Hyndman, R.J. and Athanasopoulos, G. (2021). *Forecasting: Principles and Practice*. [online] *otexts.com*. OTexts. Available at: <https://otexts.com/fpp3/>.
- 15) Hyndman, R.J. and Koehler, A.B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22(4), pp.679–688. doi: 10.1016/j.ijforecast.2006.03.001.
- 16) Hochreiter, S. and Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, [online] 9(8), pp.1735–1780. doi: 10.1162/neco.1997.9.8.1735.
- 17) Chimmula, V.K.R. and Zhang, L. (2020). Time series forecasting of COVID-19 transmission in Canada using LSTM networks. *Chaos, Solitons & Fractals*, 135, p.109864. doi: 10.1016/j.chaos.2020.109864.
- 18) Butkevych, M., Chumachenko, T. and Chumachenko, D. (2024). Forecasting of Acute Upper Respiratory Tract Infections based on LSTM Model. *2024 14th International Conference on Dependable Systems, Services and Technologies (DESSERT)*, pp.1–5. doi: 10.1109/dessert65323.2024.11122125.
- 19) Butkevych, M., Yakovlev, S. and Chumachenko, D. (2025). Data-Driven Forecasting of Acute and Chronic Hepatitis B in Ukraine with Recurrent Neural Networks. *Applied Sciences*, 15(13), pp.7573–7573. doi: 10.3390/app15137573.
- 20) Zhang, R., Mi, H., He, T., Ren, S., Zhang, R., Xu, L., Wang, M. and Su, C. (2024). Trends and multi-model prediction of hepatitis B incidence in Xiamen.

Infectious Disease Modelling, 9(4), pp.1276–1288. doi: 10.1016/j.idm.2024.08.001.

- 21) Zhang, G.P. (2003). Time Series Forecasting Using a Hybrid ARIMA and Neural Network Model. *Neurocomputing*, 50, pp.159–175. doi: 10.1016/s0925-2312(01)00702-0.
- 22) Vynnycky, E. and White, R. (2010). *An Introduction to Infectious Disease Modelling*. Oxford University Press. ISBN 978-0-19-856576-5.
- 23) Diekmann, O., Heesterbeek, J.A.P. and Metz, J.A.J. (1990). On the definition and the computation of the basic reproduction ratio R_0 in models for infectious diseases in heterogeneous populations. *Journal of Mathematical Biology*, 28(4). doi: 10.1007/bf00178324.
- 24) van den Driessche, P. and Watmough, J. (2002). Reproduction numbers and sub-threshold endemic equilibria for compartmental models of disease transmission. *Mathematical Biosciences*, [online] 180(1-2), pp.29–48. doi: 10.1016/S0025-5564(02)00108-6.
- 25) Mossong, J. et al. (2008). *Social contacts and mixing patterns relevant to the spread of infectious diseases*. [online] 10.1371/journal.pmed.0050074. Available at:
<https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.0050074>
.
- 26) Prem, K., Cook, A.R. and Jit, M. (2017). Projecting social contact matrices in 152 countries using contact surveys and demographic data. *PLOS Computational Biology*, 13(9), p.e1005697. doi: 10.1371/journal.pcbi.1005697.
- 27) Balcan, D., Colizza, V., Goncalves, B., Hu, H., Ramasco, J.J. and Vespignani, A. (2009). Multiscale mobility networks and the spatial spreading of infectious diseases. *Proceedings of the National Academy of Sciences*, 106(51), pp.21484–21489. doi: 10.1073/pnas.0906910106.

- 28) Butkevych, M., Meniailov, I., Bazilevych, K., Parfeniuk, Y. and Chumachenko, D. (2024). *Simulation of influenza dynamics with LSTM deep learning model* *. [online] Available at: <https://ceur-ws.org/Vol-3892/paper9.pdf>.
- 29) Butkevych, M., Meniailov, I., Lutsiv, Y., Kraivskyi, B. and Chumachenko, D. (2025). War-driven displacement and COVID-19 in Poland: simulation study using LSTM model. *Radioelectronic and Computer Systems*, 2025(4), pp.25–34. doi: 10.32620/reks.2025.4.02.
- 30) Chumachenko, D., Bazilevych, K., Butkevych, M., Meniailov, I., Parfeniuk, Y., Sidenko, I. and Chumachenko, T. (2024). Methodology for assessing the impact of emergencies on the spread of infectious diseases. *Radioelectronic and computer systems*, 2024(3), pp.6–26. doi: 10.32620/reks.2024.3.01.
- 31) Butkevych, M. and Chumachenko, D. (2025a). Simulation of Acute Hepatitis a in Ukraine Using LSTM Deep Learning Model. *5th International Conference on Advanced Computer Information Technologies* , pp.88–91. doi: 10.1109/acit65614.2025.11185739.
- 32) Butkevych, M., Manakova, O. and Chumachenko, D. (2024). Forecasting of Salmonellosis Dynamics with LSTM Deep Learning Model. *2024 IEEE 18th International Conference on the Experience of Designing and Application of CAD Systems (TCSET)*, pp.213–216. doi: 10.1109/tcset64720.2024.10755553.
- 33) Butkevych, M. and Chumachenko, D. (2024b). Time series analysis of leptospirosis incidence for forecasting in the Baltic countries using the ARIMA model. *Radioelectronic and Computer Systems*, 2024(4), pp.5–19. doi: 10.32620/reks.2024.4.01.
- 34) Butkevych, M. and Chumachenko, D. (2024a). *Simulation of Influenza Dynamics in Kharkiv Oblast (Ukraine) using ARIMA Model*. [online] CEUR Workshop Proceedings. 2024. Vol. 3777. Available at: <https://ceur-ws.org/Vol-3777/short7.pdf>.
- 35) Hewson, P. (2014). Bayesian Data Analysis 3rd edn A.Gelman, J. B.Carlin, H. S.Stern, D. B.Dunson, A.Vehtari and D. B.Rubin, 2013 Boca Raton, Chapman

- and Hall-CRC 676 pp., ISBN 1-439-84095-4. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 178(1), pp.301–301. doi: 10.1111/j.1467-985x.2014.12096_1.x.
- 36) Doucet, A. and Johansen, A. (2009). *A Tutorial on Particle Filtering and Smoothing: Fifteen years later*. [online] pp.656–704. Available at: https://www.stats.ox.ac.uk/~doucet/doucet_johansen_tutorialPF2011.pdf.
- 37) Marino, S., Hogue, I.B., Ray, C.J. and Kirschner, D.E. (2008). A methodology for performing global uncertainty and sensitivity analysis in systems biology. *Journal of Theoretical Biology*, 254(1), pp.178–196. doi: 10.1016/j.jtbi.2008.04.011.
- 38) Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., Saisana, M. and Tarantola, S. (2007). *Global Sensitivity Analysis. The Primer*. doi: 10.1002/9780470725184.
- 39) Gillespie, D.T. (1977). Exact stochastic simulation of coupled chemical reactions. *The Journal of Physical Chemistry*, 81(25), pp.2340–2361. doi: 10.1021/j100540a008.
- 40) Allen, L.J.S. and Lahodny, G.E. (2012). Extinction thresholds in deterministic and stochastic epidemic models. *Journal of Biological Dynamics*, 6(2), pp.590–611. doi: 10.1080/17513758.2012.665502.
- 41) Keeling, M.J. and Ross, J.V. (2007). On methods for studying stochastic disease dynamics. *Journal of The Royal Society Interface*, 5(19), pp.171–181. doi: 10.1098/rsif.2007.1106.
- 42) Shumway, R.H. and Stoffer, D.S. (2017). *Time Series Analysis and Its Applications*. [online] *Springer Texts in Statistics*. Cham: Springer International Publishing. doi: 10.1007/978-3-319-52452-8.
- 43) Boshnakov, G.N. (2016). *Introduction to Time Series Analysis and Forecasting*, 2nd Edition, Wiley Series in Probability and Statistics, by D.C. Montgomery, C.L. Jennings and M. Kulahci (eds). Published by John Wiley and

- Sons, Hoboken, NJ, USA, 2015. Total number of pag. *Journal of Time Series Analysis*, 37(6), pp.864–864. doi: 10.1111/jtsa.12203.
- 44) Benvenuto, D., Giovanetti, M., Vassallo, L., Angeletti, S. and Ciccozzi, M. (2020). Application of the ARIMA model on the COVID-2019 epidemic dataset. *Data in Brief*, 29. doi: 10.1016/j.dib.2020.105340.
- 45) Brockwell, P.J. and Davis, R.A. (2002). An Introduction to Time Series and Forecasting. *Biometrics*, [online] 39(4). doi: 10.1007/978-1-4757-2526-1.
- 46) Kwiatkowski, D., Phillips, P.C.B., Schmidt, P. and Shin, Y. (1992). Testing the null hypothesis of stationarity against the alternative of a unit root. *Journal of Econometrics*, 54(1-3), pp.159–178. doi: 10.1016/0304-4076(92)90104-y.
- 47) Marquez, J. (1995). Time series analysis : James D. Hamilton, 1994, (Princeton University Press, Princeton, NJ), 799 pp., ISBN 0-691-04289-6. *International Journal of Forecasting*, [online] 11(3), pp.494–495. Available at: <https://ideas.repec.org/a/eee/intfor/v11y1995i3p494-495.html>.
- 48) Dickey, D.A. and Fuller, W.A. (1979). Distribution of the Estimators for Autoregressive Time Series with a Unit Root. *Journal of the American Statistical Association*, [online] 74(366), pp.427–431. doi: 10.2307/2286348.
- 49) Graves, A. (2012). *Supervised Sequence Labelling with Recurrent Neural Networks. Studies in Computational Intelligence*, Berlin, Heidelberg: Springer Berlin Heidelberg, p.137. doi: 10.1007/978-3-642-24797-2.
- 50) Lipton, Z.C., Berkowitz, J. and Elkan, C. (2015). A Critical Review of Recurrent Neural Networks for Sequence Learning. *arXiv (Cornell University)*. doi: 10.48550/arxiv.1506.00019.
- 51) Bengio, Y., Simard, P. and Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, [online] 5(2), pp.157–166. doi: 10.1109/72.279181.
- 52) Gers, F.A., Schmidhuber, J. and Cummins, F. (2000). Learning to Forget: Continual Prediction with LSTM. *Neural Computation*, 12(10), pp.2451–2471. doi: 10.1162/089976600300015015.

- 53) Widrow, B. and Lehr, M.A. (1990). 30 years of adaptive neural networks: perceptron, Madaline, and backpropagation. *Proceedings of the IEEE*, 78(9), pp.1550–1560. doi: 10.1109/5.58323.
- 54) Sherstinsky, A. (2020). Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) network. *Physica D: Nonlinear Phenomena*, 404, p.132306. doi: 10.1016/j.physd.2019.132306.
- 55) Kingma, D. and Ba, J. (2014). Adam: A Method for Stochastic Optimization. *Computer Science*. doi: 10.48550/arXiv.1412.6980.
- 56) Koenker, R. and Hallock, K.F. (2001). Quantile Regression. *Journal of Economic Perspectives*, 15(4), pp.143–156. doi: 10.1257/jep.15.4.143.
- 57) Babu, C.N. and Reddy, B.E. (2014). A moving-average filter based hybrid ARIMA–ANN model for forecasting time series data. *Applied Soft Computing*, 23, pp.27–38. doi: 10.1016/j.asoc.2014.05.028.
- 58) Peng, Y., Lei, M., Li, J.-B. and Peng, X.-Y. (2012). A novel hybridization of echo state networks and multiplicative seasonal ARIMA model for mobile communication traffic series forecasting. *Neural Computing and Applications*, 24(3-4), pp.883–890. doi: 10.1007/s00521-012-1291-9.
- 59) Cranmer, K., Brehmer, J. and Louppe, G. (2020). The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences*, 117(48), pp.30055–30062. doi: 10.1073/pnas.1912789117.
- 60) Evensen, G. (2006). *Data Assimilation: The Ensemble Kalman Filter*. *Digital Access to Libraries*, Harris County Public Library, p.307. doi: 10.1007/978-3-642-03711-5.
- 61) Elder, B.D., Dukic, V.M. and Dwyer, G. (2006). Uncertainty in predictions of disease spread and public health responses to bioterrorism and emerging diseases. *Proceedings of the National Academy of Sciences*, [online] 103(42), pp.15693–15697. doi: 10.1073/pnas.0600816103.

- 62) Vuorio, A., Sajantila, A., Kovanen, P.T. and Budowle, B. (2022). Maleficent Comrades: War in Ukraine and COVID-19. *Disaster Medicine and Public Health Preparedness*, 17. doi: 10.1017/dmp.2022.227.
- 63) Ramírez, C. and Durón, R.M. (2022). The Russia-Ukraine war could bring catastrophic public-health challenges beyond COVID-19. *International Journal of Infectious Diseases*, 120, pp.44–45. doi: 10.1016/j.ijid.2022.04.016.
- 64) Spiegel, P.B., Checchi, F., Colombo, S. and Paik, E. (2010). Health-care needs of people affected by conflict: future trends and changing frameworks. *The Lancet*, 375(9711), pp.341–345. doi: 10.1016/s0140-6736(09)61873-0.
- 65) Kannan, A., Chen, R., Akhtar, Z., Sutton, B., Quigley, A., Morris, M.J. and Raina MacIntyre, C. (2024). Use of Open-Source Epidemic Intelligence for Infectious Disease Outbreaks, Ukraine, 2022. *Emerging infectious diseases*, [online] 30(9), pp.1865–1871. doi: 10.3201/eid3009.240082.
- 66) Murray, J.D. (2002). *Mathematical Biology. Interdisciplinary Applied Mathematics*. New York, NY: Springer New York. doi: 10.1007/b98868.
- 67) Side, S., Abdy, M., Arwadi, F. and Sanusi, W. (2021). SEIRI Model analysis using the mathematical graph as a solution for Hepatitis B disease in Makassar. *Journal of Physics: Conference Series*, 1899(1), p.012091. doi: 10.1088/1742-6596/1899/1/012091.
- 68) Fang, K., Cao, L., Fu, Z. and Li, W. (2023). Prediction of reported monthly incidence of hepatitis B in Hainan Province of China based on SARIMA-BPNN model. *Medicine*, 102(41), p.e35054. doi: 10.1097/md.00000000000035054.
- 69) Boivin-Champeaux, C., Velez de Mendizabal, N., Jones, A., Balsitis, S., Schmidt, S., Feigelman, J.S. and Azeredo, F.J. (2025). Disease Progression Mathematical Modeling With a Case Study on Hepatitis B Virus Infection. *CPT: Pharmacometrics & Systems Pharmacology*, 14(3), pp.420–434. doi: 10.1002/psp4.13298.

- 70) Li, Y., Yang, Y., Yang, C. and Zhang, B. (2021). Predicting the Cases of Hepatitis B with the A-LSTM Model. *Journal of Physics Conference Series*, 1995(1). doi: 10.1088/1742-6596/1995/1/012007.
- 71) Xu, C., Wang, Y., Cheng, K., Yang, X., Wang, X., Guo, S., Liu, M. and Liu, X. (2023). A Mathematical Model to Study the Potential Hepatitis B Virus Infections and Effects of Vaccination Strategies in China. *Vaccines*, [online] 11(10), pp.1530–1530. doi: 10.3390/vaccines11101530.
- 72) Ma, J. and Ma, S. (2023). Dynamics of a stochastic hepatitis B virus transmission model with media coverage and a case study of China. *Mathematical Biosciences and Engineering*, 20(2), pp.3070–3098. doi: 10.3934/mbe.2023145.
- 73) Zhao, D., Zhang, H., Cao, Q., Wang, Z. and Zhang, R. (2022). The research of SARIMA model for prediction of hepatitis B in mainland China. *Medicine*, [online] 101(23), p.101. doi: 10.1097/md.00000000000029317.
- 74) de Villiers, M.J., Gamkrelidze, I., Hallett, T.B., Nayagam, S., Razavi, H. and Razavi-Shearer, D. (2020). Modelling hepatitis B virus infection and impact of timely birth dose vaccine: A comparison of two simulation models. *PLOS ONE*, 15(8), p.e0237525. doi: 10.1371/journal.pone.0237525.
- 75) El Koufi, Y. and Rao, N.S. (2022). Stochastic Hybrid Hepatitis B Epidemic Model with Markovian Switching. *Complexity*, 2022(1). doi: 10.1155/2022/2347414.
- 76) Cheng, K., Xu, C., Guo, S. and Zhao, X. (2025). Evaluation of hepatitis B vaccination strategy based on age heterogeneity model. *Journal of Applied Mathematics and Computing*. doi: 10.1007/s12190-025-02475-3.
- 77) Chumachenko, D., Butkevych, M., Chumachenko, T., Kirpich, A., Kirichenko, L. and Matsyi, O., Deep Learning Forecasts of Hepatitis A, B, and C Incidence in War-Torn Ukraine: A Multivariate LSTM Approach, Proceedings of the International Conference Applied Computing and WWW/Internet 2025, Porto, Portugal, 2025, pp. 321-323. doi: 10.33965/ac_icwi_2025_202508p047

- 78) He, Z. and Tao, H. (2018). Epidemiology and ARIMA model of positive-rate of influenza viruses among children in Wuhan, China: A nine-year retrospective study. *International Journal of Infectious Diseases*, 74, pp.61–70. doi: 10.1016/j.ijid.2018.07.003.
- 79) Tsan, Y.-T., Chen, D.-Y., Liu, P.-Y., Kristiani, E., Phuong, L. and Yang, C.-T. (2022). The Prediction of Influenza-like Illness and Respiratory Disease Using LSTM and ARIMA. *International journal of environmental research and public health/International journal of environmental research and public health*, 19(3), pp.1858–1858. doi: 10.3390/ijerph19031858.
- 80) Nsoesie, E.O., Oladeji, O., Abah, A.S.A. and Ndeffo-Mbah, M.L. (2021). Forecasting influenza-like illness trends in Cameroon using Google Search Data. *Scientific Reports*, 11(1). doi: 10.1038/s41598-021-85987-9.
- 81) Petukhova, T., Ojkic, D., McEwen, B., Deardon, R. and Poljak, Z. (2018). Assessment of autoregressive integrated moving average (ARIMA), generalized linear autoregressive moving average (GLARMA), and random forest (RF) time series regression models for predicting influenza A virus frequency in swine in Ontario, Canada. *PLOS ONE*, 13(6). doi: 10.1371/journal.pone.0198313.
- 82) Zheng, Y., Wang, K., Zhang, L. and Wang, L. (2020). Study on the relationship between the incidence of influenza and climate indicators and the prediction of influenza incidence. *Environmental Science and Pollution Research*, 28(1), pp.473–481. doi: 10.1007/s11356-020-10523-7.
- 83) Zhu, G., Li, L., Zheng, Y., Zhang, X. and Zou, H. (2021). Forecasting Influenza Based on Autoregressive Moving Average and Holt-Winters Exponential Smoothing Models. *Journal of Advanced Computational Intelligence and Intelligent Informatics*, 25(1), pp.138–144. doi: 10.20965/jaciii.2021.p0138.
- 84) Alzahrani, S.M. and Guma, F.E. (2024). Improving Seasonal Influenza Forecasting Using Time Series Machine Learning Techniques. *Journal of Information Systems Engineering and Management*, 9(4), p.30195. doi: 10.55267/iadt.07.15132.

- 85) Conceição, K.S., Andrade, M.G. and Louzada, F. (2013). Zero-modified Poisson model: Bayesian approach, influence diagnostics, and an application to a Brazilian leptospirosis notification data. *Biometrical Journal*, 55(5), pp.661–678. doi: 10.1002/bimj.201100175.
- 86) Mukdasai, K., Sabir, Z., Raja, M.A.Z., Sadat, R., Ali, M.R. and Singkibud, P. (2022). A numerical simulation of the fractional order Leptospirosis model using the supervise neural network. *Alexandria Engineering Journal*, 61(12), pp.12431–12441. doi: 10.1016/j.aej.2022.06.013.
- 87) Bhalraj, A., Azmi, A., Mohd, M., Bhalraj, A., Azmi, A. and Mohd, M. (2021). Analytical and Numerical Solutions of Leptospirosis Model. *International Journal of Mathematics and Computer Science*, [online] 16(3), pp.949–961. Available at: <https://future-in-tech.net/16.3/R-Bhalraj-Azmi-Mohd.pdf> [Accessed 19 Feb. 2025].
- 88) Minter, A., Costa, F., Khalil, A., Childs, J.E., Diggle, P.J., Ko, A.I. and Begon, M. (2019). Optimal Control of Rat-Borne Leptospirosis in an Urban Environment. *Frontiers Ecol. Evol*, 7, p.209. doi: 10.3389/fevo.2019.00209.
- 89) Qu, H., Saifullah, S., Khan, J., Khan, A., Rahman, M.U. and Zheng, G. (2022). Dynamics of leptospirosis disease in context of piecewise classical-global and classical-fractional operators. *Fractals*, 30(08). doi: 10.1142/s0218348x22402162.
- 90) Engida, H., Theuri, D.M., Gathungu, D.K., Gachohi, J. and Alemneh, H.T. (2022). A Mathematical Model Analysis for the Transmission Dynamics of Leptospirosis Disease in Human and Rodent Populations. *Comput. Math. Methods Med*, 2022, pp.1–23. doi: 10.1155/2022/1806585.
- 91) Rahmat, F., Zulkafli, Z., Juraiza Ishak, A., Mohd Noor, S.B., Yahaya, H. and Masrani, A. (2020). Exploratory Data Analysis and Artificial Neural Network for Prediction of Leptospirosis Occurrence in Seremban, Malaysia Based on Meteorological Data. *Frontiers in Earth Science*, 8, p.377. doi: 10.3389/feart.2020.00377.

- 92) Athithan, S., Shukla, V.P. and Biradar, S.R. (2015). Voting Rule Based Cellular Automata Epidemic Spread Model for Leptospirosis. *Indian Journal of Science and Technology*, 8(4), p.337. doi: 10.17485/ijst/2015/v8i4/60441.
- 93) Ikram, R., Khan, A., Khan, A., Khan, T. and Zaman, G. (2019). Analytical approximate solution of leptospirosis epidemic model with standard incidence rate. *Computational Methods for Differential Equations*, [online] 7(3), pp.370–382. Available at: https://journals.tabrizu.ac.ir/article_9033.html [Accessed 19 Feb. 2025].
- 94) Gallego, M.C. and Simoy, M.S. (2021). Mathematical modeling of leptospirosis: A dynamic regulated by environmental carrying capacity. *Solitons & Fractals*, 152. doi: 10.1016/j.chaos.2021.111425.
- 95) Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H. and Bengio, Y. (2014). Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. [online] doi: 10.3115/v1/d14-1179.
- 96) Chung, J., Gulcehre, C., Cho, K. and Bengio, Y. (2014). *Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling*. [online] arXiv.org. doi: 10.48550/arXiv.1412.3555.
- 97) Gers, F., Schraudolph, N. and Schmidhuber, J. (2002). Learning Precise Timing with LSTM Recurrent Networks. *Journal of Machine Learning Research*, [online] 3, pp.115–143. Available at: <https://www.jmlr.org/papers/volume3/gers02a/gers02a.pdf>.
- 98) Pascanu, R., Gulcehre, C., Cho, K. and Bengio, Y. (2014). How to Construct Deep Recurrent Neural Networks. *arXiv:1312.6026 [cs, stat]*. [online] Available at: <https://arxiv.org/abs/1312.6026>.
- 99) He, K., Zhang, X., Ren, S. and Sun, J. (2016). Deep Residual Learning for Image Recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, [online] pp.770–778. doi: 10.1109/cvpr.2016.90.

- 100) Zaremba, W., Sutskever, I. and Vinyals, O. (2014). Recurrent Neural Network Regularization. *arXiv (Cornell University)*. doi: 10.48550/arxiv.1409.2329.
- 101) Gal, Y. and Ghahramani, Z. (2015). *A Theoretically Grounded Application of Dropout in Recurrent Neural Networks* pp.1019–1027. [online] arXiv.org. Available at: <https://arxiv.org/abs/1512.05287>.
- 102) Ba, J.L., Kiros, J.R. and Hinton, G.E. (2016). *Layer Normalization*. [online] arXiv. doi: 10.48550/arXiv.1607.06450.
- 103) Wen, R., Torkkola, K., Narayanaswamy, B. and Madeka, D. (2017). A Multi-Horizon Quantile Recurrent Forecaster. *arXiv (Cornell University)*. doi: 10.48550/arxiv.1711.11053.
- 104) Bengio, Y. (2012). Practical recommendations for gradient-based training of deep architectures. *arXiv (Cornell University)*, pp.249–256. doi: 10.48550/arxiv.1206.5533.
- 105) He, K., Zhang, X., Ren, S. and Sun, J. (2015). Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. *2015 IEEE International Conference on Computer Vision (ICCV)*, [online] pp.1026–1034. doi: 10.1109/iccv.2015.123.
- 106) Jozefowicz, R., Zaremba, W. and Sutskever, I. (2015). *An Empirical Exploration of Recurrent Network Architectures*. [online] proceedings.mlr.press. Available at: <https://proceedings.mlr.press/v37/jozefowicz15.html>.
- 107) Karpathy, A., Johnson, J.C. and Fei-Fei, L. (2015). Visualizing and Understanding Recurrent Networks. *arXiv (Cornell University)*. doi: 10.48550/arxiv.1506.02078.
- 108) Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R. and Samek, W. (2015). On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. *PLOS ONE*, 10(7), p.e0130140. doi: 10.1371/journal.pone.0130140.

- 109) Sundararajan, M., Taly, A. and Yan, Q. (2017). *Axiomatic Attribution for Deep Networks*. [online] proceedings.mlr.press. Available at: <https://proceedings.mlr.press/v70/sundararajan17a.html>.
- 110) Ljung, G.M. and Box, G.E.P. (1978). On a measure of lack of fit in time series models. *Biometrika*, 297–303(2), pp.297–303. doi: 10.1093/biomet/65.2.297.
- 111) Breusch, T.S. and Pagan, A.R. (1979). A Simple Test for Heteroscedasticity and Random Coefficient Variation. *Econometrica*, 47(5), pp.1287–1294. doi: 10.2307/1911963.
- 112) Williams, B.M. and Hoel, L.A. (2003). Modeling and Forecasting Vehicular Traffic Flow as a Seasonal ARIMA Process: Theoretical Basis and Empirical Results. *Journal of Transportation Engineering*, 129(6), pp.664–672. doi: 10.1061/(asce)0733-947x(2003)129:6(664).
- 113) Lütkepohl, H. (2005). *New Introduction to Multiple Time Series Analysis*. Berlin, Heidelberg: Springer Berlin Heidelberg, p.764. doi: 10.1007/978-3-540-27752-1.
- 114) Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31(3), pp.307–327. doi: 10.1016/0304-4076(86)90063-1.
- 115) McKenzie, E. (2003). Ch. 16. Discrete variate time series. *Handbook of statistics*, pp.573–606. doi: 10.1016/s0169-7161(03)21018-x.
- 116) Cleveland, R.B., Cleveland, W.S., McRae, J.E. and Terpenning, I. (1990). *STL: A Seasonal-Trend Decomposition Procedure Based on Loess*. 1st ed. [online] Journal of Official Statistics. Statistics Sweden, pp.3–73. Available at: <https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/stl-a-seasonal-trend-decomposition-procedure-based-on-loess.pdf>.
- 117) Heaton, J. (2017). I. Goodfellow, Y. Bengio, and A. Courville: Deep learning. *Genetic Programming and Evolvable Machines*, [online] 19(1-2), pp.305–307. doi: 10.1007/s10710-017-9314-z.

- 118) Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6), pp.716–723. doi: 10.1109/tac.1974.1100705.
- 119) Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, 6(2), pp.461–464. doi: 10.1214/aos/1176344136.
- 120) Hewamalage, H., Ackermann, K. and Bergmeir, C. (2022). Forecast evaluation for data scientists: common pitfalls and best practices. *Data Mining and Knowledge Discovery*, 37, pp.788–832. doi: 10.1007/s10618-022-00894-5.
- 121) Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. and Salakhutdinov, R. (2014). Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, [online] 15(56), pp.1929–1958. Available at: <https://jmlr.org/papers/v15/srivastava14a.html>.
- 122) Prechelt, L. (1998). Early Stopping - But When? *Lecture Notes in Computer Science*, 1524, pp.55–69. doi: 10.1007/3-540-49430-8_3.
- 123) Burns, B., Beda, J., Hightower, K. and Evenson, L. (2022). *Kubernetes: Up and Running, 3rd Edition [Book]*. [online] www.oreilly.com. Available at: <https://www.oreilly.com/library/view/kubernetes-up-and/9781098110192/>.
- 124) Kumar, A. (2016). Architectural Styles and the Design of Network-based Software Architectures. *Academia.edu*. [online] doi: 10361460/3159219/3719046/s200_amit.
- 125) Nishtala, R., Fugal, H., Grimm, S., Kwiatkowski, M., Lee, H., Li, H.C., McElroy, R., Paleczny, M., Peek, D., Saab, P., Stafford, D., Tung, T. and Venkataramani, V. (2013). *Scaling Memcache at Facebook*. [online] www.usenix.org. Available at: <https://www.usenix.org/conference/nsdi13/technical-sessions/presentation/nishtala>.
- 126) Turnbull, J. (2016). *The Art of Monitoring*. [online] Google Books. Available at:

- https://books.google.com/books/about/The_Art_of_Monitoring.html?id=w5QfDAAAQBAJ [Accessed 20 Feb. 2025].
- 127) Newman, S. (2021). *Building Microservices, 2nd Edition [Book]*. [online] www.oreilly.com. Available at: <https://www.oreilly.com/library/view/building-microservices-2nd/9781492034018/>.
- 128) Rosso, J., Lander, R., Brand, A. and Harris, J. (2021). *Production Kubernetes: Building Successful Application Platforms*. [online] Amazon.com. Available at: <https://www.amazon.com/Production-Kubernetes-Successful-Application-Platforms-ebook/dp/B0912M62XT> [Accessed 20 Feb. 2025].
- 129) Gneiting, T. and Raftery, A.E. (2007). Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association*, 102(477), pp.359–378. doi: 10.1198/016214506000001437.
- 130) Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M. and Bouchachia, A. (2014). A survey on concept drift adaptation. *ACM Computing Surveys*, 46(4), pp.1–37. doi: 10.1145/2523813.
- 131) Izonin, I., Tkachenko, R., Yemets, K. and Havryliuk, M. (2024). An interpretable ensemble structure with a non-iterative training algorithm to improve the predictive accuracy of healthcare data analysis. *Scientific Reports*, 14(1). doi: 10.1038/s41598-024-61776-y.
- 132) Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), pp.5–32.
- 133) Hollingsworth, T.D. (2009). Controlling infectious disease outbreaks: Lessons from mathematical modelling. *Journal of Public Health Policy*, 30(3), pp.328–341. doi: 10.1057/jphp.2009.13.
- 134) Haque, U., Bukhari, M.H., Fiedler, N., Wang, S., Korzh, O., Espinoza, J., Ahmad, M., Holovanova, I., Chumachenko, T., Marchak, O., Chumachenko, D., Ulvi, O., Sikder, I., Hubenko, H. and Barrett, E.S. (2024). A Comparison of Ukrainian Hospital Services and Functions Before and During the Russia-Ukraine War. *JAMA Health Forum*, [online] 5(5), p.e240901. doi: 10.1001/jamahealthforum.2024.0901.

- 135) Petakh, P., Tymchyk, V. and Kamyshnyi, O. (2024). Communicable diseases in Ukraine during the period of 2018–2023: Impact of the COVID-19 pandemic and war. *Travel medicine and infectious disease*, 60, pp.102733–102733. doi: 10.1016/j.tmaid.2024.102733.
- 136) Ozerova, M.I., Zhigalov, I.E. and Vershinin, V.V. (2020). Comparison of Document Generation Algorithms Using the Docs-as-Code Approach and Using a Text Editor. *Advances in intelligent systems and computing*, 247–257, pp.315–326. doi: 10.1007/978-3-030-63322-6_25.
- 137) Butkevych, M. and Chumachenko, D. (2025b). Simulation of Acute Upper Respiratory Tract Infections in Kharkiv Oblast (Ukraine): The ARIMA Model Approach. *Lecture notes in networks and systems*, pp.153–161. doi: 10.1007/978-3-031-94845-9_13.
- 138) Lopez Bernal, J., Cummins, S. and Gasparrini, A. (2016). Interrupted time series regression for the evaluation of public health interventions: a tutorial. *International Journal of Epidemiology*, [online] 46(1), pp.348–355. doi: 10.1093/ije/dyw098.
- 139) Tegegne, M.D., Melkam, M., Adane, T. and Getawa, S. (2022). COVID-19 preventive practice and associated factors in Ethiopia: A systematic review and meta-analysis. *Public Health in Practice*, 4, pp.100–329. doi: 10.1016/j.puhip.2022.100329.
- 140) Chumachenko, D., Butkevych, M., Lode, D., Frohme, M., Schmailzl, K.J.G., Nechyporenko, A. (2022). Machine Learning Methods in Predicting Patients with Suspected Myocardial Infarction Based on Short-Time HRV Data. *Sensors*, 22(18), 7033. doi: 10.3390/s22187033.
- 141) Butkevych, M., Vasiuk, K., Trofymova, I., Chumachenko, T. (2026). Forecasting of Acute Upper Respiratory Tract Infection in Ukraine during the War Conditions using LSTM Model. *CEUR Workshop Proceedings*, Vol. 4164, pp.478–486. doi: 10.5281/zenodo.19435654.

- 142) Chumachenko, T., Butkevych, M. (2023). Information technologies for the assessment of the impact of the russian war in Ukraine on public health. *V Міжнародна науково-практична конференція IT-професіоналів і аналітиків комп'ютерних систем «ProfIT Conference 2023»*, С. 19–20.
- 143) Буткевич, М.В., Чумаченко, Д.І. (2021). Інформаційна система збору даних захворюваності на COVID-19. *IV Міжнародна науково-практична конференція IT-професіоналів і аналітиків комп'ютерних систем «ProfIT Conference 2021»*, С. 16–17.



ДОДАТКИ**ДОДАТОК А****СПИСОК ПУБЛІКАЦІЙ ЗДОБУВАЧА ЗА ТЕМОЮ ДИСЕРТАЦІЇ**

- 1) Butkevych, M., Yakovlev, S., Chumachenko, D. (2025). Data-Driven Forecasting of Acute and Chronic Hepatitis B in Ukraine with Recurrent Neural Networks. *Applied Sciences*, 15 (13), 7573. doi: 10.3390/app15137573. (Scopus, Q2)
- 2) Chumachenko, D., Butkevych, M., Lode, D., Frohme, M., Schmailzl, K.J.G., Nechyporenko, A. (2022). Machine Learning Methods in Predicting Patients with Suspected Myocardial Infarction Based on Short-Time HRV Data. *Sensors*, 22(18), p.7033. doi: 10.3390/s22187033. (Scopus, Q1)
- 3) Butkevych, M., Meniailov, I., Lutsiv, Y., Kraivskyi, B., Chumachenko, D. (2025). War-driven displacement and COVID-19 in Poland: simulation study using LSTM model. *Radioelectronic and Computer Systems*, 2025(4), pp.25–34. doi: 10.32620/reks.2025.4.02. (Scopus, Q3, Категорія А)
- 4) Butkevych, M., Chumachenko, D. (2024). Time series analysis of leptospirosis incidence for forecasting in the Baltic countries using the ARIMA model. *Radioelectronic and Computer Systems*, 2024(4), pp.5–19. doi: 10.32620/reks.2024.4.01. (Scopus, Q3, Категорія А)
- 5) Chumachenko, D., Bazilevych, K., Butkevych, M., Meniailov, I., Parfeniuk, Y., Sidenko, I., Chumachenko, T. (2024). Methodology for assessing the impact of emergencies on the spread of infectious diseases. *Radioelectronic and Computer Systems*, 2024(3), pp.6–26. doi: 10.32620/reks.2024.3.01. (Scopus, Q3, Категорія А)
- 6) Butkevych, M., Chumachenko, D. (2025). Simulation of Acute Upper Respiratory Tract Infections in Kharkiv Oblast (Ukraine): The ARIMA Model Approach. In: *Lecture notes in networks and systems*. Vol. 1473,

- Integrated Computer Technologies in Mechanical Engineering - 2024. Springer International Publishing, pp.153–161. doi: 10.1007/978-3-031-94845-9_13. (Scopus)
- 7) Butkevych, M., Vasiuk, K., Trofymova, I., Chumachenko, T. (2026). Forecasting of Acute Upper Respiratory Tract Infection in Ukraine during the War Conditions using LSTM Model. *CEUR Workshop Proceedings*. Vol. 4164. 5th International Workshop of IT-professionals on Artificial Intelligence (ProfIT AI 2025) 2025. pp.478–486. doi: doi.org/10.5281/zenodo.19435654 (Scopus)
 - 8) Butkevych, M., Chumachenko, D. (2024). Simulation of Influenza Dynamics in Kharkiv Oblast (Ukraine) using ARIMA Model. *CEUR Workshop Proceedings*. Vol. 3777. 4th International Workshop of IT-professionals on Artificial Intelligence (ProfIT AI 2024) 2024, pp. 223-237. doi: doi.org/10.5281/zenodo.19413876 (Scopus)
 - 9) Butkevych, M., Meniailov, I., Bazilevych, K., Parfeniuk, Y., Chumachenko, D. (2024). Simulation of influenza dynamics with LSTM deep learning model. *CEUR Workshop Proceedings*. Vol. 3892. 7th International Conference on Informatics & Data-Driven Medicine. pp.115–125. doi: doi.org/10.5281/zenodo.19414265. (Scopus)
 - 10) Butkevych, M., Chumachenko, D. (2025). Simulation of Acute Hepatitis a in Ukraine Using LSTM Deep Learning Model. 2025 15th International Conference on Advanced Computer Information Technologies (ACIT). pp. 88–91. doi: 10.1109/acit65614.2025.11185739. (Scopus)
 - 11) Butkevych, M., Chumachenko, T. and Chumachenko, D. (2024). Forecasting of Acute Upper Respiratory Tract Infections based on LSTM Model. 2024 14th International Conference on Dependable Systems, Services and Technologies (DESSERT). pp.1–5. doi: 10.1109/dessert65323.2024.11122125. (Scopus)

- 12) Butkevych, M., Manakova, O., Chumachenko, D. (2024). Forecasting of Salmonellosis Dynamics with LSTM Deep Learning Model. 2024 IEEE 17th International Conference on Advanced Trends in Radioelectronics, Telecommunications and Computer Engineering, TCSET 2024. pp.213–216. doi: 10.1109/tcset64720.2024.10755553. (Scopus)
- 13) Chumachenko, D., Butkevych, M., Chumachenko, T., Kirpich, A., Kyrychenko, L., Matsyi, O. (2025). Deep Learning Forecasts of Hepatitis A, B, and C incidence in war-torn Ukraine: a Multivariate LSTM Approach. International Conferences on Applied Computing and WWW/Internet 2025. pp.321–323. doi: 10.33965/ac_icwi_2025_202508p047. (Scopus)
- 14) Chumachenko T., Butkevych M. Information technologies for the assessment of the impact of the russian war in Ukraine on public health. V Міжнародна науково-практична конференція ІТ-професіоналів і аналітиків комп'ютерних систем «ProfIT Conference 2023», С. 19-20.
- 15) Буткевич М.В., Чумаченко Д.І. (2021) Інформаційна система збору даних захворюваності на COVID-19. IV Міжнародна науково-практична конференція ІТ-професіоналів і аналітиків комп'ютерних систем «ProfIT Conference 2021», С. 16-17.
- 16) Чумаченко Д.І., Буткевич М.В., Комп'ютерна програма «*Модель динаміки групи на базі ARIMA*», Свідоцтво про реєстрацію авторського права на твір №132251, 20 грудня 2024.
- 17) Чумаченко Д.І., Буткевич М.В., Чумаченко Т.О., Комп'ютерна програма «*Інтелектуальна модель динаміки гострих інфекцій верхніх дихальних шляхів на базі архітектури LSTM*», Свідоцтво про реєстрацію авторського права на твір №132250, 20 грудня 2024.
- 18) Чумаченко Д.І., Буткевич М.В., Комп'ютерна програма «*Інтелектуальна модель динаміки групи на базі архітектури LSTM*», Свідоцтво про реєстрацію авторського права на твір №132249, 20 грудня 2024.

- 19) Чумаченко Д.І., Буткевич М.В., Чумаченко Т.О., *Комп'ютерна програма “Модель динаміки гострих інфекцій верхніх дихальних шляхів на базі ARIMA”*, Свідоцтво про реєстрацію авторського права на твір №131623, 25 листопада 2024.
- 20) Буткевич М.В., Чумаченко Д.І., *Комп'ютерна програма «Нейромережева модель епідемічного процесу COVID-19 з архітектурою LSTM»*, Свідоцтво про реєстрацію авторського права на твір №112762 від 27.04.2022.
- 21) Буткевич М.В., Чумаченко Д.І., *Комп'ютерна програма “Модель Prophet для моделювання динаміки захворюваності на COVID-19 в Україні”*, Свідоцтво про реєстрацію авторського права на твір №112634 від 11.04.2022.
- 22) Буткевич М.В., Чумаченко Д.І., *Комп'ютерна програма “Модель авторегресійного ковзного середнього для прогнозування епідемічного процесу COVID-19”*, Свідоцтво про реєстрацію авторського права на твір №112605 від 7.04.2022.
- 23) Буткевич М.В., Чумаченко Д.І., *Комп'ютерна програма «Інформаційна технологія збору, обробки та аналізу даних захворюваності на COVID-19»*, Свідоцтво про реєстрацію авторського права на твір №110034 від 03.12.2021.



ДОДАТОК Б
ВІДОМОСТІ ПРО АПРОБАЦІЮ РЕЗУЛЬТАТІВ ДИСЕРТАЦІЇ

- 1) IV Міжнародна науково-практична конференція IT-професіоналів і аналітиків комп'ютерних систем «ProfIT Conference 2021» (м. Харків, Україна, 2021) – очна участь
- 2) V Міжнародна науково-практична конференція IT-професіоналів і аналітиків комп'ютерних систем «ProfIT Conference 2023» (м. Харків, Україна, 2023) – дистанційна участь
- 3) 4th International Workshop of IT-professionals on Artificial Intelligence (ProfIT AI 2024) 2024 (м. Кембрідж, Массачусетс, США, 2024) – дистанційна участь
- 4) 7th International Conference on Informatics & Data-Driven Medicine (м. Бірмінгем, Велика Британія, 2024) – дистанційна участь
- 5) 2024 14th International Conference on Dependable Systems, Services and Technologies (DESSERT) (м. Афіни, Греція, 2024) – дистанційна участь
- 6) 2024 IEEE 17th International Conference on Advanced Trends in Radioelectronics, Telecommunications and Computer Engineering (м. Львів – Славське, Україна, 2024) – дистанційна участь
- 7) International Conference “Integrated Computer Technologies in Mechanical Engineering – 2024” (м. Харків, Україна, 2024) – дистанційна участь
- 8) 5th International Workshop of IT-professionals on Artificial Intelligence (ProfIT AI 2025) 2025 (м. Ліверпуль, Велика Британія, 2025) – дистанційна участь
- 9) 15th International Conference on Advanced Computer Information Technologies (ACIT) (м. Сібернік, Хорватія, 2025) – дистанційна участь
- 10) International Conferences on Applied Computing and WWW/Internet 2025 (м. Порту, Португалія) – дистанційна участь

ДОДАТОК В
АКТИ ВПРОВАДЖЕННЯ РЕЗУЛЬТАТІВ ДОСЛІДЖЕНЬ



ЗАТВЕРДЖУЮ

Генеральний директор
ДУ «Харківський ОЦКПХ МОЗ»
Любов МАХОТА

АКТ ПРО ВПРОВАДЖЕННЯ

1. **Найменування пропозиції для впровадження:** матеріали дисертаційного дослідження на здобуття наукового ступеня доктора філософії «Моделі, методи та інформаційна технологія прогнозування часових рядів епідемічних процесів».
2. **Найменування установи:** Державна установа «Харківський обласний центр контролю та профілактики хвороб Міністерства охорони здоров'я України», Помірки, Харків, Харківська область, 61000, Україна.
3. **Ким і коли запропонований:** Буткевич Микола Віталійович, 26.11.2025 р.
4. **Джерело інформації:** матеріали дисертаційного дослідження на здобуття наукового ступеня доктора філософії «Моделі, методи та інформаційна технологія прогнозування часових рядів епідемічних процесів».
5. **Де і коли впроваджено:** впроваджено в роботу Державної установи «Харківський обласний центр контролю та профілактики хвороб Міністерства охорони здоров'я України» в листопаді 2025 року.
6. **Результати застосування методу за період з 01.09.2022 року по 25.11.2025 року –** позитивні.
7. **Ефективність впровадження –** 100 %.
8. **Зауваження, пропозиції –** немає.

Відповідальний за впровадження:

завідувач відділу

епідеміологічного нагляду

та профілактики інфекційних захворювань

Карлова Т.О.

«ЗАТВЕРДЖУЮ»

Директор

КНП «Міська клінічна лікарня №13»
Харківської міської ради

Бобейко А.Є.

(підпис)

2025 р.

АКТ ПРО ВПРОВАДЖЕННЯ

матеріалів дисертаційної роботи Буткевича Миколи Віталійовича
на здобуття наукового ступеня доктора філософії

«Моделі, методи та інформаційна технологія прогнозування часових рядів епідемічних процесів»

1. **Найменування пропозиції для впровадження:** - метод циклічного кодування сезонних ознак для LSTM моделей епідеміологічного прогнозування, заснований на тригонометричному відображенні дискретного часового індексу на одиничне коло у двовимірному просторі за допомогою синусоїдальної та косинусоїдальної функцій від номера місяця.
2. **Найменування установи:** Національний аерокосмічний університет «Харківський авіаційний інститут», вул. Вадима Машька, 17, Харків, Україна.
3. **Автор:** Буткевич Микола Віталійович.
4. **Джерело інформації:** Буткевич М.В., Чумаченко Д.І., *Комп'ютерна програма «Нейромережева модель епідемічного процесу COVID-19 з архітектурою LSTM»*, Свідоцтво про реєстрацію авторського права на твір №112762 від 27.04.2022.
5. **Де і коли впроваджено:** впроваджено в роботу КНП «Міська клінічна лікарня №13» в жовтні 2022 року – листопаді 2025 року.
6. **Результати застосування методу** за період з 01.10.2022 року по 10.11.2024 року – позитивні.
7. **Ефективність впровадження** – 94 %.
8. **Зауваження, пропозиції** – немає.

Цей акт складено для подачі до спеціалізованої вченої ради для захисту дисертації і не є підставою для фінансових розрахунків.

Дата «17» 11 2025 р.

Микола Віталійович Буткевич
Відповідальний за впровадження

А.Є. Бобейко
Підпис



«ЗАТВЕРДЖУЮ»

Директор

КНП «Міська клінічна лікарня №13»
Харківської міської ради

Бобейко А.Є.

(підпис)

2025 р.

АКТ ПРО ВПРОВАДЖЕННЯ

матеріалів дисертаційної роботи Буткевича Миколи Віталійовича
на здобуття наукового ступеня доктора філософії
«Моделі, методи та інформаційна технологія прогнозування часових рядів епідемічних процесів»

1. **Найменування пропозиції для впровадження:** - метод кумулятивного представлення епідеміологічних часових рядів для прогнозування мережами довгої короткочасної пам'яті, заснований на принципі заміни рівневих значень кількості випадків захворювання наростаючим підсумком від початку періоду спостереження з подальшою min-max нормалізацією, параметри якої обчислюються виключно на навчальній вибірці.
2. **Найменування установи:** Національний аерокосмічний університет «Харківський авіаційний інститут», вул. Вадима Манька, 17, Харків, Україна.
3. **Автор:** Буткевич Микола Віталійович.
4. **Джерело інформації:** Butkevych, M., Yakovlev, S., Chumachenko, D. (2025). Data-Driven Forecasting of Acute and Chronic Hepatitis B in Ukraine with Recurrent Neural Networks. *Applied Sciences*, 15 (13), 7573.
5. **Де і коли впроваджено:** впроваджено в роботу КНП «Міська клінічна лікарня №13» в червні 2025 року – листопаді 2025 року.
6. **Результати застосування методу за період з 25.06.2025 року по 15.11.2025 року –** позитивні.
7. **Ефективність впровадження –** 99,15 %.
8. **Зауваження, пропозиції –** немає.

Цей акт складено для подачі до разової спеціалізованої вченої ради для захисту дисертації і не є підставою для фінансових розрахунків.

Дата «17» 11 2025 р.

Микола Віталійович Буткевич
Відповідальний за впровадження

А.Є. Бобейко
Підпис

ЗАТВЕРДЖУЮ:

Проректор з наукової роботи
Харківського національного медичного
університету

Олександр КРИВОШАПКА

**АКТ ПРО ВПРОВАДЖЕННЯ**

1. **Найменування пропозиції для впровадження:** матеріали дисертаційного дослідження на здобуття наукового ступеня доктора філософії «Моделі, методи та інформаційна технологія прогнозування часових рядів епідемічних процесів».
2. **Найменування установи:** Харківський національний медичний університет, пр. Науки, 4, Харків, 61022, Україна.
3. **Ким і коли запропонований:** Буткевич Микола Віталійович, 05.01.2026 р.
4. **Джерело інформації:** матеріали дисертаційного дослідження на здобуття наукового ступеня доктора філософії «Моделі, методи та інформаційна технологія прогнозування часових рядів епідемічних процесів».
5. **Де і коли впроваджено:** впроваджено в роботу кафедри епідеміології Харківського національного медичного університету, в рамках науково-дослідних робіт за проектом «Моделювання та прогнозування поширення інфекцій у воєнних та повоєнних умовах з використанням даних епідеміологічного, поведінкового та геномного спостереження» (фінансування Національною академією наук США (US NAS) через Науково-технологічний центр в Україні (STCU), № 7136) - березень 2024 року - грудень 2025 р.
6. **Результати застосування методу** за період з 01.09.2022 року по 31.12.2025 року - позитивні.
7. **Ефективність впровадження** - 100 %.
8. **Зауваження, пропозиції** - немає.

Відповідальний за впровадження:

Завідувачка кафедри епідеміології
професор

Тетяна ЧУМАЧЕНКО

ЗАТВЕРДЖУЮ:

Проректор з науково-педагогічної
роботи Харківського національного
медичного університету

Олександр КРИВОШАПКА

**АКТ ПРО ВПРОВАДЖЕННЯ**

1. **Найменування пропозиції для впровадження:** матеріали дисертаційного дослідження на здобуття наукового ступеня доктора філософії «Моделі, методи та інформаційна технологія прогнозування часових рядів епідемічних процесів».
2. **Найменування установи:** Харківський національний медичний університет, пр. Науки, 4, Харків, 61022, Україна.
3. **Ким і коли запропонований:** Буткевич Микола Віталійович, 10.01.2026 р.
4. **Джерело інформації:** матеріали дисертаційного дослідження на здобуття наукового ступеня доктора філософії «Моделі, методи та інформаційна технологія прогнозування часових рядів епідемічних процесів».
5. **Де і коли впроваджено:** впроваджено у навчальний процес кафедри епідеміології Харківського національного медичного університету, вересень 2022 року – грудень 2025 р.
6. **Результати застосування методу** за період з 01.09.2022 року по 31.12.2025 року – позитивні.
7. **Ефективність впровадження** – 100 %.
8. **Зауваження, пропозиції** – немає.

Відповідальний за впровадження:

Завідувачка кафедри епідеміології
професор

Тетяна ЧУМАЧЕНКО

«ЗАТВЕРДЖУЮ»
 Директор
 КНП ММР «Мереф'янська центральна районна лікарня»



Чеведа В.М.
 (підпис)
 2025 р.

АКТ ПРО ВПРОВАДЖЕННЯ

матеріалів дисертаційної роботи Буткевича Миколи Віталійовича
 на здобуття наукового ступеня доктора філософії
 «Моделі, методи та інформаційна технологія прогнозування часових рядів епідемічних процесів»

1. **Найменування пропозиції для впровадження:** - метод циклічного кодування сезонних ознак для LSTM моделей епідеміологічного прогнозування, заснований на тригонометричному відображенні дискретного часового індексу на одиничне коло у двовимірному просторі за допомогою синусоїдальної та косинусоїдальної функцій від номера місяця.
2. **Найменування установи:** Національний аерокосмічний університет «Харківський авіаційний інститут», вул. Вадима Манька, 17, Харків, Україна.
3. **Автор:** Буткевич Микола Віталійович.
4. **Джерело інформації:** Буткевич М.В., Чумаченко Д.І., *Комп'ютерна програма «Нейромережева модель епідемічного процесу COVID-19 з архітектурою LSTM»*, Свідоцтво про реєстрацію авторського права на твір №112762 від 27.04.2022.
5. **Де і коли впроваджено:** впроваджено в роботу КНП ММР «Мереф'янська центральна районна лікарня» в жовтні 2022 року – листопаді 2024 року.
6. **Результати застосування методу за період з 01.10.2022 року по 10.11.2024 року** – позитивні.
7. **Ефективність впровадження** – 94 %.
8. **Зауваження, пропозиції** – немає.

Цей акт складено для подачі до спеціалізованої вченої ради для захисту дисертації і не є підставою для фінансових розрахунків.

Дата «01» грудня 2025 р.

Мікало Віталійович Буткевич
 Відповідальний за впровадження

Чеведа В.М.
 Підпис

«ЗАТВЕРДЖУЮ»

Директор
КНП ММР «Мереф'янська центральна районна лікарня»Чеведа В.М.
(підпис)

2025 р.

**АКТ ПРО ВПРОВАДЖЕННЯ**

матеріалів дисертаційної роботи Буткевича Миколи Віталійовича
на здобуття наукового ступеня доктора філософії

«Моделі, методи та інформаційна технологія прогнозування часових рядів епідемічних процесів»

1. **Найменування пропозиції для впровадження:** - метод кумулятивного представлення епідеміологічних часових рядів для прогнозування мережами довгої короткочасної пам'яті, заснований на принципі заміни рівневих значень кількості випадків захворювання наростаючим підсумком від початку періоду спостереження з подальшою min-max нормалізацією, параметри якої обчислюються виключно на навчальній вибірці.
2. **Найменування установи:** Національний аерокосмічний університет «Харківський авіаційний інститут», вул. Вадима Манька, 17, Харків, Україна.
3. **Автор:** Буткевич Микола Віталійович.
4. **Джерело інформації:** Butkevych, M., Yakovlev, S., Chumachenko, D. (2025). Data-Driven Forecasting of Acute and Chronic Hepatitis B in Ukraine with Recurrent Neural Networks. *Applied Sciences*, 15 (13), 7573.
5. **Де і коли впроваджено:** впроваджено в роботу КНП ММР «Мереф'янська центральна районна лікарня» в червні 2025 року – листопаді 2025 року.
6. **Результати застосування методу** за період з 25.06.2025 року по 15.11.2025 року – позитивні.
7. **Ефективність впровадження** – 99,15 %.
8. **Зауваження, пропозиції** – немає.

Цей акт складено для подачі до разової спеціалізованої вченої ради для захисту дисертації і не є підставою для фінансових розрахунків.

Дата «02» грудня 2025 р.

Мікало Віталійович Буткевич

Відповідальний за впровадження

В.М. Чеведа

Підпис