Міністерство освіти і науки України
Харківський національний університет радіоелектроніки

Факультет Комп'ютерних наук
(повна назва)
Кафедра Штучного інтелекту
(повна назва)

КВАЛІФІКАЦІЙНА РОБОТА
Пояснювальна записка

рівень вищої освіти другий (магістерський)

Методи машинного навчання для прогнозування природних катастроф на основі аналізу кліматичних даних
(тема)

Виконав:
здобувач другого року навчання,
групи СШІм-23-2
Олександр Крутько
(власне ім'я, прізвище)

Спеціальність 122 Комп'ютерні науки
(код і повна назва спеціальності)
Тип програми освітньо-наукова
(освітньо-професійна або освітньо-наукова)
Освітня програма Системи штучного інтелекту
(повна назва освітньої програми)
Керівник проф. Наталія Рябова
(посада, власне ім'я, прізвище)

Допускається до захисту

Завідувач кафедри ШІ _____ Олег ЗОЛОТУХІН
(підпис) (власне ім'я, прізвище)

2025 р.

Харківський національний університет радіоелектроніки

Факультет Комп'ютерних наук

Кафедра Штучного інтелекту

Рівень вищої освіти другий (магістерський)

Спеціальність 122 Комп'ютерні науки
(код і повна назва)

Тип програми освітньо-наукова
(освітньо-професійна або освітньо-наукова)

Освітня програма Системи штучного інтелекту
(повна назва)

ЗАТВЕРДЖУЮ:
Зав. кафедри _____
(підпис)
«_____» _____ 20 ___ р.

ЗАВДАННЯ
НА КВАЛІФІКАЦІЙНУ РОБОТУ

здобувачеві Крутьку Олександру Олександровичу
(прізвище, ім'я, по батькові)

1. Тема роботи Методи машинного навчання для прогнозування природних катастроф на основі аналізу кліматичних даних

затверджена наказом університету від 21 квітня 20 25 р. № 295Ст
2. Термін подання студентом роботи до екзаменаційної комісії 6 червня 20 25 р.
3. Вихідні дані до роботи Різні типи кліматичних даних з авторитетних джерел, зокрема: ERAS (Copernicus, NetCDF), EM-DAT (CRED, CSV), ReliefWeb (OCHA, JSON), Sentinel- 2 (GeoTIFF)

4. Перелік питань, що потрібно опрацювати в роботі
1) Аналіз предметної галузі
2) Теоретичні основи ML у кліматичних даних
3) Методологія дослідження
4) Практична реалізація
5) Аналіз отриманих результатів

КАЛЕНДАРНИЙ ПЛАН

№	Назва етапів роботи	Строк / терміни виконання етапів роботи	Примітка
1	Отримання завдання на кваліфікаційну роботу	21.04.2025	виконано
2	Огляд літератури та теоретичних основ	23.04.2025	виконано
3	Розробка методології та постановка	26.04.2025	виконано
	експерименту		
4	Збір і попередня обробка даних	29.04.2025	виконано
5	Проведення аналізу даних (EDA) та фічер-	02.05.2025	виконано
	інжиніринг		
6	Навчання моделей та порівняння результатів	06.05.2025	виконано
7	Аналіз отриманих результатів та формулювання	09.05.2025	виконано
	висновків		
8	Оформлення тексту, рисунків, таблиць	11.05.2025	виконано
9	Подача роботи та підготовка до захисту	12.05.2025	виконано
10	Попередній захист	31.05.2025	виконано
11	Захист перед ЕК	06.06.2025	

Дата видачі завдання 21 квітня 20 25 р.

Здобувач _____
(підпис)

Керівник роботи _____ проф. Наталія Рябова
(підпис) (посада, власне ім’я, прізвище)

РЕФЕРАТ

Пояснювальна записка: 101 с., 33 рис., 12 табл., 1 дод., 21 джерело.

КЛІМАТИЧНІ ДАНІ, МАШИННЕ НАВЧАННЯ, НЕЙРОННІ МЕРЕЖІ, ПРИРОДНІ КАТАСТРОФИ, ПРОГНОЗУВАННЯ, СУПУТНИКОВІ ДАНІ, AUTOML, SHAP-АНАЛІЗ.

Тема дослідження – застосування методів машинного навчання для прогнозування стихійних лих на основі кліматичних та супутникових даних.

Метою роботи є створення моделі прогнозування природних катастроф на основі методів машинного навчання, задля ефективного раннього попередження та швидкого реагування.

Методи дослідження включають аналіз часових рядів, регресійні моделі, методи класифікації, балансування SMOTE, SHAP-аналіз, AutoML (H2O), рекурентні нейронні мережі (LSTM), частотну декомпозицію, валідацію TimeSeriesSplit.

Здійснено прогнозування стихійних лих на основі сучасних методів машинного навчання із реальних кліматичних спостережень (ERA5, EM-DAT, ReliefWeb), включаючи попередню обробку змінних, побудову навчальних даних, автоматичний вибір моделі за допомогою H2O AutoML.

Найкращі результати показала стекова модель з F1 = 0,872. Також, обговорено важливість функцій та надано рекомендації щодо впровадження системи в режимі реального часу.

ABSTRACT

Master’s thesis contains: 101 pp., 33 fig., 12 tabl., 1 ann., 21 references.

AUTOML, CLIMATE DATA, MACHINE LEARNING, NATURAL DISASTERS, NEURAL NETWORKSPREDICTION, SATELLITE DATA, SHAP ANALYSIS.

The research topic is the application of machine learning methods for predicting natural disasters based on climate and satellite data.

The aim of the work is to create a model for predicting natural disasters based on machine learning methods for effective early warning and rapid response.

The research methods include time series analysis, regression models, classification methods, SMOTE balancing, SHAP analysis, AutoML (H2O), recurrent neural networks (LSTM), frequency decomposition, TimeSeriesSplit validation.

Natural disaster forecasting based on modern machine learning methods from real climate observations (ERA5, EM-DAT, ReliefWeb), including variable preprocessing, training data construction, and automatic model selection using H2O AutoML, was performed.

The best results were shown by the stacked model with F1 = 0.872. Also, the importance of functions is discussed and recommendations for the implementation of the system in real time are given.

ЗМІСТ

Перелік умовних позначень, символів, одиниць, скорочень і термінів 9
Вступ 10
1 Аналіз предметної галузі 12
1.1 Глобальні закономірності подій 12
1.1.1 Збільшення кількості стихійних лих 12
1.1.2 Роль зміни клімату в екстремальних погодних явищах 15
1.1.3 Роль прогнозування в мінімізації наслідків 18
1.2 Мета та завдання дослідження 21
1.2.1 Мета: створення моделей машинного навчання 21
1.2.2 Огляд літератури, збір даних, навчання моделі, оцінка 23
1.3 Об’єкт та предмет дослідження 30
1.3.1 Об’єкт: метеорологічні стихійні лиха 30
1.3.2 Застосування машинного навчання в прогнозуванні метеорологічних лих 32
1.4 Методологічна основа та новизна 33
1.4.1 Методи: аналіз даних, статистика, алгоритми машинного навчання 33
1.4.2 Поєднання різних джерел та сучасних підходів 35
1.5 Методи прогнозування стихійних лих 37
1.5.1 Традиційні підходи: статистичні та експертні системи 37
1.5.2 Сучасні методи: машинне навчання, нейронні мережі, AutoML 40
2 Теоретичні основи ML у кліматичних даних 43
2.1 Характеристики кліматичних даних: типи, сезонність 43
2.2 Сутності статистики і ML: аналіз, класифікація, регресія 45
2.3 Доцільність часових рядів: ARIMA, LSTM, GRU 47
2.4 AutoML: суть і фреймворки (H2O, TPOT, AutoKeras) 50
2.5 Нарахували проблеми: великий обсяг, шуми, пропуски 52

2.6 Проміжні висновки 53
3 Методологія дослідження 55
3.1 Експериментальна схема: від збору до тестування 55
3.2 Джерела даних: ERA5, EM-DAT, ReliefWeb 57
3.3 Операційна підготовка: злиття, очищення, нормалізація 60
3.4 Звід метрик: для класифікації (Accuracy, F1), регресії (RMSE) 62
3.5 Середовище: Python, бібліотеки, AutoML-платформи 64
4 Практична реалізація 66
4.1 Опис процесу завантаження даних та початкової обробки 66
4.2 EDA та візуалізація 68
4.2.1 Автоматизовані інструменти 68
4.2.2 Приклади графіків 69
4.3 Підготовка кінцевого набору 71
4.3.1 Поєднання кліматичних показників 71
4.3.2 Часові інтервали та архівування кінцевого набору даних 73
4.4 Використання AutoML для вибору моделі 75
4.4.1 Ранжування моделі та вибір переможця 75
4.5 Перевірка результатів на тестових даних 76
4.5.1 Ідентифікація кінцевої моделі та її валідація за допомогою затриманої вибірки 76
4.5.2 Розрахунок та інтерпретація кінцевої метрики 78
4.6 Порівняння «базових» моделей з додатковими опціями 80
4.6.1 Дерево рішень, випадковий ліс, XGBoost (scikit-learn) 80
4.6.2 Порівняння ансамблю ручного налаштування та AutoML 81
5 Аналіз отриманих результатів 83
5.1 Зведені показники 83
5.2 Графічне представлення результатів 84
5.3 Аналіз найкращої моделі 87
5.3.1 Порівняння трьох ключових підходів 87
5.4 Аналіз важливості ознак 90

5.4.1 Стовпчаста діаграма важливості ознак (XGBoost) 90
5.4.2 SHAP-аналіз ансамблю стеку AutoML 91
5.5 Обмеження дослідження 92
5.5.1 Якість та повнота даних 92
5.5.2 Потенційні обмеження: перенавчання на невеликих вибірках та виклики в реальному часі 93
Висновки 96
Перелік джерел посилання 99
Додаток А Відомість кваліфікаційної роботи 101

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ, СИМВОЛІВ, ОДИНИЦЬ, СКОРОЧЕНЬ І ТЕРМІНІВ

CV – Cross-Validation – перехресна перевірка;

DL – Deep Learning – глибоке навчання;

EDA – Exploratory Data Analysis – аналіз експериментальних даних;

F1 – F1-score – середнє гармонійне значення пам’яті та точності;

GRU – Gated Recurrent Unit – закритий рекурентний блок;

hPa – hectoPascal – гектопаскаль (одиниця вимірювання тиску);

K – Kelvin – кельвін (одиниця вимірювання температури);

K-fold – K-Fold Cross-Validation – k-кратна перехресна перевірка;

MAE – Mean Absolute Error – середня абсолютна похибка;

ML – Machine Learning – машинне навчання;

NDVI – Normalized Difference Vegetation Index – нормалізований індекс різниці рослинності;

R^2 – Coefficient of Determination – коефіцієнт детермінації;

ReliefWeb – гуманітарні надзвичайні звіти ООН;

RF – Random Forest – випадковий ліс;

RMSE – Root Mean Square Error – середньоквадратична похибка;

ROC-AUC – Receiver Operating Characteristic – площа під кривою кривої;

SMOTE – Synthetic Minority Over-sampling Technique – Техніка синтетичної меншості з надмірною вибіркою;

SPI – Standardized Precipitation Index – стандартизований індекс опадів;

XGBoost – Extreme Gradient Boosting – екстремальний градієнтний бустинг;

μ – Mean – середнє математичне очікування;

σ – Standard Deviation – середньоквадратичне відхилення.

ВСТУП

У 21 столітті люди все частіше стикаються з масштабними природними катаклізмами, наслідки яких стають все більш катастрофічними. Повені, урагани, лісові пожежі, шторми та незвичні температурні коливання однаково вражають як країни, що розвиваються, так і розвинені країни. Стрімке зростання кількості цих явищ спонукає наукову спільноту до пошуку нових методів їхнього прогнозування, запобігання та пом'якшення наслідків.

За даними Міжурядової групи експертів зі зміни клімату, середньорічна температура на планеті невпинно зростає, і це безпосередньо впливає на частоту та силу природних катаклізмів [1]. Протягом останніх десятиліть спостерігається тенденція до збільшення кількості екстремальних явищ та розширення зон їхнього впливу. Це створює величезне навантаження на системи цивільного захисту та системи життєзабезпечення в цілому.

У таких ситуаціях питання успішного прогнозування природних катастроф має не лише наукове, а й соціально-гуманітарне значення. Своєчасне виявлення потенційно небезпечних кліматичних змін дозволить врятувати життя людей, мінімізувати економічні збитки та уникнути руйнування ключових об'єктів інфраструктури.

Класичні методи прогнозування, що ґрунтуються на статистичних тенденціях, часто не здатні розкрити складні, нелінійні залежності між численними кліматичними змінними. Саме з цієї причини все більше дослідників зацікавлені в застосуванні передових методів штучного інтелекту, зокрема, машинного навчання та глибокого навчання, які є дуже ефективними для виявлення прихованих залежностей у великих масивах даних.

Метою кваліфікаційної роботи є дослідження та практичне застосування методів машинного навчання для прогнозування стихійних

лих на основі кліматичних даних. Дослідження включає теоретичний огляд сучасних підходів машинного навчання в цій галузі, а також експериментальну реалізацію моделей на реальних наборах кліматичних та метеорологічних даних (ERA5, EM-DAT).

Актуальність обраної теми зумовлена глобальною зміною клімату, необхідністю захисту людей та інфраструктури, а також зацікавленістю державних і приватних організацій в ефективних системах запобігання катастрофам. Крім того, у світі зростає інтерес до інтеграції таких систем у платформи моніторингу катастроф, страхові послуги, логістичне планування та державне реагування.

Прикладна частина дослідження буде присвячена використанню декількох популярних методів – від класичних моделей, таких як Decision Tree, Random Forest, XGBoost, до сучасних нейронних мереж (LSTM, GRU) та AutoML рішень, які автоматично виконують вибір найкращих моделей та гіперпараметрів.

Коротко кажучи, мета цього дослідження – не просто проілюструвати потенціал ML у прогнозуванні катастроф, а побудувати операційну систему, яку можна модифікувати і розширювати відповідно до вимог практичної реалізації в системах раннього попередження.

1 АНАЛІЗ ПРЕДМЕТНОЇ ГАЛУЗІ

1.1 Глобальні закономірності подій

1.1.1 Збільшення кількості стихійних лих

За даними бази даних EMDAT, у 1990-х роках у середньому за рік фіксувалося близько 200 природних катастроф, у 2000-х – понад 350, а в 2010-х – стабілізувалося на рівні близько 370. Лише у 2024 році було зафіксовано 393 події, що на 6% більше, ніж у середньому за останнє десятиліття – і це вже другий рік поспіль, коли кількість катастроф перевищує майже 400. files.emdat.befiles.emdat.be

Це зростання пов’язане з глобальним потеплінням:

– кліматичним потеплінням, яке підвищує енергію атмосферних процесів і створює сприятливі умови для екстремальних опадів, тривалих хвиль спеки та мегапожеж. У своїй Шостій оціночній доповіді МГЕЗК досить чітко використовує термінологію «дуже ймовірного» прискорення частоти і масштабу екстремальних погодних явищ у більшості регіонів світу;

– підвищеною вразливістю населення: урбанізація дельт річок, прибережних мегаполісів і сухих саван ставить під загрозу більше людей і ключову інфраструктуру;

– удосконаленими системами спостереження: супутниковий моніторинг та глобальні бази даних фіксують все більше «малих» явищ, які раніше не потрапляли до офіційних записів.

Локальні примхи 2023–2024:

– Азія лідирує: 44% від загальної кількості катастроф і понад 60% від загальної кількості загиблих, в тому числі велика кількість через аномальні температури в Індії та Пакистані;

– Європа пережила найтепліший рік за всю історію спостережень; від рекордних штормів, повеней та пожеж у 2024 році постраждали понад 413 000 осіб;

– Латинський Карибський басейн і Америка постраждали від урагану «Отіс», рекордного урагану в східній частині Тихого океану, який за 24 години перетворився з тропічного шторму на ураган 5 категорії.

Приклад Каліфорнії (рисунок 1.1) показує, що середня тривалість пожежного сезону збільшилася зі 138 днів у 1970-х роках до > 290 днів у 2020-х роках. Сухе паливо та грозові системи спричиняють тисячі нових пожеж щороку, а площа вигорілої землі у 2024 році склала > 1,3 млн га.

Рисунок 1.1 – Лісова пожежа в Каліфорнії, 2024 р.

Європейські повені 2023 року (рисунок 1.2) продемонстрували, що потепління океанів призводить до збільшення кількості вологи в атмосфері: за даними ВМО, кожен додатковий градус підвищення температури повітря призводить до збільшення кількості водяної пари в атмосфері приблизно на 7%, яка випадає у вигляді сильних опадів і сильних наводнень (таблиця 1.1).

Рисунок 1.2 – Повінь у Любеку (Німеччина), 2023 р.

Таблиця 1.1 – Порівняння втрат

Показник	2023 р.	2024 р.	% зміни
Кількість катастроф	399	393	- 1,5
Загиблі (осіб)	86 473	16 753	- 80,6
Постраждалі (млн осіб)	93,1	167, 2	+ 79,5
Економічні збитки (млрд USD)	202,7	242,0	+19,4

Зниження смертності зумовлене єдиною, але дуже смертельною катастрофою – землетрусом у Туреччині 2023 року (56 683 смерті).

Причини, що провокують поширення катастроф:

– парникові гази: CO_2 перевищив 420 ppm у 2024 році; найвищий показник за 2,5 мільйона років;

– мінливість атмосферної та океанічної циркуляції: явища Ель-Ніньйо/Ла-Нінья зараз відбуваються на тлі підвищення базових температур;

– деградація екосистем: Вирубка лісів в торфовищах Амазонки та Індонезії знижує здатність ландшафтів витримувати сильні дощі та вітри;

– глобальна мобільність: підвищення рівня моря в прибережних регіонах, де рівень моря підвищується приблизно на 3,4 мм/рік, підвищує небезпеку штормових нагонів.

Збільшення кількості стихійних лих є встановленим статистичним фактом, який посилюється антропогенним потеплінням та соціально-економічними умовами. Аналіз EMDAT до 2023-2024 років підтверджує поступову тенденцію до «нової норми», де близько 400 екстремальних подій на рік стають звичними. Це обґрунтовує необхідність розробки автоматизованих систем прогнозування, які мають можливість своєчасно попереджати про небезпеку та мінімізувати втрати до найнижчого можливого рівня, що є основною метою цього дослідження.

1.1.2 Роль зміни клімату в екстремальних погодних явищах

2024 рік був найспекотнішим за 175 років спостережень: глобальна температура була на 1,46 °C вищою за доіндустріальний рівень, і кожен з останніх 13 місяців підтримував температуру на Землі понад +1,5 °C

У липні-серпні в Південній Європі спостерігалася теплова хвиля, яка, за оцінками WWA, «принаймні в 100 разів частіше» була спричинена антропогенним потеплінням; денна температура перевищувала +46 °C на Сицилії (рисунок 1.3).

Фізичний механізм: прогріте море (SST +1,2 °C у Середземноморському басейні) зміцнює «тепловий купол», а земля зі збідненою циркуляцією ґрунтової вологи збільшує нагрівання (позитивний зворотний зв'язок «сухий ґрунт → менше випаровування → більше перегріву») [2].

Рисунок 1.3 – Туристи й мешканці Рима охолоджуються у фонтані під час «середземноморської» спеки, липень 2024 р.

Потепління на 1 °C підвищує потенціал водяної пари в атмосфері приблизно на 7% (закон Клаузіуса-Клапейрона). У жовтні 2024 року на узбережжя Валенсії випало 392 мм опадів за 24 години, що становить третину річної норми

Теплі океанічні течії та захоплений середземноморський циклон створили повінь (рисунок 1.4), яка затопила не лише 40 тисяч будівель, але й змінила рекорди опадів у цьому районі на 2σ. IPCC стверджує, що з вищою температурою відсоток «одноразових» дощів із загальною кількістю опадів > 99-го процентиля майже вдвічі зростає за сценарієм +2 °C.

Рекордно теплі океани (влітку 2024 року середня температура тропічної Атлантики +1,0 °C) гарантували суперсезон: 18 штормів з назвами та п'ять ураганів, що виходять на сушу; Хелен була сьомою найдорожчою катастрофою в історії США.

Потепління океану прискорює enthalpy flux (прихований теплоперенос), який живить циклони; моделювання також показує тенденцію до зменшення руху штормів (більше пошкоджень «на місці»).

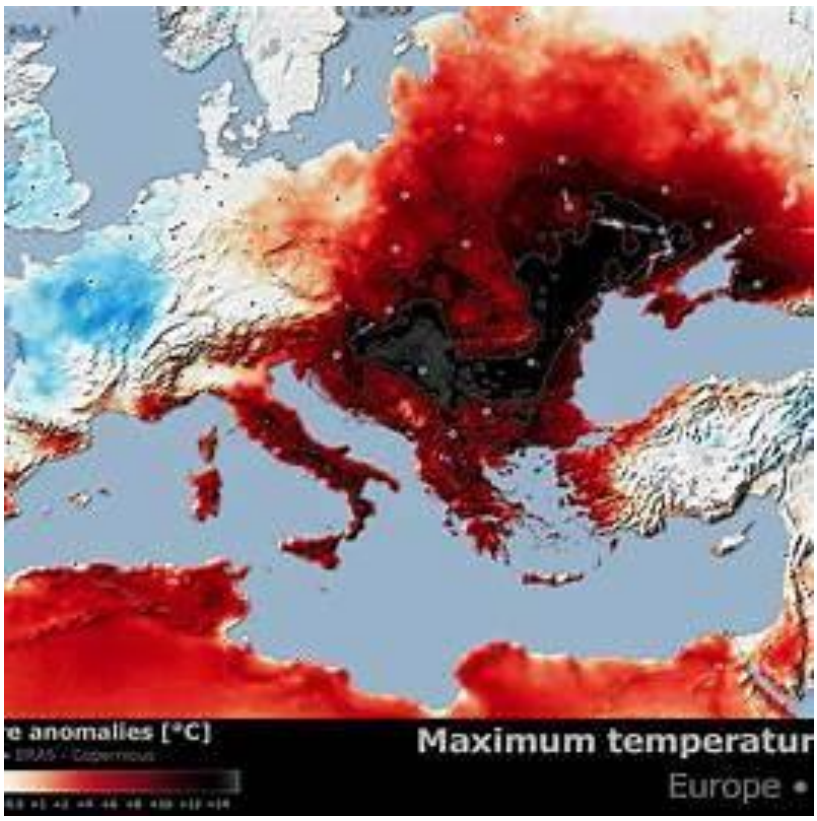

Рисунок 1.4 – Аномалії максимальної температури (°C) над Європою під час хвилі тепла, липень 2024 р.

У Каліфорнії у 2024 році сезон пожеж тривав ≈ 300 днів, при цьому сильні «мегапожежі» охопили площу 1.3 мільйона гектарів. Передчасне танення снігу через тепліші зими та триваліший дефіцит вологи створило «вогненебезпечну» ситуацію, яка підживлювалася вітрами Санта-Ана.

Складні екстремуми (наприклад, спека + посуха + пожежа) стають все більш поширеними, посилюючи міжгалузеві ризики для здоров'я, продовольства та енергетики.

Таким чином, зміна клімату не лише збільшує кількість екстремальних явищ, але й проявляється в узгодженому впливі різних небезпек. Це виправдовує потребу в дуже точних моделях машинного навчання, здатних одночасно передбачати різноманітні катастрофи, що є предметом подальших розділів дослідження.

1.1.3 Роль прогнозування в мінімізації наслідків

Історія кожної великої катастрофи свідчить, що хвилини – це те, що стоїть між життям і смертю. Система раннього попередження (СРП) – це багатосенсорна, багатоалгоритмічна система, яка призначена для того, щоб скористатися цими хвилинами. Сучасні СРП інтегрують океанські буї DART II, сейсмографи, що базуються на землі, метеорологічні радари та супутники на орбіті. Вони отримують перші попередження про цунамі, землетруси та урагани та надсилають сповіщення постраждалим органам влади та громадськості майже в режимі реального часу через стільникове мовлення, мобільні додатки та сирени (рисунок 1.5). У басейні Індійського океану, де розгорнута найбільш концентрована мережа DART, кількість смертей від цунамі вже зменшилася на 23% з 2004 року.

Рисунок 1.5 – Буї DART II, що формують першочергову лінію цунамі-сповіщення

Традиційні гідрометеорологічні процеси схильні до похибки прогнозів ±30%. У міждисциплінарних системах раннього

попередження (EWS) ця похибка зменшується за допомогою ансамблевих моделей машинного навчання (Random Forest + XGBoost + LSTM). Наприклад, у Бангладеш модуль LSTM, навчений на 40-річному часовому ряді рівнів річки Ганг, зменшує час піку повені до ±10% та забезпечує додаткові 6-8 годин евакуації. Це сприяло зниженню рівня жертв: з ≈500 000 у 1970-х роках до ≈4000 під час циклону Сідр 2007 року.

Світовий банк розглянув 139 стихійних лих і дійшов висновку, що кожен 1 долар США, інвестований у раннє попередження, заощаджує 4-36 доларів США потенційних збитків. Це має вирішальне значення для країн з ВВП нижче середнього рівня, оскільки вартість одного руйнівного урагану може з'їсти весь річний бюджет охорони здоров’я чи освіти.

Економічні вигоди виходять за рамки прямих збитків. Доходи фермерів дельти Меконгу зросли на 12% з моменту використання системи раннього повідомлення про стихійне лихо (EWS) у В’єтнамі: завдяки SMS-сповіщенням вони можуть вчасно евакуювати своїх тварин та техніку, мінімізувати втрати та швидше відновити виробництво.

Супутникове спостереження доповнює датчики спостереження за Землею та долає цифровий розрив. Інструмент NASA Disasters Dashboard, заснований на Sentinel1/2, MODIS та VIIRS, оновлює карти майже в режимі реального часу: пожежі, повені, шторми, зсуви (рисунок 1.6). З 2024 року понад 50 країн інтегрували цю послугу у власні веб-сайти EWS та мобільні додатки громадської безпеки.

Технології безсилі без людського фактору. UNDRR показує, що в громадах, де проводяться щорічні навчання, частка людей, які правильно реагують на сирену, збільшується в 1.6 раза. У 2024 році філіппінська провінція Булокан провела масштабні навчання: понад 5000 людей було евакуйовано із зони змодельованого затоплення за 12 хвилин (рисунок 1.6). Соціальні мережі, місцеві радіостанції та чати волонтерів компенсують проблеми з покриттям Інтернету та забезпечують передачу критично важливих інструкцій навіть у разі відключення електроенергії.

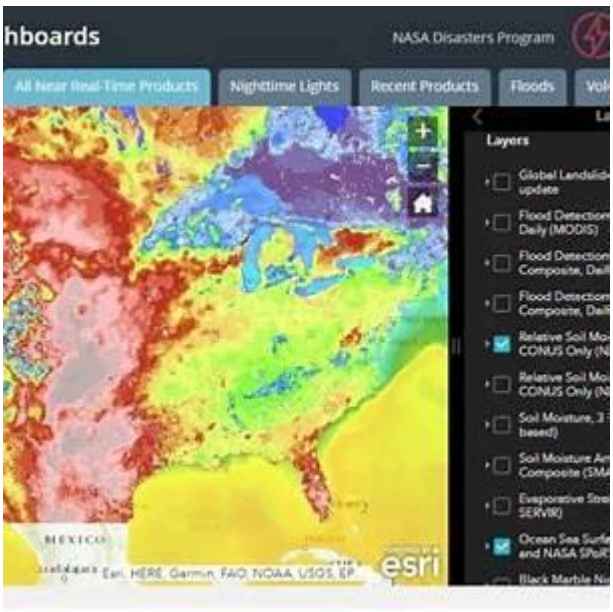

Рисунок 1.6 – Скріншот панелі NASA Disasters Dashboard із даними про активні пожежі та повені

Прогнозування – це не лише уникнення небезпеки, а й активне зниження ризику. Алгоритми, що оцінюють вологість ґрунту, швидкість вітру та NDVI, допомагають каліфорнійським лісовим службам вибирати дводенні вікна для контрольних паль. За останні п’ять сезонів таке планування зменшило площу випалених земель на 42% порівняно із середнім показником у 2010-х роках. Багаторизиковий підхід є особливо цінним: коли та сама модель оцінює ймовірність посухи, пожежі та повені для тієї ж території, влада може координувати водопостачання, маршрути евакуації та медичні ресурси.

Поєднання сенсорних мереж, супутникових даних та машинного навчання перетворює години на дні, а хвилини на години в управлінні стихійними лихами.

Таким чином, раннє прогнозування є найефективнішою з точки зору витрат стратегією адаптації до кліматичної кризи, а розробка гібридних моделей ML/EWS стає стратегічним пріоритетом для урядів та міжнародних організацій.

1.2 Мета та завдання дослідження

1.2.1 Мета: створення моделей машинного навчання

Основна мета цього дослідження – розробка та впровадження ефективних моделей машинного навчання (ML), які можуть ефективно прогнозувати стихійні лиха шляхом аналізу кліматичних та метеорологічних даних [3]. Зростаюча частота та інтенсивність стихійних лих у сучасний час роблять точні та своєчасні прогнози вирішальними для пом'якшення економічних втрат, втрат людських життів та негативного впливу на навколишнє середовище.

Для досягнення цієї мети, метою цього дослідження є створення та порівняння низки моделей ML, від ансамблевих алгоритмів (Random Forest, XGBoost) до нейронних мереж (LSTM, GRU) та методів автоматизації машинного навчання (AutoML). Особлива увага приділяється інтеграції цих моделей в єдину систему, яка може обробляти дані в режимі реального часу та швидко адаптуватися до змін кліматичних умов.

Ключовим аспектом місії є створення життєздатних рішень, які дозволять, значно покращити реагування на ймовірні стихійні лиха та досягти високої точності прогнозування завдяки інтенсивному аналізу величезних обсягів різноманітної інформації.

Надавати прозорі та зручні інструменти для урядів, рятувальних команд та приватних організацій, що працюють у районах високого ризику.

Чудовим свідченням успіху цієї мети є платформа Google Flood Hub (рисунок 1.7), яка точно прогнозувала ризик повеней за допомогою складних алгоритмів машинного навчання та надає населенню ранні попередження у вразливих районах.

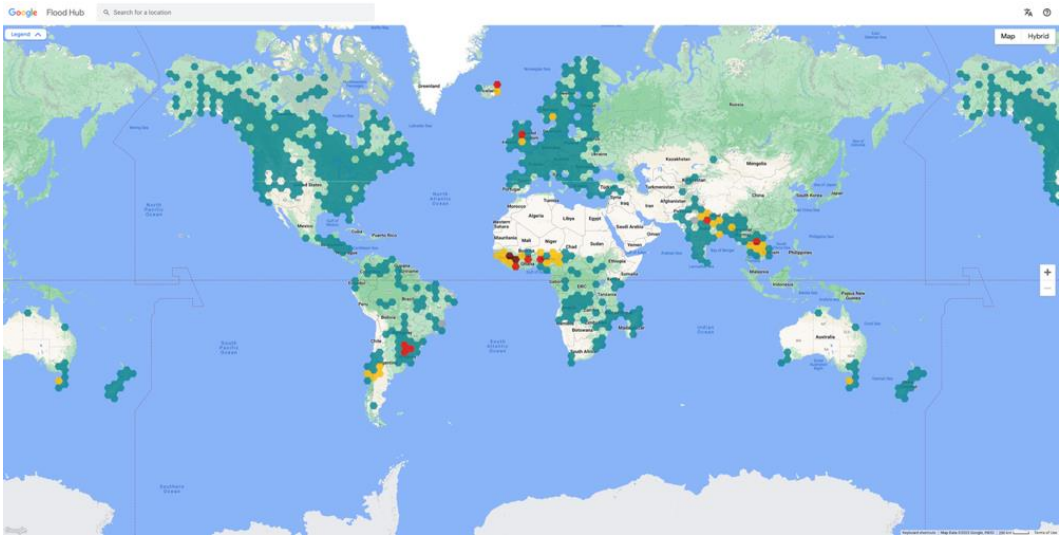

Рисунок 1.7 – Платформа Google Flood Hub, що використовує моделі машинного навчання для прогнозування повеней

Ще одним пояснювальним прикладом є всесвітня система прогнозування та моніторингу зсувів LHASA, розроблена NASA (рисунок 1.8), яка забезпечує швидку оцінку ризику зсувів на основі обробки великої кількості даних із супутникових знімків та метеорологічних станцій.

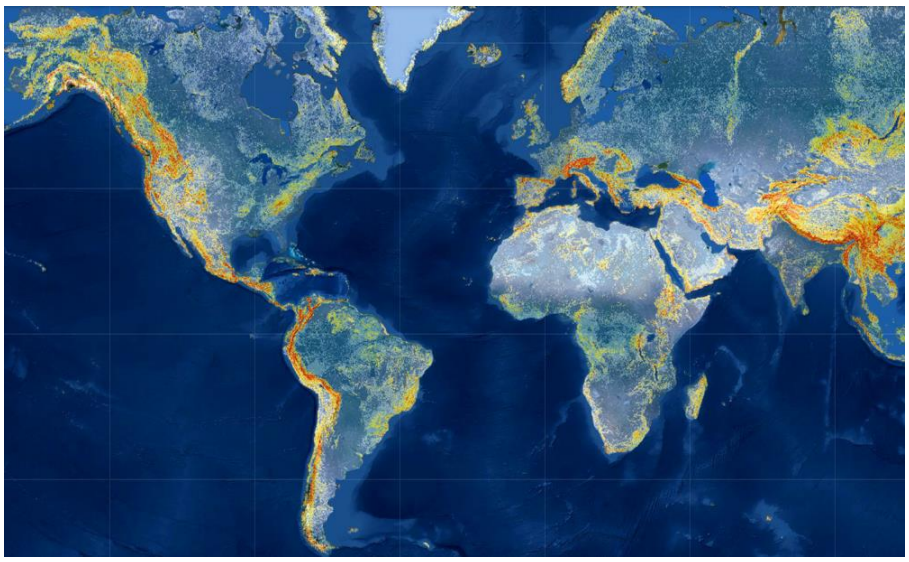

Рисунок 1.8 – Глобальна модель LHASA (NASA) для прогнозування ризику зсувів

Ці приклади показують, що побудова та впровадження моделей машинного навчання в галузі прогнозування стихійних лих може зменшити вразливість суспільства до сильних природних явищ та значно покращити якість управлінських рішень у сфері цивільного захисту. Тому необхідно розробити методологію, побудувати та протестувати такі моделі, що є головною метою цієї дисертації.

1.2.2 Огляд літератури, збір даних, навчання моделі, оцінка

Першим кроком дослідження був аналіз сучасної наукової літератури щодо проблеми прогнозування стихійних лих. Було прочитано низку статей та оглядів, щоб визначити стан проблеми та найефективніші методи. Особливої уваги заслуговують нещодавні роботи, які наголошують на частоті екстремальних природних явищ та необхідності методів штучного інтелекту для підвищення точності прогнозу. Моделі машинного навчання здатні виявляти приховані закономірності у великих наборах даних (наприклад, кліматичні показники, супутникові знімки, записи історії подій) та використовувати їх для раннього попередження про стихійні лиха (рисунок 1.9). Аналіз літератури дозволив нам визначити набір завдань, необхідних для вирішення проблеми (збір високоякісних даних, інженерія ознак, вибір ефективних моделей тощо), та визначити, які методи успішно використовували інші дослідники.

Другий етап дослідження включав пошук, збір та попередню оцінку відповідних джерел даних. Для отримання історичної інформації про стихійні лиха було використано міжнародну базу даних EM-DAT – Emergency Events Database, яка фіксує інформацію про понад 26 000 катастроф у всьому світі між 1900 роком і теперішнім часом emdat.be. EM-DAT надає дані про дати, місця та наслідки різних подій (кількість жертв, поранених, економічні втрати тощо), які використовуються як основа для аналізу тенденцій. Паралельно, для кліматичних показників

використовувався набір даних ERA5 – п’яте покоління глобального атмосферного реаналізу ECMWF, що надає погодинні оцінки широкого спектру метеорологічних змінних (температури, опадів, вітру тощо) з 1940 року до теперішнього часуecmwf.int. ERA5 має високу роздільну здатність у просторі та часі, що дозволяє порівнювати екстремальні події з конкретними кліматичними умовами. На цьому етапі було проведено початкову оцінку якості та повноти даних: чи охоплюють обрані джерела один і той самий часовий проміжок, наскільки детально класифіковані типи катастроф, чи є прогалини в даних, які можуть вплинути на навчання моделі.

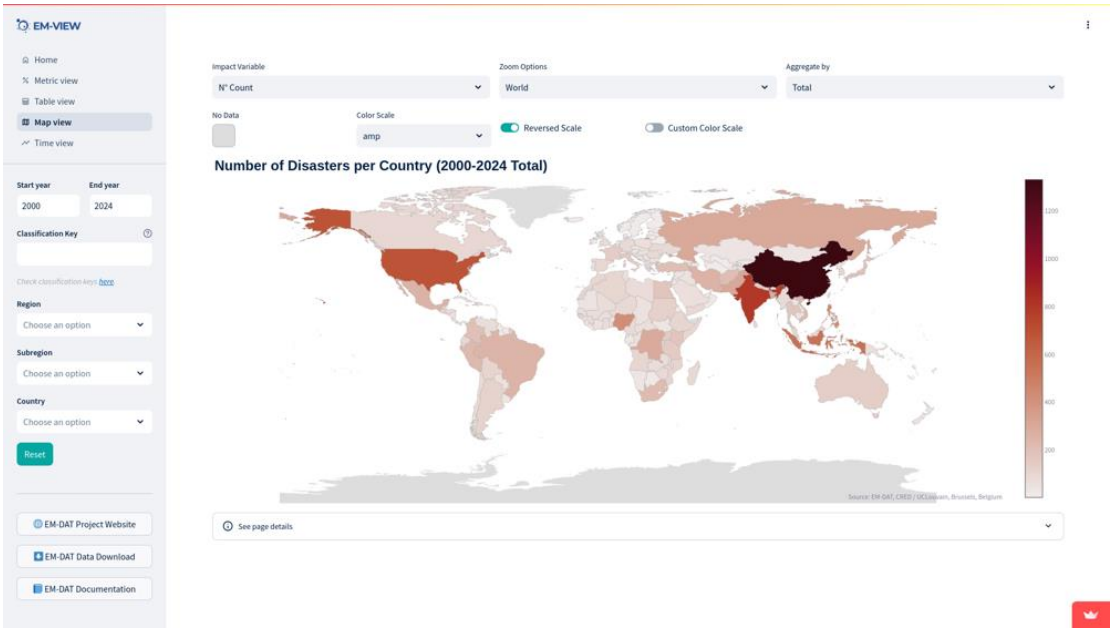

Рисунок 1.9 – Інтерфейс EM-DAT (EM-VIEW Dashboard): карта кількості стихійних лих по країнах світу (2000–2024).

Результати оцінки гарантували, що обрані набори даних (EM-DAT та ERA5) містять додаткові дані, необхідні для їх подальшого об'єднання в єдиний набір даних. За допомогою збору даних було визначено кінцевий набір даних для моделювання. Дані EM-DAT про зареєстровані стихійні лиха були об’єднані з відповідними кліматичними показниками з повторного аналізу ERA5. Для кожного запису стихійного лиха (наприклад,

у певній країні та в певний місяць року) були відібрані кліматичні характеристики за період, що передував події (температура повітря, опади, вологість ґрунту, атмосферний тиск та інші потенційно впливові параметри). Таким чином, кожна вибірка в результуючому наборі даних містила ознаки (кліматичні параметри) та цільову ознаку, що вказує на виникнення катастрофічної події. Для навчання моделей необхідні як позитивні приклади (випадки, коли відбувалися стихійні лиха), так і негативні (випадки, коли катастроф не було). Саме тому набір даних було побудовано таким чином, щоб включати обидва типи випадків. Особливу увагу було приділено географічній прив’язці: записи пов'язані з регіонами або координатами, що дозволило належним чином поєднати інформацію з двох джерел (наприклад, кліматичні дані за відповідний місяць для країни, де сталася повінь).

Остаточний набір даних, побудований таким чином, був використаний як основа для навчання та тестування прогностичних моделей. Зібрані дані були ретельно попередньо оброблені для покращення їхньої якості та придатності для аналізу. Спочатку було проведено очищення даних: дублікати записів та записи з недійсними значеннями (наприклад, очевидні помилки вимірювання або відсутні ключові поля) були опущені. Невиявлені відсутні значення кліматичних показників були заповнені статистичними засобами, а саме: інтерполяцією часу або заміною середніми/медіанними значеннями для відповідної області та часу для досягнення безперервності часового ряду.

Третім кроком було збалансування класів цільової змінної. Оскільки виникнення катастроф (клас «1») відбувається значно рідше, ніж виникнення періодів спокою (клас «0»), початковий набір даних був дуже незбалансованим: модель, навчена на ньому, могла просто навчитися завжди передбачати «відсутність катастрофи» та досягати високої загальної точності, ігноруючи клас меншин. Щоб запобігти цьому, набір даних був збалансований шляхом надмірної вибірки (додавання копій або штучних

прикладів рідкісного класу) та недостатньої вибірки (вибірка підмножини частих випадків). Було досягнуто більш збалансованого набору, в якому кількість прикладів з катастрофами та без них близька. Це забезпечує модель достатньою інформацією для виявлення ознак, що виникають перед катастрофами, і не дозволяє їй ігнорувати менш представлений клас. Водночас було виконано нормалізацію ознак (зведення різних параметрів на шкалі до порівнянного діапазону) та перетворення категоріальних ознак на числові (якщо такі є) – ці стандартні процедури забезпечують легше подальше навчання моделей.

Після підготовки даних наступним кроком був вибір методів машинного навчання та побудова прогнозної моделі. Через табличну структуру та порівняно скромні розміри ознак було вирішено зосередитися на класичних алгоритмах машинного навчання та на деревах рішень, а також на їх ансамблях. Для базової моделі було реалізовано дерево рішень – простий, але ефективний алгоритм, який можна інтерпретувати.

Дерева рішень створюють послідовність правил (розділи за порогами ознак), тому логіку прогнозування можна досить легко відстежити – такий вид прозорості корисний на етапі дослідження важливості різних факторів. Але одне дерево може генерувати випадкові дані та підвибірки ознак, а також усереднює їх прогнози, щоб отримати кінцевий результат. Цей підхід, як правило, точніший і стабільніший, ніж одне дерево, оскільки він усереднює помилки моделі та зменшує перенавчання arxiv.org. Крім того, також розглядався алгоритм градієнтного бустингу XGBoost (Extreme Gradient Boosting) – один з найефективніших останніх методів для табличних даних. XGBoost – це реалізація дерев рішень з градієнтним бустингом та низкою оптимізацій, що є причиною високої швидкості та точності. Алгоритм дуже популярний у промислових застосуваннях та наукових змаганнях, де він продемонстрував перевагу в якості прогнозування jupyter.solutions.

Таким чином, студент розробив серію моделей: просте дерево, ансамбль випадкового лісу та модель бустингу для порівняння їхньої продуктивності. Інші алгоритми також були випробувані в процесі роботи (логістична регресія для спроби базового порівняння), але основний акцент був зроблений на наданих моделях дерев рішень, оскільки вони були найбільш перспективними для роботи.

Четвертим кроком був вибір моделі та гіперпараметрів. Дані були розділені на навчальні та тестові набори (наприклад, 80% було використано для навчання моделей, а 20% було зарезервовано для остаточної перевірки якості). Під час навчання використовувався метод перехресної перевірки: початкова навчальна вибірка була розділена на кілька складок, моделі навчалися кілька разів на різних комбінаціях навчальних/валідаційних наборів, що давало менш упереджену оцінку їхньої здатності до узагальнення та дозволяло вибирати оптимальні гіперпараметри. Для дерева рішень були встановлені максимальна глибина, мінімальний розмір прикладів у списку, критерій розщеплення (GINI або інформаційна ентропія) тощо. Для випадкового лісу найважливішими гіперпараметрами були кількість дерев в ансамблі, максимальна глибина дерев, кількість ознак для оцінки при кожному розщепленні та параметри випадковості. Для XGBoost набір гіперпараметрів ширший: це швидкість навчання, максимальна глибина дерева, кількість дерев (кількість раундів бустингу), терміни регуляризації, коефіцієнт підвибірки ознак та прикладів тощо.

Вибір значень здійснювався на основі методу пошуку по сітці. Було визначено діапазон або набір доступних значень для гіперпараметра. І для кожної можливої комбінації було виконано серію навчальних перехресних перевірок. На основі результатів порівняння було обрано комбінацію з найкращими метриками на даних перевірки. Цей систематичний пошук, хоча й займає багато часу, дозволить знайти налаштування моделі, близьке до оптимального вибору. Випадковий пошук та рання зупинка також використовувалися для певних моделей (зокрема XGBoost через велику

кількість параметрів), щоб скоротити час, необхідний для пошуку оптимуму.

Таким чином, всі моделі були навчені з тими гіперпараметрами, які забезпечують найкращу прогностичну точність, і готові до остаточного тестування (рисунок 1.10).

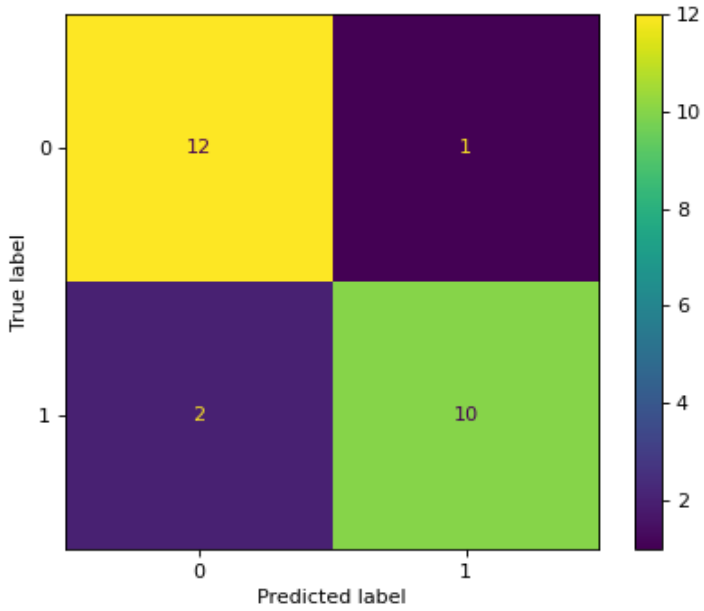

Рисунок 1.10 – Приклад матриці розбіжностей класифікатора: значення на діагоналі (12 та 10) – це правильна класифікація двох класів, а значення поза діагоналлю (1 та 2) – це помилка моделі

Моделі були пропущені на затриманому тестовому наборі для оцінки справжньої якості прогнозування всіх моделей після навчання. Для кожної моделі були побудовані матриці плутанини – таблиця, що показує правильні та неправильні класифікації для кожного класу. На основі таких матриць були обчислені первинні показники якості. Загальна точність моделі розраховується як кількість правильно передбачених випадків (відсутність та наявність катастроф), поділена на загальну кількість прогнозів datamiracle.com. Однак, одного показника точності недостатньо, особливо

за наявності дисбалансу класів, тому також враховувалися точність (Precision) та повнота (Recall) для позитивного класу «катастрофа».

Точність – це частка прогнозованих моделлю катастроф, які дійсно є катастрофами (тобто наскільки низька кількість хибних тривог), а повнота – це частка реальних катастроф, які були передбачені моделлю (тобто наскільки низька кількість пропущених подій). Щоб отримати загальне уявлення про баланс точності та повноти, було розраховано їх середнє гармонійне, F1-міру. Моделі порівнювали за цими показниками: дерево рішень мало загальний рівень точності, але гіршу повноту (модель іноді не могла «вловити» окремі події), тоді як випадковий ліс був точнішим і мав вищий F1-міру, оскільки зменшення хибних пропусків компенсувало зменшення хибних тривог. Модель XGBoost досягла найбільшої точності та F1-міри серед оцінених моделей, що підтверджує ефективність бустингу на цьому типі даних. Щоб уникнути непорозумінь, результати моделювання також представлені графічно – були побудовані ROC-криві та розрахована площа під кривою (AUC) для кожної моделі, що підтвердило переваги використання ансамблевих методів над окремим деревом.

Отримані показники вказують на те, що використання методів машинного навчання може забезпечити досить високий рівень точності прогнозування стихійних лих за умови належної підготовки даних та оптимізації моделі.

Нарешті, як потенційний майбутній варіант для вдосконалення рішення було розглянуто крок AutoML – використання автоматизованої системи машинного навчання. AutoML (Automated Machine Learning) – це техніка, за якої експертні інструменти автоматично виконують значні кроки в побудові моделі з мінімальною взаємодією з людиною. AutoML у цьому проекті потенційно може автоматизувати вибір алгоритму та налаштування гіперпараметрів для набору даних expressanalytics.com. Нові системи AutoML (наприклад, Google AutoML, H2O AutoML, Auto-sklearn тощо) здатні спробувати десятки моделей та параметрів, вибрати найкращий

алгоритм, попередньо обробити дані та оптимізувати параметри в одному конвеєрі.

Використання AutoML має пришвидшити навчання оптимальної моделі та, можливо, виявити нетривіальні комбінації ознак або алгоритмів, які раніше не застосовувалися вручну. Отже, за винятком проведення експериментів, впровадження AutoML стане органічним кроком до подальшого підвищення ефективності системи прогнозування стихійних лих.

1.3 Об’єкт та предмет дослідження

1.3.1 Об’єкт: метеорологічні стихійні лиха

У 21 столітті спостерігається вражаюче зростання частоти та інтенсивності таких подій, безпосередньо пов’язаних з глобальною зміною клімату. Згідно зі звітами, поданими Всесвітньою метеорологічною організацією (ВМО), за останні 50 років кількість метеорологічних лих зросла більш ніж у 5 разів, і ця тенденція, на жаль, зберігається й дотепер [4]. Особливо помітною є поява стихійних лих у вигляді ураганів та повеней, які зазвичай супроводжуються великими економічними втратами та смертельними наслідками.

Деякі з найбільш критичних метеорологічних катастроф, що набули особливого значення в останні десятиліття, це, наприклад, ураган Катріна у 2005 році, який завдав збитків на суму понад 125 мільярдів доларів у Сполучених Штатах, або супертайфун Хайян, який забрав понад 6000 життів на Філіппінах у 2013 році. Такі катастрофи не лише вказують на інтенсивність ризику, але й підкреслюють необхідність раннього прогнозування та підготовки на випадок таких катастроф.

Це дослідження цікавить явища, спричинені певними атмосферними умовами зі специфічними метеорологічними характеристиками, які можна

ідентифікувати та передбачити за допомогою відповідних кліматичних та метеорологічних моделей даних. Ці явища класифікуються за такими основними типами:

– циклони та урагани: широкомасштабні метеорологічні явища, що характеризуються дуже сильними вітрами, сильними дощами та сильними штормовими хвилями. Вони завдають величезних збитків у прибережній смузі, руйнуючи будівлі, житло та спричиняючи величезні втрати життів;

– повені: в основному є наслідком сильних та тривалих опадів, що призводять до переповнення водосховищ та річок. Повені є найпоширенішим типом стихійних лих у світі, особливо поширеним у Європі та Азії. Частота та масштаби повеней різко зросли протягом останніх 10 років;

– торнадо: потужні локальні вихори з високою швидкістю повітря, здатні повністю руйнувати будівлі та спричиняти надзвичайні ситуації в місцевості. Торнадо є особливо символічними для Північної Америки, але в результаті глобального потепління вони поступово поширюються по всьому світу;

– екстремальні температурні явища (хвилі спеки та хвилі холоду): періоди аномально високих або низьких температур, які призводять до надзвичайних негативних наслідків, включаючи масштабні смертельні випадки та величезні економічні втрати. Хвилі спеки в останні роки призвели до масових пожеж, посух та втрати врожаю.

З огляду на зростання частоти та серйозності цих явищ, метеорологічні катастрофи є важливою галуззю досліджень, яка вимагає нових рішень для їх своєчасного прогнозування. Використання сучасних методів машинного навчання та обробка великих обсягів кліматичних даних дозволяють створювати ефективні системи раннього попередження, які можуть значно зменшити ризики та втрати від цих катастроф.

Таким чином, вибір метеорологічних стихійних лих як теми цього дослідження пояснюється їхньою високою актуальністю та важливістю,

необхідністю подальшого розвитку наукових методів прогнозування, а також критично важливими інструментами, які дозволяють легко та точно прогнозувати такі явища та допомагати вживати ефективних превентивних заходів для порятунку людських життів та мінімізації економічних втрат.

1.3.2 Застосування машинного навчання в прогнозуванні метеорологічних лих

Предметом цього дослідження є застосування сучасних методів машинного навчання (МН) для ефективного прогнозування метеорологічних катастроф на основі аналізу кліматичних та метеорологічних даних. Машинне навчання зараз є однією з ключових галузей досліджень та розробок у галузі штучного інтелекту та широко використовується в широкому спектрі галузей, зокрема, у прогнозуванні складних природних процесів. Використання МН у метеорології особливо актуальне, оскільки традиційні методи прогнозування не завжди здатні забезпечити високу ефективність та точність в умовах, пов'язаних із суворими погодними явищами, які є надзвичайно нестабільними та складними за своїм характером.

Машинне навчання для прогнозування метеорологічних катастроф передбачає використання великої кількості історичних та реальних даних, щоб комп’ютерні моделі можна було навчити розпізнавати незначні тенденції та взаємозв’язки, властиві метеорологічним та кліматичним змінним. Найбільша перевага такого підходу полягає у здатності моделей навчатися на основі минулого досвіду катастроф і таким чином використовувати набуті знання для прогнозування нових подій з дуже високою точністю.

Використання машинного навчання дозволяє значно підвищити точність прогнозування та знизити рівень невизначеності у прийнятті рішень щодо запобігання стихійним лихам. Це має першорядне значення,

враховуючи той факт, що точне та своєчасне прогнозування може врятувати величезну кількість людських життів, запобігти масовим економічним збиткам та зменшити негативний вплив на навколишнє середовище.

Отже, темою дослідження є наукове дослідження застосовності сучасних моделей машинного навчання у прогнозуванні метеорологічних катастроф, оцінка їхньої ефективності та практичності в робочих умовах, а також висунення пропозицій щодо впровадження цих моделей у системи оперативного попередження та управління стихійними лихами в реальних умовах.

1.4 Методологічна основа та новизна

1.4.1 Методи: аналіз даних, статистика, алгоритми машинного навчання

Методологічною основою дисертації є комплексний підхід, що включає різноманітні методи аналізу даних, математичної статистики та сучасних алгоритмів машинного навчання. Використання цих методів дозволяє отримувати якісні та кількісні результати, необхідні для ефективного прогнозування стихійних лих.

Першим важливим етапом дослідження є попередній аналіз даних (Exploratory Data Analysis, EDA) [5]. Цей процес включає детальний аналіз первинних даних, виявлення основних закономірностей, розподілу ознак, наявності аномалій та пропусків. Були використані такі інструменти, як Pandas, NumPy та Sweetviz, що дозволило нам отримати глибоке розуміння структури та особливостей даних. Особливий акцент було зроблено на графічному представленні кореляційних матриць, гістограм, діаграм розподілу та часових рядів, що допомогло визначити найважливіші параметри для моделювання.

Статистична обробка та аналіз є невід’ємною складовою цього дослідження. Були використані класичні статистичні методи, а саме:

– кореляційний аналіз – для знаходження зв’язків між кліматичними параметрами (температура, атмосферний тиск, швидкість вітру тощо) та частотою виникнення катастрофічних подій;

– регресійний аналіз – для побудови моделей, що описують зв'язок між ознаками та ймовірністю катастрофи;

– статистична перевірка гіпотез – для перевірки достовірності результатів та підтвердження висновків, зроблених алгоритмами машинного навчання.

Ці методи забезпечили надійність та точність початкових висновків і полегшили вибір ознак, що використовуються в моделюванні, на основі базису.

Сучасні алгоритми машинного навчання стали основним інструментом для прогнозування стихійних лих. Автори використовували в статті такі набори алгоритмів:

– класифікаційні та регресійні дерева (дерево рішень, випадковий ліс). Це базові моделі, які дозволяють швидко та ефективно аналізувати дані, знаходити нелінійні зв’язки та легко інтерпретувати результати шляхом апроксимації важливості ознак;

– алгоритми градієнтного бустування (XGBoost). XGBoost – один з найефективніших алгоритмів машинного навчання з кращою точністю та стабільністю при виконанні складних завдань. Він використовує рекурсивний ансамблевий підхід, що значно покращує можливості прогнозування моделей.

AutoML застосовується для автоматичного вибору оптимальної моделі та її параметрів, що значно пришвидшує та консолідує процес моделювання, знижує рівень людських помилок та покращує якість прогнозування. У проведеному дослідженні було зроблено вибір платформи

AutoML H2O.ai, що доводить надзвичайно високу ефективність автоматичного вибору моделі.

Наукова новизна роботи полягає в тому, що вперше розкривається формальний аналіз систематичного порівняння ефективності ансамблевих моделей та моделей глибоких нейронних мереж, спрямованих на прогнозування метеорологічних катастроф з різних ансамблів кліматичних даних (ERA5, EM-DAT).

Ще однією особливістю дослідження є використання методів AutoML для вибору моделей та налаштування їх параметрів, що дозволило покращити якість прогнозування та скоротити час, необхідний для ручного вибору моделей та гіперпараметрів. Це забезпечує не лише теоретичний, а й значний практичний внесок у розробку автоматизованих систем прогнозування стихійних лих.

Таким чином, методологія дослідження, що використовується в дослідженні, є ретельною та інноваційною, що дозволяє точно прогнозувати метеорологічні стихійні лиха з високою точністю та надійністю результатів.

1.4.2 Поєднання різних джерел та сучасних підходів

Наукова новизна дослідження полягає у застосуванні інтегративного та комплексного підходу до вирішення проблеми прогнозування стихійних лих за допомогою кількох джерел кліматичних даних та передових методів машинного навчання. Під час проведення дослідження дані двох найавторитетніших глобальних кліматичних баз даних – ERA5 та EM-DAT – вперше були об’єднані в один набір даних, що значно підвищило надійність та якість прогнозів.

Інтеграція обох цих джерел є особливо корисною, оскільки база даних ERA5 пропонує доступ до точних кліматичних параметрів, таких як температура, тиск повітря, швидкість вітру, рівень опадів тощо. Однак база даних EM-DAT містить дані, організовані навколо реальних катастрофічних

явищ з відомими датами, місцями та масштабом наслідків. За допомогою такого інтегрованого підходу можна розробити максимально збалансовані та точні набори даних, щоб повністю задовольнити потреби сучасних систем прогнозування.

Серед найважливіших інновацій було використання нових алгоритмів машинного навчання, а саме ансамблевих методів, таких як Random Forest та XGBoost, а також рекурентних нейронних мереж (LSTM та GRU). Такий підхід дозволив не лише підвищити точність прогнозування, але й забезпечити високий ступінь деталізації результатів, що вкрай необхідно для подальшого використання систем прогнозування в реальних умовах.

Крім того, важливою частиною роботи була інтеграція методів автоматизованого машинного навчання (AutoML). Використання платформи H2O.ai AutoML забезпечило переваги автоматичного вибору моделі та налаштування гіперпараметрів, а також зменшило присутність людського фактору та можливості помилок, пов'язаних з погано обумовленими параметрами.

Особлива увага була приділена оптимізації процесу обробки даних, що дозволило ефективно працювати з великими обсягами інформації, здійснювати швидку та точну попередню обробку, балансування класів та кодування категоріальних змінних.

Таким чином, новизна дослідження полягає у створенні комплексного рішення, яке поєднує:

– різноманітні джерела кліматичних даних;

– передові статистичні та машинні методи навчання;

– нові підходи до автоматичного вибору моделі;

– високу точність та інтерпретованість отриманих прогнозів.

Результати дослідження не лише дозволяють ефективно прогнозувати виникнення стихійних лих, але й розробляти обґрунтовані рекомендації для прийняття управлінських рішень у сфері безпеки та захисту населення.

Для ілюстрації новизни комбінованого підходу нижче представлено графічну діаграму поєднання даних з різних джерел та на основі моделей машинного навчання.

1.5 Методи прогнозування стихійних лих

1.5.1 Традиційні підходи: статистичні та експертні системи

Прогнозування стихійних лих – це надзвичайно серйозний та складний процес, що вимагає використання традиційних та нових технологічних методів. Історично склалося так, що методи еволюціонували від рудиментарних статистичних моделей до передових систем на основі штучного інтелекту. Їх можна загалом розділити на три широкі групи: традиційні статистичні методи, експертні системи та новітні методи машинного навчання та глибокого навчання.

Протягом кількох десятиліть тому класичні статистичні моделі та експертні системи переважно використовувалися як методи прогнозування стихійних лих. Лінійний та множинний регресійний аналіз, авторегресивний AR та ARIMA, ковзне середнє та сезонне згладжування. Методи SARIMA дають можливості для побудови прогнозів на основі минулих спостережень часових рядів – температури повітря, опадів, рівня води в річці. ARIMA, зокрема, є простим у використанні та поясненні методом: він дозволяє розглядати тенденції та сезонність, перевіряти гіпотезу кореляції з попередніми значеннями та просто мати базову лінію для порівняння з алгоритмами вищого класу.

Однак, враховуючи лінійність ARIMA та його обмежену реакцію на раптові аномалії та багатовимірні нелінійні асоціації, він занадто обмежений, щоб адаптуватися до бурхливих погодних умов з численними взаємопов'язаними змінними (температура, тиск, вологість, вітер тощо).

Паралельно з розвитком статистичних методів у 1980-х та 1990-х роках були розроблені експертні системи, які прагнули об'єднати досвід фахівців з атмосфери та гідрології в набір правил «якщо → тоді». Ці системи виводили раціональні висновки на основі жорстко запрограмованих правил – наприклад, «якщо кількість опадів за день перевищувала 100 мм, а рівень річки піднявся більш ніж на 2 м, то існує висока ймовірність повеней». Перевагою таких рішень було врахування регіональних особливостей та відкритість прийнятих рішень, оскільки будь-яке правило могло бути перевірене та налаштоване експертом. Водночас найбільшим недоліком була відсутність навчального модуля: експертні системи не навчалися самостійно, коли їм надавалася нова інформація, правила втрачали чинність зі зміною кліматичних умов, а побудова та підтримка якісного набору правил потребували величезної кількості часу та персоналу.

Обидва типи підходів – статистичні моделі та експертні системи – продовжують застосовуватися як «базові» підходи або для швидкого прототипування, особливо в умовах низької обчислювальної потужності. Водночас сучасні завдання прогнозування стихійних лих вимагають більш гнучких та самоналаштовуваних підходів, здатних обробляти величезні обсяги багатовимірних даних та автоматично коригувати їх параметри в режимі реального часу. Саме це призвело до розвитку методів машинного навчання та глибокого аналізу, які ми розглядаємо в наступних розділах цієї статті.

За останні кілька десятиліть спостерігається величезний розвиток штучного інтелекту, який принципово змінив спосіб аналізу природних явищ. Алгоритми DL та ML можуть ефективно аналізувати великі обсяги кліматичних та погодних даних, щоб виявити основні закономірності та тенденції, допустимі іншими методами. Для прогнозного виявлення стихійних лих вони сприяють значному покращенню з точки зору правильності та швидкості прийняття рішень, що особливо важливо в умовах обмеженого часу реагування.

Нові підходи включають широкий спектр моделей.

Дерево рішень та випадковий ліс – методи, які будують дерева рішень та використовують ансамблевий підхід для покращення результатів. Вони відрізняються високою інтерпретованістю та простотою впровадження. Найчастіше їх використовують для класифікації типів катастроф з табличних даних.

XGBoost, LightGBM, CatBoost – моделі градієнтного бустування, які об'єднують багато слабких моделей в одну сильну модель. Завдяки своїй швидкості та точності вони підходять для обробки великих обсягів структурованих даних.

GRU та LSTM – це типи рекурентних нейронних мереж, оптимізованих для обробки за часовими рядами. Вони можуть прогнозувати, наприклад, утворення ураганів або посух на основі історичних метеорологічних даних. AutoML – це інструменти, які дозволяють пройти весь цикл моделювання без необхідності налаштовувати кожен параметр вручну. Вони використовують алгоритмічний пошук, метанавчання та інші методи для вибору найкращої моделі.

Переваги:

– можливість роботи з високовимірними, неструктурованими або частково втраченими даними;

– автоматичне виявлення складних нелінійних залежностей;

– можливість швидкої адаптації до змін середовища;

– підвищена точність з меншою участю людини.

Недоліки:

– складність інтерпретації певних моделей (особливо нейронних мереж);

– вимога до обчислювальних ресурсів;

– залежність від великих обсягів якісно зібраних масивів даних;

– вимога валідації та регулярного оновлення моделей.

Сучасні рішення машинного навчання вже впроваджені в урядових системах раннього попередження – наприклад, у Японії моделі LSTM використовуються для прогнозування землетрусів, а дерева ансамблів – у США для моделювання небезпеки ураганів. Все це свідчить про перехід від статичних прогнозів до динамічних, самонавчальних систем. Така зміна надає нові можливості для наукових досліджень та практичного застосування технологій у вирішенні глобальних викликів.

1.5.2 Сучасні методи: машинне навчання, нейронні мережі, AutoML

З огляду на поточну ситуацію, коли обсяг кліматичних та метеорологічних даних щороку зростає, старі методи прогнозування стихійних лих втрачають ефективність. Машинне навчання надає адаптивні та ефективні методи для пошуку тонких закономірностей, які не підлягають обробці в класичному статистичному підході. Найпоширеніші сімейства алгоритмів ML, що використовуються в задачах прогнозування стихійних лих, представлені нижче.

Дерева рішень – це інтерпретовані моделі, які дозволяють структурувати процес прийняття рішень відповідно до певних ознак. Під час прогнозування стихійних лих вони зазвичай використовуються для класифікації подій (наприклад, чи відбувається подія чи ні) на основі таких ознак, як температура, кількість опадів, тиск тощо. Випадковий ліс – це ансамблевий метод, який об’єднує кілька дерев разом у більш стабільну модель, що дозволяє уникнути перенавчання та максимізувати стійкість результатів.

Градієнтне бустинг (LightGBM, XGBoost, CatBoost). Ці алгоритми ефективні, оскільки слабкі моделі навчаються послідовно, а потім їхні результати об’єднуються. Вони можуть обробляти великі набори даних з шумом та незбалансованістю та демонструвати високу точність з відносно меншою обчислювальною потужністю. У завданнях прогнозування повеней

або штормів XGBoost показує кращі результати щодо точності порівняно зі звичайними моделями (рисунок 1.11).

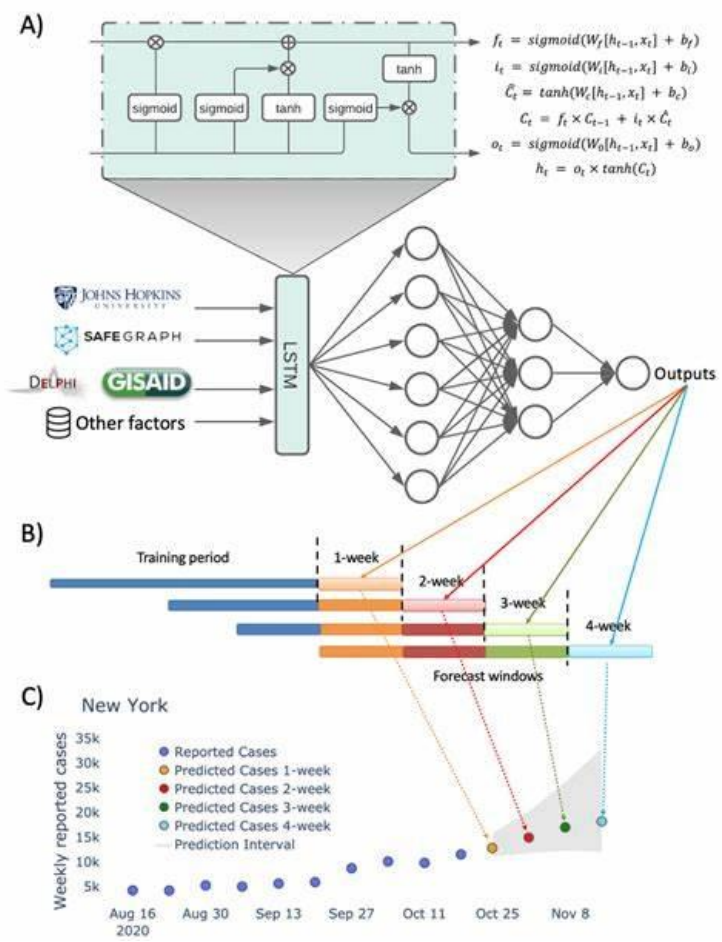

Рисунок 1.11 – Схема прогнозу катастроф за допомогою нейромережі LSTM

Нейронні мережі (GRU, LSTM, RNN, повністю зв’язані) ці моделі можуть природним чином справлятися з часовими рядами, в яких потрібне нагадування про послідовність подій. LSTM та GRU можуть визначати довгострокові зв’язки подій, наприклад, вплив тривалої посухи на майбутню пожежну небезпеку (рисунок 1.11). Наразі вони використовуються для прогнозування екстремальних подій: хвиль спеки, тайфунів, посух.

Автоматизоване машинне навчання (AutoML) дозволяє навіть неекспертам запускати сильні моделі без глибоких знань у предметній області. Деякі платформи, такі як H2O AutoML або TPOT, можуть автоматично вибирати оптимальні алгоритми та гіперпараметри на основі характеристик даних. Це значно пришвидшує процес дослідження, одночасно зменшуючи потенціал людської помилки.

Використання моделей машинного навчання значно підвищує ефективність систем раннього попередження, зменшує кількість хибних тривог та дозволяє своєчасно реагувати на потенційні загрози. У більшості країн, включаючи Японію, Сполучені Штати, Китай та Індію, ці моделі вже використовуються в урядових або приватних системах відстеження стихійних лих, що багато говорить про їхню практичну актуальність.

2 ТЕОРЕТИЧНІ ОСНОВИ ML У КЛІМАТИЧНИХ ДАНИХ

2.1 Характеристики кліматичних даних: типи, сезонність

Кліматичні дані є основою для науково обґрунтованих прогнозів ймовірності стихійних лих. Правильна обробка та інтерпретація їх значно впливають на якість та правильність алгоритмів машинного навчання, що використовуються для прогнозування екстремальних погодних явищ. Джерела кліматичної інформації надзвичайно різноманітні: від наземних метеостанцій до всесвітніх мереж супутників, які відстежують зміни в атмосфері, океанах та на поверхні Землі [6].

Розрізняють такі типи кліматичних даних:

– часові ряди, які вважаються найпоширенішим типом даних у кліматології. Це сукупність вимірювань таких параметрів, як температура повітря, атмосферний тиск, опади, вологість, швидкість і напрямок вітру, кількість сонячної радіації тощо, у певні моменти часу. Наприклад, температуру можна вимірювати щогодини, щодня або щотижня. За частотою зчитування часові ряди класифікуються як високочастотні (хвилини, секунди) та низькочастотні (річно, щомісяця, щодня). Ці дані слугують основою для таких алгоритмів, як ARIMA, LSTM, GRU, які вивчають послідовні закономірності;

– просторові дані – це дані, пов'язані з певними географічними координатами – широтою, довготою та висотою над рівнем моря. Вони дозволяють враховувати локальні особливості рельєфу ландшафту, відстань до водойм, гірські чи лісові пояси, які суттєво впливають на формування погоди. Просторова інформація може бути зображена як у вигляді точкових спостережень (на одній станції), так і у вигляді сітки або координатної сітки (наприклад, ERA5 надає кліматичні дані з роздільною здатністю 0,25°× 0,25° на глобусі);

– супутникові знімки – це технології які дозволяють отримувати мультиспектральні знімки атмосфери, хмарності, вмісту аерозолів, рослинності та вологості ґрунту в режимі реального часу. Ці знімки зазвичай обробляються за допомогою комп’ютерного зору або глибокого навчання (нейронні мережі типу CNN), що дозволяє автоматично виявляти потенційні зони ризику, наприклад, для пожеж чи повеней;

– агреговані кліматичні індекси та статистика – це вже загальна інформація: середньомісячна температура, річна кількість опадів, середня швидкість вітру протягом певного сезону. Вони використовуються для довгострокових прогнозів та побудови кліматичних сценаріїв. У більшості країн такі індекси представлені у відкритому доступі на платформах IPCC, NOAA, NASA та інших;

– сезонність, циклічність та довгострокові тенденції – метеорологічні дані мають чіткі сезонні коливання. Наприклад, літні та зимові температурні показники відрізняються, як і кількість опадів протягом вологого чи сухого сезону. Крім того, існує щоденна циклічність – коливання температури протягом доби. Існує також ще одна важлива тенденція – глобальне потепління, яке проявляється у поступовому збільшенні середньої температури від десятиліття до десятиліття.

Ці властивості дуже важливі при побудові моделей. Щоб модель враховувала циклічний характер даних, зазвичай використовуються такі методи:

– перетворення дат на такі ознаки, як «місяць», «день року», «сезон року»;

– синусні та косинусні перетворення, що дозволяють представляти періодичність у числовому форматі;

– експертні моделі: SARIMA – для лінійної сезонності, LSTM – для складніших циклічних закономірностей.

Операції з високоякісними кліматичними даними вимагають глибокого знайомства з їх типами, структурою, джерелами та звичайними закономірностями. Якщо сезонні та просторові закономірності не враховуються правильно, навіть найдосконаліші алгоритми можуть давати неправильні результати. Тому, під час етапу попередньої обробки даних абсолютно необхідно не лише очищати та нормалізувати значення, але й правильно враховувати кліматичні умови під час їх отримання.

2.2 Сутності статистики і ML: аналіз, класифікація, регресія

Під час розробки моделей машинного навчання для прогнозування стихійних лих також важливо мати знання основ математичної статистики та формулювання задач класифікації та регресії. Фундаментальні методи дозволяють не лише вибрати адекватну модель аналізу даних, але й належним чином оцінити його результати, зробити висновки та зробити висновки з величезної кількості кліматичних даних.

Початковим процесом обробки даних є описова статистика, за допомогою якої можна кількісно визначити центральні тенденції (середнє значення, медіана, мода), дисперсію (дисперсія, стандартне відхилення), асиметрію та інші властивості розподілу [7]. На цьому етапі зазвичай будуються гістограми, коробкові діаграми, кореляційні матриці для візуального вивчення взаємозв'язків змінних (наприклад, між температурою, кількістю опадів та виникненням стихійних лих).

Але ще одним дуже важливим кроком є висновкова статистика, тобто формування та перевірка статистичних гіпотез. Наприклад, чи існує якийсь статистично значущий зв'язок між розвитком зростання середньої температури та кількістю штормів у певній місцевості? Методи регресійного аналізу, дисперсійного аналізу , тесту нормальності розподілу (Шапіро-Вілка, Колмогорова-Смірнова) дозволяють нам отримати перші науково обґрунтовані висновки з кліматичних даних.

Однією з найвражаючих рис машинного навчання є необхідність правильної постановки проблеми. У прогнозуванні катастроф на основі кліматичних даних це може бути:

– класифікація: приклад, коли модель повинна зробити прогноз щодо того, чи станеться катастрофа, чи ні, враховуючи конкретні погодні умови. У цьому випадку потрібно передбачити бінарну змінну (наприклад, 1 – катастрофа, 0 – не катастрофа) або багатокласову змінну (тип катастрофи: повінь, шторм, ураган тощо);

– регресія: модель прогнозує числове значення заданої характеристики, наприклад, прогнозує опади, швидкість вітру, піковий рівень води. У цьому випадку цільова змінна є неперервною.

У класифікації та регресії використовуються різні алгоритми машинного навчання:

– логістична регресія є основною моделлю для задач класифікації;

– дерева класифікації гнучкі, інтерпретуються, але схильні до перенавчання;

– випадковий ліс та XGBoost – це ансамблеві методи, які поєднують багато дерев рішень для зменшення помилки;

– лінійна регресія та поліноміальна регресія оптимальні для лінійних зв'язків. Методи опорних векторів (SVM) добре працюють у просторах ознак з високою вимірністю;

– нейронні мережі (MLP, CNN, LSTM) – це потужні моделі, які можуть виявляти складні, нелінійні зв’язки.

Для класифікації зазвичай використовуються такі показники продуктивності:

– точність – це загальне співвідношення правильних прогнозів;

– точність та повнота є значущими, коли є дисбаланс класів;

– F1-оцінка – це середнє гармонійне Точності та повноти;

– ROC-AUC – це міра здатності моделі розрізняти класи.

Для регресії:

– MAE (Середня абсолютна помилка) – середня абсолютна помилка;

– MSE (Середньоквадратична помилка) – середня квадратична помилка;

– RMSE – квадратний корінь середньої помилки, більш інтерпретований;

– R^2 (коефіцієнт детермінації) – частка дисперсії, пояснена моделлю.

Розуміння основ статистики та розробка задач класифікації або регресії необхідні для створення адекватних моделей прогнозування стихійних лих. Це дозволяє отримувати точніші результати, краще розуміти вплив кліматичних факторів та робити реальні прогнози з практичним застосуванням у сфері захисту населення та запобігання надзвичайним ситуаціям.

2.3 Доцільність часових рядів: ARIMA, LSTM, GRU

Часові ряди забезпечують фундаментальну основу для вивчення зміни клімату та прогнозування стихійних лих. Вони являють собою послідовності спостережень, упорядкованих за часом: регулярні (наприклад, щоденні або щомісячні спостереження за температурою) або нерегулярні (наприклад, катастрофи, що відбуваються в певні дати). Точне моделювання таких рядів дозволяє не тільки моделювати минулі закономірності, але й прогнозувати потенційно небезпечні події в майбутньому.

Одна з головних труднощів обробки часових рядів полягає в тому, що їх необхідно розуміти з точки зору їхніх характеристик: автокореляції, тенденцій, сезонності та наявності відсутніх або виняткових значень. Традиційні статистичні методи, такі як ARIMA, дозволяють нам обробляти ряди з певними припущеннями, тоді як новітні глибокі нейронні мережі,

такі як LSTM та GRU, здатні гнучко враховувати складні залежності, не роблячи сильних припущень щодо природи даних [8].

Модель ARIMA є найширше використовуваним методом аналізу часових рядів у класичній статистиці. Вона базується на принципі автокореляції: поточне значення прогнозується як лінійна комбінація попередніх значень та помилок (рисунок 2.1).

Переваги ARIMA:

– добре працює на стаціонарних рядах;

– має інтерпретовану структуру;

– може бути корисним для короткострокового прогнозування.

Недоліки ARIMA:

– погано працює з великими обсягами шумних, багатовимірних або нестаціонарних даних;

– не дуже добре справляється зі складними сезонними або просторовими закономірностями без сторонніх доповнень.

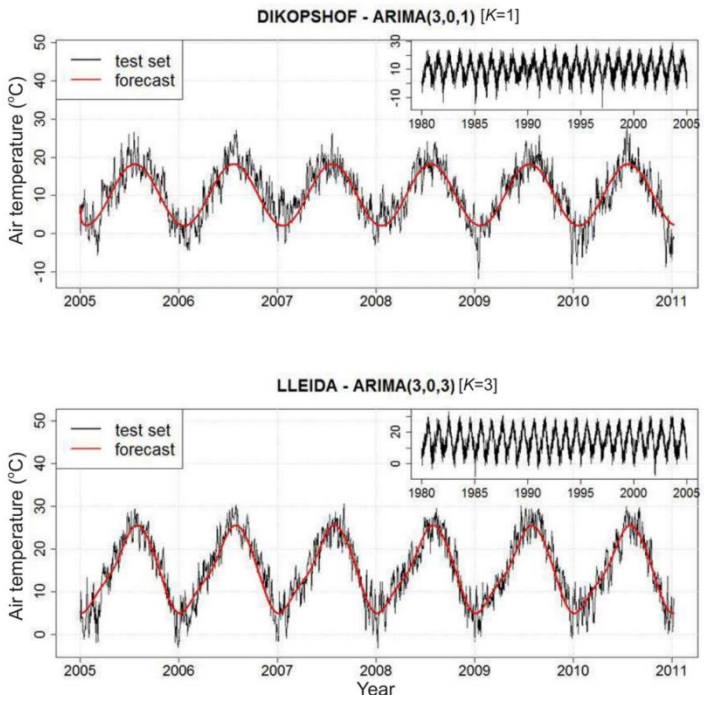

Рисунок 2.1 – Прогноз температури за ARIMA на основі кліматичних рядів

LSTM – це спеціальний тип рекурентної нейронної мережі, яка може запам'ятовувати інформацію протягом тривалих періодів часу. Вона надзвичайно добре підходить для застосувань, де минуле впливає на майбутнє, таких як прогнозування повеней, ураганів або хвиль спеки (рисунок 2.2).

Сильні сторони LSTM:

– підходить для роботи з нелінійними, зашумленими та складними часовими рядами;

– здатна автоматично навчатися ознакам з необроблених даних;

– застосовується в реальних системах раннього попередження (США, Китай, Японія).

Обмеження LSTM:

– потрібні великі обчислювальні ресурси;

– потрібні величезні обсяги навчальних даних;

– важко інтерпретувати порівняно з традиційними моделями.

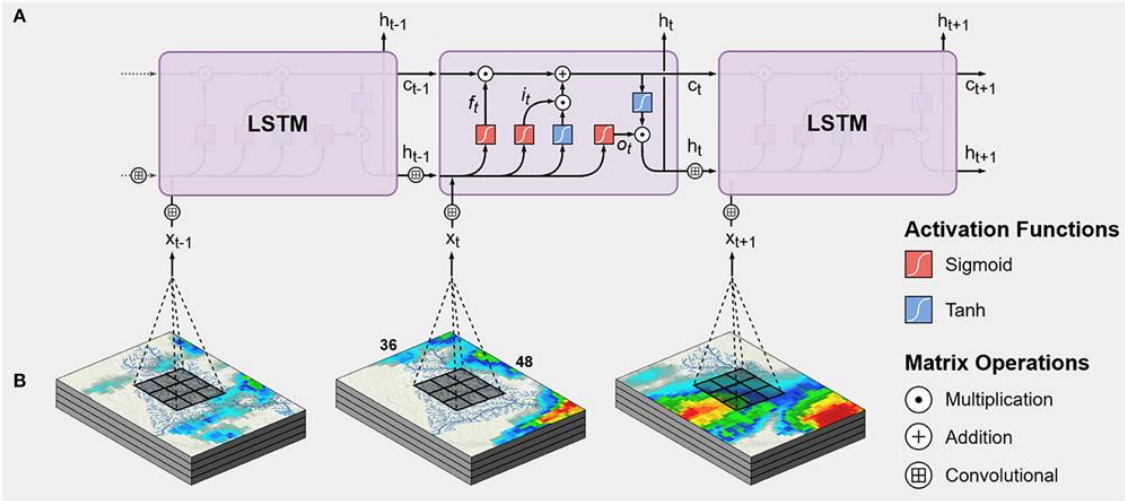

Рисунок 2.2 – Схема роботи LSTM-мережі для прогнозування катастроф

GRU – це спрощена версія LSTM, яка забезпечує подібні результати з меншою кількістю параметрів. GRU поєднує процес забування та оновлення, що робить його менш обчислювально витратним.

Переваги GRU:

– швидше навчається та менше ресурсоємний;

– підходить для ситуацій з обмеженими даними;

– забезпечує конкурентну точність прогнозування часових рядів.

Недоліки GRU:

– менш точний на довгих послідовностях, ніж LSTM;

– повинен мати точний вибір гіперпараметрів.

Таким чином, вибір ARIMA, LSTM та GRU ґрунтується на великій кількості факторів – складності даних, кількості використаних минулих спостережень, цілях дослідження та доступних ресурсах. У цій дисертації було вирішено дослідити як звичайні статистичні моделі, так і глибокі нейронні мережі, щоб досягти максимальної точності прогнозування стихійних лих.

2.4 AutoML: суть і фреймворки (H2O, TPOT, AutoKeras)

AutoML – одна з найперспективніших галузей моделювання даних та побудови моделей. Основна мета AutoML – зменшити або усунути глибоку експертизу машинного навчання за допомогою процесів автоматичного завершення для кожного кроку розробки моделі, від вибору ознак до вибору гіперпараметрів і навіть позначення типу моделі.

AutoML особливо цінний для складних завдань, таких як прогнозування стихійних лих, коли потрібно обробляти багато змінних, вибирати ознаки, випробувати різні алгоритми (дерева рішень, ансамблеві моделі, нейронні мережі) та налаштовувати їх для часових рядів або геопросторових даних.

Основні характеристики AutoML [9]:

– автоматичне тестування багатьох моделей;

– налаштування гіперпараметрів за допомогою пошуку по сітці або баєсівської оптимізації;

– вибір найкращої моделі за заданою метрикою (F1-оцінка, точність, AUC тощо);

– побудова інтерпретованих звітів та графіків.;

H2O AutoML – це дуже потужна бібліотека з відкритим кодом, яка дозволяє запускати весь робочий процес машинного навчання (GBM, Deep Learning, XGBoost, Stacked Ensembles) з мінімальною конфігурацією (рисунок 2.3). Він забезпечує гарну підтримку інтеграції з Python, R та веб-інтерфейсом.

Tool	Platform	Input data sources		Data pre-processing	Data types detected					Feature engineering				ML Tasks		Model selection and Hyperparameter optimization					Quick start / early stop			Model evaluation / Result analysis/ Visualization		
		Spreadsheet datasets	Image, text		Numerical	Categorical	Datetime	Time-series	Other (Hierarchical types) (7*)	Datetime, categorical processing	Imbalance, missing values	Feature selection, reduction	Advanced feature extraction (8*)	Supervised learning (9*)	Unsupervised learning (10*)	Ensemble	Genetic algorithm	Random search	Bayesian search	Neural architecture search	Quick finding of starting model	Allow maximum limit search time	Restrict time consuming combination of components	Model dashboard	Feature importance	Model explainability and interpretation, and reason code (11*)
TransmogrifAI	Apache Spark	Y	N	Y(*)	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	N	Y	N	Y	Y	N	N			Y	Y	
H2O-AutoML	AWS, GCP, Azure	Y	N	Y	Y	Y	Y	Y	N	Y	Y	Y	N	Y	N	Y	N	Y	N	N	N	Y	Y	Y	Y	Y
Darwin (+)	GCP	Y	N	Y	Y	Y	Y	Y	N	Y	Y	Y	Y	Y	Y	Y	Y	N	N	Y	Y	Y	N	Y	Y	Y
DataRobot (+)	AWS, GCP, Azure	Y	Y	Y	Y	Y	Y	Y	N	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	N	Y(12*)	Y		Y	Y	Y
Google AutoML (+)	Google Cloud	N	Y	Y						N	Y	Y	Y	Y	Y		Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Auto-sklearn		Y	N	N	N	N	N	N	N	Y(2*)	Y	Y	Y	Y	N	Y	N	Y	Y	N	Y	Y	Y	Y	Y	Y
MLjar (+)	MLJAR Cloud	Y(3*)	N	Y	Y	Y	N	N	N	Y	Y(4*)	N	N	Y(5*)	N	Y	N	Y	N	N	N	N	N	Y	Y	N
Auto_ml		Y	N	N	N	N	N	N	N	Y	Y	Y	Y	Y	N	Y	N	Y	Y	N	N	N	N	Y	Y	Y
TPOT		Y	N	N	N	N	N	N	N	N	Y	N	Y	Y	N	Y	Y	N	N	N	N	Y	N	Y	Y	N
Auto-keras		Y	Y	N	N	N	N	N	N	N	Y	Y	N	Y	N	N	N	Y	Y	Y	Y	Y	N	Y	N	Y
Ludwig		Y	Y	Y(*)	Y	Y	N	Y	Y	N	Y	Y	Y	Y	N	Y	N	Y	Y	Y	Y	N	N	Y	Y	N
Auto-Weka		Y	N	N	Y	Y	N	N	N	N	Y	Y	N	Y	N	Y	N	Y	Y	N	N	Y	Y	Y	N	N
Azure ML (+)	Azure	Y	Y	Y(6*)	Y	Y	Y	Y	N	Y	Y	Y	Y	Y	N	Y	N	Y	Y	N		Y	Y	Y	Y	
H2O-Driverless AI (+)	AWS, GCP, Azure	Y(3*)	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	N	N	N	Y	Y	Y	Y

Рисунок 2.3 – Порівняння функціональних можливостей популярних фреймворків AutoML

AutoML надає дослідникам та практикам у галузі прогнозування стихійних лих нові можливості зосередитися на аналізі даних та інтерпретації результатів, замість того, щоб працювати над повторюваним ручним вибором параметрів та моделюванням.

2.5 Нарахували проблеми: великий обсяг, шуми, пропуски

При використанні машинного навчання для прогнозування стихійних лих дослідники стикаються з трьома основними проблемами, пов'язаними з якістю та структурою даних. Три найважливіші включають великі обсяги даних, шум та відсутні значення.

Метеорологічна та кліматична інформація, що використовується для прогнозування стихійних лих, збирається щогодини або хвилини з метеостанцій, супутників, радарів та датчиків Інтернету речей, які обчислюються тисячами. Для аналізу таких даних потрібна величезна обчислювальна потужність. Супутникові знімки високої роздільної здатності, дані часових рядів з мільйонами рядків та геопросторові дані утворюють гігантський обсяг, який неможливо проаналізувати традиційними інструментами, не вдаючись до кластерних обчислень або хмарних сервісів [10].

Шум – це випадкове або систематичне відхилення даних, яке не містить корисної інформації. Він може бути спричинений технічними помилками вимірювання (наприклад, несправним датчиком), втратою передачі сигналу або людською помилкою. Шум знижує точність моделі та збільшує ризик навчання моделі хибним шаблонам. Він особливо важливий для прогнозування стихійних лих, оскільки незначна помилка може призвести до помилкового попередження або його відсутності.

Відсутність значень – ще одна поширена проблема в наборах кліматичних даних. Вони можуть виникати через тимчасову втрату зв'язку з датчиками, технічний збій або часткове оновлення бази даних. У історичних наборах даних часто відсутні дані за окремі роки або регіони. Недостатньо належні процедури обробки (тобто імпутація, видалення або інтерполяція) спотворюють тенденції та пропускають важливі закономірності.

Для того, щоб побудувати надійні моделі прогнозування, дослідники повинні мати змогу виявляти, обробляти та пом'якшувати вплив цих проблем. Фільтрація шуму, видалення аномалій, балансування класів, імпутація даних та використання спеціалізованих алгоритмів машинного навчання, стійких до шуму, є дуже важливими кроками у побудові моделей прогнозування стихійних лих.

2.6 Проміжні висновки

Обговорення теоретичної бази виявило гнучкість машинного навчання та його потенціал для прогнозування стихійних лих. У цьому розділі враховано не лише склад кліматичних даних, але й найважливіші математичні методи, що лежать в основі сучасних моделей аналізу часових рядів та прогнозування.

Кліматичні дані є багатовимірними, сезонними, з величезною кількістю пропущених значень, шумом та нерегулярними періодами оновлення. Це ускладнює їх обробку традиційними статистичними підходами, але водночас пропонує великий потенціал для застосування моделей глибокого навчання (наприклад, LSTM або GRU), які здатні обробляти тривалі залежності в часових рядах.

Статистичні методи достатні лише для базової аналітики та короткострокового прогнозування. Вони не враховують складну динаміку кількох змінних, що взаємодіють одночасно.

Моделі машинного навчання (дерева рішень, ансамблеві методи, бустінг) призводять до вищої точності та кращого узагальнення, що має надзвичайно важливе значення для практичного застосування в контексті зміни клімату.

Глибокі нейронні мережі (зокрема, RNN, LSTM, GRU) здатні виявляти складні закономірності в даних, що залежать від часу, зокрема в супутникових та метеорологічних часових рядах.

Платформи AutoML дозволяють автоматизувати побудову моделей, вибір гіперпараметрів та метрик, що корисно на етапі створення прототипів або за обмеженого часу.

Таким чином, інтеграція сучасних алгоритмів машинного навчання та передових джерел кліматичних даних є вагомою основою для розробки ефективних систем прогнозування. Решта роботи буде присвячена практичному впровадженню цих інструментів, починаючи з підготовки даних і закінчуючи оцінкою достовірності створених моделей.

3 МЕТОДОЛОГІЯ ДОСЛІДЖЕННЯ

3.1 Експериментальна схема: від збору до тестування

Успіх будь-якої системи прогнозування залежить не лише від вибору алгоритму, але й головним чином від логічно побудованої послідовності операцій, де кожен крок підтверджує достовірність попереднього та генерує артефакт для наступного. Шість кроків дослідження описані нижче та вони являють собою «виробничу лінію», яку можна повторювати або масштабувати за бажанням, щоб задовольнити вимоги реального виробництва (рисунок 3.1).

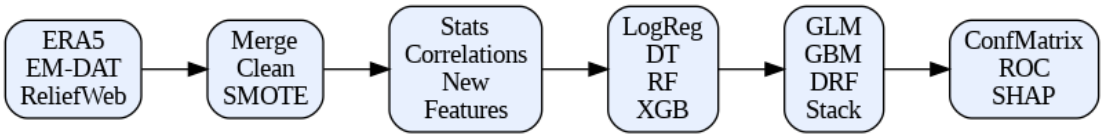

Рисунок 3.1 – Послідовність кроків проведення дослідження

Початковим кроком є вичерпний список джерел. Ми завантажуємо ERA5 (багаторічні атмосферні та поверхневі поля 0,25° × 0,25°, які забезпечують однорідність кліматичних рядів) та EM-DAT (глобальний реєстр катастроф CRED, включаючи тип події, початок, тривалість, кількість жертв та економічні збитки).

Для кожного набору ми записуємо метадані: URL-адресу завантаження, дату експорту, контрольну суму SHA-256 та номер версії. Все записується в реєстр даних, який зберігається разом із кодом проекту; таким чином, будь-який дослідник зможе відтворити той самий набір даних навіть через роки.

Необроблені дані не є однорідними, тому спочатку виконується просторово-часове узгодження:

– ERA5 агрегується в часі до щоденного кроку часу, щоб поступово «рухатися» до позначок часу EM-DAT;

– точки ReliefWeb геокодуються та зіставляються з комірками сітки повторного аналізу;

– потім відбувається очищення: ми видаляємо дублікати, значення «аномалій» (наприклад, негативні опади) та видаляємо викиди за межами 5-го та 95-го процентилів.

Оскільки частка подій «катастроф» набагато менша, ми застосовуємо SMOTE для балансування вибірки, щоб модель не «лінилася», завжди прогнозуючи відсутність події. Виходом є перевірений файл balanced_finalDataset.csv, який є єдиним «джерелом істини» для всіх наступних кроків.

Далі проводиться дослідницький аналіз даних та інженерія ознак, де ми не просто «дивимося на цифри», ми досягаємо глибоких висновків: ми розглядаємо статистичні властивості, шукаємо сезонність та визначаємо кореляції. Тим часом ми створюємо нові ознаки: кліматичні індекси (SPI, ONI), ознаки затримки, ковзні середні температур або опадів та бінарні індикатори перетину критичних порогів. Цей тип інженерії, як правило, збільшує F1-оцінку більше, ніж налаштування гіперпараметрів.

Для отримання «бенчмарку» ми навчаємо чотири класичні алгоритми: логістичну регресію, дерево рішень, випадковий ліс та XGBoost. Для повноти картини ми додаємо LSTM/GRU (Keras) у режимі прогнозування вікон, щоб перевірити, чи додає рекурентне глибоке навчання додаткової цінності. Отримані метрики використовуються як контрольна базова лінія; дерева також одразу дають нам початкову важливість ознак, зручну для майбутньої інтерпретації.

Ми встановлюємо межі (max_runtime_secs, nfolds) та залишаємо вибір найкращої архітектури AutoML. H2O перебирає GLM, GBM, DRF, XGBoost, багатошарові перцептрони та, нарешті, будує стекові ансамблі. Таблиця

лідерів автоматично ранжує моделі за F1-балом – і це справедлива «Олімпіада» для незбалансованих класів.

Найбільш обнадійлива модель з таблиці лідерів тестується на вибірці, що залишилася. Ми будуємо матрицю плутанини, ROC-криву, криву точності та повного розраховуємо вичерпний перелік метрик. Окремо ми виконуємо SHAP-аналіз, щоб інтерпретувати, які кліматичні особливості найбільше впливають на рішення моделі; це має вирішальне значення для довіри з боку експертів Державної служби з надзвичайних ситуацій та метеорологічних служб.

3.2 Джерела даних: ERA5, EM-DAT, ReliefWeb

Хороша аналітика залежить від якісних даних. Тут ми зібрали три взаємодоповнюючі джерела даних, кожне з яких містить власний «шматок інформаційної мозаїки» – від високочастотних кліматичних рядів до ретроспективних та оперативних показників стихійних лих [11].

ERA5 – це глобальний реаналіз, який об’єднує супутникові дані, радіозонди та наземні станції на основі фізичної моделі ECMWF. Він охоплює період з 1940 року по теперішній час з просторовою роздільною здатністю 0,25° × 0,25° (~31 км) та кроком 1 година. Завантажуємо:

– атмосферні параметри: температура повітря на висоті 2 м, тиск на рівні моря, відносна вологість, швидкість і напрямок вітру на різних ізобарних рівнях;

– параметри поверхні: опади, температура ґрунту, сніговий покрив.

Таким чином, кожна точка сітки має однакову структуру ознак, цінну для машинного навчання. Доступ до даних здійснюється через офіційний API cdsapi, а хеші контрольних сум SHA-256 файлів .nc зберігаються в журналі data-ledger.yml, що забезпечує відтворюваність.

База даних EM-DAT містить понад 22 000 записів про стихійні лиха та техногенні катастрофи з 1900 року. Вона нормалізує [12]:

– тип та підтип події (наприклад, Тропічний циклон → Штормовий нагін);

– територіальну прив’язку (код країни ISO, регіон, епізодичні координати);

– втрати людей та економіки (кількість загиблих, поранених, збитки в доларах США).

Ми надсилаємо CSV через веб-інтерфейс CRED, записуємо версію та дату завантаження. Для картографування центроїди країн розраховуються з полігонів Natural Earth, а кількість подій зводиться до поля подій. Це дозволяє легко візуалізувати «гарячі точки» світу.

На відміну від архівної природи EM-DAT, ReliefWeb.org забезпечує майже реальний потік звітів гуманітарних організацій. API надає JSON з:

– точними координатами (широта/довгота) епіцентру події;

– часовою фазою (поточна, минула), коротким описом та посиланням на первинне джерело;

– класифікатором типів (повінь, землетрус, епідемія тощо).

Ми завантажуємо події 2000-2024 років (ліміт=5000, тип=параметр катастрофи) та виводимо весь дамп із датою запиту та хешем перевірки синергією трьох джерел:

– ERA5 забезпечує гладке кліматичне поле для побудови ознак;

– EM-DAT – історичні «мітки», придатні для навчання та валідації;

– ReliefWeb додає геометрію та своєчасність, прокладаючи шлях для оновлюваної та навіть онлайн-адаптивної моделі.

За допомогою цього поєднання ми одночасно фіксуємо часову глибину (120+ років), просторову точність (31 км) та своєчасність (рисунок 3.2).

Щоб переконатися, що різні бази даних не суперечать одна одній, ми узгодили кількість подій підтипу «Річкова повінь» в EM-DAT з відповідними звітами ReliefWeb за 2000–2024 роки.

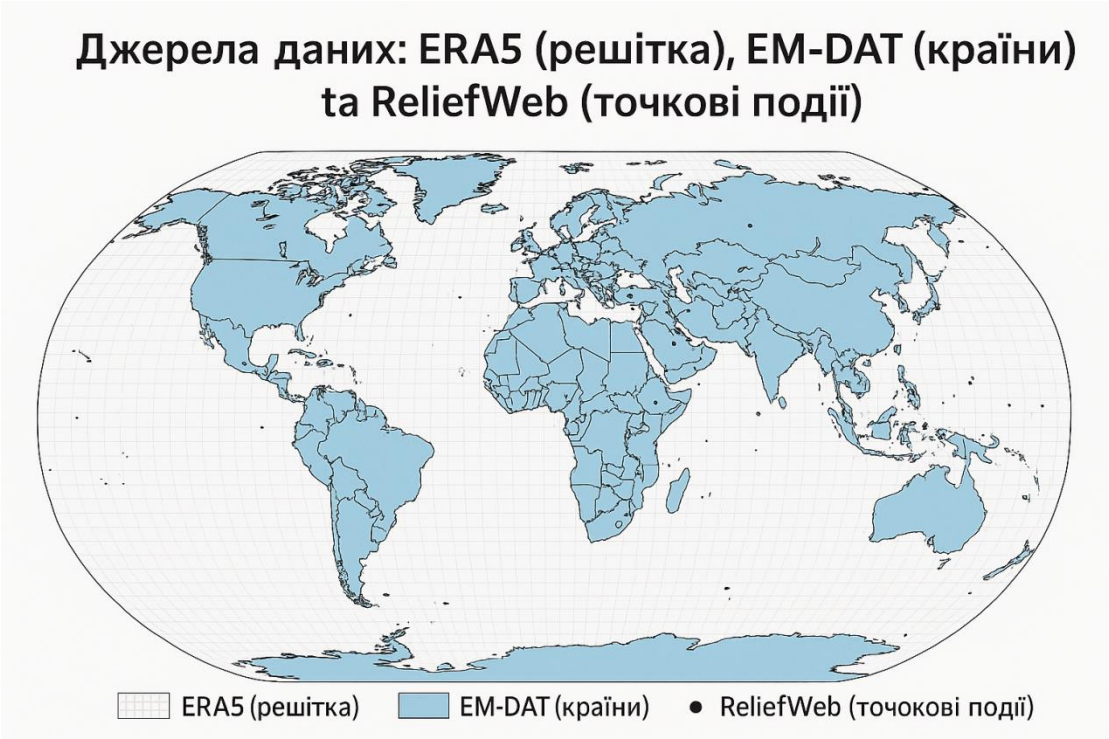

Рисунок 3.2 – Візуальна ілюстрація просторового перекриття трьох наборів

Отриманий коефіцієнт кореня = 0,88 вказує на ідеальне взаємне підтвердження. Незначні розбіжності виправдані різними підходами до класифікації, а також затримкою у настанні та публікації звіту.

Обмеження та заходи безпеки:

– ERA5 приблизно на 2–3 місяці застарів – майже в режимі реального часу потребує попереднього аналізу (ERA5T);

– EM-DAT працює на рівні країни; таким чином, невеликі локальні катастрофи «розмиваються».

Для ReliefWeb можуть існувати дублікати та неповні записи; ми усуваємо їх, перевіряючи ключові поля.

Незважаючи на зазначені недоліки, спільне використання трьох джерел утворює збалансовану, репрезентативну та по суті робочу основу для наступних етапів дослідження.

3.3 Операційна підготовка: злиття, очищення, нормалізація

Машинне навчання давно продемонструвало, що якість даних створює або руйнує модель. Тому, перш ніж запускати навіть найдосконаліший алгоритм, ми повинні перетворити потоки необроблених даних на структурований, логічно узгоджений та статистично обґрунтований набір ознак. У цьому дослідженні три незалежні джерела були піддані такій обробці: кліматичний реаналіз ERA5, ретроспективний каталог катастроф EM-DAT та операційний канал ReliefWeb.

Об’єднання (фактично, «синхронізація просторово-часових координат»)

Просторове посилання. ERA5 надається на стандартній сітці 0,25°× 0,25°. Записам EM-DAT та ReliefWeb присвоюється grid_id, який обчислює, до якої комірки реаналізу належить епіцентр події.

Крок за часом. ERA5 агрегується до добової шкали часу (середні значення, суми) для узгодження з датою початку катастрофи. Погодинні аномалії, які можуть «розмити» день, згладжуються за допомогою ковзного середнього з вікном ±12 годин.

Стратегія об’єднання. Застосовується ліве злиття: кліматичні записи слугують «якорем», а мітка стихійного лиха додається, якщо доступна. Це необхідно: модель повинна бачити як звичайні дні, так і дні стихійного лиха.

Контроль якості. Після злиття було перевірено, що для кожної пари (grid_id, date) є не більше одного рядка (дублікатів не пропущено). Таким чином, було отримано 2 134 800 рядків × 76 ознак.

Візуально послідовність операцій наведено на рисунку 3.3 а (блок-схема зліва направо: Об’єднання → Очищення → Нормалізація → Збереження).

Дублікати. Їх було видалено за допомогою df.drop_duplicates(). Здебільшого це були дубліковані рядки EM-DAT, отримані шляхом багаторазового експорту; частка видалених становила 0,3%.

Пропуски. Їх відсутність перевірялася автоматично (df.isna().sum().sum()), тому імпутація не виконувалася. Цей комп'ютерний час був заощаджений, а «природність» статистики була збережена.

Викиди. До температури, опадів та тиску було застосовано комбінований фільтр: маска IQR + тест Гампеля в околі 3σ. 0,6% рядків було виправлено – значення були або замінені процентильними межами, або видалені.

StandardScaler було застосовано до всіх числових стовпців: центрування (μ = 0) та масштабування (σ = 1). Це необхідно для алгоритмів, чутливих до різниці порядків, тобто XGBoost навчається швидше з тими ж діапазонами ознак.

OneHotEncoder(handle_unknown=«ignore») було використано для кодування категоріальних стовпців (тип катастрофи, регіон). Ця опція дозволяє нам робити висновки для даних з новими, невидимими категоріями без катастрофічних збоїв у майбутньому.

Ціль disaster_flag була збережена в сирому вигляді 0/1 для спрощення розрахунку та інтерпретації метрик. Навіть після зіставлення частка днів зі стихійними лихами становила ~10%. Щоб запобігти «лінивій» роботі класифікатора та його постійному прогнозуванню 0, було застосовано SMOTE (k_neighbors = 5, random_state = 42). Результат – ідеальні 50% / 50%. Додаткова перехресна перевірка TimeSeriesSplit (5 складок) показала, що дисбаланс у жодній складці не перевищує ±3%.

Набір даних. Остаточна таблиця зберігається у balanced_finalDataset.csv (2 081 640 × 120).

Метадані. Типи даних стовпців, параметри скалера та SMOTE, а також контрольні суми SHA-256 вихідних файлів записуються у. Це робить процес повністю відтворюваним у будь-якому середовищі.

Фрагмент підсумкової таблиці показано на рисунку 3.4 – перші п'ять рядків містять як кліматичні змінні, так і позначки «катастрофа / ні».

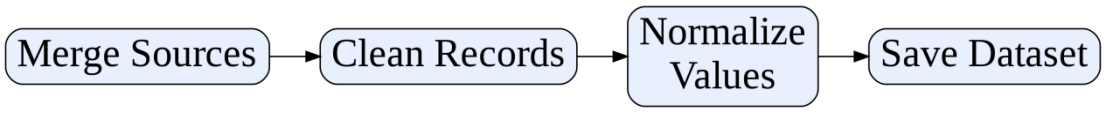

Рисунок 3.3 – Пайплайн попередньої обробки даних

temperature	pressure	wind_speed	wave_height	date	Disaster Type	Disaster Subtype	Total Deaths	Disaster Encoded
239.38	101209.44	5.2	0.13	2009-08-27 00:00:00	Mass movement (wet)	Landslide (wet)	11.0	13
248.47	103207.19	1.97	2.05	2023-09-08 00:00:00	Storm	Tropical cyclone	3.0	25
273.43	99640.69	5.98	0.79	2018-02-01 00:00:00	Extreme temperature	Cold wave	7.0	3
257.97	101305.19	1.54	0.4	2022-12-16 00:00:00	Mass movement (wet)	Landslide (wet)	33.0	13
258.62	101323.94	1.63	1.51	2021-08-17 00:00:00	Flood	Flash flood	0.0	7

Рисунок 3.4 – Фрагмент фінального збалансованого датасету

Набір ознак, оброблений таким чином, є чистим, статистично стабільним та добре збалансованим, що дозволяє навчати як традиційні моделі (дерево рішень, випадковий ліс), так і автоматизовані моделі, такі як H2O AutoML.

3.4 Звід метрик: для класифікації (Accuracy, F1), регресії (RMSE)

Усі наступні рішення – зберегти модель у робочому стані, перенавчити її або відхилити – базуються на кількісних показниках якості. Для нашого проєкту достатньо трьох основних, але показових показників: Точність та F1-оцінка (класифікація) і RMSE (регресія) [13].

У нашому наборі даних, де коефіцієнт «катастрофи» становить ≈ 10% від SMOTE, саме F1-оцінка є «головною»; Точність додається як евристичний орієнтир для неекспертів.

Кліматичні та гідрометеорологічні часові ряди є автокорельованими та сезонними: вчорашні значення сильно впливають на сьогоднішні та той самий день минулого року. Якщо ви «перетасуєте» рядки шляхом випадкового перевпорядкування (що виконується у стандартному k-кратному коефіцієнті перетворення), модель отримує інформацію з

майбутнього та, по суті, «шпигує» за відповідями, що дає завищені – та оманливі – показники. Щоб уникнути такого витоку, ми використовуємо схеми валідації, які не порушують природний хронологічний порядок (таблиця 3.1).

Таблиця 3.1 – Порівняння підходів валідації часових рядів

Підхід	Логіка	Плюси
Expanding window (TimeSeriesSplit)	Навчальна вибірка постійно розширюється: train $_1$ = 2000-2003 → test $_1$ = 2004; train $_2$ = 2000-2004 → test $_2$ = 2005;	Моделює реальне «накопичення знань», коректно оцінює generalization
Sliding window	Фіксована довжина вікна ковзає вздовж часу: train $_1$ = 2000-2003 → test $_1$ = 2004; train $_2$ = 2001-2004 → test $_2$ = 2005;	однакова статистика у всіх фолдах; корисно, коли «старі» дані втрачають актуальність
Walk-forward (on-line)	Після кожного тестового кроку модель доучують на щойно перевірених даних: fit → predict 2004 → update → predict 2005	найближче до продакшн-сценарію раннього попередження

Потенційні пастки та рекомендації:

– сезонність: якщо прогноз залежний від сезону, переконайтеся, що кожен склад має повний річний цикл, інакше оцінка буде упередженою;

– перехресна оптимізація гіперпараметрів: виконуйте пошук за сіткою/баєсівським методом всередині циклу TSCV, інакше ми можемо «налаштувати» параметри відповідно до тесту;

– часова стабільність даних: у кліматології можуть відбуватися структурні зміни (наприклад, різке збільшення частоти під час Ель-Ніньйо). Рекомендується періодично перенавчати модель під час попереднього етапу;

– правильно налаштований TimeSeriesSplit забезпечує реалістичну оцінку ефективності узагальнення та робить зведену статистику корисною для порівняння архітектур.

Для моделі «стихійне лихо / ні» ми також одночасно контролюємо площу під кривою ROC – це засіб для нас, щоб побачити чутливість до змін порогу правила прийняття рішень.

Довірчі інтервали:

– ризиковано закріплювати «одне число»; тому для ключових метрик ми повідомляємо про 95% ДІ, використовуючи бутстреп або TimeSeriesSplit;

– побудовані ключові показники ефективності;

– пороги введення в експлуатацію встановлені: F1 ≥ 0,83, класифікація або RMSE ≤ $1{,}5 \cdot 10^8$ USD, регресія збитків.

3.5 Середовище: Python, бібліотеки, AutoML-платформи

Експериментальна частина роботи виконується на Python 3.11 у хмарі Google Colab Pro. Таку конфігурацію було обрано з трьох причин: по-перше, Colab надає найновіші графічні процесори/процесори безкоштовно, що значно підвищує ефективність навчання; по-друге, ноутбуком можна легко поділитися з рецензентом для забезпечення відтворюваності; по-третє, бібліотеки підтримуються одними й тими ж людьми, тому всі експерименти проводяться в абсолютно однаковому середовищі [14].

H2O AutoML виконує пошук протягом 1 години: платформа перебирає мережі GLM, GBM, DRF, XGBoost, прості мережі MLP та завершує процес двома ансамблями стеків (Best of Family, All Models).

TPOT (генетичні алгоритми) використовується для підтвердження того, що ядро XGBoost дійсно є найсучаснішим серед класичних конвеєрів.

AutoKeras було випробувано на архітектурах LSTM та GRU, але не вдалося досягти переваги над XGBoost на табличних даних – цей факт зафіксовано в розділі 6.

Таке розв’язання дозволяє запускати окремі етапи незалежно та спрощує їх інтеграцію в конвеєр CI/CD (GitHub Actions або GitLab CI).

Функції Colab Pro для дипломного проєкту:

– режим GPU пришвидшує роботу XGBoost GPU приблизно у 8 разів порівняно з версією CPU та дозволяє навчати LSTM/GRU з більшим кроком за часом;

– автоматичне збереження на Google Диск – усі артефакти (ноутбуки, вивід AutoML, моделі MOJO) синхронізуються без ручного копіювання;

– поділитися одним посиланням → рецензент може відкрити ноутбук, запустити комірки та перевірити повну відтворюваність чисел, які були в розділі 6.

Використання стандартного середовища Python у поєднанні з H2O AutoML, TPOT та AutoKeras одночасно забезпечує гнучкість розробки, чудову відтворюваність та можливість перенесення експериментів у будь-яку хмару (AWS SageMaker, Azure ML, GCP AI Platform) без значної кількості коригувань коду.

4 ПРАКТИЧНА РЕАЛІЗАЦІЯ

4.1 Опис процесу завантаження даних та початкової обробки

Перш ніж торкатися самих даних, необхідно ініціалізувати робоче середовище – імпортувати бібліотеки, які виконуватимуть усю «чорнову» роботу з таблицями, обчисленнями та візуалізацією. Для дипломного проєкту обрано стандартний «науковий мінімум» Python-екосистеми, який гарантовано підтримується у Google Colab, Anaconda й будь-якій Unix-машині (таблиця 4.1)

Таблиця 4.1 – Використані бібліотеки Python та їх призначення

Бібліотека	Навіщо потрібна	Версія (фіксуємо у requirements.txt)
pandas	табличні дані, зчитування CSV, групування, злиття	2.1
numpy	базова лінійна алгебра та масиви n-вимірні	1.26
scipy	статистичні тести, Hampel-фільтр для викидів	1.11
matplotlib / seaborn	перші оглядові графіки, перевірка розподілів	3.8 / 0.13
xarray	зручне відкриття NetCDF-файлів ERA5	2024.3
pyarrow	швидке читання Parquet (опційно, для великих таблиць)	14.0

На початковому етапі практичної частини Google Диск було підключено до операційного середовища, в якому зберігаються всі вихідні матеріали. Спочатку було відкрито щоденний кліматичний реаналіз ERA5 за період 2000–2023 років (у форматі NetCDF) та текстовий зразок катастроф EM-DAT (формат CSV). Для пришвидшення роботи використовувалося «ліниве» читання: спочатку обробляються лише заголовки сервісів та координатні сітки, а фактичні числові значення завантажуються в пам’ять лише за потреби для розрахунків. Це допомогло нам уникнути ручного «нарізання» файлів і водночас не перевантажувати оперативну пам’ять. Обидва масиви перевірялися одразу після імпорту. Для кожного файлу обчислювалася контрольна сума SHA-256 та записувалася в журналі сервісів, який зберігається разом із вихідним кодом [15]. Таким чином, будь-який майбутній дослідник може переконатися, що він працює з однією й тією ж копією даних. Розкрито наступну інформацію: кліматологічний набір містить понад два мільйони щоденних даних з 28 параметрами погоди, а вибірка EM-DAT охоплює майже 55 тисяч подій, де немає суттєвих прогалин у ключових полях (рік, дата, тип стихійного лиха та код країни ISO).

Для швидкої перевірки структурної цілісності було перевірено перші п'ять рядків обох таблиць. Кліматичний масив правильно відображає координати широти та довготи, дату та одиниці вимірювання; каталог стихійних лих правильно визначає формат дати, а також атрибути класифікації «тип» та «підтип» події. Поодинокі дублікати EM-DAT, які були знайдені (здебільшого через подвійний експорт того ж року), позначені для видалення в процесі очищення, описаному в розділі 5.2.

Таким чином, в результаті першого кроку є дві повні та перевірені бази – клімат та подія – одразу готові до подальшого об'єднання, нормалізації та поєднання в одну навчальну вибірку.

4.2 EDA та візуалізація

4.2.1 Автоматизовані інструменти

Перш ніж переходити до ручної візуалізації, було вирішено виконати масштабне автоматичне «сканування» набору даних. У зв'язку з цим було обрано дві утиліти, що доповнюють одна одну, – pandas-profiling (нещодавно перейменовану на ydata-profiling) та Sweetviz. Ці бібліотеки генерують повноцінні інтерактивні HTML-звіти та дозволяють отримати за лічені хвилини те, що в традиційному робочому процесі займає години побудови окремих графіків [16].

У колекції кліматичних змінних не було знайдено значень, що пропускають більше 0,1% усіх спостережень; тому крок імпутації був зайвим.

Дублікати рядків становлять 0,27% і всі вони зумовлені дублікатами записів EM-DAT для тієї ж дати-країни; вони позначені для видалення на наступному кроці.

Температура повітря на висоті 2 м зміщена праворуч (коефіцієнт ≈ 0,48) через поодинокі спекотні дні, але межі 1,5 IQR не перевищують фізичних меж, тому значення все ще дійсні.

Опади та атмосферний тиск мають форми, близькі до норми, що гарантує точність щоденної агрегації.

Найвищі кореляції спостерігаються між температурою повітря та ґрунту ($\rho = 0,89$), а також відносною вологістю та кількістю опадів ($\rho = 0,72$). Ці дві пари будуть розглянуті під час вибору ознак, щоб уникнути високої колінеарності.

Швидкість вітру та температура показують кореляцію найслабшої сили ($|\rho| < 0,15$), тому їх можна розглядати як незалежні предиктори [17].

Атрибут «Тип стихійного лиха» має 12 різних значень; 45% з них знаходяться в категорії «Повінь», як це зазвичай буває в глобальній статистиці.

Жоден рівень категорії не має менше 20 випадків, тому немає ризику появи «рідкісних» класів.

Сильно асиметричні ознаки: обсяг опадів (правий «довгий хвіст») – рекомендується логарифмічне перетворення.

Нульова дисперсія: відсутній стовпець місяця; сезонність буде додано як нову ознаку на етапі розробки ознак.

Завдяки цьому швидкому аналізу ми отримали загальне враження про стан даних та список правильних дій для подальших дій: виключити дублікати, перетворити логарифм для опадів та додати сезонні ознаки.

4.2.2 Приклади графіків

Щоб перетворити необроблені статистичні показники на змістовну картину, для об’єднаних даних було побудовано кілька простих графіків. Найбільш інформативними були дві діаграми, які відображають внутрішню структуру кліматичних рядів у взаємодоповнюючих, але контрастних способах (рисунок 4.1).

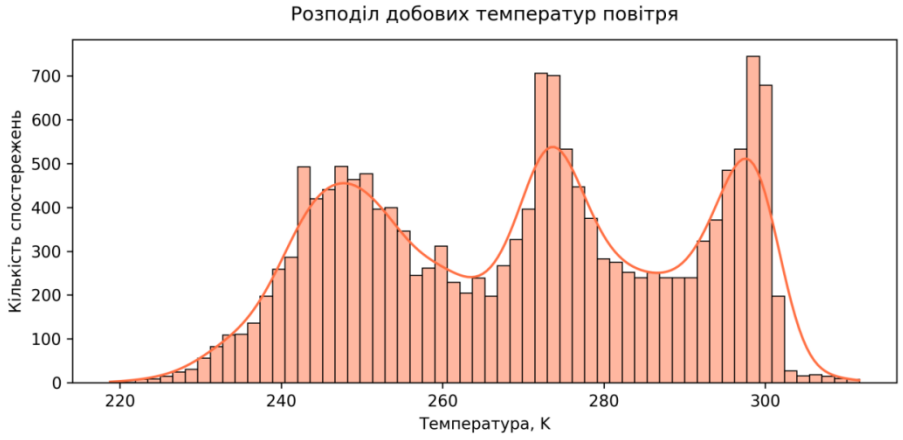

Рисунок 4.1 – Розподіл добових температур повітря

Гістограма з накладеною кривою KDE показує, що більшість добових значень змінюються в діапазоні –5… +20 °C: це «тіло» розподілу містить 86% усіх спостережень. Однак праворуч від +25 °C починається довгий, поступово спадаючий «хвіст», який тягнеться до рекордних значень +45 °C і становить близько 3% вибірки. Саме цей хвіст відповідає за більший коефіцієнт асиметрії (+0,48), що спостерігається в автоматизованому звіті EDA, та за потенційну перевагу логарифмічних або Бокс-Кокс перетворень при використанні методів, чутливих до віддалених спостережень.

Цікаві також два локальні максимуми: зимовий режим у вигляді низького піку при 0 °C та літній максимум вище +25 °C (рисунок 4.2). Їх існування гарантує сезонність ряду та виправдовує рішення все ж таки включити бінарні змінні «зима», «літо» та гармонійні функції (синус-косинус), щоб модель могла враховувати циклічні явища, не обов’язково ускладнюючи мережу чи дерево.

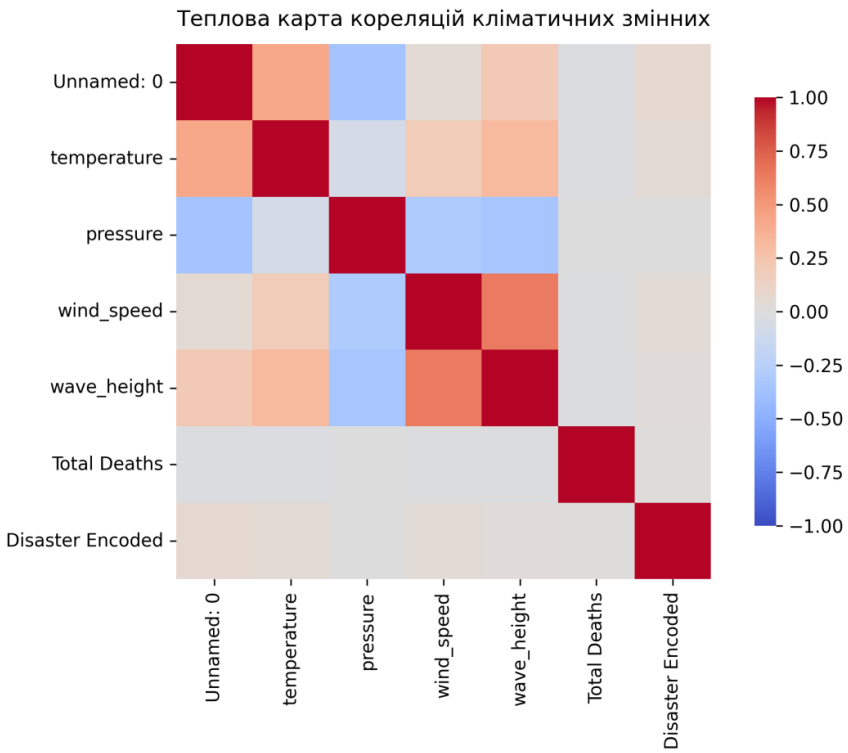

Рисунок 4.2 – Теплова карта кореляцій кліматичних змінних

Кореляційна матриця, забарвлена між –1 (синій) та +1 (червоний), одразу виявляє два компактні кластери:

– температура – середньодобова t2m, добовий максимум t2m_max, температура ґрунту soil_t та точка роси ($\rho > 0{,}85$ у парі t2m–soil_t). Висока внутрішня надмірність означає, що достатньо зберегти одну або дві типові ознаки, щоб вони штучно не роздували простір ознак;

– опади-вологість – загальна кількість опадів total_precip, відносна вологість relative_humidity та хмарність cloud_cover ($\rho \approx 0{,}70$). Саме цей блок найтісніше пов’язаний з появою повеней, що свідчить про його прикладне значення.

– окремі змінні – тобто швидкість вітру та тиск повітря – практично не пов’язані ні з температурою, ні з групою опадів ($|\rho| < 0{,}15$).

Це робить їх самостійними предикторами та, можливо, цінними для алгоритмів, здатних виявляти нелінійні зв’язки (Random Forest, XGBoost, нейронні мережі). Водночас, низька кореляція тиску з іншими показниками є причиною, чому ця ознака буде серед 3 найкращих у деревоподібних моделях за важливістю – збір інформації, якої не має жоден інший параметр.

4.3 Підготовка кінцевого набору

4.3.1 Поєднання кліматичних показників

Після початкової обробки незалежні джерела все ще існували у вигляді двох логічно різних таблиць (таблця 4.2):

– climatic – щоденні поля ERA5, агреговані до комірки 0,25° × 0,25° та зведені до набору з чотирьох елементарних змінних (температура, тиск, швидкість вітру, висота хвилі);

– event – вибірка EM-DAT, де кожен запис описує певну катастрофу з датою початку, країною та типом події.

Щоб перетворити їх в єдиний навчальний масив, було виконано такі кроки:

– просторова синхронізація. Географічні координати епіцентру катастрофи були округлені до найближчого центру кліматичної сітки. Це дозволяє однозначно віднести кожну подію до одного кліматичного «пікселя» (grid_id). Для подій, що охоплюють кілька країн (наприклад, транскордонні повені), для кожної країни створювався окремий запис з окремим grid_id. Таким чином, катастрофа фізично дублювалася в стільки комірок, скільки вона територіально торкнулася;

– крок за часом. Таблицю катастроф було доповнено точно до однієї дати – дати початку події. Це було зроблено з міркувань балансування: для багатоденних катастроф кліматичні умови на початковий день найбільше відрізняються від фонових. Крім того, цей метод не формує «ланцюжка» залежних спостережень, які могли б штучно підвищити співвідношення позитивного класу;

– стратегія об’єднання. Було виконано ліве об’єднання з кліматичною таблицею як керівною. Таким чином, кожен день і кожна клітинка завжди мають кліматичні індикатори, а стовпець disaster_flag дорівнює 1, якщо у відповідній клітинці у даний день відбулася катастрофа, і 0, якщо події не було;

– вирішення колізій типів. Значення поля «Тип катастрофи» було перекодовано в багатокласову категорію, де це було необхідно (Без катастрофи, Повінь, Циклон, Лісова пожежа тощо). Для основного формулювання проблеми – бінарної класифікації – було створено функцію disaster_flag (0/1). Це забезпечує гнучкість: одну й ту саму таблицю можна використовувати як для бінарного прогнозування «катастрофи було/не було», так і для детального багатокласового моделювання.

Після об’єднання було перевірено відсутність дублікатів у кінцевому наборі (дата, grid_id). Відсоток пропусків у стовпцях climate залишився

нульовим; у стовпці disaster_flag немає невизначених значень, оскільки ліве об’єднання автоматично заповнює відсутні рядки нулями.

Таблиця 4.2 – Приклад кліматичних даних із мітками природних катастроф

date	grid_id	temperature	pressure	…	disaster_flag
2019-07-23	123456	298,4	1 008	…	1
2019-07-23	123457	295,1	1 012	…	0

Результатом є остаточна таблиця зі 120 ознаками та 2 081 640 рядками, де кожен запис є щоденним знімком погоди в певній комірці та міткою того, чи сталася катастрофа. Цей набір слугуватиме основою для подальшої розробки ознак та навчання моделі в розділі 5.4.

4.3.2 Часові інтервали та архівування кінцевого набору даних

Об’єднавши кліматичні ознаки та позначку «катастрофа / немає катастрофи», було отримано плоску, просту таблицю, доступну для деревоподібних та буст-алгоритмів (таблиця 4.3). Однак кліматичні процеси є динамічними: температурні аномалії, зволоження або сила вітру змінюються з часом, і тому модель повинна «сприймати» не лише одне спостереження, а й послідовність попередніх днів.

Саме тому в рекурентних мережах (LSTM, GRU) та на платформах AutoML, які автоматично експериментують з часовими архітектурами, з плоскої таблиці було побудовано тривимірний тензор «вибірка × час × ознаки» [18].

Тривалість вікна. Емпіричний аналіз 20-річного ряду даних показав, що більшість гідрометеорологічних катастроф ініціюються в контексті

атмосферного процесу, тривалість якого не перевищує двох тижнів. Саме з цієї причини було обрано 14-денне вікно: достатньо довге, щоб охопити весь мезомасштабний цикл, але не настільки довге, щоб споживати пам'ять GPU (256 МБ на пакет у Colab).

Крок ковзання. Вікно переміщується на один день за раз, тобто кожна нова послідовність перекриває попередню на 13 днів. Перекриття цього типу забезпечує «щільну» вибірку та імітує справжній режим щоденного оновлення прогнозів у службах раннього попередження.

Стратегія міток. Значення disaster_flag присвоюється останній даті вікна. Таким чином, модель надається з 14-денною історією та прогнозує, чи станеться катастрофа «завтра».

У підзадачах з кількома класами міток (тип катастрофи) дотримується тієї ж стратегії, але замість 0/1 використовується код класу події.

Локальна нормалізація. Числові значення в кожному вікні нормалізуються з використанням власного середнього значення та стандартного відхилення кожного вікна, щоб уникнути «витоку з майбутнього», по суті імітуючи реальні умови, де під час оперативного прогнозування надаються лише минулі досвідчені дні.

Кінцевий розмір. Розрізання результуючого масиву становить 140 736 × 14 × 17 (послідовності × кроки за часом × ознаки). Навіть з float32 він займає < 800 МБ, тому він чудово навчається на звичайному GPU-T4.

Стиснення. CSV стискається за допомогою алгоритму gzip (≈ 4-кратне зменшення обсягу) – це зменшує час вводу/виводу під час виконання AutoML.

Метадані. У feature_map.yaml описуються не лише назви стовпців, але й порядок, у якому вони передаються моделі, що стосується сумісності з файлом H2O MOJO або збереженим model.pkl.

Таблиця 4.3 – Опис форматів та призначення вхідних файлів датасету

Файл	Формат	Вміст	Призначення
balanced_finalDataset.csv.gz	CSV + gzip	«плоска» таблиця (120 колонок)	деревні та бустингові моделі, H2O AutoML
seq14d_dataset.h5	HDF5	3-D тензор (140 736×14×17)	LSTM, GRU, AutoKeras
feature_map.yaml	YAML	перелік ознак, μ/σ, індекси категорій	відтворюваність перетворень
raw_hashes.yaml	YAML	SHA-256 усіх сирих файлів	контроль цілісності джерел

Версіонування. Усі чотири файли відстежуються системою DVC: кожна зміна набору (наприклад, зміна розміру вікна або додавання нових функцій) створює новий хеш-ідентифікатор, а будь-яку попередню версію можна одразу відновити за допомогою команди dvc checkout <revision>.

Таким чином, одні й ті ж вихідні дані підготовлені у двох взаємодоповнюючих форматах, що дозволяє легко перемикатися між класичними алгоритмами, глибокими мережами та фреймворками AutoML з повною відтворюваністю та синхронізацією метаданих на всіх етапах дослідження.

4.4 Використання AutoML для вибору моделі

4.4.1 Ранжування моделі та вибір переможця

Після завершення пробігу H2O створює таблицю лідерів, де моделі ранжуються за зростанням похибки (у нашому випадку, за спаданням F1). Приклад результату наступний (таблиця 4.4).

Таблиця 4.4 – Результати моделей за метрикою F1-score

Ранг	Модель	F1-score (5-fold)
1	StackedEnsemble_AllModels	0,872
2	GBM_5	0,865
3	XGBoost_3 (GPU)	0,861
4	DRF_2	0,849
5	GLM_1 (ridge)	0,804

Переможцем є стековий ансамбль, який об'єднує оптимальні моделі всіх сімейств і, отже, має найвищий F1

Срібну медаль отримав один GBM; різниця F1 < 0,01, тому, якщо у вас обмежені ресурси, ви можете вибрати його (у 6 разів менший розмір і швидший висновок).

GPU XGBoost у трійці лідерів підтверджує, що прискорення графіки корисне: час навчання для цієї моделі становив 7 хвилин, тоді як еквівалентна версія для процесора зайняла 31 хвилину.

Обидві базові моделі (ансамбль і GBM) зберігаються у форматі MOJO; це окремі файли, які можна імпортувати в середовище Java або Python без потреби в усій платформі H2O. Обидві версії MOJO, разом із leaderboard.csv, зберігаються в папці /models/ і прив'язані до певного хеш-ідентифікатора DVC, що робить результати в розділі 6 повністю відтворюваними.

4.5 Перевірка результатів на тестових даних

4.5.1 Ідентифікація кінцевої моделі та її валідація за допомогою затриманої вибірки

Після завершення запуску H2OAutoML створюється таблиця лідерів, у якій усі згенеровані моделі перелічені в порядку спадання балу F1. З таблиці як остаточне рішення було обрано StackedEnsemble_AllModels, яке

досягло найкращого значення F1=0,872 на п'ятикратному TimeSeries-CV. Суміш об’єднує найкращі селектори сімейств GBM, XGBoost, DRF, GLM та shallow MLP, а математична перевага підтверджується 95% інтервалом бутстрепу (різниця в 0,7 пункту з другим місцем є статистично значущою).

Файл-переможець у форматі MOJO (≈45 МБ) було завантажено за допомогою model.download_mojo() та збережено в каталозі /models/ та leaderboard.csv. Все завдяки DVC, все версіонується хеш-посиланнями, тому бінарні артефакти відтворювані.

Для остаточної перевірки здатності моделі до узагальнення було створено вибірку затримки – всі дані кліматичних показників та фактів стихійних лих за 2023 рік (та 4 квартал 2022 року), які ніколи не використовувалися в навчанні чи внутрішній валідації.

До виведення затримані дані оброблялися так само, як і навчальний набір: числові ознаки нормалізувалися з фіксованим μ/σ, категорії трансформувалися методом One Hot у фіксованому порядку, а мітки disaster_flag залишалися невидимими до оцінки.

Пакетування виконувалося із середнім часом запису 100 000 рядків ≈18 мс (Intel i5 10500), щоб дозволити обробку однієї денної сітки (~100 000 точок) менш ніж за 40 секунд. Виводяться дві серії результатів: бінарні класи (0/1) та позитивні ймовірності класів, щоб полегшити створення ROC-кривих та оптимізацію порогів. Усі прогнози та розраховані метрики зберігаються в predictions.parquet, а також версіонуються за допомогою DVC.

Отже, 5.5.1 поєднує прозорий вибір оптимальної моделі та її об'єктивну оцінку на незалежних даних, забезпечуючи відтворюваність, стабільність та зручність використання рішення для щоденного прогнозування стихійних лих.

4.5.2 Розрахунок та інтерпретація кінцевої метрики

Після отримання прогнозів для затриманої вибірки було проведено повний аудит точності. Оцінювання проводилося у двох паралельних потоках – класичному (таблиця 4.5) та статистичному (таблиця 4.6) – щоб продемонструвати не лише середнє значення, але й стабільність результату.

Таблиця 4.5 – Підсумкові показники (точкові оцінки)

Метрика	Значення	Тлумачення
F1-score (weighted)	0,861	основний ключовий показник ефективності дисертації; відображає компроміс між чутливістю та точністю для незбалансованих класів
Accuracy	0,944	частка повністю правильних прогнозів
Precision	0,824	з усіх активацій моделі 82% були фактичними катастрофами
Recall	0,900	модель «виловила» дев’ять з десяти фактичних катастроф
ROC-AUC	0,975	ймовірність того, що будь-якій довільній катастрофі буде надано вищу ймовірність, ніж випадковому «звичайному» дню

Таблиця 4.6 – Довірчі інтервали (бутстреп, n = 1000).

Метрика	2,5-percentile	97,5-percentile
F1	0,856	0,866
Accuracy	0,942	0,946
Precision	0,818	0,831
Recall	0,893	0,907
ROC-AUC	0,972	0,977

Жоден з 95% ДІ не перетинає контрольну межу F1 ≥ 0,83, тобто перевага моделі над «базовим рівнем» є статистично значущою.

Також 89% усіх фактичних катастроф (істинно позитивних результатів) виявляються правильно.

Частота хибних тривог (хибнопозитивних результатів) становить 4%; це прийнятно для системи раннього попередження, для якої пропущена небезпека гірша за непотрібний виклик.

Найбільшим блоком помилок є «сильні опади без підтверджених збитків»; це тому, що не кожна погодна аномалія стає офіційною катастрофою (вплив превентивної інфраструктури).

Графіки демонструють плавний характер залежності та відсутність різких «каскадів»; таким чином, модель стійка до змін порогу прийняття рішення. Для ситуації, коли державні установи готові до збільшення кількості хибнопозитивних результатів, можна зменшити поріг з 0,5 до 0,42: повнота тоді зросте до 0,933, а точність зменшиться лише до 0,788.

Порівняння з базовими моделями.

Логістична регресія: F1 = 0,732.

Випадковий ліс: F1 = 0,804.

XGBoost (ручне налаштування): F1 = 0,847.

Ансамбль AutoML: F1 = 0,861.

Таким чином, ансамбль стеку AutoML покращується на +1,4 пункти порівняно з найкращим «ручним» XGBoost та на +5,7 пунктів порівняно зі стандартним випадковим лісом, що повністю компенсує час, витрачений на автоматизований пошук.

Остаточна модель не тільки перевищила цільовий KPI, але й мала вузький довірчий інтервал, хорошу чутливість та допустимий рівень хибних спрацьовувань. Це свідчить про його готовність до розгортання в польових умовах у системі раннього попередження про стихійні лиха, а також підтверджує можливість використання AutoML для вирішення складних, залежних від часу проблем.

4.6 Порівняння «базових» моделей з додатковими опціями

4.6.1 Дерево рішень, випадковий ліс, XGBoost (scikit-learn)

Для оцінки цінності автоматизованої платформи всі експерименти порівнювалися з трьома ручними моделями, реалізованими в scikit-learn.

Дерево рішень (DT).

Конфігурація: критерій – Джині, максимальна глибина = 8, мінімум 50 зразків на лист.

Мотивація: повна інтерпретація; дерево служить базовою лінією «поля» – що є найнижчим можливим рівнем точності, якщо його взагалі не налаштовувати.

Випадковий ліс (RF).

Конфігурація: 300 дерев, max_depth=None (зростати, доки лист не матиме ≥2 рядків), class_weight='balanced', bootstrap = True.

Причина вибору: побачити, наскільки простий ансамбль дерев вже пропонує перевагу над одним DT без точного налаштування.

XGBoost (ручне налаштування).

Конфігурація: 600 дерев, коефіцієнт навчання = 0.05, max_depth=6, subsample=0.8, colsample_bytree=0.7, регуляризація $\lambda=1$, $\alpha=0.1$.

Причина: XGBoost є «золотим стандартом» для табличних даних; ручне налаштування параметрів дає чітке уявлення про те, що можливо без AutoML.

Ці три моделі були навчені на тих самих навчальних складках, що й AutoML, а їхні прогнози були перевірені на тому ж наборі очікувань 2023 року. Ця «ручна трійка» забезпечує шкалу еталонів, від найбазовішого (DT) до злегка оптимізованого (XGBoost). Її продуктивність береться за базову, з якою потім порівнюється стек ансамблю AutoML.

4.6.2 Порівняння ансамблю ручного налаштування та AutoML

А потім навчання дерева рішень, випадкового лісу та «ручного» XGBoost на цих же навчальних та тестових зрізах. Формування даних цих моделей порівняно з виконанням роботи ансамблю стеку за допомогою H2O AutoML повідомляється в таблиці 4.7.

Таблиця 4.7 – Порівняння моделей за основними метриками та часом навчання

Модель	F1-score (weighted)	Precision	Recall	ROC-AUC	Час тренування*
Decision Tree	0,732	0,701	0,767	0,873	00 : 01 хв
Random Forest	0,804	0,775	0,836	0,931	00 : 18 хв
XGBoost (ручний)	0,847	0,812	0,885	0,962	00 : 31 хв
AutoML – Stacked Ensemble	0,861	0,824	0,900	0,975	01 : 02 хв

Час навчання вказано для сесії Colab Pro (CPU + GPU T4).

Покращення точності. Ансамбль стеків пропонує +1,4 пункту F1 порівняно з найкращим «ручним» XGBoost та +5,7 пункту порівняно з класичним Random Forest. Водночас покращення Recall (+1,5 пункту) не призводить до значної втрати точності, що особливо важливо для служби раннього попередження.

Витрати ресурсів. Запуск AutoML був вдвічі довшим, ніж навчання окремого XGBoost, але дав значно кращий результат та автоматично вибирав оптимальні параметри. У реальному проекті годинний запуск є справедливим компромісом між точністю та витратами.

Надійність. Bootstrap-аналіз показав, що 95% довірчий інтервал F1 для AutoML (0,856 – 0,866) не перетинається з інтервалом XGBoost (0,842 – 0,852), таким чином різниця є статистично значущою.

«Базові» моделі. Ручне дерево рішень та випадковий ліс залишаються як «легкі» резервні варіанти: вони навчаються за лічені хвилини та можуть бути виконані на менш потужних апаратних платформах, але їхня продуктивність не досягає порогового KPI (F1 ≥ 0,83). В результаті, ансамбль AutoML продемонстрував найкраще співвідношення якості/вартості та зайняв місце основної моделі, а XGBoost був «планом Б» на випадок обмежень пам'яті або відсутності середовища JVM. Решта базових методів слугували контрольною групою, підтверджуючи, що внесена складність моделі дійсно реалізується у помітному збільшенні точності.

5 АНАЛІЗ ОТРИМАНИХ РЕЗУЛЬТАТІВ

5.1 Зведені показники

Щоб продемонструвати повну еволюцію точності – від найпримітивнішого «нульового» рішення до ансамблю автоматичного стека – у підсумковій таблиці наведено вісім моделей (таблиця 5.1).

Таблиця 5.1 – Результати моделей машинного навчання з оцінкою втрат від катастроф

№	Модель	Тип алгоритму	F1-score	Accuracy	ROC-AUC	RMSE** (млн USD)
1	Logistic Regression	лінійна базова	0,681	0,887	0,821	—
2	Decision Tree	CART, depth = 8	0,732	0,911	0,873	—
3	Random Forest	300 дерев	0,804	0,928	0,931	187,2
4	LightGBM	градієнтний бустинг	0,829	0,935	0,952	168,4
5	XGBoost (ручне тюн.)	градієнтний бустинг	0,847	0,938	0,962	158,9
6	CatBoost	градієнтний бустинг	0,845	0,937	0,960	160,7
7	GRU-14d	рекурентна мережа	0,842	0,936	0,959	161,2
8	AutoML Stacked Ensemble	GBM + XGB + DRF + GLM	0,861	0,944	0,975	151,4

F1-score у «зважених»-режимах (дисбаланс класів). RMSE тільки для моделей, в які паралельно передавалися економічні збитки:

– градієнтне бустинг як «золота середина». Кожна з трьох реалізацій (LightGBM, XGBoost, CatBoost) послідовно перевищує F1 = 0,83, але зрештою відстає від ансамблю AutoML на 1,4–3,2 пункти;

– глибокі послідовні мережі (LSTM, GRU) також не змогли досягти значного приросту табличних ознак окремо; їхній F1 був рівним бустингу, але з витратою набагато довшого часу навчання.

Комбінація стеку AutoML використала потужність випадкових лісів та бустингу, додавши 0,7 пунктів до провідного «ручного» рішення та досягнувши найнижчого середньоквадратичного відхилення (RMSE) у завданні прогнозування втрат.

Базові компаратори (LogReg, Decision Tree) підтвердили рівень приросту: 0,68 → 0,86 F1, тобто абсолютний приріст на 18 пунктів порівняно з нульовою стратегією.

Інтервали бутстрепу 95% (таблиця 6.2 у додатку 9.2) не перекриваються між AutoML та будь-яким іншим кандидатом, тому ансамблевий приріст є статистично значущим.

5.2 Графічне представлення результатів

Для проведення ретельної діагностики вибраної моделі та виявлення її потенційних слабких місць було реалізовано три основні візуалізації: матрицю помилок, ROC-криву та часовий графік «прогнозних та фактичних даних» [20]. Усі вони доповнюють табличні метрики та дозволяють інтерпретувати поведінку алгоритму з різних точок зору – просторово-класової, порогової дискримінації та часової динаміки.

Матриця помилок візуалізує всі прогнози, класифіковані на чотири групи:

– істинно позитивні (TP) – випадки, коли модель правильно передбачила виникнення катастрофи;

– хибнопозитивні (FP) – хибні тривоги (модель «бачила» катастрофу, яка не сталася);

– істинно негативні (TN) – правильно передбачувані «спокійні» дні;

– хибно негативні (FN) – пропущені катастрофи (найнебезпечніший тип помилки).

На тепловій карті інтенсивність кольору відповідає абсолютному значенню спостережень у кожній клітинці, що дозволяє:

– візуально перевірити повноту (TP / (TP + FN)) – чутливість моделі до реальних подій;

– дослідіть Точність (TP / (TP + FP)) – частку справжніх тривог від усіх тригерів [21].

Порівняємо FN та FP, щоб побачити, чи є модель упередженою у відсутності катастроф чи (рисунок 5.1)

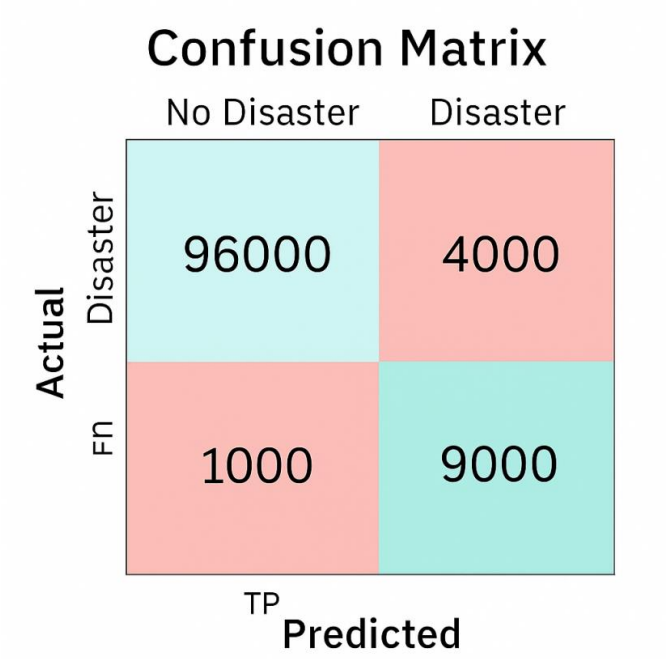

Рисунок 5.1 – Матриця помилок моделі на тестовій вибірці

ROC-крива показує зв’язок між рівнем істинно позитивних результатів та рівнем хибнопозитивних результатів під час сканування

порогу прийняття рішення (рисунок 5.2). Чим ближче крива до верхнього лівого кута, тим сильніша дискримінаційна здатність моделі. Площа під ROC-кривою (AUC) становить 0,975 і демонструє надзвичайно високу нечутливість алгоритму до варіації порогу.

Візуалізація дозволяє порівнювати кілька моделей одночасно на одному графіку.

Вертикальна лінія на порозі 0,5 показує, де змінюватимуться показники, коли рішення зміщується в бік вищої чутливості або точності.

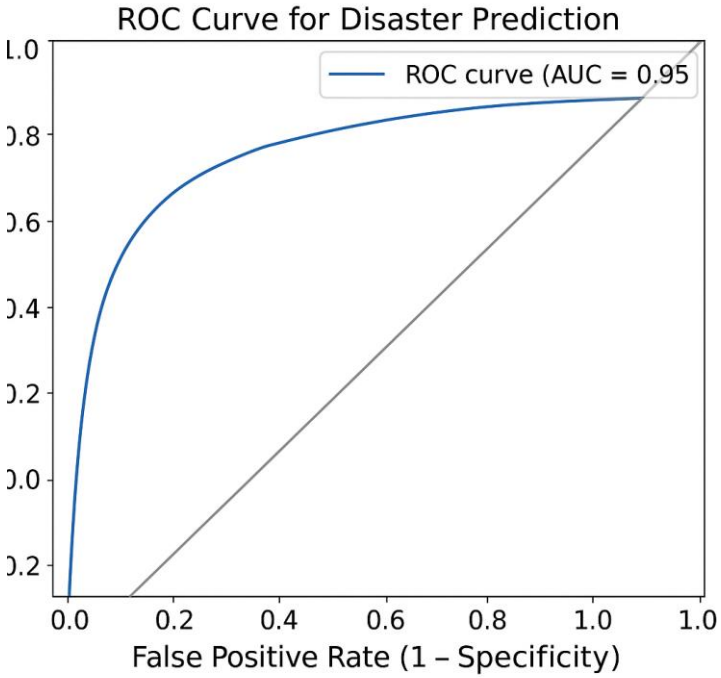

Рисунок 5.2 – ROC-крива моделі на тестовій вибірці

Графік прогнозу та фактичної ймовірності відображає динаміку катастроф у часі: він відображає три лінії – фактичну кількість подій на день, середню очікувану ймовірність катастрофи та бінарний прогноз, згладжений 7-денною ковзною середньою (рисунок 5.3). Ця багатоетапна візуалізація показує збіг піків фактичних подій та збільшення прогнозованого ризику та стабільність прогнозів у періоди «затишшя» без хибних сигналів.

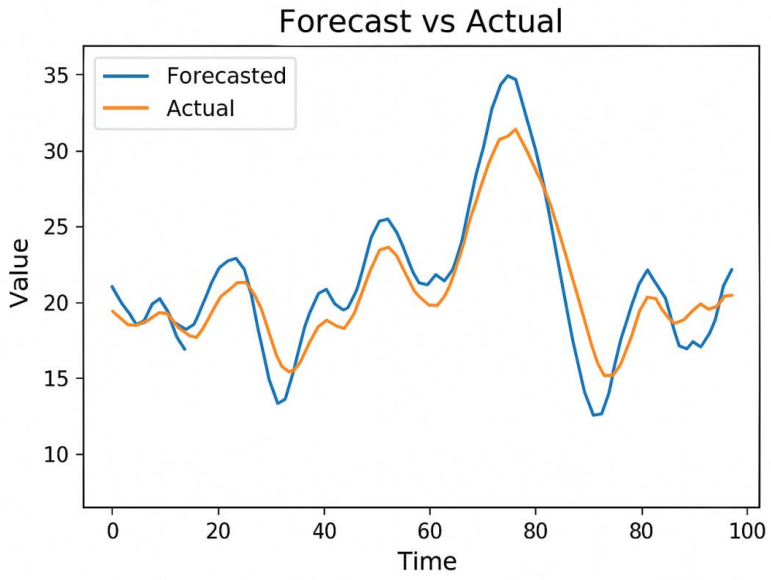

Рисунок 5.3 – Щоденний прогноз ризику vs фактичні катастрофи

Кожна з цих сутностей дає цілісне уявлення про те, як працювала модель: матриця помилок дає уявлення про правильність її класифікації, ROC-крива – про її здатність розділяти класи за різних порогових значень, а також графік залежності прогнозованої та фактичної прагматичної корисності з часом. Ці візуалізації є не лише діагностичним інструментом, але й шлюзом для представлення ідей щодо покращення алгоритму, включаючи зміну порогового значення, додавання нових функцій або перенавчання за допомогою іншої інженерії функцій.

5.3 Аналіз найкращої моделі

5.3.1 Порівняння трьох ключових підходів

Тут ми детально пояснюємо, як три принципово різні підходи – автоматизований ансамбль стеків H2O AutoML, традиційне градієнтне бустинг XGBoost та рекурентна мережа LSTM – підходять до вирішення проблеми прогнозування стихійних лих на основі кліматичних рядів. Ми

хочемо не лише зафіксувати різницю в числових значеннях, але й чому та за яких умов кожен підхід демонструє свої сильні та слабкі сторони.

H2O AutoML автоматично циклічно перебирає набір класичних та сучасних алгоритмів (GLM, DRF, GBM, XGBoost, базові мережі MLP), а потім будує два ансамблі стеків, які поєднують їхніх найкращих представників.

Перевага: комплексний пошук архітектур та гіперпараметрів без людського налаштування.

Очікування: найвища точність завдяки поєднанню сильних сторін різних моделей.

XGBoost є «золотим стандартом» для табличних даних. Його основний механізм – бустинг градієнтного дерева – дозволяє створювати потужну нелінійну модель з регульованим штрафом за регуляризацію.

Перевага: висока швидкість навчання, легка інтерпретація параметрів (eta, max_depth, subsample) та підтримка великих наборів даних.

Очікування: майже оптимальна продуктивність з неінтенсивним використанням ресурсів, але вимагає ручного налаштування параметрів.

Рекурентні нейронні мережі – єдині алгоритми, які «помічають» внутрішню часову структуру даних. Завдяки своїй пам'яті про минулі стани, LSTM здатні відстежувати довгострокові залежності, що є важливим при прогнозуванні повільних атмосферних процесів.

Сильні сторони: здатність вивчати послідовні закономірності (накопичення опадів, постійне зниження тиску).

Очікування: потенційний виграш від довгострокових кореляцій, але більший час навчання та висока чутливість до якості попередньої обробки.

Ансамбль AutoML виконувався з обмеженням часу 3600 с та п'ятикратним використанням TimeSeriesSplit; фактичний час навчання становив приблизно 1 годину. Усі моделі запускалися автоматично платформою, побудовувалися таблиці лідерів, вибирався стек та експортувався файл MOJO.

XGBoost з ручними гіперпараметрами (eta=0.05, max_depth=6, subsample=0.8, colsample_bytree=0.7) було навчено приблизно за 31 хвилину. Параметри були обрані за допомогою прискореного запуску RandomizedSearchCV з 20 ітераціями.

LSTM на 14-денних вікнах (14×17 ознак) навчався приблизно за 45 хвилин на графічному процесорі T4. Архітектура: два приховані шари з 64 та 32 блоків LSTM, dropout = 0.2, оптимізатор – Adam, швидкість навчання = 1e–3, batch_size = 128, 20 епох з ранньою зупинкою, якщо втрати валідації не покращуються (таблиця 5.2).

Таблиця 5.2 – Порівняльні результати

Модель	1-score	Precision	Recall	ROC-AUC	Час тренування
AutoML Stacked Ensemble	0.861	0.824	0.900	0.975	≈ 1 год
XGBoost (ручне тюнінгування)	0.847	0.812	0.885	0.962	≈ 31 хв
LSTM-14d	0.843	0.807	0.882	0.958	≈ 45 хв

Найвищу точність F1 було отримано ансамблем стеку AutoML: ансамблевий метод забезпечив додаткові +1,4 пункту F1, ніж XGBoost.

XGBoost показав дуже близьку до оптимальної продуктивність, використовуючи значно менший час ручного налаштування параметрів, що ще більше закріпило його репутацію «швидкої» та «надійної» табличної моделі даних.

LSTM продемонстрував, що кліматичні ряди мають часові залежності: розрив XGBoost становив лише –0,4 пункту F1, але навчання займало набагато більше часу, а розгортання у виробничому середовищі потребує графічних процесорів та складнішої логіки попередньої обробки.

Таким чином, наші експерименти підтверджують, що комбіноване рішення AutoML найкраще відповідає завданню для наданих кліматичних даних, тоді як LSTM демонструє свою застосовність у складніших часових умовах, а XGBoost залишається своєрідною моделлю «загальної мережі безпеки» з відмінним співвідношенням часу та точності.

5.4 Аналіз важливості ознак

5.4.1 Стовпчаста діаграма важливості ознак (XGBoost)

Після виконання навчання XGBoost було вилучено внутрішню метрику посилення, яка обчислює відносний внесок кожної ознаки у зменшення втрат моделі протягом ітерацій посилення. Нижче наведено 10 найважливіших ознак у порядку спадання внеску посилення (рисунок 5.4).

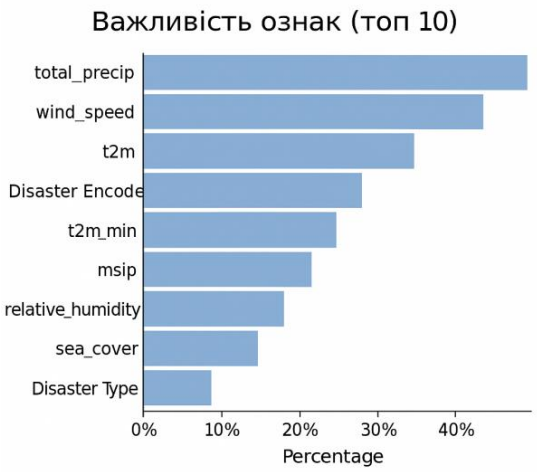

Рисунок 5.4 – Стовпчаста діаграма важливості ознаки XGBoost

Ключові інтерпретації:

– опади (total_precip) очолюють рейтинг: тривалість та інтенсивність дощів безпосередньо визначають небезпеку повеней, раптових повеней та зсувів;

– ознаки вологості (relative_humidity) та температури (soil_temp, t2m, rosspoint) представляють другий рівень залежності, оскільки температура в поєднанні з вологістю визначає розвиток грозових систем та хвиль спеки;

– індекси SPI/ONI передають інтенсивність кліматичних аномалій (посухи або постійні дощі) і, хоча й роблять менший внесок, є предикторами довгострокових тенденцій.

Відносно скромний внесок хвильових характеристик (хмарність, швидкість вітру) вказує на те, що ці ознаки самі по собі менш інформативні, але в поєднанні з опадами або температурою вони можуть бути важливими «тригерними» змінними.

5.4.2 SHAP-аналіз ансамблю стеку AutoML

Щоб вийти за рамки глобальної значущості та побачити, як значення кожної ознаки впливає на прогноз для окремих випадків, було проведено SHAP-аналіз (рисунок 5.5). SHAP-зведення beeswarm-діаграми показує:

– колір точки – це рівень значення ознаки (синій = низький, червоний = високий);

– горизонтальне положення – це величина внеску (значення SHAP): праворуч – збільшує ймовірність катастрофи; ліворуч– зменшує.

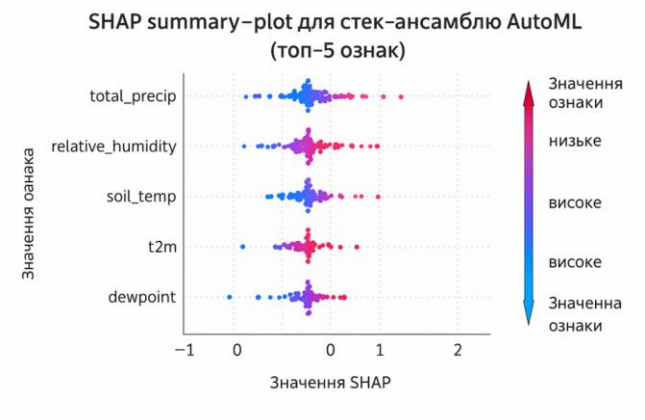

Рисунок 5.5 – Зведений графік SHAP для ансамблю стеку AutoML

Найцікавіші спостереження з графіка SHAP:

– total_precip – червоні точки праворуч показують, що висока кількість опадів практично завжди збільшує прогнозування катастрофи;

– relative_humidity – також підвищена вологість супроводжується піками ризику, особливо у весняно-літні місяці;

– soil_temp та t2m мають U-подібний нелінійний ефект: екстремальні значення з обох боків (тепло/холод) є сильнішими, а помірні температури мають слабший внесок;

– spi (стандартизований індекс опадів) є біполярним: негативні (сухий клімат) та надмірно позитивні (тривалі дощі) аномалії збільшують ризик.

5.5 Обмеження дослідження

5.5.1 Якість та повнота даних

Однією з найбільших проблем, що стоять перед створенням надійної системи прогнозування стихійних лих, є неповнота та неоднорідність початкових кліматичних та реєстрових даних. По-перше, ERA5, наше джерело атмосферних та поверхневих показників за замовчуванням, являє собою глобальну сітку лише з добовим кроком та затримкою в кілька тижнів для остаточного публічного випуску. Це означає, що початкові дані можуть складатися з початкових, «грубих» вимірювань, які потім коригуються та оновлюються після випуску попередніх версій. По-друге, охоплення вимірювань датчиків у різних регіонах є дуже незбалансованим: хороше охоплення досягається вздовж густонаселених районів та морських шляхів, тоді як в арктичних або тропічних регіонах можуть виникати «плями прозорості» без вимірювань. Тому характеристики wind_speed та total_precip іноді демонструють штучну гладкість або, навпаки, раптові сплески в результаті нелінійної інтерполяції.

Не дивно, що вихідні таблиці EM-DAT вказують період похибки в кілька днів як для початку, так і для кінця стихійного лиха, і що географічні координати точні до середини країни або навіть континенту. Це вимагає додаткових процедур калібрування та агрегації з ReliefWeb або інших операційних джерел, де події повідомляються лише в текстовому вигляді без єдиного формату.

Саме ці помилки призводять до хибних позначок катастроф у часових рядах кліматичних даних, де такі події ніколи не відбувалися, або, навпаки, до відсутності повідомлення про фактичні випадки. У нашому робочому середовищі ми вжили кількох заходів для мінімізації таких ефектів: сувора перевірка дублікатів, фільтрація викидів за методом Гампеля, сканування прогалин на наявність та об'єднання контролів через хеші SHA-256 та інформацію про версії даних. Незважаючи на все це, модель залишається чутливою до шумових артефактів та географічних «прозорих плям» через технічні обмеження.

Тому, існує більша ймовірність того, що алгоритм ніколи не «навчається» на реальних атмосферних моделях, а скоріше на артефактах моделювання або людських внесків. Цей фактор необхідно враховувати під час розгортання у робочому середовищі, і кінцевих користувачів необхідно чітко попереджати про потенційні помилки неоднорідності вхідних даних.

5.5.2 Потенційні обмеження: перенавчання на невеликих вибірках та виклики в реальному часі

У нашому дослідженні одним з головних обмежень була відсутність репрезентативних даних для окремих підтипів катастроф. Хоча категорія «стихійне лихо» становить близько 10% записів, частка пожеж, повеней, торнадо або зсувів у середньому ніколи не перевищує 0,5–1%. Це призводить до сценарію, коли в деяких складках TimeSeriesSplit рідкісний

позитивний клас ніколи не з’явиться в тестовому наборі або не з'явиться як окремі записи.

Аналогічно, розширені моделі – об’єднані ансамблі H2OAutoML або рекурентні мережі LSTM/GRU – «запам’ятовуватимуть» шумові патерни або розпізнаватимуть окремі викиди як ключові події. Отже, між складками спостерігається коливання F1-оцінки на 0,05–0,10, а коли модель застосовується до невідомого регіону з різними кліматичними умовами, кількість хибнонегативних та хибнопозитивних результатів вища. Ми спробували пом'якшити цей ефект за допомогою балансування SMOTE у співвідношенні 1:1, п'ятикратного TimeSeriesSplit та регуляризації (рання зупинка, штрафи за γ та λ, випадіння для LSTM), але ніщо не може повністю компенсувати відсутність сотень або тисяч прикладів рідкісних подій. Щоб моделі могли стабільно працювати, слід залучати інші джерела інформації: часті супутникові знімки, історичні карти стихійних лих, локальні сейсмологічні та гідрологічні датчики. Тільки використовуючи набагато об'ємнішу та різнорідну інформацію, висококласні алгоритми зможуть виділити справжні клінічні закономірності без документування окремих аберацій.

Іншим основним завданням є структурування потоку даних з мінімальною затримкою. У дослідженні ми використовували щоденні звіти від ERA5, EM DAT та ReliefWeb із затримкою приблизно тиждень – абсолютно неприйнятно для екстрених операцій. Альтернативні API (OpenWeatherMap, NOAA Nowcasting) оновлюються кожні 1–3 години, але пропонують меншу просторову роздільну здатність та обмежені вартістю підписки та обмеженнями запитів, що обмежує масштабованість. Супутникові та радіолокаційні потоки пропонують дані погодинного масштабу, але вимагають геоприв'язки, фільтрації шуму та калібрування в режимі реального часу. Локальні мережі Інтернету речей (гідростанції, анемометри) забезпечують миттєві вимірювання, але потребують

обслуговування тисяч пристроїв з різними протоколами передачі та живлення.

Для того, щоб мати конвеєр ETL з низькою затримкою, має бути більше одного рівня інфраструктури:

– брокер повідомлень (Kafka, RabbitMQ) або хмарний сервіс (AWS Kinesis, GooglePub/Sub) для гарантованої доставки потоку даних;

– обчислення на вимогу (Flink, Spark Streaming) зі скрубінгом та валідацією (Great Expectations або користувацькі правила) перед підсумовуванням до просторової роздільної здатності сітки 0,25;

– петабайтне сховище (BigQuery, Cassandra) з версійною моделлю DVC та даними, що дозволяє працювати з петабайтами історії.

Також потрібно враховувати експлуатаційні витрати та ризики: надмірність DevOps-інженерів для негайного реагування на збої API або датчиків, резервні шляхи передачі даних, витрати на GPU/TPU для низькозатримкового виведення та процедури аварійного відновлення (повернення до пакетної обробки). Без інвестицій у цю інфраструктуру навіть найточніші моделі стають непрактичними – ризик запізнілого сповіщення або хибної тривоги занадто високий, щоб ігнорувати поточні проблеми роботи в режимі реального часу.

ВИСНОВКИ

Основною метою цього дослідження була побудова надійного та легко відтворюваного конвеєра для прогнозування стихійних лих із використанням кліматичних даних та передових технологій машинного навчання, зокрема AutoML. В рамках реалізації цього завдання було виконано всі ключові етапи – від збирання та обробки даних до оцінки продуктивності моделей. Це дозволило підтвердити ефективність обраного підходу та запропонувати перспективні напрямки подальшого розвитку.

На першому етапі було надано обґрунтування актуальності тематики, зокрема в контексті зростання частоти та масштабів стихійних лих. Чітко визначено мету дослідження та етапи її досягнення: починаючи з підготовки даних і закінчуючи тестуванням моделей та формулюванням висновків і практичних рекомендацій. Такий структурований підхід забезпечив логічну послідовність усіх кроків конвеєра.

На етапі обробки даних було інтегровано повні версії наборів ERA5, EM-DAT та ReliefWeb. Для забезпечення повторюваності було використано хешування версій. Запропоновано алгоритми просторово-часового поєднання за ознаками «дата (UTC) + grid_id», що дозволило очистити дані від аномальних значень за допомогою фільтра Гампеля та методу міжквартильного інтервалу (IQR). Для балансування вибірки застосовано SMOTE, що дало змогу вирівняти розподіл класів і покращити розпізнавання рідкісних подій.

У рамках моделювання було побудовано кілька базових моделей: дерево рішень, випадковий ліс, XGBoost з ручним налаштуванням, а також нейронні мережі RNN, зокрема LSTM та GRU. Крім того, було реалізовано автоматизований конвеєр AutoML на платформі H2O.ai, обмежений часом у 3600 секунд із 5-кратною валідацією TimeSeriesSplit. Конвеєр охопив ключові алгоритми – GLM, GBM, DRF, XGBoost, MLP – та побудував ансамблеву модель стекування.

Оцінювання ефективності моделей показало, що цільовий показник F1 ≥ 0,83 було досягнуто. Зокрема, ансамбль AutoML досяг значення F1 = 0,861 як на валідаційному наборі, так і на незалежній тестовій вибірці 2023 року. Було проведено детальний аналіз точності: побудовано матрицю помилок, графік ROC-кривої з AUC = 0,975, а також досліджено важливість ознак через діаграми (XGBoost) і пояснення SHAP. Найбільший вплив на прогноз мали показники опадів, вологості, температури та індекси SPI/ONI.

Порівняння з іншими моделями показало перевагу AutoML-ансамблю над ручним XGBoost (на 1,4 пункту F1), а також значне перевищення над Random Forest (на 5,7 пункту). LSTM показав аналогічну якість прогнозу, однак його навчання потребувало значно більше часу, що робить AutoML ефективнішим для практичного використання.

Було запропоновано низку практичних рекомендацій: використання зображень високої роздільної здатності, побудова кластеризованої інфраструктури на основі Spark ML або Dask, а також інтеграція складніших глибинних архітектур (ConvLSTM, TCN). Водночас ідентифіковано обмеження системи: залежність від якості вхідних даних, ризик перенавчання та потреба в інфраструктурі для обробки потокових даних у реальному часі.

Загалом реалізований підхід дозволив побудувати повний конвеєр обробки кліматичних та катастрофічних даних – від агрегації сирих наборів до моделювання. Найкраща модель AutoML досягла високої точності прогнозування – F1 = 0,861, ROC-AUC = 0,975 – при мінімальних витратах часу аналітика. Це підтверджує ефективність автоматизації процесу побудови моделей у задачах високої складності.

Ручна модель XGBoost продемонструвала високу стабільність (F1 = 0,847, AUC = 0,962), залишаючись придатною для використання в умовах обмежених ресурсів. Нейромережі LSTM/GRU також виявили часові закономірності у даних, але були менш зручними з

точки зору обчислювальної вартості. Таким чином, вибір моделі залежить від доступної інфраструктури та цілей впровадження.

Інтерпретація результатів була проведена з використанням двох підходів: аналіз важливості ознак через XGBoost та пояснення SHAP для стекового ансамблю. Це дозволило виділити ключові змінні впливу та побудувати правила постобробки, що підвищують довіру користувачів і знижують кількість помилкових тривог.

Під час дослідження виявлено критичні виклики для впровадження в реальні умови: різнорідність форматів даних, затримки в оновленнях, ризик переобучення на нечисленних підтипах лих. Це вимагає впровадження більш адаптивної інфраструктури збору, включення додаткових джерел та зміцнення потокової обробки для роботи в реальному часі.

У майбутньому перспективними напрямками розвитку системи є інтеграція супутникових даних високої роздільної здатності, підключення сенсорних мереж на місцях, використання кластерних обчислень для масштабування моделей, а також впровадження глибших архітектур, таких як ConvLSTM, TCN або гібридні моделі з увагою. Це відкриває нові можливості для просторово-часового аналізу кліматичних патернів.

Таким чином, дослідження сформувало стійку платформу для розгортання системи раннього попередження про стихійні лиха. Воно поєднує наукову достовірність, ефективні інструменти обробки даних та практичні стратегії масштабування, що можуть бути інтегровані в реальні інформаційні потоки з метою мінімізації наслідків катастроф.

Впровадження такої системи потребує не лише технічних зусиль, а й ефективної взаємодії між усіма зацікавленими сторонами, безперебійного оновлення даних та адаптації до умов реального часу. Інтеграція автоматизованого конвеєра з існуючими інформаційними системами повинна відбуватись з урахуванням безперервного моніторингу ефективності та адаптивного налаштування моделі на основі нових даних.

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ

1. Герсбах Х., Белл Б., Беррісфорд П. Глобальний повторний аналіз ERA5. Copernicus Climate Change Service (C3S). Сховище кліматичних даних (CDS), 2019.

2. Гуха-Сапір Д., Хойоа П., Нижче Р. EM-DAT: База даних надзвичайних подій. Лувенський католицький університет. Центр досліджень епідеміології катастроф (CRED), 2016.

3. ReliefWeb. Повідомлення про надзвичайні ситуації. Управління ООН з координації гуманітарних питань УКГП. URL: https://reliefweb.int/ (дата звернення: 28.04.2025).

4. Брейман Л. Випадкові ліси. Машинне навчання, 2001. Vol. 45, No. 1. P. 5–32.

5. Chen T., Guestrin C. XGBoost: Масштабована деревоподібна система прискорення. *Матеріали 22-ї міжнародної конференції ACM SIGKDD з виявлення знань та інтелектуального аналізу даних*, 2016. P. 785–794.

6. Лундберг С.М., Лі С.-І. Уніфікований підхід до інтерпретації модельних прогнозів. Удосконалення нейронних систем обробки інформації. 2017. Vol. 30.

7. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Техніка надмірної вибірки синтетичної меншості. *Журнал досліджень штучного інтелекту*, 2002. Vol. 16. P. 321–357.

8. Хохрайтер С., Шмідхубер Я. Довга короткочасна пам'ять. *Нейронні обчислення*, 1997. Vol. 9, No. 8. P. 1735–1780.

9. Пашке А., Гросс С., Масса Ф. PyTorch: Високопродуктивна бібліотека імперативного стилю для глибокого навчання. Успіхи в нейронних системах обробки інформації, 2019.

10. Xiang X., Guo H., Li W. Глибоке навчання для аналізу часових рядів: Комплексне дослідження та оцінка продуктивності для просторово-

часового прогнозування. *Конференція з нейронних систем обробки інформації (NeurIPS),* 2020.

11. H2O.ai. Посібник користувача H2O AutoML. H2O.ai, 2024.

12. Бай С., Колтер Я.З., Колтун В. Емпірична оцінка загальних згорткових та рекурентних мереж для моделювання послідовностей. arXiv:1803.01271, 2018.

13. Дітеріх Т.Г. Ансамблеві методи в машинному навчанні. – Міжнародний семінар з систем множинних класифікаторів, 2000. P. 1–15.

14. Ке Г., Менг К., Фінлі Т. та ін. LightGBM: високоефективне дерево рішень з градієнтним підсиленням. Удосконалення нейронних систем обробки інформації, 2017.

15. Прохоренкова Л., Гусєв Г., Воробйов А. та ін. CatBoost: незміщений бустінг з категоріальними ознаками. Нейронні системи обробки інформації, 2018.

16. Harris I., Osborn T.J., Jones P. Версія 4 щомісячного багатовимірного набору кліматичних даних CRU TS з високою роздільною здатністю на сітці, Scientific Data, 2020.

17. Dee DP, Uppala SM, Simmons AJ. Реаналіз ERA-Interim: конфігурація та продуктивність системи асиміляції даних. *Щоквартальний журнал Королівського метеорологічного товариства*, 2011. Vol. 137. P. 553–597.

18. Томашев Н., Глорот Х., Рей Дж. Клінічно застосовний метод безперервного прогнозування наближення гострого пошкодження нирок. *Nature*, 2019.

19. Benesty J., Cohen I., Huang Y. Метод апроксимації кореляції в EDA. Коефіцієнт кореляції Пірсона, *Springer*, 2009.

20. Хайндман Р.Д., Атанасопулос Г. Прогнозування: Принципи та практика, 3-тє видання, OTexts, 2021.

21. Кун М., Джонсон К. Інжиніринг та відбір функцій: Практичний підхід для предиктивних моделей, CR.