Міністерство освіти і науки України
Харківський національний університет радіоелектроніки

Факультет _____Комп'ютерних наук_____
(повна назва)

Кафедра _____Програмної інженерії_____
(повна назва)

АТЕСТАЦІЙНА РОБОТА
Пояснювальна записка

_____другий (магістерський)_____
(рівень вищої освіти)

Дослідження авторегресійних й нейромережевих моделей для короткострокового прогнозування забруднення повітря
(тема)

Виконав: студент 2 курсу, групи ІПЗм-18-2___
спеціальності 121- Інженерія програмного забезпечення
(код і повна назва спеціальності)
Освітньо-професійної програми
_Програмне забезпечення систем

_____Таламанова І.С._____
(прізвище, ініціали)
Керівник _____проф. Смеляков К.С._____
(посада, прізвище, ініціали)

Допускається до захисту

Зав. кафедри, проф. _____ З.В.Дудар

2020 р.

Харківський національний університет радіоелектроніки

Факультет комп'ютерних наук

Кафедра програмної інженерії

Рівень вищої освіти другий (магістерський)

Спеціальність 121– Інженерія програмного забезпечення
(код і повна назва)

Освітньо-професійна програма Програмне забезпечення систем
(повна назва)

ЗАТВЕРДЖУЮ:
Зав. кафедри _____
(підпис)
«_____»_____20 ___ р.

ЗАВДАННЯ
НА АТЕСТАЦІЙНУ РОБОТУ

студентові Таламановій Ірині Сергіївні
(прізвище, ім'я, по батькові)

1. Тема роботи Дослідження авторегресійних й нейромережевих моделей для короткострокового прогнозування забруднення повітря

затверджена наказом по університету від «27» березня 2020 р № 473 Ст
заповнюється вручну після отримання наказу

2. Термін подання студентом роботи до екзаменаційної комісії «13» травня 2020 р.

3. Вихідні дані до роботи *Теоретичні відомості про методи аналізу та прогнозовування часових рядів, програмна реалізація та порівняння результатів прогнозування*

4. Перелік питань, що потрібно опрацювати в роботі *аналіз проблемної галузі і постановка задачі, опис проблем методів прогнозування, використовувані методи та алгоритми, опис розробленої програмної системи, аналіз можливих застосувань*

5 Консультанти розділів роботи

Найменування розділу	Консультант (посада, прізвище, ім’я, по батькові)	Позначка консультанта про виконання розділу	
		підпис	дата
Спецчастина			

КАЛЕНДАРНИЙ ПЛАН

№	Назва етапів роботи	Терміни виконання етапів роботи	Примітка *
1.	Аналіз предметної галузі	3 квітня 2020 р.	
2.	Огляд існуючих методів	5 квітня 2020 р.	
3.	Підготовка програми	7 квітня 2020 р.	
4.	Підготовка пояснювальної записки	17 квітня 2020 р.	
5.	Спецчастина	23 квітня 2020 р.	
6.	Підготовка презентації та доповіді	25 квітня 2020 р.	
7.	Попередній захист	4 травня 2020 р.	
8.	Нормоконтроль, рецензування	7 травня 2020 р.	
9.	Занесення диплома в електронний архів	8 травня 2020 р.	
10.	Допуск до захисту у зав. кафедри	11 травня 2020 р.	
* заповнюється вручну після виконання чергового пункту			

Дата видачі завдання _____ _____2020 р.

Студент _____

(підпис)

Керівник роботи _____ __проф. Смеляков К.С_____

(підпис) (посада, прізвище, ініціали)

РЕФЕРАТ / ABSTRACT

Пояснювальна записка до атестаційної роботи: 88 с., 16 рис., 9 табл., 32 джер.

ПЕРЕДБАЧЕННЯ ЗАБРУДНЕННЯ ПОВІТРЯ, ПРОГНОЗУВАННЯ ЧАСОВИХ РЯДІВ, ЕКСПОНЕНЦІАЛЬНЕ ЗГЛАДЖУВАННЯ, АВТОРЕГРЕСІЙНА ІНТЕГРОВАНА МОДЕЛЬ КОВЗНОГО СЕРЕДНЬОГО (ARIMA), ДОВГА КОРОТКОЧАСНА ПАМ’ЯТЬ (LSTM)

Об’єктом дослідження є існуючі методи, підходи та інструментарій для короткострокового прогнозування забруднення повітря.

Метою роботи є дослідження існуючих методів, підходів та інструментів для короткострокового прогнозування забруднення повітря, а також порівняння їх ефективності для реальних даних.

В результаті роботи здійснена програмна реалізація системи для прогнозування забруднення повітря авторегресійними та нейромережевими методами та порівняння цих результатів.

AIR POLLUTION PREDICTION, TIME SERIES PREDICTION, EXPONENTIAL SMOOTHING, AUTO-REGRESSIVE INTEGRATED MOVING AVERAGE(ARIMA), LONG SHORT TERM MEMORY(LSTM)

The subject of the study is investigation of existing methods, approaches and tools for air pollution prediction.

The purpose of the research is to study existing methods, approaches and tools for short-term forecasting of air pollution and compare their effectiveness for real data.

As a result of the work, the implementation of the software for air pollution prediction was created. The software is using autoregressive and neural network methods and comparing results of the prediction.

ЗМІСТ

ВСТУП

У 1950-х близько 34% населення світу жило у містах, проте за прогнозами вже до 2050 року цей показник зросте до 80%. Варто зауважити що висока густота населення у містах часто призводить до високого рівня різного виду забруднень, а саме: повітряне [1], світлове [2], та шумове [3] забруднення.

Забруднення повітря це одна з найголовніших ознак рівня якості життя для кожного міста, адже саме воно уражає людське здоров'я хворобами серця і легень [4]. Крім того, забруднення повітря є найбільшою причиною зміни клімату [5]. Дані ВООЗ показують, що 9 з 10 людей дихають повітрям, що містить високий рівень забруднюючих речовин. Від смогу, що нависає над містами, до куріння всередині будинку, забруднення повітря становить велику загрозу. Сукупний вплив забруднення навколишнього середовища (на відкритому повітрі) та побутового повітря спричиняє близько семи мільйонів передчасних смертей щороку, значною мірою внаслідок збільшення смертності від інсульту, хвороб серця, хронічної обструктивної хвороби легень, раку легенів та ГРВІ.

Основними джерелами забруднення на вулиці є транспортні засоби, виробництво електроенергії, будівництво систем опалення, сільське господарство, спалювання відходів та промисловість.

Проте навіть у приміщенні можуть існувати забружнювачі повітря. Вплив забруднювачів повітря в будівлі може призвести до широкого спектру несприятливих наслідків для здоров'я як у дітей, так і у дорослих−від респіраторних захворювань до раку до очних проблем.

Сьогодні спеціально обладнані станції вимірюють рівень забруднення повітря, а також роблять його прогноз на майбутнє. Оскільки згадані станції розташовані

далеко не по всій території країни, чи хоча б конкретного міста, достовірність отриманих ними прогнозів ставиться під питання.

У найближчому майбутньому ідентифікатори забруднення повітря стануть більш точними, меншими, і доступнішими для кожного жителя. Джудіт Су [6] повідомляє що прилади для вимірювання забруднення повітря можуть створити мережу раннього попередження забруднення, яка буде вміщувати в собі дуже багато інформації. До цієї мережі попередження забруднення входять портативні засоби, які можна носити з собою будь-куди. Оскільки дані поповнюються постійно завдяки сенсорам то для таких даних найбільш підходить саме короткострокове прогнозування забруднення яке враховує найновіші дані і будує прогноз саме за ними.

Передбачити значення забрудення повітря можна використовуючи мотоди передбачення часових рядів. Різні методи для прогнозування дають різну точність передбачення.

Для побудови прогнозу можно використовувати статистичні або методи або методи машинного навчання. Головна різниця полягає у тому що за допомогою нейромережевих методів можна виокремити нелінійні залежності у часовому ряді. Було вирішено дослідити які з методів надають найкращу точність передбачення.

Отже, метою роботи є дослідження найефективніших методів короткострокового передбачення забруднення повітря, а також дослідження застосування цих методів для практичних задач.

Об’єктом дослідження є дані щодо забруднення повітря та моделі для їх прогнозування.

Предметом дослідження даної роботи є дослідження ефективності авторегресійних та нейромережевих моделей для знаходження найкращого методу для короткострокового передбачення забруднення повітря.

Методами дослідження є проведення експерементів для визначення найбільш ефективного методу для короткострокового забруднення повітря.

Наукова новизна полягає у знаходженні найефективнішого методу для короткострокового передбачення забруднення повітря. На відміну від інших робіт що розглядають забруднення повітря, у даній роботі експерименти проводяться з постійним оновленням даних.

На практиці результати роботи можуть бути використані для побудови розумних систем моніторингу якості повітря у містах.

1. АНАЛІЗ ПРЕДМЕТНОЇ ГАЛУЗІ ТА ПОСТАНОВКА ЗАДАЧІ

1.1 Забруднення повітря

Якість атмосферного повітря це важливий показник рівня якості життя у кожній країні. Усім відомо що високий рівень захворюваності спостерігається саме у людей, які дихають забрудненим повітрям.

Забруднення повітря впливає на всі системи людського тіла [7]. Таблиця 1.1 розкриває як забруднене повітря впливає на організм людини.

Таблиця 1.1 – Наслідки впливу забрудненого повітря на людський організм

Забрудник	Наслідки
CO	Негативно впливає на серцево-судинну система і призводить до нестачі кисню в організмі людини. Як наслідок виникає розсіювання уваги, уповільнення рефлексів.Крім того, може призвести до запалення легенів.
O3	Уражає легені та дихальну систему, що може призвести до запалення легенів. Більш того, суттєво зменшує функції легенів, збільшує ризик астми і раку легенів.
NO2, SO2	Наносить важкий удар дихальній системі, знижуючи її стійкість до респіраторних інфекцій.
ТЧ	Потрапляють в легені і викликають систематичні запальні зміни. Невеликі розміри дозволяють твердим частинкам потрапляти до серця та мозку і викликати там запалення.

Дана наукова робота розглядає передбачення забруднення твердими частинками (ТЧ), адже це один з найнебезпечніших забрудників, який майже неможливо відфільтрувати.

Забруднення твердими частинками відбувається через потрапляння частинок пилу в атмосферу. Науковці [8] виділяють ультратонкі, тонкі та грубі ТЧ. Рисунок 1.1 демонструє основні критерії цієї класифікації.

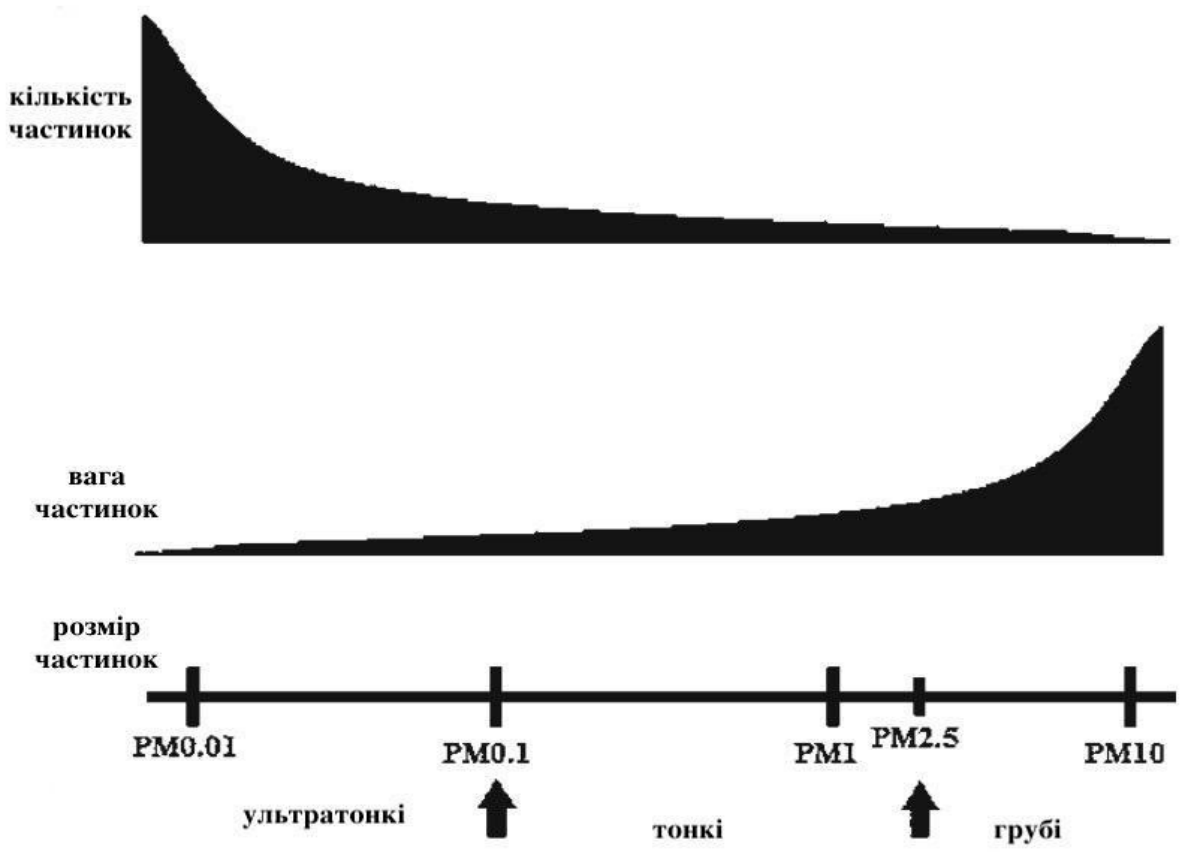

Рисунок 1.1 − Класифікація ТЧ. Ультратонкі, тонкі і грубі ТЧ

А. Зека та ін. [9] наголошують що навіть короткотривале забруднення ТЧ викликає смертельні серцево-судинні захворювання, зокрема астму. Крім того, Клог та ін. [10] доводять що короткотривале забруднення ТЧ підвищує рівень смертності

на 2.5%. Саме тому аналіз і прогноз короткотривалого забруднення ТЧ може зберегти багато життів.

До причин утворення ТЧ входять полум'я, машити та інші засобу пересування, які мають мотор, заводи, природній пил, тощо. ТЧ зазвичай складаються з металів, органічних компонентів, заліза і хімічних газів. Варто зауважити що ці частинки настільки дрібні, що можуть легко потрапити до будь-якої частини людського тіла і завдати до непоправної шкоди здоров'ю. Що менше ця частинка то більшу загрозу для організму вона несе. Саме тому для зменшення впливу забруднення на життя і здоров'я людей необхідно вимірювати і аналізувати рівень забруднення.

1.2 Прогноз часових рядів

У цьому підрозділі розкрито основні засади прогнозу часових рядів.

Часовий ряд−це дані, що реєструються з регулярним інтервалом у часі. Бауер [11] писав що метою прогнозної моделі є оцінка майбутнього значення невідомої змінної. У даному випадку у нас є незалежна змінна величина t (час) і залежна змінна y_t, яка є нашою цільовою змінною. Вихід моделі−це передбачуване значення для у за час t.

Прогноз часових рядів поділений на категорії за допомогою величини горизонту прогнозування. Відповідно до Хайндмана [12], це коротко-, середньо- та довгострокові прогнозування. Різні джерела мають різні пояснення точного розміру горизонту, але зазвичай категорії мають схожі пояснення. А саме, горизонт для короткострокового прогнозування становить від 1 до 24 годин. Горизонт у середньостроковій перспективі становить від 1 до 7 днів, в той час як горизонт довгострокового прогнозування становить понад 7 днів.

Наше дослідження зосереджено на прогнозуванні до 24 годин, тобто на короткостроковому прогнозуванні.

Коли лише одна змінна фіксується за конкретну часову одиницю, часовий ряд називається одновимірним. Багатовимірний часовий ряд має кілька змінних, записаних у часі.

Часові ряди мають кілька компонентів:

−тренд відображає довгострокове просування серії, він існує, коли в даних є постійний напрямок збільшення або зменшення;

− сезонна компонента існує за умови впливу сезонних факторів на часовий ряд; сезонність відбувається протягом визначеного та відомого періоду (наприклад, чверть року, місяць чи день тижня);

−циклічна компонента відображає повторювані, але неперіодичні коливання;

− нерегулярна компонента включає в себе всі фактори, які залишилися після усунення тенденції, сезонності та циклу від структури часових рядів.

Стаціонарність – дуже важлива властивість часових рядів. Часовий ряд називається стаціонарним, якщо його статистичні властивості (середнє значення, дисперсія, автокореляція тощо) постійні у часі. Для прогнозування за допомогою статистичних методів часовий ряд повинен бути нерухомим.

Для досягнення стаціонарності можна використовувати кілька методів зведення часового ряду до стаціонарного, а саме:

− трансформація Бокса-Кокса − метод, яка дозволяє перетворити дані таким чином, щоб вони стали рівномірно розподілятися. Ця методика стабілізує дисперсію часового ряду.

− різницеві оператори − диференціація видаляє компонент тренду з часового ряду. Це реалізовується відніманням попереднього спостереження від поточного спостереження.

− сезонні різницеві оператори − сезонне розмежування вилучає сезонність із часового ряду. Воно формується шляхом віднімання спостереження і етапів від поточного спостереження. У цьому випадку сезон часового ряду містить n етапів.

Для вибору найкращої моделі з набору використовується інформаційний критерій Акайке (ІКА). Цей критерій використовується для оцінки ймовірності моделі прогнозування майбутніх значень. Що менший ІКА, то краще підходять дані для моделі, а тому і більша прогнозована потужність моделі.

Для оптимізації ефективності алгоритмів машинного навчання та оцінки методики використовується модель перехресної перевірки. Основна ідея - розділити навчальний набір на дані горизонтального наведення і валідацію для додаткової перевірки.

1.3 Огляд методів аналізу прогнозування часових рядів

Описані нижче методи є найпопулярнішими методами прогнозування часових рядів. Для того щоб обрати методи для дослідження було проведено аналіз сильних та слабких сторін методів.

1.3.1 Лінійна регресія

Лінійний підхід до моделювання взаємозв'язку між результатом (або залежною змінною) та одним або кількома наслідками (або незалежними змінними). Іншими словами, це спосіб описати зв’язок між безперервними змінними.

Плюси лінійної регресії полягають в тому ща це простий метод, нескладний у використанні. Він може обробляти різні компоненти та функції часового ряду.

Мінуси полягають в тому що метод передбачає лінійне співвідношення між залежними та незалежними змінними, яке може бути хибним та чутливим до випадаючих показників.

1.3.2 Експоненціальне згладжування

Техніка згладжування даних прогнозування часових рядів за допомогою функції експоненціального вікна. Існують різні типи експоненційного згладжування, які включають одно-, дво- та потрійне експоненційне згладжування.

Плюси експоненційного згладджування такі що його легко використовувати. З його допомогою можна обробити рівень, тред і сезонність змінних компонентів.

Мінусами є чутливість до випадаючих показників та дуже малий довірчий інтервал.

1.3.3 Інтегрована модель авторегресії та ковзного середнього

Інтегрована модель авторегресії та ковзного середнього або ARIMA – це клас моделей прогнозування часових рядів, який розпізнає тред і сезонність. ARIMA використовує попередні спостереження для обчислення наступних термінів беручи

до уваги зв’язок з відсталими термінами. Цей метод складається з двох більш простих методів: AR що розшифровується як авторегресія та MA що розшифровується як ковзне середнє.

Плюси ARIMA полягають у тому що як правило метод забезпечує точний і надійний прогноз, широкі довірчі інтервали.

Мінусом є те що метод вимагає більше спостережень, ніж інші статистичні методи.

1.3.4 Фільтр Калмана

Метод прогнозування систем майбутнього стану на основі попередніх станів. Він оцінює спільний розподіл ймовірності точок даних для кожного часового періоду і повертає оцінене прогнозування. Він намагається усунути помилку за статичних величин.

Плюсом є те що метод націлений на передбачення. Крім того, він не потребує стаціонарних даних, що допомагає скоротити час попередньої обробки даних. Також метод легкий у застосуванні.

Мінусом є те що фільтр Калмана передбачає лінійні залежності і найкраще працює, коли часовий ряд має розподіл Гауса.

1.3.5 Тета метод

Тета модель – це метод прогнозування, який застосовується до одновимірного часового ряду. В основу лежить розкладання часового ряду на дві криві, використовуючи коефіцієнти Тета, які застосовуються до другої різниці даних. Криві називаються тета-лініями і мають те саме середнє значення та нахил, що і вихідні дані. Прогноз виконується за допомогою комбінації різних тета-ліній.

Плюсами є те він легкий у використанні а також має сезонні адаптації для прогнозування сезонних даних.

Мінусом є те що прогноз може бути неточним, так як використовується велика кількістю рядів.

1.3.6 Лінійна динамічна система

Лінійна динамічна система представляє ще один клас методів прогнозування часових рядів. Ключова відмінність цієї моделі від моделі статичної регресії полягає в тому, що для кожного етапу коефіцієнт регресії змінюється. Лінійна динамічна система зазвичай використовується для короткострокового прогнозування та моніторингу.

Плюсом є те що її легко використовувати і розшифровувати.

Мінусом є те що для прогнозування потрібно більше часу, ніж для моделі статичної регресії.

1.3.7 Нейронна мережа

Нейронні мережі – це обчислювальні системи, основні принципи яких були сформовані на основі роботи людського мозку. Основна перевага полягає в тому, що НМ може «вчитися» та приймати рішення, не будучи безпосередньо запрограмованою на певну дію. НМ можна було б описати як основу багатьох алгоритмів машинного навчання для обробки даних складних структур. Існує багато типів нейронних мереж, зокрема, БП (багатошаровий персептрон), РНМ (рекурентна нейронна мережа), LSTM (довга короткочасна пам’ять). LSTM – найпопулярніша структура нейронної мережі для прогнозування часових рядів.

Плюсами є те що для нейронної мережі менше обмежень та припущень. Вона здатна обробляти складні нелінійні залежності в часовому ряді. Має високу прогнозовану силу та можливість автоматизації.

Мінусами є те що НМ має низьку інтерпретацію. Також для отримання інтервалів довіри для прогнозу потрібно багато даних.

1.3.8 Метод найближчого сусіда (KNN)

KNN – це простий непараметричний класифікаційний метод, який ще називають ледачим алгоритмом навчання. Метод заснований на обчисленні відстані від одного об’єкта до усіх інших об'єктів. Коефіцієнт K пояснює, скільки найближчих точок даних слід використовувати для обчислення прогнозу.

Плюсами є те що для методу найближчого сусіда легко інтерпретувати результати. Також він розраховує прогноз за короткий час.

Мінусом є те що метод вимагає багато пам’яті та зберігає всі дані, через що сповільнюється прогнозування.

1.3.9 Регресія опорних векторів (SVR)

Керований алгоритм машинного навчання SVR розшифровується як регресія опорних векторів. Основна ідея полягає у використанні методики під назвою хитрість ядра, яка дозволяє зробити дуже складну трансформацію даних, через що доводиться описувати дані у багатовимірному просторі. Після цього алгоритм намагається знайти найкращу криву під назвою гіперплан, яка найкращим чином розділяє дані. Прогнозування здійснюється шляхом перевірки того, наскільки близько до гіперплану знаходиться точка даних.

Плюси його такі що метод точний в багатовимірному просторі та продуктивно використовує пам’ять.

Мінусом є те що використання регресії опорних векторів може призвести до ідеалізації даних, тому прогноз не буде точним. Для великих наборів даних обчислювальна вартість дуже висока.

1.4 Постановка задачі

Оскільки дані про забруднення повітря є часовим рядом то було прийняте рішення порівняти принципово різні підходи до прогнозування. На основі виконаного аналізу предметної галузі метою роботи стає дослідження авторегресійних та нейромережевих моделей для короткострокового прогнозування забруднення повітря.

Оскільки для короткострокового передбачення дуже важливо оновлювати передбачення, було прийняте рішення порівнювати методи для постійно оновлюваних даних, так, як це було б реалізовано на практиці.

Однією з основних задач дослідження є проектування та реалізація системи для короткострокового прогнозування повітря а також проведення експерементів на реальних даних.

Дані для тестування роботи системи є відкритим набором даних про забруднення повітря міста Скоп’є з 2008 по 2018 роки.

Результатом проведеного дослідження буде порівняння результатів прогнозування методами експоненційного згладжування, ARIMA та LSTM на проміжках часу різної довжини. Для оцінки точності прогнозу буде використано метод RMSE як найбільш точний показник правильності прогнозу.

Буде розроблена комп’ютерна програма для прогнозування даних на короткі проміжки часу. Програма зможе оцінювати точність прогнозу за допомогою показника та будувати графіки для візуальної їх оцінки.

Щоб дослідити на порівняти авторегресійні та нейромережеві моделі необхідно вірішити такі задачі:

−провести аналіз даних;

−проаналізувати методи для передбачення часових рядів;

−реалізувати обрані методи;

− порівняти результати обраних алгоритмів для короткострокового прогнозування.

2. МОДЕЛІ ПРОГНОЗУВАННЯ ЧАСОВИХ РЯДІВ

З усіх представлених в підрозділі 1.3 моделей для прогнозування часових рядів вирішено було обрати три. Оскільки темою роботи є дослідження авторегресійних та нейромережевих можелей, то було прийняте рішення порівняти методи ARIMA та LSTM. Для того щоб мати базис для порівняння прогнозів також до експериментну було включено метод експоненційного згладжування.

Більш детальний розгляд обраних методів представлено нижче.

2.1 Експоненційне згладжування

Експоненціальне згладжування – це метод прогнозування часових рядів для одновимірних даних. У цьому методі прогнозування є зваженою сумою минулих спостережень. Кожне значення серії набирає ваги, і ця вага експоненціально зменшується для минулих значень.

Існує три види експоненціального згладжування: просте експоненціальне згладжування, подвійне експоненціальне згладжування, також відоме як метод Хольта, і потрійне експоненціальне згладжування, також відоме як метод Хольт-Вінтерса. Різниця полягає в тому, що подвійне ES може враховувати тренд часових рядів, а потрійне згладжування стане в пригоді для ряду з трендом та сезонністю.

2.1.1 Просте експоненційне згладжування

Просте експоненційне згладжування використовується для передбачення таких часових рядів де немає ані тренду ані сезонностіб але середнє(або рівнь) часового ряду y_t повільно змінюється з часом.

Рівняння простого експоненційного згладжування наведено нижче.

$$S_t = \alpha y_{t-1} + (1 - \alpha) S_{t-1} \quad (1)$$

де α це костанта згладжування, значення якої лежить між 0 та 1

Фактично більш старі значення часового ряду мають меншу вагу тобто менше впливають на передбачення. На практиці параметр альфа вибирають за допомогою сіткового метода, де підбираються значення, наприклад, від 0.1 до 0.9 із зааданим кроком.

2.1.2 Подвійне експоненційне згладжування

Подвійне експоненційне згладжування застосовується до часових рядів що мають тренд. Беручи до уваги тренд, маємо наступне рівняння для побудови прогнозу. Рівняння 2 показує як розраховується передбачення $\hat{y}_{t+p}$.

$$\hat{y}_{t+p}(T) = S_t + pb_t,$$

$$S_t = \alpha y_t + (1 - \alpha)(S_{t-1} + b_{t-1}),$$

$$b_t = \beta(S_t - S_{t-1}) + (1-\beta)b_{t-1} \quad (2)$$

де $0 \leq \alpha \leq 1$, $0 \leq \beta \leq 1$

Перше рівняння згладжування коригує S_t безпосередньо для тренду для попереднього періоду, b_{t-1}, додаючи його до останнього згладженого значення, S_{t-1}. Це допомагає усунути відставання та приводить St до відповідної бази поточного значення.

Потім друге рівняння згладжування оновлює тренд, який виражається як різниця між останніми двома значеннями. Рівняння подібне до основної форми одиничного згладжування, але тут воно застосовується для оновлення тренду.

2.1.3 Потрійне експоненційне згладжування

Потрійне експоненційне згладжування було винайдене Вінтерсом у 1960 році. Вінтерс запропонував враховувати не тільки тренд а й сезонність для побудови прогнозу. Рівняння 3 для потрійного експоненційного згладжування предсставлено нижче.

$$F_{t+m} = (S_t + mb_t)I_{t-L+m},$$
$$S_t = \alpha \frac{y_t}{I_{t-L}} + (1-\alpha)(S_{t-1} - b_{t-1}),$$
$$b_t = \gamma(S_t - S_{t-1}) + (1-\gamma)b_{t-1},$$
$$I_t = \beta \frac{y_t}{S_t} + (1-\beta)I_{t-L} \quad (3)$$

де $0 \leq \alpha \leq 1$, $0 \leq \beta \leq 1$, $0 \leq \gamma \leq 1$

2.2 Авторегресійна інтегрована модель ковзного середнього (ARIMA)

ARIMA широко використовується для аналізу та прогнозування часових рядів. акронім ARIMA розшифровується як авторегресійна інтегрована модель ковзного середнього.

Авторегресія (АР або англ. AR) відноситься до моделі, в якій поточне значення часового ряду лінійно залежить від попередніх значень цих рядів. Інтегрованість (I) дозволяє зробити часові ряди нерухомими, використовуючи диференціювання. Ковзне середнє (КС або англ. MA) обчислює залежність між поточним значенням та залишковою помилкою від моделі ковзної середньої, застосованої до відстаючих спостережень.

Якщо об'єднати моделі АР та КС отримаємо модель ARMA. Цю модель застосовують коли у ряді немає ані тренду ані сезонності.

Якщо ж треба проаналізувати часовий ряд із трендом то використовують модель ARIMA, а якщо до того додається також сезонність то тоді слід використовувати модель (S)ARIMA. На практиці модель із сезонною компонентою називають ARIMA, тобто надалі під ARIMA буде матися на увазі метод з урахуванням сезонності.

У ARIMA питома вага часових рядів вважається лінійною функцією кількох минулих спостережень та залишкової помилки. [13]

ARIMA має 7 параметрів, на які слід налаштувати прогноз часових рядів:

−p – порядок відставання, що вказує на кількість відстаючих спостережень у моделі;

− d – ступінь диференціації який сигналізує про кількість розбіжностей у первинних спостереженнях;

− q – порядок ковзаючого середнього вікна, який вказує на розмір вікна ковзного середнього.

Для несезонної моделі ARIMA(p,d,q) застосовуємо наступну формулу:

$$y'_t = c + \varphi_1 y'_{t-1} + ... + \varphi_p y'_{t-p} + \theta_1 \varepsilon_{t-1} + ... + \theta_q \varepsilon_{t-q} + \varepsilon_t \ , \quad (4)$$

де y'_t – продиференційований часовий ряд, параметри моделей AR та MA знаходяться у правій частині.

Для сезонного компоненту додаються наступні параметри:

−літера P – сезонний авторегресійний порядок;

−літера D – порядок сезонних відмінностей;

−літера Q – сезонне ковзне середнє;

−літера m – кількість часових кроків що входять в сезонний період.

Загальна форма рівняння сезонної моделі ARIMA(p,d,q)(P,D,Q)m є таким:

$$\phi_P(L^m)\varphi(L)\nabla_m^D\nabla^d y_t = \Theta_Q(L^m)\theta(L)\varepsilon_t \ , \quad (5)$$

де φ(L) та θ(L) −параметри AP та КС порядку p та q, а $\phi_P(L^m)$ та $\Theta_Q(L^m)$−сезонні параметри AP та КС порядку P та Q. ∇_m^D та ∇^d – параметри диференціювання звичайних та сезонних даних. L – лаговий оператор, m – сезонність.

2.3 Нейромережеві моделі

У цьому розділі описані неромережеві моделі які можна застосувати для передбачення часових рядів.

2.3.1 Штучна нейронна мережа

Штучна нейронна мережа – це послідовність з'єднаних між собою нейронів. Структура нейронних мереж прийшла до програмування з біології. Використовуючи цю структуру, машина могла аналізувати та навіть запам'ятовувати різну інформацію.

Нейрон виступає обчислювальною одиниця, яка отримує інформацію, виконує прості обчислення цієї інформації та передає її далі. Нейрони складаються в шари. Вхідний шар необхідний для отримання інформації з вибірки, приховані шари обробляють цю інформацію, а результат отримують через вихідний шар.

Один перехід вперед і один зворотний прохід всіх навчальних даних через мережу називається періодом. Однак одночасне використання великої кількості періодів може призвести до перенавчання мережі. Надмірне навчання свідчить про те, що нейронна мережа не вивчила загальну схему, а натомість просто запам’ятала задані значення та відповіді. Це може статися, якщо дані надто багато разів проходили через мережу, або іншими словами, якщо на етапі навчання було використано занадто багато періодів. Однією з методик запобігання перенавчання є перехресна перевірка.

2.3.2 Рекурентна нейронна мережа

Рекурентна нейронна мережа (РНМ) це тип нейронної мережі де нейрон може не тільки обчислити якесь значення, але й може запам'ятати стан мережі. Проте з іншого боку це вдосконалення призводить до зникнення або вибуху градієнту, якщо мережа тренується на довгій послідовності [14]. На рисунку 2.1 представлено схему рекурентної нейронної мережі.

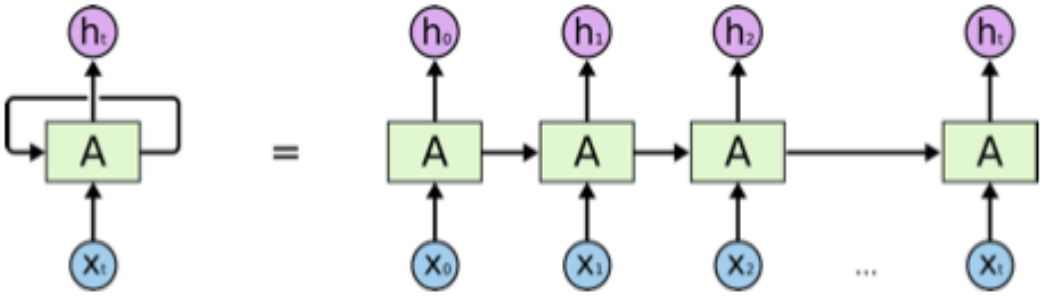

Рисунок. 2.1−Розгорнута РНМ

2.3.3 Довга короткочасна пам’ять (LSTM)

LSTM або довга короткочасна пам’ять [14] використовується у випадку потреби в історичному контексті матеріалів. Цю варіацію періодичної нейронної мережі запропонували Хохрейтер та Шмідхубер для вирішення проблеми вибуху і зникнення градієнта. Клітина LSTM має більш складну структуру, ніж клітина РНМ. Ідея полягає у запам’ятовуванні лише корисної інформації та забуванні даних, які нам більше не потрібні. Будова вузла така що він утримує два стани: стан вузла та прихований стан.

Вузли часто мають три вентилі:

−забувальний вентиль керує мірою, до якої значення залишається в пам'яті;

−входовий вентиль керує мірою, до якої нове значення входить до пам'яті;

−виходовий вентиль керує мірою, до якої значення в пам'яті використовується для обчислення активування виходу блоку.

Залежно від вентилів вузли стан вузла і прихований стан оновлюються для кожної комірки. Основна перевага LSTM полягає в тому, що стан клітин запобігає

зникненню та вибуху градієнта для довгих послідовностей. На рисунку 2.2 представлено схему вузла LSTM.

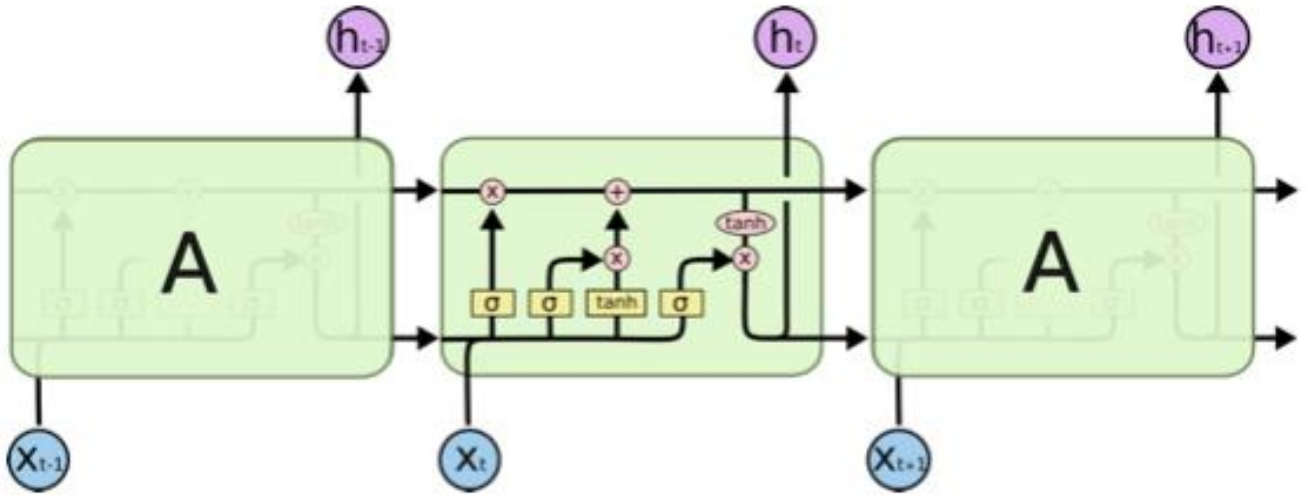

Рисунок. 2.2–LSTM вузол

2.4 Метрики оцінки точності прогнозу часового ряду

Найпопулярнішими показниками для вимірювання похибки є середня абсолютна похибка (MAE), середня похибка (MSE), коренева середньоквадратична похибка (англ. RMSE), середня абсолютна процентна помилка (MAPE).

В основі цього дослідження ми вирішили враховувати кореневу середньоквадратичну похибку так як вона не така чутлива до випадаючих показників а також має такі ж одиниці виміру як і вимірювана одиниця, тобто забруднення повітря.

Коренева середньоквадратична помилка – це квадратний корінь середнього квадратичного різниці між передбачуваним значенням і фактичним значенням ряду. Рівняння RMSE представлено формулою нижче.

$$RMSE = \sqrt{\frac{\sum_{t=1}^{n}(y_i - x_i)^2}{n}} \quad (6)$$

2.5 Огляд існуючих рішень та датасетів

2.5.1 Огляд існуючих рішень

Було проведено огляд літературних джерел для того щоб визначити досліджуваність цієї проблеми в наш час.

У статті «Система моніторингу забруднення атмосферного повітря з моделями прогнозування» [15] йдеться про систему прогнозування повітряного забруднення. Автори досліджують 3 алгоритми машинного навчання: метод опорних векторів, модельні дерева M5P та штучні нейронні мережі. Отримані дані можна використовувати у випадках високого забруднення повітря. Майбутні доопрацювання даної роботи передбачають вивчення довготривалих прогнозів забруднення у реальному часі.

Автори «Порівняльного дослідження методів прогнозування часових рядів для короткотермінового прогнозування споживання електричної енергії у розумних будинках» [16] порівнюють найпопулярніші алгоритми прогнозування споживання енергії, такі як авторегресійне рівняння з інтегрованим ковзним середнім (ARIMA) та штучна нейронна мережа (ШНМ). Порівняння було проведено між методами

статистичного та машинного навчання. Дослідження показало, що алгоритми машинного навчання працюють краще, ніж статистичні алгоритми.

У статті «Аналіз даних для прогнозування концентрації забруднюючих речовин у атмосферному повітрі в Розумному місті Уппсала» [17] увага привертається до вивчення підходящої методики видобутку даних, яка допоможе краще прогнозувати концентрацію забруднення. Для вилучення даних використовуються методи це множинної лінійної регресії та нейронні мережі.

Ле та ін. [18] використовують модель прогнозування забруднення повітря, в основі якої лежить часопросторовий набір даних. Дані зібрані з датчиків якості повітря, встановлених на таксі, яке їздить містом Тегу, Корея. Обсяг даних величезний (інтервал в одну секунду) і в просторовому, і в часовому форматі. В основі алгоритму прогнозування лежить згорткова нейронна мережа (ЗНМ) для зображення просторового забруднення повітря. Тимчасова інформація в даних обробляється за допомогою комбінації блоку довгої короткочасна пам’ять (LSTM) для даних часових рядів і моделі нейронної мережі для інших факторів впливу забруднення повітря, таких як погодні умови для побудови гібридної моделі прогнозування.

У статті «Airvlc: застосунок для прогнозування забруднення атмосферного повітря в реальному часі» [19], Очандо та ін. прогнозують якість повітря в межах міста за допомогою різних моделей прогнозування. Це відбувається завдяки збору даних про дорожній рух та погоду, на основі яких складається прогноз забруднення у трьох районах. Згідно отриманих ними результатів, найкраще працює модель випадкового лісу (англ. Random Forest model). Однак автори зосереджуються на наданні загальної інформації про місто, але не для чітко визначеного місця (району, мікрорайону).

Рецензовані наукові роботи в основному зосереджені на аналізі існуючих часових рядів, але не на постійно оновлюваних серіях. Проте для реальних рішень,

таких як прогнозування отриманих від датчиків даних, завжди можна отримати нові значення. Врахування цих значень суттєво покращить прогноз, оскільки доступною буде найновіша інформація.

2.5.2 Огляд існуючих датасетів

Класичний підхід до проведення дослідження полягає в тому, щоб спочатку зібрати дані. Можна придбати датчики забруднення повітря, проте вони досить дорогі. Якщо ж придбати дешеві датчики, то точність вимірювальних даних може бути незадовільною. Саме ці фактори створили необхідність пошуку існуючих наборів даних.

Доступні в Інтернеті дані можна структурувати. Перша група містить набори даних з історичними даними. Зазвичай такі набори даних містять інформацію з щомісячною періодичністю і не можуть бути використані для короткострокового прогнозування в режимі реального часу. Прикладом таких наборів даних можуть бути «Дані про якість повітря в Індії» [20], у які входять лише 3 спостереження за кожен місяць.

Друга група включає в себе набори даних із погодинною частотою, але невеликим обсягом даних. Ці набори даних можуть використовуватися для короткострокового прогнозування за допомогою статистичних методів, але нейронним мережам потрібно набагато більше даних для вивчення структури даних. Прикладом такого набору даних може бути «Набір даних про якість повітря» [21], який має однорічні дані, зібрані в Італії.

Третя група складається з наборів даних, які слід використовувати для цього конкретного дослідження. Набори даних мають щонайменше погодинну частоту

записуваних даних і принаймні 2 роки кількість даних. Прикладами таких наборів даних є «Пекінський набір даних PM2.5» [22] та «Забруднення повітря в Скоп'є з 2008 по 2018 рік» [23].

У цьому дослідженні реальні дані з пристрою моніторингу забруднення повітря емулюються за допомогою наявних даних. Частина даних, що називається даними для горизонтального наведення, використовується для побудови моделі. Інша частина, що називається тестовими даними, імітує невідомі майбутні дані та використовується для оцінки прогнозів моделі.

3. ПРОГРАМНА РЕАЛІЗАЦІЯ

3.1 Обґрунтування вибору середовища та мови програмної реалізації

Усі методи реалізовані в Python3 з використанням бібліотек та фреймворків із відкритим кодом. Кожен метод потребує різного програмного забезпечення для його реалізації.

Всі алгоритми реалізовані в Python за допомогою бібліотек та фреймворків з відкритим кодом (statsmodels [24], pmdarima [25], keras [26], tensorflow [27]). Вихідний код реалізації можна знайти за адресою [28].

Так як нас цікавлять короткострокові прогнози (на 24 години вперед), тому інтервал тестування становить 24 години. За даними Siami [29], інтервал формування і тестування розділений відповідно на 70% та 30%. Саме такий розподіл ми вирішили використовувати для вибору належного розміру тесту. Але, як відомо, що більше даних нейронна мережа отримує, то точніше прогноз, який вона виробляє. Так що для моделі LSTM навчальний набір значно більший, ніж для статистичних моделей.

Необхідно зауважити що всі експерименти проводилися кілька разів на різних часових інтервалах і було взято середнє значення результатів. Тож для кожного методу був розроблений багатоетапний прогноз для подальшого порівняння моделей.

3.2 Програмна реалізація додатку

3.2.1 Первинна обробка даних

Для експериментальної оцінки взятий вже існуючий набір даних. Дані збирали в Скоп'є з 2008 по 2018 рік. Цей набір даних містить вимірювання від семи станцій, на який виміряно найбільш шкідливі забруднювачі, такі як CO, NO2, O3, PM10 і PM2.5. Ці забруднювачі мають досить схожі ознаки, тому було вирішено проілюструвати підготовку даних лише на разовій серії. Обраний набір даних містить значення PM2.5.

На рисунку 3.1 видно, що дані в наборі даних не повні, адже вони містять відсутні значення та видатки.

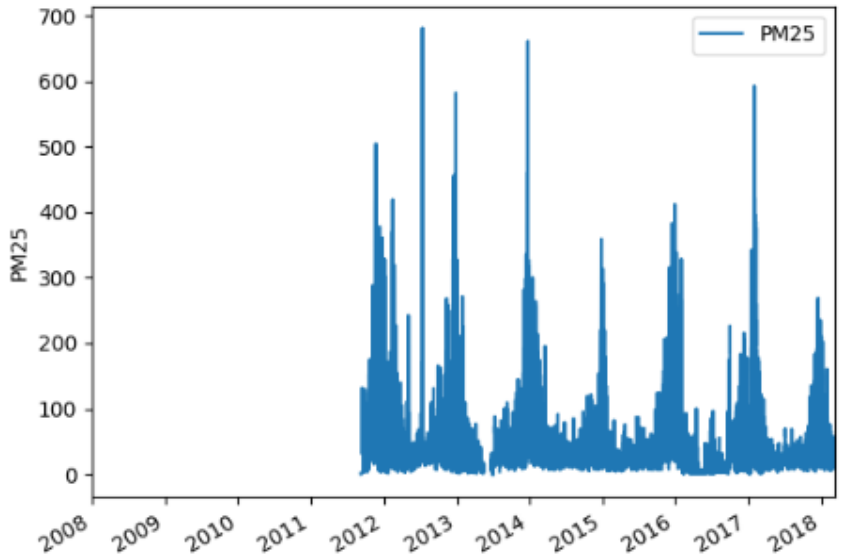

Рисунок 3.1–Початковий набір даних що показує рівень PM25 в Скоп'є з 2008 по 2018 рік

Також було виявлено, що в наборі даних відсутні дати, тому ми включимо їх, але будемо вважати ці значення як відсутні. На рисунку 3.1 ми бачимо, що набір даних має кілька випадаючих показників. На графіку розподілу (рисунку 3.2) видно, що майже всі значення лежать між 0 і 100.

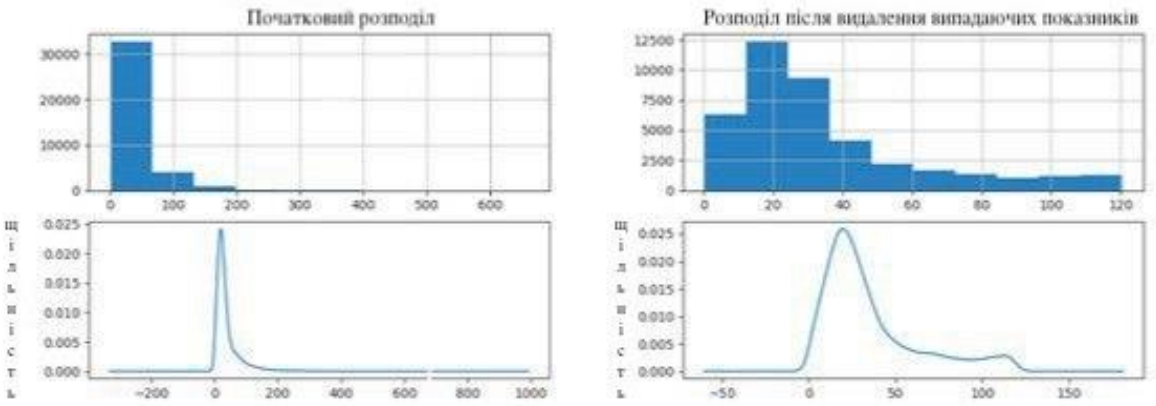

Рисунок 3.2–Порівняння розподілу початкового набору даних та набору даних після вилучення випадаючих показників

Це свідчить про те, що більш високі значення слід вважати вибуховими та видаляти. Було вирішено видалити значення, що перевищують 2 σ.

Набір даних після застосування цього методу показано на рисунку 3.3.

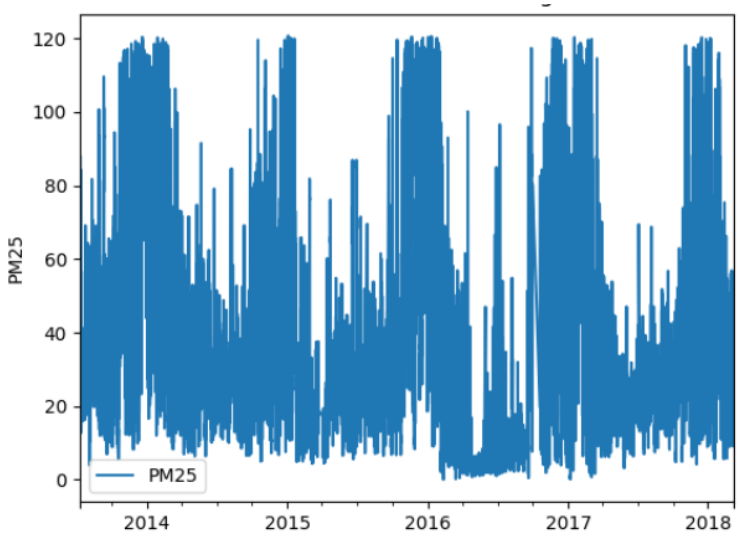

Рисунок 3.3–Набір даних після заповнення пропущених значень

Для заповнення пропущених значень використовувався метод лінійної інтерполяції, який передбачає, що відсутні значення знаходяться на лінії, яку можна намалювати за допомогою набору відомих точок.

3.2.2 Аналіз даних

Для прогнозування даних нам потрібно зрозуміти, які шаблони вони містять. Чехович [30] зазначав, що значення забруднювача залежить від дня тижня та місяця. Було виявлено, що влітку значення вимірюваного забруднювача нижчі, ніж взимку. Це означає, що ми маємо щорічну сезонність у нашій серії.

Також було виявлено, що часовий ряд має ще одну сезонність на кожен день. Складаючи графік сезонного розкладу, який показаний на рисунку 3.4, ми можемо це спостерігати за допомогою короткого періоду.

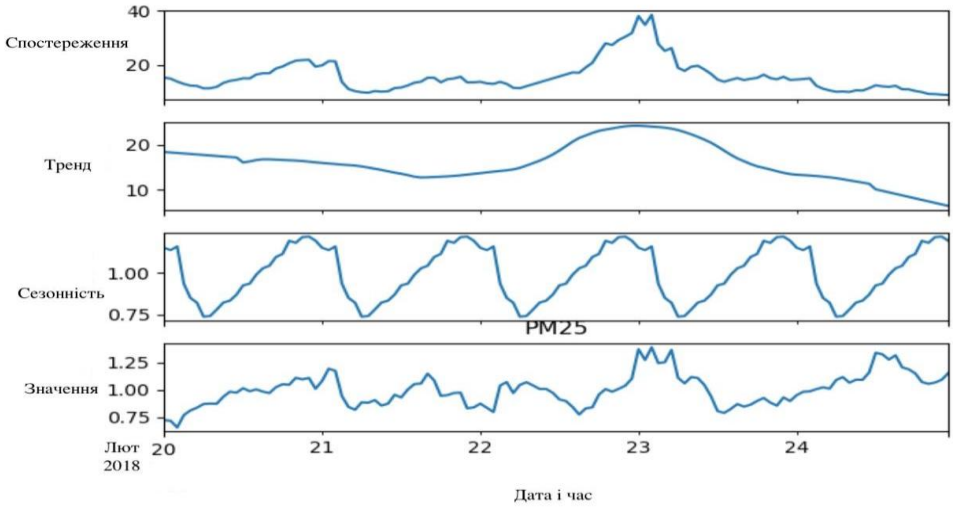

Рисунок 3.4–Сезонне розкладання протягом 5 днів

Наш часовий ряд не є стаціонарним. Ми можемо перевірити це, застосувавши до нього розширений тест Дікі-Фуллера (рДФ). Результат тесту показаний у таблиці 3.1.

Як ми бачимо, р-значення перевищує 0,05, тому нульова гіпотеза про те, що часові періоди стаціонарні, може бути відхилена. Це означає, що дані слід трансформувати для того щоб вони стали стаціонарними.

Таблиця 3.1–Розширений тест Дікі-Фуллера для підготовлених даних

Статистика тесту	-1.279960e+01
р-значення	6.798855e-24
Критичне значення (1%)	-3.430510e+00
Критичне значення (5%)	-2.861611e+00
Критичне значення (10%)	-2.566808e+00

Один із способів зменшити мінливість даних – це застосувати до неї трансформацію Бокса-Кокса. Ця трансформація, описана у підрозділі 2.4, зменшує розповсюдження даних. Але після тестування рДФ перетворення було застосовано до трансформованого ряду, і серія виявилася непостійною (нестатичною). Ця трансформація корисна в будь-якому випадку, оскільки вона зменшує дисперсію часових рядів.

Для виявлення стаціонарності серій у більш короткі періоди ми застосовували тест рДФ у тижневі періоди. Порівняний часовий ряд показаний на рисунку 3.5. Досліджений часовий ряд триває 7 днів, оскільки це дослідження зосереджено на короткотерміновому прогнозуванні.

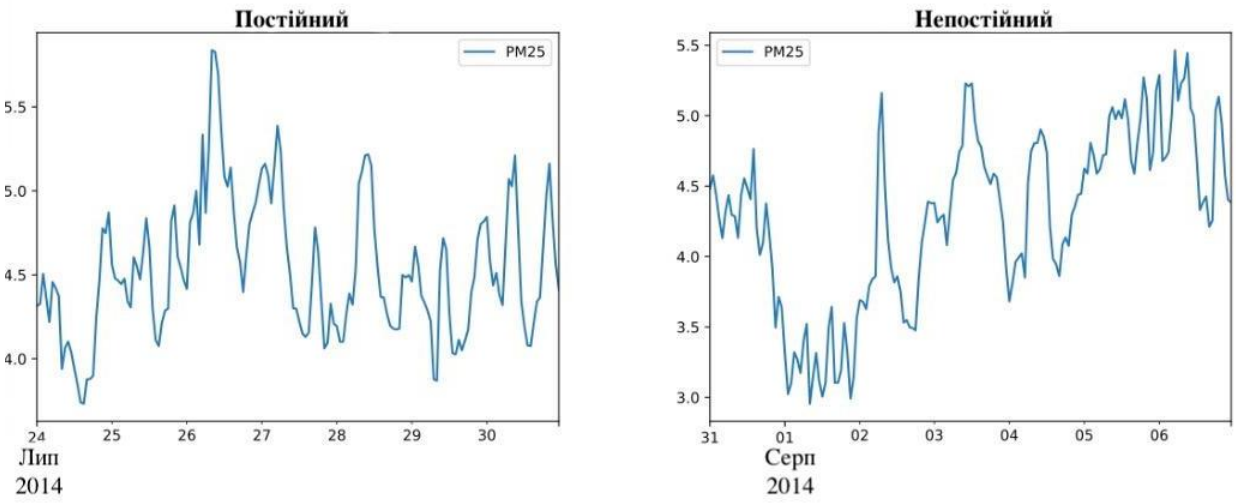

Рисунок 3.5–Порівняння двох часових рядів кожні 7 днів. За допомогою тестів виявлено, що один постійний, а інший–ні.

Результат застосування тестів рДФ до цієї серії представлений у таблиці 3.2. Як видно, одна із серій є нерухомою згідно тесту, а інша - нерухомою.

Таблиця 3.2–Порівняння двох тестів тижневих часових рядів

	Перший тиждень	Другий тиждень
Статистика тесту	-1.279960e+01	-1.781917
p-значення	6.798855e-24	0.389475
Критичне значення (1%)	-3.430510e+00	-3.477262
Критичне значення (5%)	-2.861611e+00	-2.882118
Критичне значення (10%)	-2.566808e+00	-2.577743

Зазвичай для досягнення стаціонарності даних застосовуються методи диференціації та сезонність. Поки дані іноді мають непередбачувані коливання, дуже важко знайти єдиний спосіб зробити серію нерухомою. Прогноз робиться лише на 24 години вперед, тому не варто використовувати сезонні розрізнення, так як добова сезонність є слабкою. Для усунення тренду в часових рядах потрібна звичайна диференціація, але тренд не завжди присутній в серії. Також обрані методи ES та ARIMA вже мають інструменти для усунення тренду та сезонності.

3.2.3 Реалізація методів

Для відповіді на дослідницькі завдання ми провели наступний експеримент. Він спрямований на порівняння того, як би поводилися різні методи в режимі реального часу та наскільки стабільними будуть їх прогнози. Ми будемо вимірювати це за допомогою показника RMSE.

Для емуляції навколишнього середовища в режимі реального часу використовується метод ковзного вікна. Після побудови моделі та складання прогнозу дані переміщуються на одну годину вперед, і ті ж дії повторюються.

На рисунку 3.6 описано ключову концепцію експерименту, яку називають ковзним прогнозом.

Рисунок 3.6–Ковзний прогноз

Прогноз робиться кілька разів, використовуючи різні дані, тому помилка підраховується шляхом усереднення всіх помилок, які ми отримали. Цей прийом дозволяє бути незалежним від даних та отримувати правильні результати.

3.2.4 Статистичні методи

У підрозділі 3.2.2 йшлося про те що дані мають щоденну сезонність, тому ми будемо використовувати ці знання для створення відповідної моделі.

Так як часовий ряд слід оцінювати щоразу, коли надходять нові дані, було прийнято рішення використовувати автоматичне прогнозування. Основна проблема автоматичного передбачення полягала у виборі правильної конфігурації моделі.

Зазвичай кращу модель вибирають за допомогою інформаційного критерію. У цьому конкретному випадку ми використовуємо інформаційний критерій Акайке, описаний у підрозділі 1.2. Реалізація бібліотек, що використовуються в цій роботі,

ґрунтується на роботі Хайдмана та ін. [31]. У цій роботі описані основні поняття подібної бібліотеки в R. Точні алгоритми реалізації автоматичного прогнозування більш детально розглядаються далі.

Для реалізації методу експоненціального згладжування було використано функцію ES форми статистичних моделей. Реалізація дозволяє уникнути ручної настройки коефіцієнтів α, β, γ і φ.

На рисунку 3.7 описано алгоритм автоматичного прогнозування за допомогою експоненціального згладжування.

Рисунок 3.7−Алгоритм автоматичного прогнозування за допомогою ES

За допомогою методу максимальної вірогідності оптимізовані параметри та початкові значення моделі. Після вибору найкращої моделі за допомогою ІКА складається прогноз.

На рисунку 3.3 видно, що часовий ряд дуже зашумлений, тому важко знайти модель, яка могла б точно описати весь ряд. Через це існуюча реалізація була доповнена мережевим пошуком параметрів тенденції та сезонності.

Впровадження методу експоненціального згладжування має наступні параметри:

−тренд - правильно встановлений за умови якщо він існує у часових рядах;

−сезонність - може бути додатковою або мультиплікативною;

−спад - правильно встановлений за умови якщо тред має йти на спад.

Наступний статистичний метод ARIMA реалізується за допомогою Python, використовуючи функцію auto_arima з відкритою бібліотеки Pmdarima. Ця реалізація створена для вибору найкращих параметрів та побудови моделі ARIMA. Наші дані мають морський характер, тому в цьому експерименті було використано сезонне ARIMA (SARIMA). Змінні вибираються автоматично із заданих інтервалів. На рисунку 3.8 представлений автоматичний алгоритм вибору ARIMA.

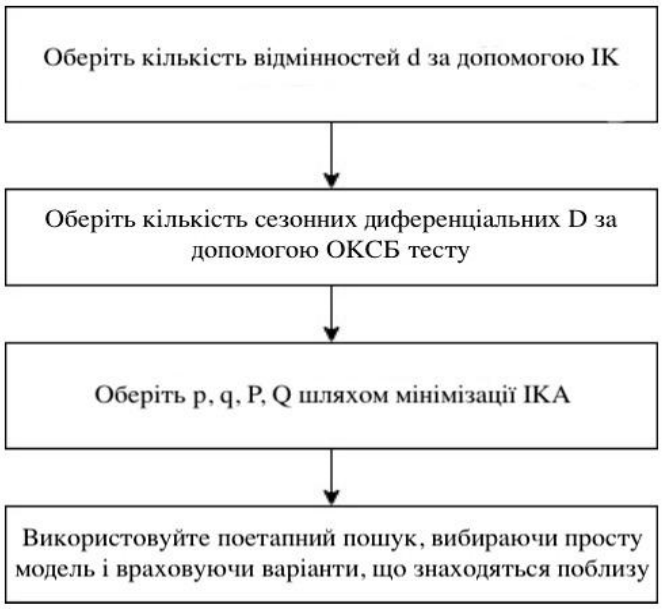

Рисунок 3.8−Алгоритм автоматичного прогнозування методу ARIMA

Основна ідея полягає у виборі найкращої моделі серед інших, що використовують ARMA.

3.2.5 Нейромережеві методи

Метод глибинного навчання, а саме LSTM, вимагає різних підходів і статистичних методів. По-перше, потрібно перегрупування структури даних. Цей крок необхідно зробити, оскільки нейронна мережа може вивчити шаблон часового ряду лише за допомогою спеціального формату даних. По-друге, налаштування змінних є складнішим завданням, ніж будь-які статистичні методи. Деякі зі змінних можна вибрати за допомогою сіткового пошуку, але інші параметри слід вручну вибирати науковцем. По-третє, слід використовувати належну будову мережі.

Для прогнозування часових рядів за допомогою нейронної мережі дані слід реструктурувати. Для прогнозування вперед поточні і минулі дані слід об'єднанти для того, щоб НМ змогла вивести залежність. Загальна структура даних представлена на рисунку 3.9.

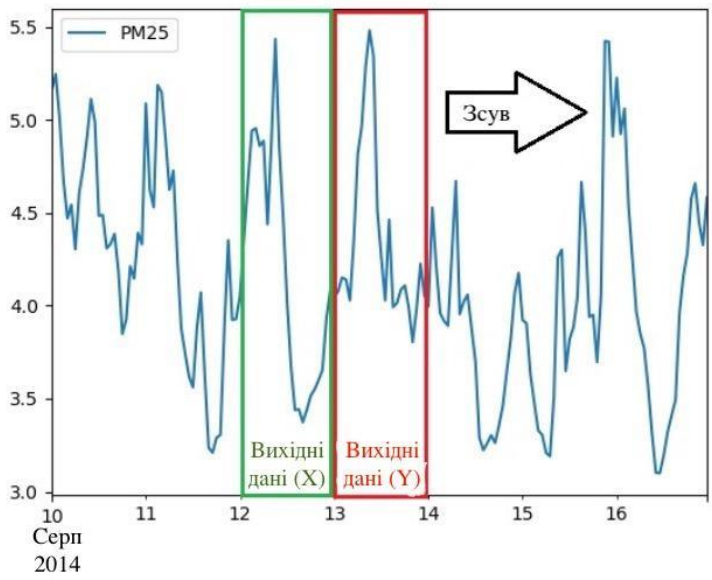

Рисунок 3.9 –Реструктуризація даних для використання нейронної мережі: поєднання поточного кроку часового ряду та попереднього масиву навчальних зразків.

Реалізація нейронної мережі разом з LSTM була здійснена за допомогою парадигм Keras і Tensorflow. Keras вимагає упаковки даних так, як це показано на рисунку 3.10. Дані слід зберігати у вигляді тривимірного масиву, який називається тензором.

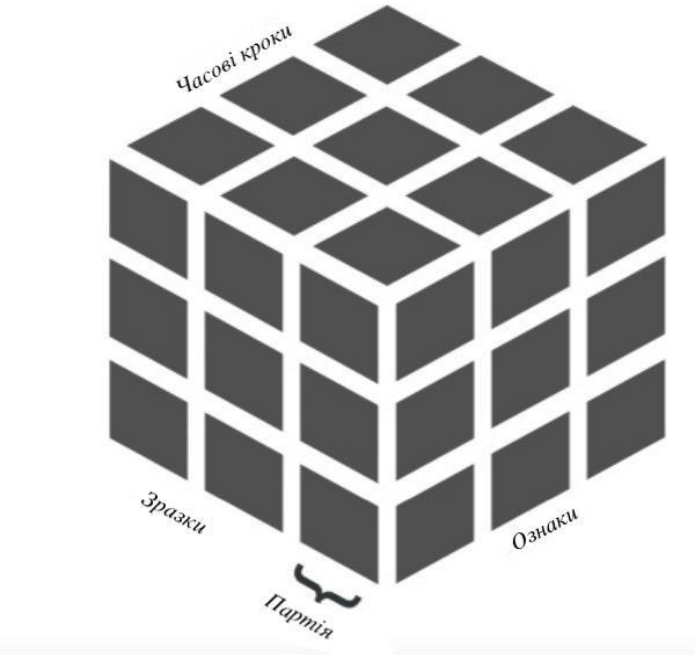

Рисунок 3.10 −LSTM у парадигмі Keras

Тензор повинен містити:

−зразки – початковий часовий ряд, можна упакувати в партії;

−ознаки – кількість функцій (у нашому випадку 1);

−часові кроки – як далеко прогнозувати.

Партія містить один або кілька зразків і допомагає нейронній мережі дізнатися більше з даних. На рисунку 3.10 партія містить лише один зразок, тому її розмір один. Якщо партія має більше зразків, вона називається міні-партія. Під час використання міні-партії НМ отримує інформацію порційно, а навчання стає більш продуктивною. Цей параметр відіграє важливу роль у прогнозуванні.

Змінні це ключовий етап побудови конфігурації нейронної мережі. Правильна настройка може зробити прогноз більш точним і точним.

Так як дослідження концентрується на короткотерміновому прогнозуванні, прогноз був зроблений на 24 години вперед. Параметр довжини прогнозу

називається step_out і є ключовим числом при підготовці структури даних. Параметр був випробуваний на горизонті прогнозування від 1 до 24 годин.

Наступні налаштування змінних зроблені на основі найменшого значення показника RMSE. Для кожного можливого значення змінних побудована мережа і зроблений прогноз. Діапазон значень можливих параметрів пояснюється для кожної змінної окремо.

Кількість періодів відповідає кількості навчальних пропусків за даними інтервалів. Чим більше періодів пройшло, тим більш навченим є НМ. Однак занадто багато періодів можуть призвести до проблеми перенавчання мережі, поясненої в підрозділі 2.3.1. Діапазон тестування становить від 100 до 1000 з кроком 100.

Витримка в Keras запобігає надмірним проблемам. Вона впливає на кількість періодів і вказує, скільки період може пройти без будь-якого прогресу точності прогнозування. Витримка підраховується за допомогою рівняння 7.

$$\text{витримка} = \max\{1, \text{період} * \text{коефіцієнт\_витримки}\} \qquad (7)$$

Коефіцієнт витримки знаходиться в межах від 0,1, до 0,3 з кроком 0.1. Поки значення цього параметра пов'язане з кількістю періодів, вибір цих двох періодів проводиться разом.

Розмір даних, що використовуються для перехресної перевірки, пояснено у підрозділі 1.2. Значення цього коефіцієнта має бути кратним розміру тестових даних. Отже, це може бути зараховано як test_size помножена на k для якої k знаходиться в діапазоні від 1 до 5 з кроком 1.

Параметри одноразова та багаторазова втрата даних вказують, які відносини слід упустити у НМ. Втрата інформації призводить до швидкого обчислення ваг і

відповідно швидших тренувань у НМ [32]. Втрата даних знаходиться в межах від 0 до 0,3 з кроком 0.1. Поточна втрата знаходиться в межах від 0 до 0.3 з кроком 0.1.

Коефіцієнт «розмір партії» використовується для групування зразків разом у групи. Підхід дозволяє НМ краще вчитися, використовуючи ті самі дані, оскільки додаткове навчання можна проводити після кожної партії. Кожна партія може містити одну або кілька зразків.

Потік навчання LSTM може проходити різними шляхами. Для LSTM без стану ваги скидаються після кожної партії, а для наступної партії зазвичай ініціалізуються нулями. Статичний LSTM використовує вихідні ваги, отримані з партії, для ініціалізації наступної матриці вагових партій.

Кількість одиниць відповідає кількості нейронів, які будуть використані для шару LSTM. Параметр обчислюється за допомогою наступного рівняння 8.

$$amount\_of\_units = 2 * size(input\_series)/coef/(steps\_in + steps\_out) \qquad (8)$$

Units_coef знаходиться в діапазоні від 2 до 10 з кроком 1

Кількість часових одиниць або параметр n_steps_in позначає початкову кількість часових кроків, що надходять на НМ. Структура даних для навчання містить початкову інформацію та правильний вихід, який нейронна мережа вивчить та використовуватиме для майбутнього прогнозування.

Вибір архітектури НМ є важливим кроком у побудові моделі прогнозування. Ретельний вибір архітектури може зробити прогноз більш точним. Загальна архітектура мережі складається з шару LSTM і твердого шару.

Рівень LSTM включає одиниці LSTM і випадання даних. Цей шар потрібен для навчання та прогнозування. Метод випадання допомагає зробити навчання

більш ефективним, скинувши невикористані одиниці в НМ. Твердий шар використовується для перетворення виходу шару LSTM у певну форму.

Шар LSTM також може мати різну архітектуру. У цьому дослідженні ми розглянули чотири найпопулярніші архітектури для прогнозування часових рядів:

−проста мережева архітектура має один прихований LSTM шар.

−складена архітектура має два шари LSTM, розміщені один за одним. Вихід першого шару є входом для другого шару.

− двонаправлена архітектура НМ, де послідовність даних аналізується не тільки з починаючи з кінця, але і навпаки.

− послідовність до послідовності архітектура також відома як архітектура кодера-декодера. НМ містить два LSTM, один кодер і один декодер, які відображають дані.

Складені та двонаправлені архітектури складніші та більш трудомісткі, тому було вирішено перевірити, яка архітектура дає найкращу точність для цього конкретного дослідження. Результати дослідження представлені у розділі 3.3.

3.3 Порівняння результатів прогнозування

У цьому розділі ми досліджуємо найкращі інтервали тренувань для ES, ARIMA та LSTM для прогнозування в режимі реального часу. Крім того, ми порівняли похибки кожного з алгоритмів.

3.3.1 Експоненціальне згладжування

Для вибору найкращого інтервалу формування ми провели наступний експеримент. Ми встановили період навчання від 24 до 360 і обрали той, який має найнижчий RMSE. Такі межі були встановлені, оскільки наші прогнози є короткотерміновими і не потрібно приймати більші інтервали формування.

Результати експерименту наведені в Таблиці 3.3, де представлено середнє значення RMSE за різні інтервали формування. У таблиці вказується, що найкращий інтервал формування для експоненціального згладжування становить 96 годин, оскільки RMSE для цього періоду є найнижчим.

Таблиця 3.3–RMSE для ES з різним інтервалом формування

Інтервал	24	48	72	96	120	144	168	196	220	360
RMSE	9.28	10.54	9.76	6.39	8.26	10.47	12.6	10.12	9.79	9.73

На рисунку 3.11 показаний повний результат цього експерименту.

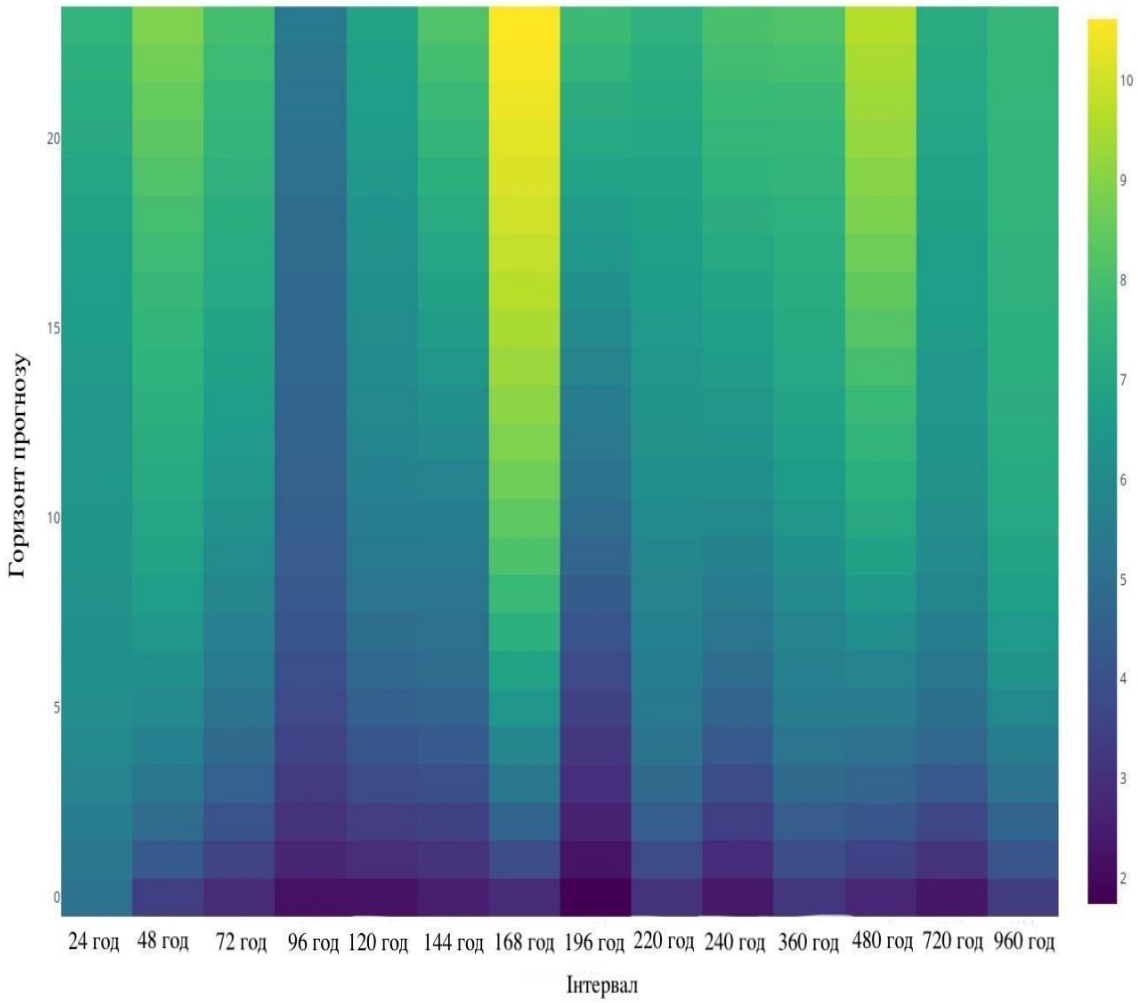

Рисунок 3.11−Теплова карта точності згладжування експонентів як функція розміру навчального набору та розміру горизонту прогнозу.

Ця теплова карта представляє RMSE для різних періодів навчання та горизонтів прогнозу. Чим темніший колір стовпця, тим нижча помилка прогнозу.

Як ми бачимо, найтемніша колонка відповідає 96 годин навчального періоду.

Щодо прогнозів, зроблених у режимі реального часу, ми хочемо знати, зміни RMSE під час збільшення горизонту. Як ми бачимо, з часом похибка зростає. Час передбачення - 20,5 секунди.

3.3.2 ARIMA

Для прискорення процесу прогнозування і скорочення часу вибору моделі, ми провели експеримент вибору інтервалів параметрів. Ми вибрали різні часові інтервали, щоб з’ясувати мінімальні і максимальні показники. Результати цього експерименту представлені в таблиці 3.4.

Таблиця 3.4 −Інтервали і коефіцієнти ARIMA

Змінна	Мінімальне значення	Максимальне значення
p	0	6
d	0	2
q	0	5
P	0	3
D	0	1
Q	0	2

Після обрання меж змінних ми помітили значне поліпшення швидкості побудови моделі ARIMA.

Час побудови моделі ARIMA та прогнозування − 19,8 секунди. Час на створення моделі менше, ніж для експоненціального згладжування, але похибка прогнозування вища.

Рисунок 3.12 відповідає результатам експерименту з моделлю ARIMA.

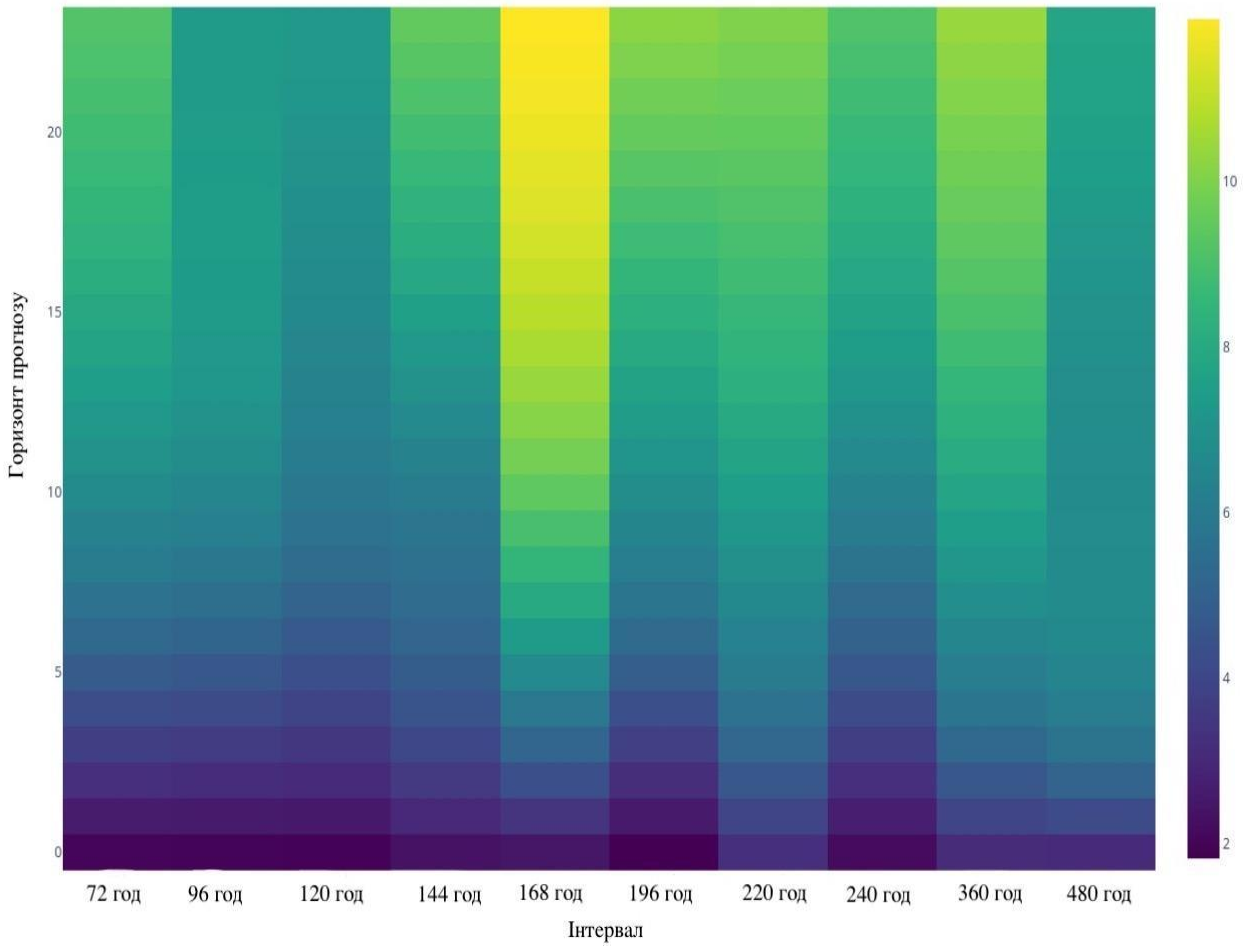

Рисунок 3.12 –Карта точності ARIMA як функція розміру навчального набору та розміру горизонту прогнозу.

Як ми бачимо, найтемніший стовпчик відповідає інтервалу 120 годин прогнозу, що означає, що цей інтервал дає найменше значення RMSE.

Вибір найкращого періоду прогнозування показаний у таблиці 3.5.

Таблиця 3.5–RMSE для ARIMA із різним часовим інтервалом

Період	72	96	120	144	168	196	220	360	480
RMSE	11.24	8.89	8.59	12.1	14.32	12.83	11.56	12.4	10.71

Результат цього експерименту доводить, що найкращий інтервал поїздів для прогнозування моделі ARIMA - 120 годин.

3.3.3 LSTM

LSTM реалізовано за допомогою системи Keras, яка підходить для виконання запланованих експериментів. На першому етапі експерименту перевірені різні конфігурації LSTM для вибору найкращої. Результати представлені в таблиці 3.6.

Таблиця 3.6 –Порівняння різних конфігурацій мережі

Конфігурація мережі	Точність (RMSE)	Час
Проста	3.265	1196.32
Складена	3.328	1253.21
Двонаправлена	4.190	1372.65
Послідовність до послідовності	3.753	1476.87

Після запуску експерименту з представленими вище мережевими структурами найкраща точність була досягнута за допомогою простої конфігурації LSTM. Для подальшого експерименту було вирішено сконцентруватися на цій мережевій архітектурі.

Наступним кроком експерименту є вибір найкращих коефіцієнтів для моделі. У таблиці 3.7 показані результати експерименту.

Таблиця 3.7–Результати налаштування параметрів LSTM

Коефіцієнт	Значення
Епохи	800
Витримка	0.1
Розмір перевіркового сету	72 год
Одноразова втрата даних	0.1
Багаторазова втрата даних	0.3
Розмір партії	12
Тип	Статичний
Коефіцієнт підрахунку одиниць	3
Розмір тренувального набору	8000 год

Результати помилки прогнозування та час виконання наведено в таблиці 3.8.

Таблиця 3.8–LSTM: Горизонт і RMSE

Горизонт	RMSE	Час
1	3.265	1282.46
2	4.47	953.4
3	5.357	1140.425
4	6.178	1185.942
5	7.07	1101.652
6	7.314	967.065
7	8.24	1203.88
8	6.215	900.618
9	9.068	1111.178
10	9.146	1055.00
11	9.672	1354.134
12	10.415	1081.185
13	10.444	1258.37
14	11.061	1195.12
15	10.714	936.59
16	11.299	1107.04

Кінець таблиці 3.8

17	11.881	1213.67
18	13.409	1155ю57
19	12.027	977.91
20	12.556	966.46
21	12.717	1199.055
22	12.936	1078.39
23	11.749	1003.031
24	14.629	117.89

Видно, що час на виконання прогнозу навіть на одну годину вперед набагато більший, ніж для інших методів.

З таблиці 3.8 видно, що результат прогнозування погіршується при збільшенні горизонту прогнозування.

3.4 Порівняння роботи методів

Як помітно на рисунку 3.13, що ES дає найнижчу RMSE, тоді як LSTM дає найвищу похибку. Крім того, крива для ES зростає з найнижчою швидкістю, що означає, що навіть за 24 години прогнозування дає меншу помилку, ніж ARIMA та LSTM.

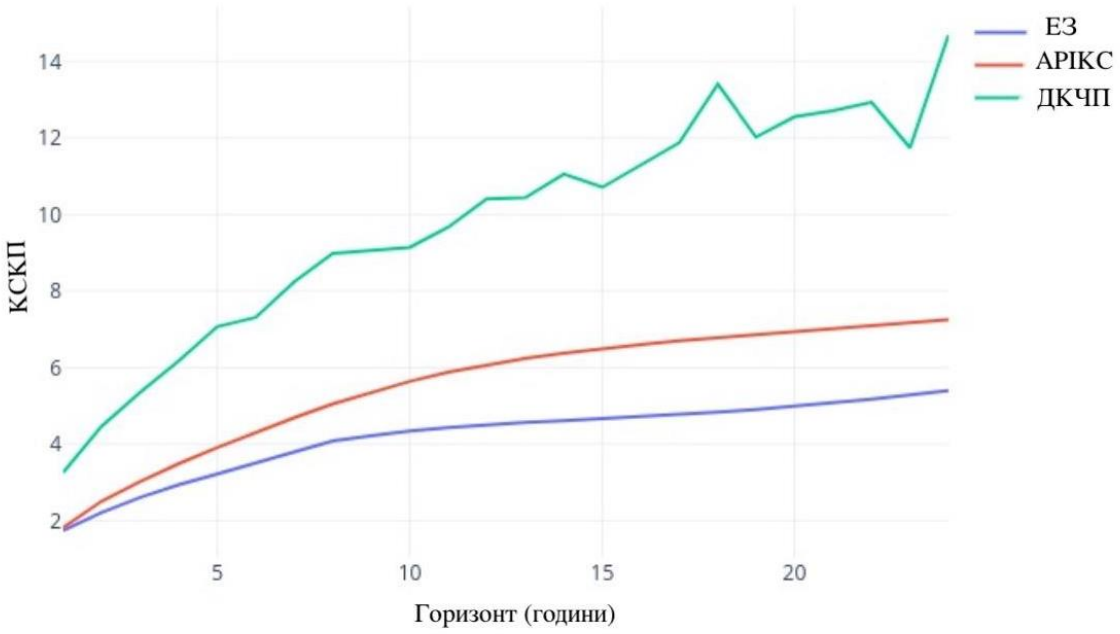

Рисунок 3.13–Залежність RMSE та горизонту прогнозування для трьох методів

Порівняння показує, що експоненційне згладжування найкраще працює для короткострокового прогнозування забруднення повітря. Навіть на один крок вперед прогноз ES ефективніший як з точки зору точності, так і часу виконання. Для прогнозу на кілька кроків вперед ми можемо побачити величезний розрив між помилками прогнозування статистичних методів та прогнозами нейронної мережі.

Час побудови моделі та прогнозування був найкращим для ARIMA (19,8 секунд), але ES (20,5 секунд) також був порівняно швидким. Однак для LSTM знадобилося набагато більше часу, щоб побудувати модель (мінімум 936 секунд), що можна побачити в таблиці 5.9.

У підсумку ми бачимо що для короткострокового прогнозування повітря авторегресійні, як більш прості методи, підходять більше аніж нейромережеві методи.

ВИСНОВКИ

Метою даної роботи було дослідження найефективніших методів короткострокового передбачення забруднення повітря, а також дослідження застосування цих методів для практичних задач. Пошук найкращих методів, налаштування моделей та підрахунок помилок прогнозування були ключовими темами для вивчення. Щодо цих тем були складені дослідницькі завдання і проведені експерименти. Були порівняні та проаналізовані три методи, зокрема, експоненціальне згладжування, ARIMA та нейронна мережа LSTM.

Були проведені експерименти з прогнозування забруднення повітря у майбутньому і виявлено найефективніші параметри моделей прогнозування. Ефективність методів було порівняно за допомогою методу середньквадратичної похибки.

Порівнюючи вибрані методи прогнозування, виявлено, авторегресійні методи, зокрема ARIMA для короткострокового прогнозування дає кращий прогноз аніж нейронна мережа. Результати показують, що навіть на одну годину вперед прогноз LSTM показує не лише набагато більшу похибку прогнозування, ніж статистичні методи, а й витрачається більше часу на те, щоб зробити прогноз.

Для багатокрокового прогнозування розрив між похибками прогнозування цих методів зростає зі збільшенням кількості кроків.

Незважаючи на те що більшість досліджень вказують що нейронні мережі це більш точний метод прогнозування, наші результати показують, що це не завжди вірно для коротокострокового прогнозування.

Подальшим напрямом дослідження є аналіз інших методів прогнозування забруднення повітря та перевірка їх ефективності на практиці.

За результатами досліджень опублікована стаття на міжнародній науковій інтернет-конференції «Інформаційне суспільство: технологічні, економічні та технічні аспекти становлення»(випуск 48).

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ

1. B. Brunekreef and S. T. Holgate, “Air pollution and health,” The lancet, vol. 360, no. 9341, pp. 1233–1242, 2002.

2. T. Longcore and C. Rich, “Ecological light pollution,”Frontiers in Ecology and the Environment, vol. 2, no. 4, pp. 191–198, 2004.

3. Y. Alsouda, S. Pllana, and A. Kurti, “Iot-based urban noise identification using machine learning: Performance of svm, knn, bagging, and random forest,” in Proceedings of the International Conference on Omni-Layer Intelligent Systems, ser. COINS ’19. New York, NY, USA: ACM, 2019, pp. 62–67.

4. N. Künzli, R. Kaiser, S. Medina, M. Studnicka, O. Chanel, P. Filliger, M. Herry,F. Horak Jr, V. Puybonnieux-Texier, P. Quénelet al., “Public-health impact ofoutdoor and traffic-related air pollution: a european assessment,” The Lancet, vol. 356, no. 9232, pp. 795–801, 2000.

5. J. H. Seinfeld and S. N. Pandis, Atmospheric chemistry and physics: from air pollution to climate change. John Wiley & Sons, 2016.

6. J. Su, “Portable and sensitive air pollution monitoring,” Light, science & applications, vol. 7, 2018.

7. M. Kampa and E. Castanas, “Human health effects of air pollution,” Environmental pollution, vol. 151, no. 2, pp. 362–367, 2008.

8. J. O. Anderson, J. G. Thundiyil, and A. Stolbach, “Clearing the air: a review of the effects of particulate matter air pollution on human health,” Journal of Medical Toxicology, vol. 8, no. 2, pp. 166–175, 2012.

8. A. Zeka, A. Zanobetti, and J. Schwartz, “Short term effects of particulate matter on cause specific mortality: effects of lags and modification by city characteristics,” Occupational and environmental medicine, vol. 62, no. 10, pp.718–725, 2005

10. I. Kloog, B. Ridgway, P. Koutrakis, B. A. Coull, and J. D. Schwartz, “Long-andshort-term exposure to pm2. 5 and mortality: using novel exposure models,” Epidemiology (Cambridge, Mass.), vol. 24, no. 4, p. 555, 2013.

11. A. Bauer, M. Züfle, N. Herbst, and S. Kounev, “Best practices for time series forecasting (tutorial paper),” in 2019 IEEE 4th International Workshops on Foundations and Applications of Self Systems (FAS*W), Umea, Sweden, June2019, pp. 255–256.

12. R. J. Hyndman and G. Athanasopoulos, Forecasting: principles and practice. OTexts, 2018.

13. G. E. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, Time series analysis:forecasting and control. John Wiley & Sons, 2015.

14. S. Hochreiter, “The vanishing gradient problem during learning recurrent neural nets and problem solutions,” International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, vol. 6, no. 02, pp. 107–116, 1998.

15. K. B. Shaban, A. Kadri, and E. Rezk, “Urban air pollution monitoring system with forecasting models,” IEEE Sensors Journal, vol. 16, no. 8, pp. 2598–2606,April 2016.

16. F. Divina, M. G. Torres, F. A. Goméz Vela, and J. L. Vázquez Noguera, “Acomparative study of time series forecasting methods for short term electric energy consumption prediction in smart buildings,” Energies, vol. 12, no. 10,p. 1934, 2019.

17. V. N. Subramanian, “Data analysis for predicting air pollutant concentrationin smart city Uppsala,” 2016.

18. D. Le, “Real-time air pollution prediction model based on spatiotemporal bigdata,” arXiv preprint arXiv:1805.00432, 2018

19. L. C. Ochando, C. I. F. Julián, F. C. Ochando, and C. F. Ramirez, “Airvlc: Anapplication for real-time forecasting urban air pollution,” 2015.

20. S. Bhargava. (2018) India Air Quality Data. URL: https://www.kaggle.com/shrutibhargava94/india-air-quality-data (Дата звернення: Травень 25, 2019)

21. S. D. Vito. (2016) Air Quality Data Set. URL: https://archive.ics.uci.edu/ml/datasets/Air+quality (Дата звернення: Липень 16, 2019)

22. S. X. Chen. (2017) Beijing PM2.5 Data Data Set. [Online]. Available:https://archive.ics.uci.edu/ml/datasets/Beijing+PM2.5+Data

23. S. Petrushevski. (2018) Air pollution in Skopje from 2008to 2018. [Online]. Available:https://www.kaggle.com/cokastefan/pm10-pollution-data-in-skopje-from-2008-to-2018

24. L. Massaron and A. Boschetti,Regression Analysis with Python. Packt Publishing Ltd, 2016.

25. T. G. Smith et al., “pmdarima: Arima estimators for Python,” 2017, [Online; accessed 2019-06-20]. [Online]. Available: http://www.alkaline-ml.com/pmdarima33

26. A. Gulli and S. Pal,Deep Learning with Keras. Packt Publishing Ltd, 2017.

27. M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isardet al., “Tensorflow: A system for large-scale ma-chine learning,” in12th{USENIX}Symposium on Operating Systems Designand Implementation ({OSDI}16), 2016, pp. 265–283.

28. I. Talamanova, “Data-driven air pollution short-term prediction in real-time,” May 2019 URL: https://doi.org/10.5281/zenodo.3585519 (Дата звернення: Травень 10, 2019)

29. S. Siami-Namini and A. S. Namin, “Forecasting economics and financial time series: Arima vs. lstm,” arXiv preprint arXiv:1803.06386, 2018.

30. R. Cichowicz, G. Wielgosiński, and W. Fetter, “Dispersion of atmospheric air pollution in summer and winter season,” Environmental monitoring and assessment, vol. 189, no. 12, p. 605, 2017

31. R. J. Hyndman, Y. Khandakar et al., Automatic time series for forecasting: the forecast package for R. Monash University, Department of Econometrics and Business Statistics 2007, no. 6/0

32. A. Viebke, S. Memeti, S. Pllana, and A. Abraham, “Chaos: a parallelization scheme for training convolutional neural networks on intel xeon phi,” The Journal of Supercomputing, vol. 75, no. 1, pp. 197–227, Jan 2019. [Online]. Available: https://doi.org/10.1007/s11227-017-1994-x