Міністерство освіти і науки України

Національний університет «Києво-Могилянська академія»

Факультет інформатики

Кафедра інформатики

Кваліфікаційна робота

освітній ступінь – бакалавр

на тему: «**ВИЯВЛЕННЯ АНОМАЛІЙ У ТРАЄКТОРІЯХ РУХУ ТРАНСПОРТНИХ ЗАСОБІВ**»

Виконала: студентка 4-го року навчання
Освітньої програми «Комп'ютерні науки»,
спеціальності 122 Комп'ютерні науки

Болдескул Софія Вікторівна

Керівник Швай Н. О.,
кандидат фіз.-мат. наук, доцент

Рецензент _____
(прізвище та ініціали)

Кваліфікаційна робота захищена
з оцінкою _____

Секретар ЕК _____
«____» _____20____ р.

Київ – 2024

Міністерство освіти і науки України

НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ «КИЄВО-МОГИЛЯНСЬКА АКАДЕМІЯ»

Кафедра інформатики факультету інформатики

ЗАТВЕРДЖУЮ

Зав.кафедри інформатики,

проф., д.ф-м.н.

_____ Гороховський С. С. (підпис)

„____”_____2024 р.

ІНДИВІДУАЛЬНЕ ЗАВДАННЯ

для кваліфікаційної роботи

студентці 4-го курсу, факультету інформатики Болдескул Софії Вікторівні

ТЕМА: Виявлення аномалій у траєкторіях руху транспортних засобів

Зміст текстової частини до кваліфікаційної роботи:

Анотація

Вступ

Розділ 1: Методологія та техніки пошуку аномалій в траєкторіях

Розділ 2: Вибір та аналіз даних для детекції аномалій

Розділ 3: Практична частина: Розробка алгоритму для виявлення аномалій в траєкторіях

Висновки

Список використаних джерел

Дата видачі „___” _____ 2024 р. Керівник _____ (підпис)

Завдання отримав _____ (підпис)

Графік підготовки кваліфікаційної роботи до захисту

Графік узгоджено «_____» _____ 2024 р.

No з/п	Перелік робіт	Термін виконання етапу	Підпис наукового керівника	Дата ознайомлення наукового керівника	Примітка
1.	Вибір теми, затвердження її на засіданні кафедри та закріплення наукового керівника	вересень			
2.	Ознайомлення з темою кваліфікаційної роботи.	жовтень-листопад			
3.	Розробка плану та структури роботи.	листопад-грудень			
4.	Робота з науковою літературою, опис основних означень.	грудень-січень			
5.	Постановка практичного завдання, аналіз отриманих результатів наукового дослідження.	лютий-березень			
6.	Робота над текстовим оформленням результатів.	березень-квітень			
7.	Попередній аналіз кваліфікаційної роботи. Виправлення помилок.	травень			
8.	Попередній захист кваліфікаційної роботи.	травень			
9.	Захист кваліфікаційної роботи.	травень			

Науковий керівник _____(ПІБ)
Виконавець кваліфікаційної роботи _____(ПІБ)

ЗМІСТ

АНОТАЦІЯ

Метою цієї кваліфікаційної роботи є застосування методів машиного навчання для пошуку аномалій у траєкторіях руху транспортних засобів. В роботі було розглянуто статистичні методи та сучасні методи машинного навчання, які використовуються для вирішення задачі з пошуку аномалій в траєкторіях. Також, було порівняно різні набори даних, які можуть бути використано в конктексті даної задачі.

Нарешті, було практично застосовано два методи машинного навчання та порівняно їх ефективність у пошуку аномалій на основі таких показників як швидкість, легкість масштабування, точність та простота у використанні.

ВСТУП

У світі наук про дані та комп'ютерних наук завдяки сучасним пристроям, таким як дрони, камери спостереження, та радари, можна зібрати велику кількість даних про рух різних транспортних засобів, таких як автомобілі, автобуси, мотоцикли. Наприклад, дрони і камери спостереження активно застосовують для запису відеоматеріалу з високою роздільною здатністю, який потім аналізують за допомогою спеціалізованих алгоритмів, націлених на роботу з високорозмірними типами даних. В свою чергу, радари забезпечують точні дані про швидкість і відстань, що дає змогу якісніше відстежувати рух транспортних засобів та ідентифікувати аномальні траєкторії.

Обробка та подальший аналіз даних про рух транспортних засобів стали важливою та невід'ємною частиною створення програмного забезпечення, що використовується для забезпечення громадської безпеки на дорожніх трасах, як у міській місцевості, так і на швидкісних шосе.

Поява технологій машинного навчання, здатних аналізувати рух автомобілів та ідентифікувати потенційно небезпечні траєкторії, стала поштовхом для того, щоб розробити та впровадити алгоритм з пошуку аномалій в інтелектуально транспортні системи. Сучасні інтелектуально транспортні системи представляють програмне забезпечення, націлене на збільшення безпеки на дорогах. Інтеграція алгоритму з пошуку аномалій могла б збільшити ефективність подібних систем і, відповідно, зменшити кількість аварій.

Актуальність теми цієї дипломної роботи також підтверджується офіційними даними 2023 року, згідно з якими кількість дорожньо-транспортних пригод (ДТП) в Україні зросла на 26,9% порівняно з попереднім 2022 роком [1]. Найчастішими причинами ДТП з летальними наслідками стали

перевищення безпечної швидкості (50 %) і порушення правил маневрування (16 %) .

Дорожньо-транспортні пригоди за причинами
за період з 01.01.2023 по 31.12.2023

Причини	ДТП з загиблими та/або травмованими		
	Усього ДТП	Загинуло осіб	Травмовано осіб
ПЕРЕВИЩЕННЯ БЕЗПЕЧНОЇ ШВИДКОСТІ	9215	1570	11564
ПОРУШЕННЯ ПРАВИЛ МАНЕВРУВАННЯ	5191	476	6465
ПОРУШЕННЯ ПРАВИЛ ПРОЇЗДУ ПЕРЕХРЕСТЬ	2014	73	2879
ПОРУШЕННЯ ПРАВИЛ ПРОЇЗДУ ПІШОХІДНИХ ПЕРЕХОДІВ	1642	106	1643
НЕДОТРИМАННЯ ДИСТАНЦІЇ	1148	84	1436
КЕРУВАННЯ ТРАНСПОРТНИМ ЗАСОБОМ У СТАНІ СП'ЯНІННЯ	942	96	1226
ПЕРЕХІД ПІШОХОДІВ У НЕВСТАНОВЛЕНОМУ МІСЦІ	781	167	648
ПЕРЕВИЩЕННЯ ВСТАНОВЛЕНОЇ ШВИДКОСТІ	642	146	800
ВИЇЗД НА СМУГУ ЗУСТРІЧНОГО РУХУ	470	145	816
ПОРУШЕННЯ ПРАВИЛ НАДАННЯ БЕЗПЕРЕШКОДНОГО ПРОЇЗДУ	290	20	422
НЕВИКОНАННЯ ВОДІЯМИ ВИМОГ СИГНАЛІВ РЕГУЛЮВАННЯ	273	12	407
ПОРУШЕННЯ ПРАВИЛ ОБГОНУ	267	38	434

Рис. 1: Статистика ДТП в Україні

Таким чином, розробка програмного забезпечення націленого на пошук аномалій у траєкторіях транспортних засобів з використанням машинного

навчання **є актуальним і необхідним кроком в удосконаленні загальної системи безпеки руху.**

Важливим етапом у створенні подібного програмного забезпечення є створення алгоритму, який би міг розпізнати нетиповий рух транспортного засобу. Відповідно, **метою цієї дипломної роботи** є застосування алгоритмів машинного навчання, здатних виявляти аномалії в траєкторіях транспортних засобів, для підвищення безпеки на автошляхах.

Мета роботи зумовила наступне наукове завдання:

- проаналізувати існуючі методи виявлення аномалій в траекторіях транспортних засобів;
- дослідити, який тип мають набори даних на яких проводиться тренування існуючих алгоритмів;
- розширити аналіз задачі з пошуку аномалій в траєкторіях та проаналізувати існуючі методи в пошуку аномалій у русі людей;
- знайти набір даних на якому буде проведено тренування розробленого алгоритму;
- отримати дозвіл на використання набору даних;
- провести преобробку набору даних;
- вибрати та натренувати моделі для пошуку аномалій в траекторіях;
- оцінити ефективність створених алгоритмів.

Однак, складність розробки алгоритму з пошуку аномалій в траєкторіях руху полягає в тому, що поняття аномалії не має чіткого визначення і може інтерпретуватися по-різному. Ця невизначеність посилюється в контексті траєкторій транспортних засобів, де на результат впливає багато факторів навколишнього середовища. Наприклад, місцевість, час доби, пора року, день тижня та обмеження швидкості — всі ці чинники відіграють вирішальну роль

у визначенні того, що є аномалією. Мінливість місцевості та динаміка руху лише додають складності в процесі розробки алгоритму.

Так само не менш важливим фактором у створенні алгоритму пошуку аномалій у траєкторіях транспортних засобів є те, що для кожного транспортного засобу аномальним рухом вважатимуться різні патерни в траєкторії.

Наприклад, якщо транспортний засіб класифікується як легковий автомобіль, але насправді це вантажівка або автобус, то його траєкторія може відрізнятися від типових рухів легкових автомобілів. Це може призвести до неправильного визначення аномалій, оскільки алгоритм очікуватиме патерни, характерні для легкових автомобілів, та ігноруватиме відхилення, які типові для вантажівок або автобусів.

Важливо зауважити, що загальна швидкість руху машини та вантажного транспортного засобу в умовах руху на території шосе безумовно буде різною, відповідно для цих двох транспортних засобів аномальними вважатимуться різні патерни руху.

Звернення уваги на тип транспортного засобу відіграє фундаментальну роль у пошуку аномалій і часто не береться в увагу під час розробки алгоритму.

Саме факт неправильної класифікації класу транспортного засобу як індикатор аномалії був узятий за основу під час написання алгоритму для пошуку аномалій у рамках цього наукового завдання.

У практичній частині роботи було використано набір даних, у якому вказано інформацію щодо типу транспортного засобу, а саме автомобіль це чи вантажний транспортний засіб.

Інформація про тип транспортного засобу зіграла ключову роль у навчанні моделі машинного навчання. При використанні методів контрольованого навчання, модель була навчена розпізнавати особливості руху кожного типу

транспортного засобу, що значно підвищило її точність у виявленні аномалій, залежно від типу транспортного засобу.

Високий рівень точності підтверджує ефективність використання інформації про тип транспортного засобу в розробленому алгоритмі.

РОЗДІЛ 1. МЕТОДОЛОГІЯ ТА ТЕХНІКИ ПОШУКУ АНОМАЛІЙ В ТРАЄКТОРІЯХ

1.1. Огляд та постановка задачі з пошуку аномалій

З появою науки про великі дані, пошук і виявлення аномалій стали одним з найбільш досліджуваних питань, пов'язаних з обробкою даних. Дійсно, у сучасному світі, в якому щодня відстежують велику кількість даних, таких як траєкторії руху транспортних засобів, трафік відвідування вебсервісів або моніторинг даних про використання енергії, майже в кожній сфері життєдіяльності необхідно аналізувати й вміти ідентифікувати потенційно підозрілі дані та вчасно повідомляти про них. Грамотний і точний аналіз даних, а також виявлення аномалій у використовуваних даних, можуть стати причиною модифікації різних процесів від оновлення програмного забезпечення до перепланування швидкісного шосе. Такі модифікації часто позитивно впливають на загальну працездатність і ефективність системи.

У контексті цієї дипломної роботи будуть розглянуті траєкторії транспортних засобів як дані, в яких буде виконуватися виявлення аномалій. Траєкторією транспортного засобу вважається шлях, який проходить транспортний засіб у просторі протягом певного періоду часу [1]. Найчастіше, у наборах даних про рух транспортних засобів траєкторії представлені як набір відомих положень об'єкта в певні проміжки часу. Переважно, траєкторії використовують у задачах із пошуку аномалій, а також у задачах із прогнозування поведінки водіїв. Як вже було сказано вище, досить складно дати чітке визначення аномалії. Тому на абстрактному рівні заведено вважати, що аномалії — це патерни в даних, які не відповідають чітко визначеному поняттю нормальної поведінки [2]. Таким чином, у контексті пошуку аномалій в траєкторіях руху

транспортних засобів аномалією можна вважати відхилення від стандартної для даного типу транспортного засобу поведінки. Відхиленням може бути повільна або навпаки занадто швидка швидкість руху транспортного засобу, різка зміна напрямку руху, раптові зупинки в непризначених для цього місцях. Так само важливо визначити поняття детекції аномалій. Детекція аномалій — це техніка виявлення та ідентифікації випадків, які відхиляються від норми та демонструють нестандартну поведінку [1]. У контексті траєкторій транспортних засобів, детекція аномалій полягає у виявленні нетипових для конкретного транспортного засобу патернів руху.

Аномалії заведено ділити на три типи [2] :

- точкові аномалії (анг. point anomaly);
- контекстуальні аномалії (анг. contextual anomaly);
- колективні аномалії. (анг. collective anomaly)

Точкові аномалії вважають найпоширенішими та найпростішими типами аномалій. Найчастіше це одиничні екземпляри даних, які значно відрізняються від інших записів. Прикладом подібних аномалій може бути раптова незвично велика кількість запитів на сервер, яка може вказувати на кібератаку.

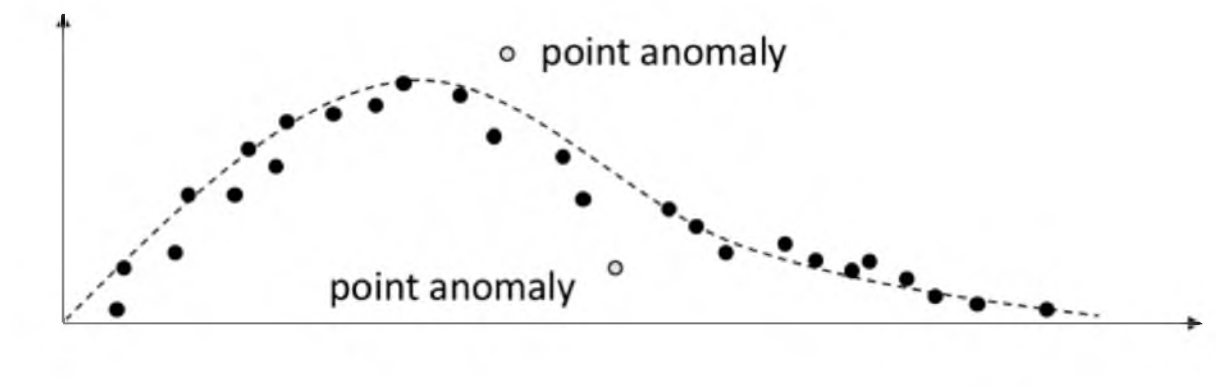

Рис. 2: Графік з прикладом точкової аномалії [5]

Контекстуальні аномалії — це приклад екземплярів даних, які можуть вважатися аномальними в одному контексті та нормальними в іншому. Цей тип аномалій найчастіше зустрічається в даних, пов'язаних із часовими рядами та географічними даними, як-от довгота і широта [2]. Контекстуальні аномалії вважаються доволі складним типом аномалій, оскільки знань про місцевість і навколишнє середовище не завжди вистачає для того, щоб точно встановити, чи є дані аномалією. Прикладом контекстуальної аномалії може слугувати високий рівень інсуліну в крові, адже після їди це нормальний екземпляр даних, а натщесерце може свідчити про наявність аномалії.

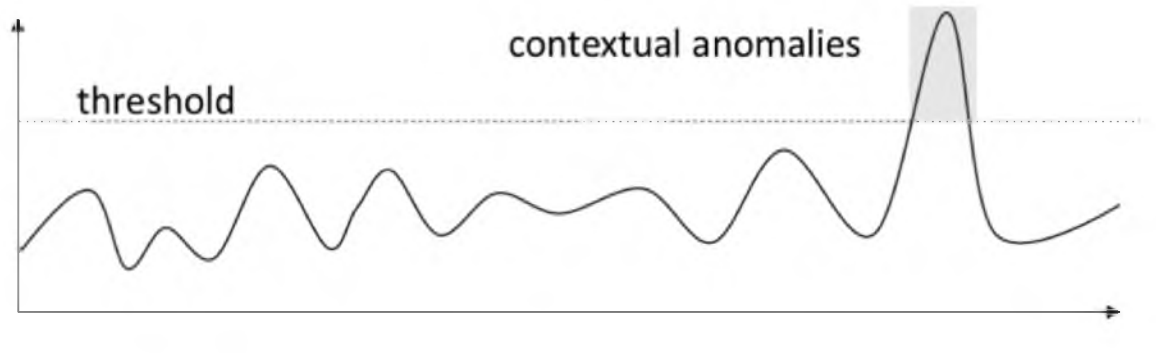

Рис. 3: Графік з прикладом контекстуальної аномалії [5]

Третім типом аномалій вважаються колективні аномалії. Колективні аномалії вказують на цілий набір даних, який є аномальним і лише поодинокі екземпляри з цього набору можуть вважатися нормальними [2]. Збій у русі великої кількості транспортних засобів на зазвичай непроблемній ділянці траси може бути яскравим прикладом колективного типу аномалії та вказувати на можливу аварію або іншу перешкоду нормального руху.

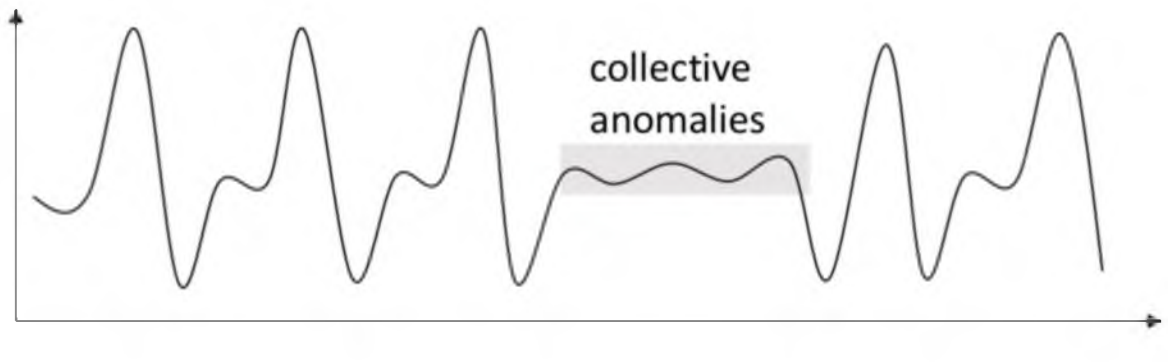

Рис. 4: Графік з прикладом колективної аномалії [5]

Знання і правильне розуміння типу аномалії в наборі даних є важливим кроком у розробці ефективного методу детекції аномалій. Саме тому часто перед створенням алгоритму з пошуку аномалій важливо провести попередній аналіз даних і ознайомитися із закономірностями, що існують у цьому наборі даних.

1.2. Аналіз методів детекції аномалій в траєкторіях транспортних засобів

У цьому підрозділі буде розказано про методи, які використовуються для детекції аномалій, а також алгоритми, представлені для пошуку аномалій. Глобально можна поділити методи детекції аномалії на статичні методи та методи з використанням машинного навчання.

Статистичні методи є досить старими алгоритмами для пошуку аномалій. Проте вони є ефективними для невеликого набору даних. Найвідомішими статистичними методами є Z-оцінка (Z-score), боксові діаграми (box diagrams), метод перцентилів (percentile method).

Z-оцінка — це статистична міра, що використовується для визначення наскільки сильно значення в наборі даних відхиляється від середнього

значення в тому ж наборі даних. Z-оцінка дає змогу оцінити, чи є значення типовим або аномальним для даного набору даних.

Боксові діаграми — це графічний метод розподілу даних для визначення аномалій. Одна з головних його переваг над рештою методів — це наочність, яка спрощує його інтерпретацію. Структура боксової діаграми складається з мінімального і максимального значення, двох квартилів (Q1 та Q3) і медіани між квартилями. Перший і третій квартилі представляють 25% і 75% кумулятивного розподілу відповідно. Метод виявлення аномалій за допомогою боксових діаграм базується на використанні правила 1.5 інтерквартильного розмаху (IQR) для визначення викидів у даних. Викидами вважаються дані, які лежать за межами 1.5 IQR від квартилів. Тобто значення, які перевищують (Q3 + 1.5 IQR) або є меншими за (Q1 - 1.5 IQR) вважаються аномаліями. Цей метод підходить для виявлення нестандартних зупинок або несподіваних точок прискорення в траєкторіях руху транспортних засобів.

Метод перцентилів, так само як правило 1.5 міжквартильного розмаху, використовує квартилі для виявлення викидів, проте є гнучкішою технікою, адже дає змогу встановити конкретні значення перцентильних порогів для ідентифікації аномалій. Наприклад, метод перцентилів дає змогу встановити, що всі дані, які перевищують 90-й перцентиль або розташовуються нижче 10-го перцентиля, вважаються викидами. Метод перцентилів може бути корисним у випадках де розподіл аномалій є несиметричним. У контексті задачі з виявлення аномалій у траєкторіях руху транспортних засобів, метод перцентилів може бути ефективним для аналізу аномально довгих або надзвичайно коротких відстаней між транспортними засобами.

Найбільш поширеними методами машинного навчання в задачі про пошук аномалій у траєкторіях руху транспортних засобів є методи класифікації,

кластеризації, нейронні мережі, які скоріше відносять до методів класифікації, хоча вони є окремим підкласом алгоритмів у контексті задачі про пошук аномалій та метод Isolation Forest. Методи LSTM та Isolation Forest будуть детально розглянуті в наступному підрозділі. В цьому підрозділі увагу буде зосереджено на класифікаційних та кластеризаційних підходах машинного навчання.

Класифікація — це метод машинного навчання, головний принцип якого полягає у віднесенні об'єкта до певної класу даних відповідно до значень його атрибутів на основі попередньо навченої моделі [3]. До алгоритмів, що базуються на основі принципу класифікації, належать SVM, Decision Tree, Bayesian Network, Logistic Regression, K-nearest Neighbours. Головним принципом класифікаційного підходу є використання міток у навчанні моделі правильно ідентифікувати клас екземпляра даних. Алгоритми класифікації швидко опрацьовують великі набори даних, однак, одним із їхніх недоліків є те, що ефективність алгоритму напряму залежить від точності та обсягу тренувальних даних.

Принцип роботи алгоритмів кластеризації, своєю чергою, базується на групуванні об'єктів у кластери на основі ступеня їхньої схожості без використання попередньо встановлених міток [3]. Прикладами алгоритмів кластеризації є K-means, DBSCAN. Ці алгоритми використовують для ідентифікації груп траєкторій зі схожими характеристиками та виявлення викидів, які формують невеликий кластер, або точок, що не належать до жодного кластера і відповідно можуть бути аномаліями. Алгоритми кластеризації допомагають виявити природні патерни в траєкторіях транспортних засобів. Цей тип алгоритмів є потужним у виявленні нетипових та неочевидних поведінкових шаблонів.

Різноманітність методів, які використовуються для вирішення задачі про виявлення аномалій у траєкторіях доволі велика, і кожен із них має свої недоліки та переваги. У даній кваліфікаційній роботі будуть більш конкретно розглянуті методи LSTM, а також Isolation forest, оскільки після детального дослідження було визначено, що саме вони найбільше підходять для типу даних часовий ряд.

1.3. Порівняння технік машинного навчання у виявленні аномалій

Як було зазначено вище, цей підрозділ присвячено детальному розгляду алгоритмів Long Short-Term Memory (LSTM) та Isolation forest для виявлення аномалій у траєкторіях транспортних засобів.

Алгоритм Long Short-Term Memory є одним із варіантом рекурентних нейронних мереж і широко використовується в різних сферах життєдіяльності, завдяки своїй ефективності в моделюванні послідовних даних, як-от годинникові ряди, текст, аудіо та відео ряди [6]. У сфері аналізу даних о траєкторіях руху транспортних засобів алгоритм першочергово використовується для виявлення аномалій у траєкторіях або для прогнозування даних про майбутні траєкторії руху. Траєкторії руху є досить складним типом даних, тому що для їхнього успішного опрацювання потрібно створювати послідовності з точок даних, у яких кожна точка відповідає певному положенню транспортного засобу в певний період часу. Однак, після правильного кодування даних у послідовності алгоритм LSTM може ефективно аналізувати динаміку цих послідовностей з урахуванням часових залежностей. Алгоритм LSTM вирізняється з-поміж інших алгоритмів керованого навчання своєю здатність запам'ятовувати та використовувати

інформацію на тривалий час, завдяки чому він може з точністю аналізувати та знаходити аномалії в даних про траєкторії. Принцип роботи рекурентної нейронної мережі LSTM головним чином полягає в наявності спеціальної структури такої як «комірка пам'яті», яка забезпечує механізми воріт. Механізми воріт - це насамперед інструменти, які запобігають втраті зникаючого градієнта, що характерна для стандартних RNN. У LSTM моделі розрізняють такі типи воріт: - вхідні ворота - цей тип воріт визначає, яка частина інформації буде збережена в комірці пам'яті; - ворота забування - ворота, які визначають, яку частину інформації слід забути та не передавати далі; - вихідні ворота - ворота, які вирішують, яку інформацію з комірки пам'яті буде передано далі на вихід мережі. Використання цього механізму роботи рекурентної мережі LSTM дало змогу аналізувати не тільки поточний стан, в якому перебуває об'єкт, а й історію руху, а отже, вловлювати комплексні залежності руху й аналізувати відхилення від стандартних шаблонів поведінки транспортних засобів. У даній кваліфікаційній роботі нейронну мережу LSTM було використано для моделювання послідовностей даних і подальшого виявлення аномалій у цих послідовностях на основі помилкової класифікації типу транспортного засобу. Саме здатність алгоритму LSTM з точністю опрацьовувати часові ряди, при цьому зберігаючи інформацію про клас транспортного засобу, стала причиною його використання в даному дослідженні. На противагу розробленому алгоритму на основі моделі LSTM, в рамках даної роботи також буде досліджено використання алгоритму Isolation forest для того, щоб виявити аномальні точки в траєкторіях руху на основі даних про швидкість і прискорення в момент руху транспортного засобу.

Алгоритм Isolation forest є прикладом алгоритму неконтрольованого навчання, тому його порівняння з рекурентною нейронною мережею LSTM, що

використовує класифікаційний підхід у пошуку аномалій у траєкторіях руху транспортних засобів, є цікавим предметом дослідження. Алгоритм Isolation forest це ефективна техніка неконтрольованого машинного навчання для пошуку аномалій, що добре підходить для роботи з комплексними розподілами даних та виявленням точок, у яких рух транспортного засобу не відповідав очікуваній поведінці [7]. Однією з головних причин широкого використання алгоритму Isolation forest у глобальному контексті пошуку аномалій є його здатність обробляти значення без попередньо заданих міток. Це особливо корисно в контексті даних, у яких аномальна поведінка не має чіткого однозначного шаблону. Робота алгоритму Isolation forest базується на використанні принципу ізоляції аномальних даних від інших. Головною ідеєю цього принципу є те, що аномальні точки заведено вважати рідкістю і їх легше ізолювати від нормальних даних. Алгоритм випадковим чином обирає атрибут, на основі якого відбуватиметься поділ точок. Після цього, він так само випадковим чином обирає точку поділу в межах значень обраної ознаки та починає поділ даних на підгрупи на основі того, менші чи більші вони за точку поділу. Процес поділу рекурсивно повторюється і дозволяє виявити ізольовану точку даних. Точки даних, які потребують меншої кількості поділів для їхньої однозначної ідентифікації, тобто ізолюються швидше за звичайні, вважаються аномаліями. Це зумовлено тим, що ці точки легше віддаляються від норми та, отже сильніше відрізняються від більшості даних. Перевагою алгоритму Isolation forest є його висока швидкість у знаходженні аномальних точок. Цей критерій може бути особливо важливим у системах опрацювання даних про траєкторії транспортних засобів у режимі реального часу, де своєчасна сигналізація про потенційну роль відіграє критично важливу роль для забезпечення безпеки всіх учасників дорожнього транспортного руху. Також алгоритм Isolation forest може бути легко масштабований, що так само

є важливим критерієм у контексті його використання в мегаполісах або на трасах із високим рівнем трафіку. Водночас алгоритм Isolation Forest ідентифікує аномальні точки без уточнення інформації про часові залежності, яка може відігравати важливу роль у детекції аномалій на основі часових рамок.

Таким чином порівняння двох алгоритмів зводиться до чотирьох ознак: швидкість алгоритму, легкість масштабування, точність виявлення аномалій, а також простота у використанні. Алгоритм Isolation forest вирізняється більш простотою у використанні та демонструє високу швидкість виконання програми, проте є менш точним за LSTM-модель, яка хоч і потребує попереднього кодування даних у послідовності, проте демонструє вищу точність та здатність брати до уваги історичні дані щодо траєкторії об'єкта.

РОЗДІЛ 2. ВИБІР ТА АНАЛІЗ ДАНИХ ДЛЯ ДЕТЕКЦІЇ АНОМАЛІЙ

2.1. Огляд доступних наборів даних

Наступним кроком після детального огляду задачі з пошуку аномалій у траєкторіях руху автівок і аналізу доступних алгоритмів для розв'язання цієї задачі, був початок пошуку набору даних, який підійшов би для тестування розроблених алгоритмів. У процесі пошуку було виокремлено чотири набори даних, які можуть використовуватися в рамках поставленого завдання, а саме ApolloScape [10], Zen Traffic Data [11], highD dataset [12], а також Next Generation Simulation (NGSIM) Open Data [13].

Для використання наборів даних Zen Traffic Data та highD dataset потрібно було пройти процес подачі та схвалення заявки на отримання доступу до баз даних з траєкторіями транспортних засобів для наукових цілей. У рамках дослідження було подано заявки для використання наборів даних та отримано схвалення на використання набору даних highD Dataset у рамках цієї наукової роботи. Набори даних Next Generation Simultation (NGSIM) Open Data та ApolloScape не вимагали подання заявки для отримання дозволу на використання, тому їх також було проаналізовано для визначення, який набір даних буде використано в роботі. Важливим фактором при виборі наборів даних була дата створення набору даних.

Доступні набори даних мали наступні дати створення:

- NGSIM -> 15 червня 2005 року;
- ApolloScape -> 4 липня 2019 рік (дата виходу статті про набір даних);
- highD dataset -> вересень 2017 - липень 2018 року.

Таким чином, перший набір даних NGSIM, який вважається фундаментальним в аналізі даних про траєкторії руху, можна вважати застарілим, адже у сфері

розроблення програмного забезпечення, націленого на пошук аномалій у траєкторіях руху, новизна даних є важливим індикатором для створення ефективного робочого алгоритму. Важливо зауважити, що використання набору даних NGSIM триває донині й нещодавно професори Університету Північної Кароліни видали статтю про прогнозування руху на основі набору даних NGSIM [15]. Тобто, попри те, що набір даних було створено майже 20 років тому, його й надалі використовують для завдань у сфері прогнозування траєкторій руху.

Attribute	**Dataset**	
	NGSIM	***highD***
Recording Duration [hours]	1.5	16.5
Lanes (per direction)	5-6	2-3
Recorded Distance [m]	500-640	400-420
Vehicles	9206	110 000
Cars	8860	90 000
Trucks	278	20 000
Driven distance [km]	5071	45 000
Driven time [h]	174	447

Рис. 5: Порівняння обсягу даних у наборах даних NGSIM та highD [12]

Набір даних ApolloScape містить великий масив даних траєкторій з міських вулиць, який використовується для планування і прогнозування траєкторій. Набір даних було зібрано в різний час доби та під різним рівнем освітлення.

Дані були представлені у вигляді зображень, а також відсканованих за допомогою технології LIDAR (Light Identification, Detection and Ranging) хмар точок. За рахунок зйомки в міському середовищі набір даних ApolloScape наповнений даними про рух не тільки транспортних засобів, а й пішоходів. Однак, для дослідження в рамках пошуку аномалій цей набір даних викликав труднощі через відсутність даних про точки прискорення в траєкторіях руху. Дані про прискорення необхідні для успішного тренування моделі, оскільки саме раптове прискорення часто є причиною дорожньо-транспортних пригод. Таким чином, з огляду на невідповідність наборів даних NGSIM і ApolloScape науковому завданню, поставленому в рамках цієї наукової роботи, було ухвалено рішення проаналізувати набір даних highD dataset, докладний опис якого буде представлено в наступному підрозділі.

2.2. Опис вибраного набору даних у рамках дослідження

Як було вказано в попередньому підрозділі, набір даних highD dataset вимагав отримання дозволу на його використання в наукових цілях. Набір даних було розроблено і опубліковано командою Інституту автомобільної інженерії Аахенського університету RWTH. Інститут автомобільної інженерії Аахенського університету RWTH виготовляє високоякісні натуралістичні сценарні дані. За офіційним описом, highD dataset - це новий набір даних натуралістичних траєкторій руху транспортних засобів, зафіксованих на німецьких автомагістралях. Завдяки використанню безпілотника типові обмеження традиційних методів збору даних про дорожній рух, такі як перешкоди, долаються за допомогою повітряної перспективи. Трафік був зафіксований у шести різних локаціях і налічує понад 110 500 транспортних засобів. Траєкторія кожного транспортного засобу, включаючи тип, розмір і

маневри, витягується автоматично. Завдяки використанню найсучасніших алгоритмів комп'ютерного зору, похибка позиціонування зазвичай не перевищує десяти сантиметрів. Хоча набір даних був створений для перевірки безпеки високоавтоматизованих транспортних засобів, він також підходить для багатьох інших завдань, таких як аналіз схем руху або параметризація моделей водіїв [12].

Отже, набір даних налічує 60 записів з шести різних локацій на автошляхах Німеччини. В полі зору безпілотника було 420 метрів автошляху. Таким чином, було записано 147 годин руху транспортних засобів, 44500 проїханих кілометрів, 110500 транспортних засобів. Також важливо зазначити, що було зафіксовано різні стани дорожнього руху. Інформація про дані з кожного запису була збережена в трьох csv файлах. Перший файл Recording Meta Information (XX_recordingMeta.csv) зберігає мета інформацію про запис, а саме про: унікальний номер запису, частоту кадрів з якою було зроблено запис, ідентифікатор місця запису, обмеження швидкості на автошляху, інформацію про день тижня та місяць в якому було зроблено запис, тривалість запису, час початку запису, загальну пройдену відстань усіх відстежуємих транспортних засобів, загальний час руху всіх відстежуємих транспортних засобів, кількість відстежуємих транспортних засобів (окремо кількість автівок та вантажівок).

Другий файл Track Meta Information (XX_tracksMeta.csv) зберігає мета інформацію про розміри транспортного засобу (ширина та довжина), його клас, максимальну та мінімальну швидкість під час поїздки, кількість змін смуги руху та напрямок руху транспортного засобу. Також файл містить інформацію щодо таких параметрів: мінімальної дистанції до попереду їдучого транспортного засобу minDHW(це значення було встановлено на -1, якщо попереду немає транспортного засобу), мінімального часового інтервалу

до попереду їдучого транспортного засобу minTHW(це значення було встановлено на -1, якщо попереду не має транспортного засобу), мінімального часу до зіткнення minTTC (це значення встановлюється на -1, якщо попереду немає транспортного засобу або не існує дійсного значення).

Третій файл Tracks (XX_tracks.csv) містить покадрові, залежні від часу значення щодо положення у просторі кожного транспортного засобу. В рамках дослідження файл надавав критично важливі дані стосовно поточної швидкості, положення в просторі, прискорення та відстані до інших транспортних засобів. Основна мета цього файлу - забезпечити детальну інформацію про динамічну поведінку кожного відстежуваного транспортного засобу в часових проміжках.

В експериментальній частині було використано наступні атрибути з даного набору даних (у квадратних дужках вказана міра вимірювання):

- frame: Поточний кадр у послідовності.
- id: Унікальний в межах одного запису ідентифікатор транспортного засобу.
- x: Позиція x верхнього лівого кута обмежувальної рамки транспортного засобу.
- y: Позиція y верхнього лівого кута обмежувальної рамки транспортного засобу.
- width: Ширина обмежувальної рамки транспортного засобу. [м]
- height: Висота обмежувальної рамки транспортного засобу. [м]
- xVelocity: Поздовжня швидкість у системі координат зображення. [м/с]
- yVelocity: Поперечна швидкість у системі координат зображення. [м/с]

- xAcceleration: Поздовжнє прискорення в системі координат зображення. [м/с²]
- yAcceleration: Поперечне прискорення в системі координат зображення. [м/с²]
- dhw: Дистанція руху, встановлюється на 0, якщо попередній транспортний засіб відсутній. [м]
- thw: Часовий інтервал, встановлюється на 0, якщо попередній транспортний засіб відсутній. [с]
- ttc: Час до зіткнення, встановлюється на 0, якщо не існує попереднього транспортного засобу або дійсного TTC. [с]
- laneId: Ідентифікатор смуги руху, починаючи з 1 і призначається у порядку зростання.
- frontSightDistance: Відстань до кінця записаної ділянки дороги у напрямку руху від центру транспортного засобу. [м]
- backSightDistance: Відстань до кінця записаної ділянки дороги у напрямку, протилежному напрямку руху від центру транспортного засобу. [м]

Таким чином, файл «XX_tracks(XX_tracks.csv)» містить детальні дані, необхідні для аналізу руху транспортних засобів та створення закодованих у послідовності траєкторій [12].

Підсумовуючи, у другому розділі було розглянуто чотири набори даних, які теоретично можуть використовуватись в контексті задачі з пошуку аномалій в траєкторіях. З огляду на такі особливості як обмеження доступу, застарілість даних та інші фактори, було прийнято рішення використовувати набір даних HighD dataset. Детальний огляд цього набору даних довів, що HighD dataset

має важливі в контексті задачі з пошуку аномалій характеристики, такі як висока якість запису траєкторій, велика кількість записаних транспортних засобів та різновид локацій, на яких проводилися записи. Важливо зазначити, що в наступному розділі буде опрацьовано набір даних HighD dataset для розробки алгоритму виявлення аномалій у траєкторіях руху транспортних засобів. Першочергово, цей процес передбачає підготовку даних, їх дослідження та аналіз, а також розробку додаткових функцій для підвищення продуктивності моделі. Успішність роботи алгоритму багато в чому залежатиме саме від набору даних, на якому здійснюватиметься навчання моделі машинного навчання, тому детальний і прискіпливий вибір набору даних був дуже важливим у межах цього наукового дослідження.

РОЗДІЛ 3. ПРАКТИЧНА ЧАСТИНА: РОЗРОБКА АЛГОРИТМУ ДЛЯ ВИЯЛЕННЯ АНОМАЛІЙ У ТРАЄКТОРІЯХ

3.1. Підготовка даних та EDA (Розвідковий аналіз даних)

У цьому підрозділі буде описано кроки з підготовки набору даних для подальшого тренування моделі, а також надано висновки, які було зроблено після розвідкового аналізу метаданих і даних про траєкторії руху транспортних засобів. Саме завдяки попередньому детальному аналізу даних, було з'ясовано на основі яких атрибутів буде проведено закодування траєкторій у послідовності для надання на вхід алгоритму LSTM та які атрибути будуть використовуватись у тренуванні алгоритму Isolation Forest. Процес підготовки даних був спрямований на створення чистого і точного набору даних для майбутнього аналізу. Було виконано такі кроки:

- агрегація метаданих: усі метадані про записи було скомпоновано в один набір даних для зручного маніпулювання та проведення подальшого аналізу метаданих;
- перевірка набору даних на наявність пропущених значень: під час дослідження було виявлено, що набір даних highD dataset не має пропущених значень, що знову вказує на його високу якість у збірці даних;
- агрегація даних з файлів XX_tracks (XX_tracks.csv) та Track Meta Information (XX_tracksMeta.csv): дані з цих файлів були скомпоновані завдяки унікальному номеру транспортного засобу та унікальному номеру запису. Це дозволило об'єднати інформацію про транспортний засіб, таку як його клас (машина чи вантажівка), з даними про траєкторію руху в один файл. Це об'єднання забезпечує більш повну

картину для подальшого аналізу та тренування моделі, полегшуючи процес обробки та інтерпретації даних.

Після обробки набору даних було проведено спочатку розвідувальний аналіз метаданих про записи з траєкторіями.

Розвідувальний аналіз даних допоміг виявити ключові патерни у метаданих записів:

1. Обмеженість даних за локаціями (результати аналізу представлені на рис. 6):

   Лише з першої локації дані було зібрано в різні дні тижня, відповідно, алгоритм може бути недостатньо адаптований до змін у патернах руху протягом тижня. Для усунення упередженості алгоритму також рекомендовано зібрати дані в різні дні тижня.

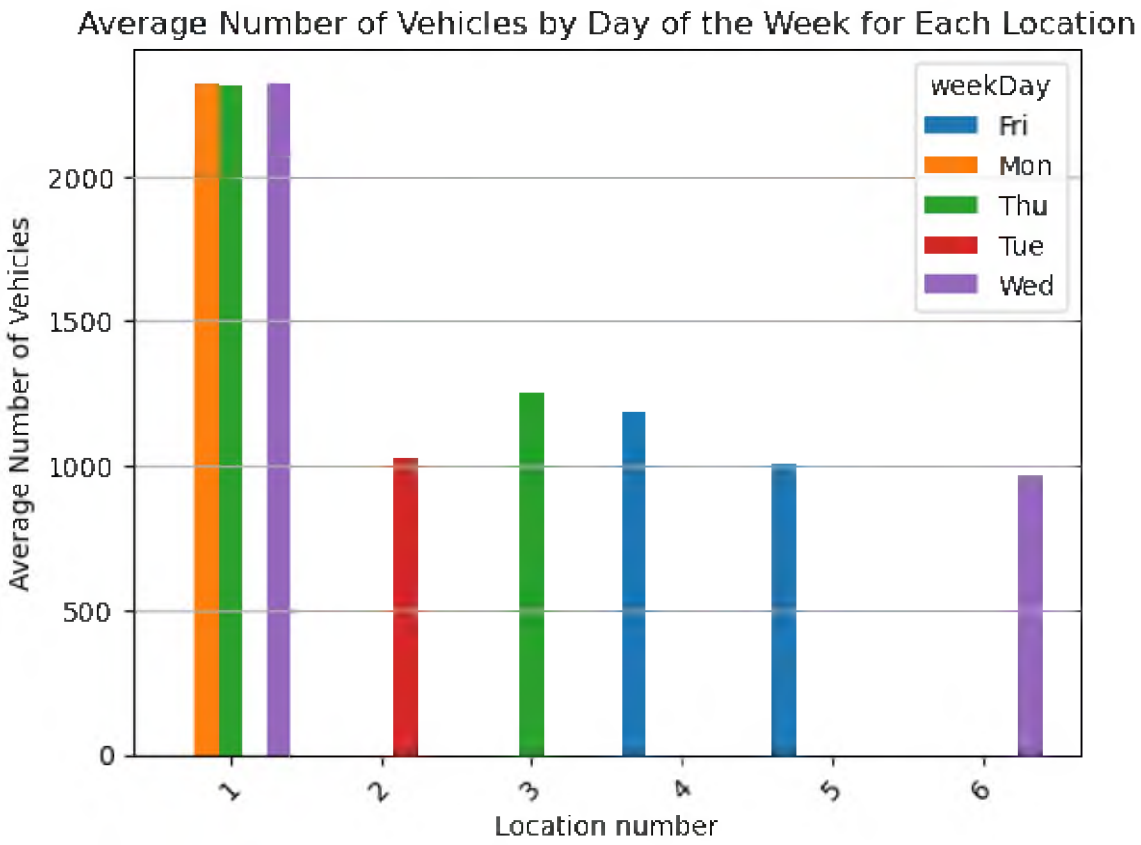

Рис. 6: Середня кількість транспортних засобів за днями тижня для кожної локації

2. Переважна більшість записів була зібрана в ранковий час доби (результати аналізу представлені на рис. 7)

   Оскільки суттєва кількість даних була зібрана с 05:00 до 12:00, тренування алгоритму може бути упереджено до виявлення аномалій характерних для цього часу. Для покращення узагальненості алгоритму рекомендується зібрати додаткові дані протягом усього дня.

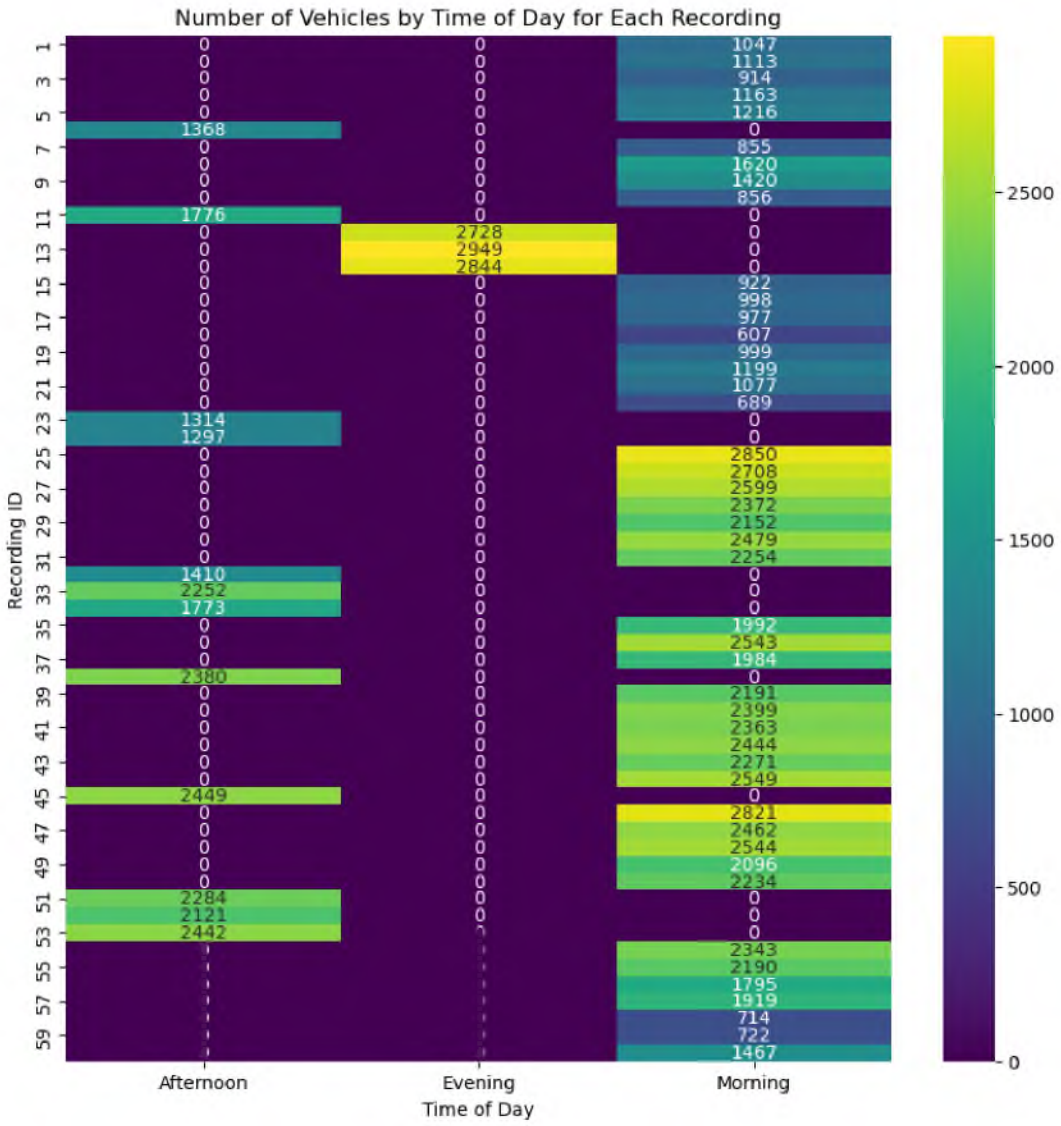

Рис. 7: Аналіз кількості транспортних засобів за часом доби для кожного запису

3. Значні піки в кількості транспортних засобів спостерігаються в вечірні години (результати аналізу представлені на рис. 8)

   Виявлені піки можуть вказувати на підвищену активність у вечірні години, що є важливим фактором для аналізу. Всі вечірні записи демонструють високу інтенсивність трафіку, але таких записи лише три.

Рекомендується розширення часових рамок збору даних аби забезпечити більш рівномірне покриття та, відповідно, покращити здатність моделі до виявлення аномалій у різний час доби.

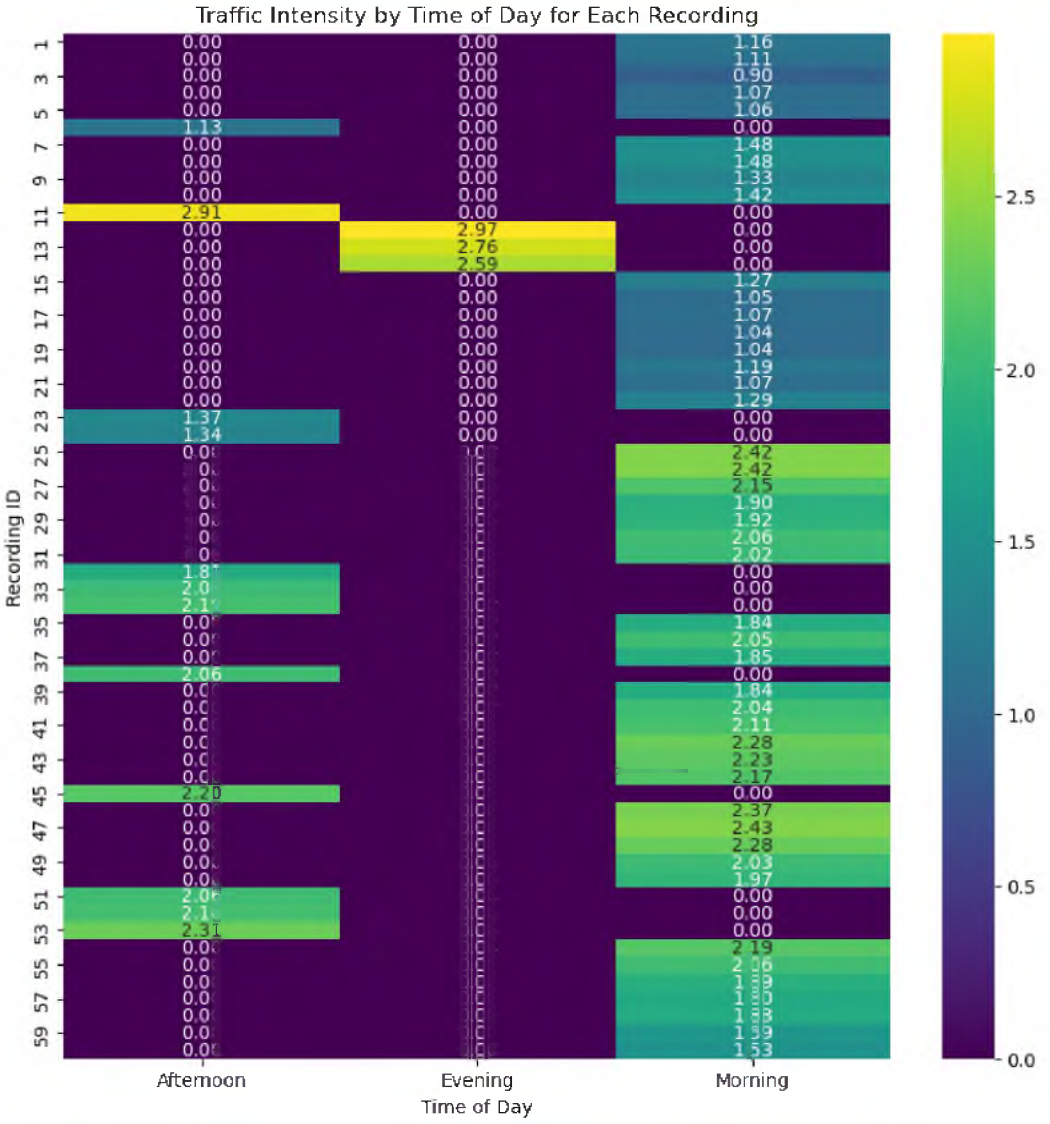

Рис. 8: Інтенсивність руху за часом доби для кожного запису

3.2. Розробка додаткових ознак

Наступним кроком у рамках наукового дослідження була розроблення додаткових ознак на основі атрибутів, які були доступні у наборі даних. Враховуючи, що попередній аналіз показав наявність підозрілих значень швидкості транспортних засобів, було прийнято рішення обрати швидкість та відповідно прискорення транспортних засобів як основні індикатори аномалій. У наборі даних були доступні значення xVelocity та yVelocity, а також xAcceleration та yAcceleration. Відповідно, для розрахунку загальної швидкості було використано наступну формулу [16]:

$$totalVelocity = \sqrt{xVelocity^2 + yVelocity^2}$$

Для розрахунку загального прискорення було використано наступну формулу:

$$totalAcceleration = \sqrt{xAcceleration^2 + yAcceleration^2}$$

Ці ознаки були додані до моделі для покращення точності ідентифікації аномалій. Також було проведено додатковий аналіз розподілу цих ознак на кожному записі. Це допомогло визначити, наскільки ефективними є нові ознаки для виявлення аномалій у траєкторіях транспортних засобів. Для кращого розуміння розподілу нових ознак було побудовано гістограми для швидкості та прискорення транспортних засобів. Гістограми допомагають візуально виявити аномальні значення та підтвердити важливість включення цих ознак у алгоритми. Для наочності в роботі буде продемонстровано приклади кожної з гістограм.

На першій гістограмі (Рис. 9) зображено розподіл загальної швидкості транспортних засобів для одного з записів (Локація 1, Запис 25). Цей графік показує кількість транспортних засобів для кожного інтервалу швидкостей.

Відповідно до графіка видно, що існує декілька піків, які можуть вказувати на стандартні швидкості руху для різних категорій транспортних засобів. Два основні піки на графіку, швидше за все, відображають швидкості легкових автомобілів і вантажівок.

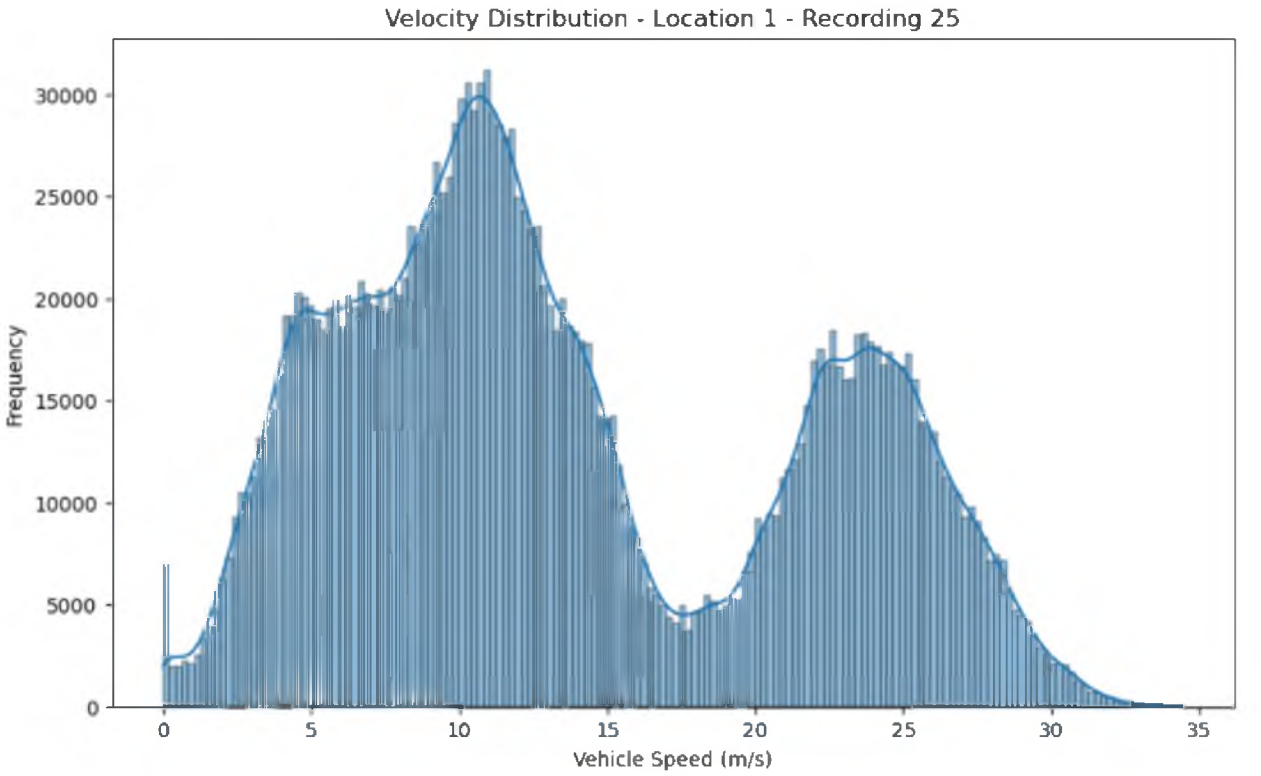

Рис. 9: Розподіл швидкості - Локація 1 - Запис 25

На другій гістограмі (Рис. 10) зображено розподіл загального прискорення транспортних засобів для того ж запису (Локація 1, Запис 25). Цей графік показує кількість транспортних засобів для кожного інтервалу прискорення. З графіка видно, що більшість транспортних засобів мають незначне прискорення, що є очікуваним для більшості транспортних засобів під час руху. Проте, існують деякі випадки надто високих значень прискорення, які можуть бути ознакою аномальної поведінки.

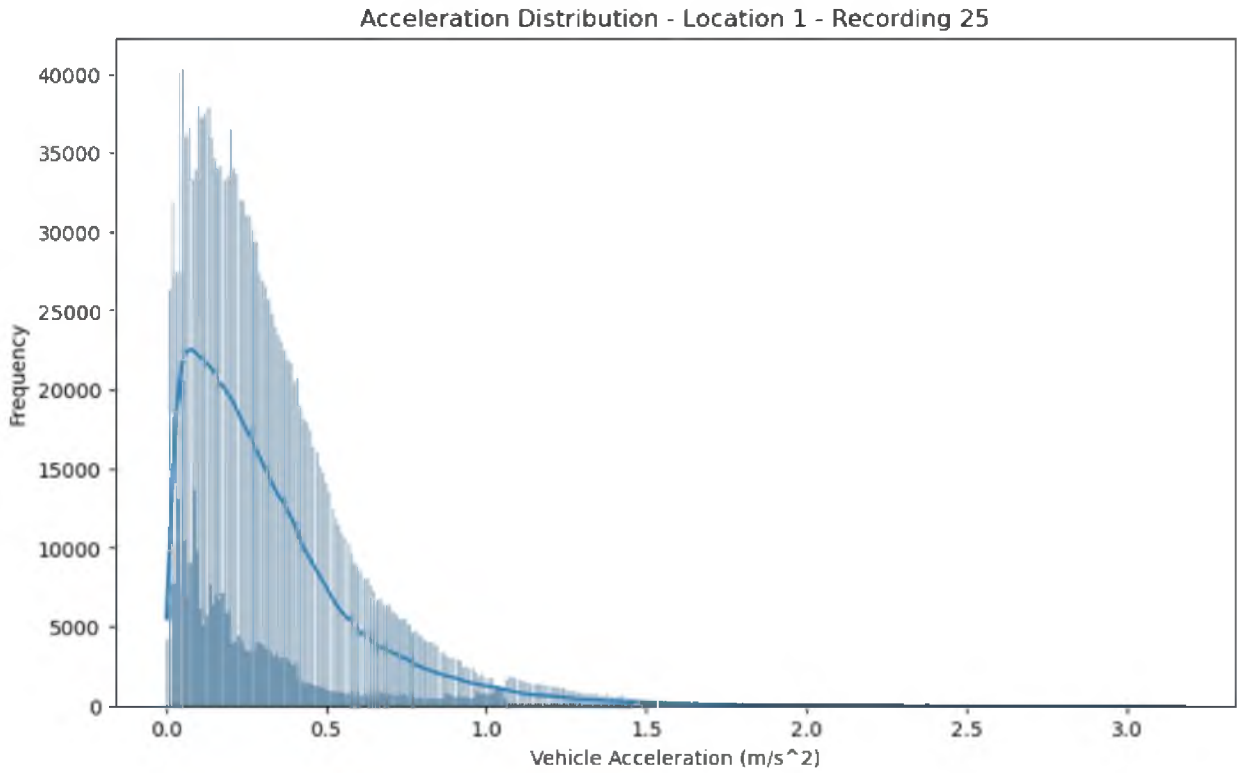

Рис. 10: Розподіл прискорення - Локація 1 - Запис 25

Використовуючи методи аналізу та візуалізації, було підтверджено важливість розробки таких ознак як швидкість та прискорення для ідентифікації відхилень від нормальної поведінки. Надалі ці ознаки будуть використовуватися для тренування алгоритмів LSTM та Isolation Forest, що дозволить поліпшити точність ідентифікації аномалій у траєкторіях транспортних засобів.

3.3. Застосування алгоритмів Long Short-Term Memory та Isolation Forest

В підрозділі 1.3. було детально розглянуто принципи роботи алгоритмів, які будуть використовуватись в експериментальній частині цієї кваліфікаційної

роботи, а саме: алгоритми LSTM та Isolation Forest для пошуку аномалій в траєкторіях транспортних засобів.

3.3.1. Застосування алгоритму LSTM

Як було згадано, алгоритм Long-Short-Term-Memory (LSTM) є підтипом рекурентних нейронних мереж, який вирізняється здатністю ефективно оброблювати послідовні дані, а саме числові ряди.

Етапи розробки LSTM алгоритму:

1. Вибір ознак та групування даних:
   - було обрано такі ознаки для закодовування: 'totalVelocity', 'acceleration', 'frontSightDistance', 'backSightDistance', 'dhw', 'thw', 'ttc',' class', 'recordingId';
   - дані було згруповано за унікальним номер транспортного засобу ('id') та номером запису ('recordingId').
2. Кодування класів:

- мітки класів транспортних засобів('car' or 'truck') було закодовано за допомогою LabelEncoder;

3. Закодовування даних у послідовності:

- для кожного транспортного засобу було створено послідовності з фіксованою довжиною 'n_steps' рівне 35. Значення 35 було обрано після аналізу розподілу довжини часових рядів (рис.11). Важливо зазначити, що значення 35 є компромісом між мінімізацією втрати інформації та збереженням достатнього обсягу даних для аналізу. Проте, у майбутньому рекомендується використати крос-валідацію для систематичного оцінювання різних довжин послідовностей та визначення оптимального значення 'n_steps'.

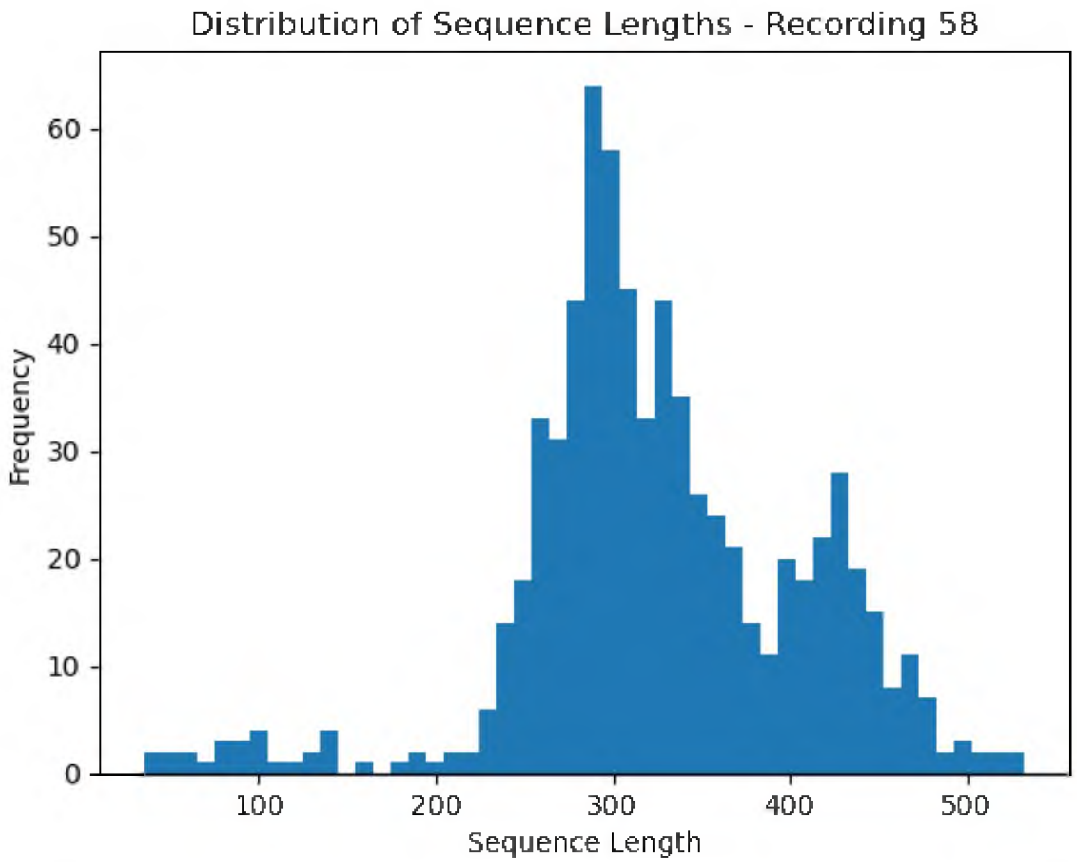

Рис. 11: Розподіл довжин послідовностей – Запис 58

4. Розбиття та підготовка даних:

- дані було розподілено на навчальну, валідаційну та тестові вибірки;
- було проведено перетворення типу даних у масиви NumPy, аби підготувати їх до навчання моделі.

5. Масштабування даних:

- розподілені дані було нормалізовано за допомогою StandardScaler для забезпечення продуктивності моделі.

6. Розробка архітектури моделі:

- створено модель LSTM, яка складається з одного LSTM - шару з 64 нейронами та вихідного шару з одним нейроном з активацією 'sigmoud' для класифікації типу транспортного засобу;

- модель було скомпільовано з використання оптимізатора ´adam’ та функцією втрат 'binary_crossentropy'.

7. Тренування моделі:

- модель будо натреновано на навчальному наборі даних з використанням валідаційношл набору даних для оцінки продуктивності протягом тренування.

8. Оцінювання моделі:

- модель було оцінено на тестовому наборі даних за допомогою наступних показників: 'Validation Loss', 'Validation Accuracy', 'Test Loss', 'Test Accuracy'.

3.3.2. Застосування алгоритму Isolation Forest

Для розширення та доповнення аналізу з виявлення аномалій в траекторіях транспортних засобів було також застосовано Isolation Forest як альтернативний підхід до пошуку аномалій.

Алгоритм Isolation Forest виявляє аномалії в траєкторіях транспортних засобів, аналізуючи кожну окрему точку в багатовимірному просторі ознак і визначаючи, чи є точка аномальною на основі ступеня її відхилення від нормальних точок. Детальний механізм роботи алгоритму Isolation Forest був описаний у підрозділі 1.3 "Порівняння технік машинного навчання у виявленні аномалій".

Етапи розробки алгоритму Isolation Forest:

1. Вибір ознак:

- було вибрано ознаки, які можуть вказувати на аномалію в траєкторії, а саме: 'x', 'y', 'totalVelocity', 'acceleration', 'frontSightDistance', 'backSightDistance', 'dhw', 'thw', 'ttc'.

2. Нормалізація даних:
- дані було нормалізовано для забезпечення рівномірності ознак.
3. Розбиття даних:
- дані було розподілено на навчальну, валідаційну та тестову вибірки за допомогою функції 'train_test_split'.
4. Тренування моделі:
- було використано алгоритм Isolation Forest з параметром contamination=0.05:
5. Виявлення та оцінка результатів:
- модель було оцінено на валідаційній та тестовій вибірках;
- кількість знайдених аномалій було вираховано.
6. Візуалізація отриманих результатів:
- було побудовано графіки для візуального оцінювання виявлених результатів у валідаційній та тестовій вибірках (рис. 12 та рис. 13). На графіках вісі X та Y представляють нормалізовані координати положення транспортних засобів. Вісь X відповідає за нормалізовану координату x, а вісь Y - за нормалізовану координату y. Ці координати є частиною просторових характеристик траєкторій транспортних засобів. На графіках значення 1.0 позначає точки жовтого кольору, визначені як аномальні, а значення 0.0 позначає фіолетові точки, визначені як нормальні.

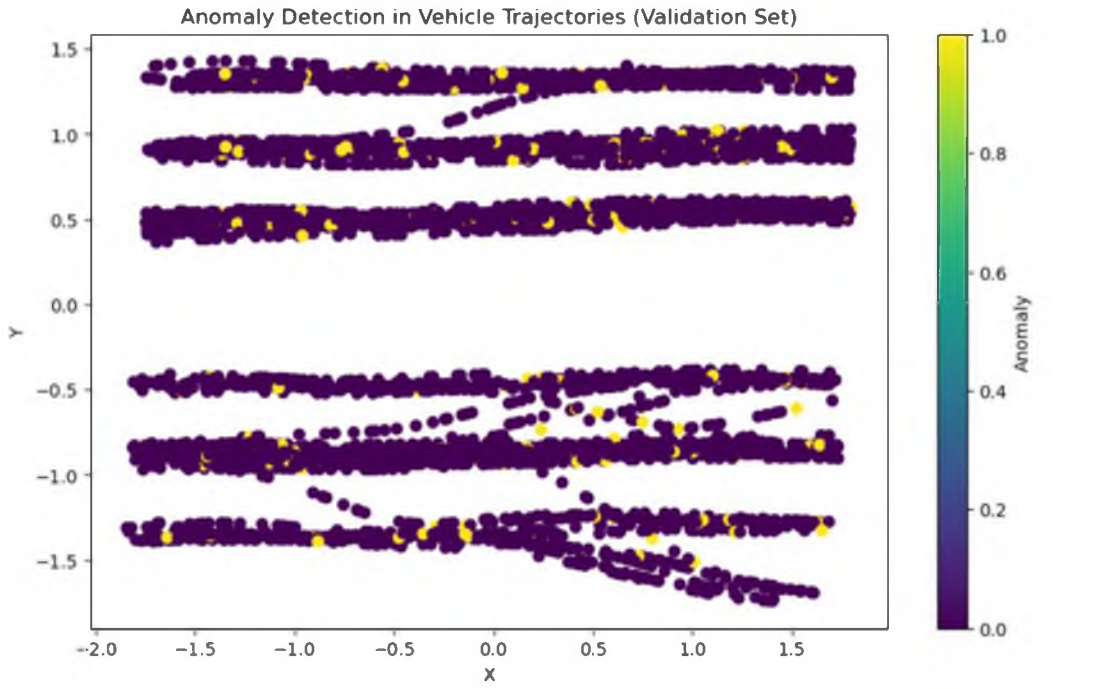

Рис. 12: Виявлення аномальних точок у траєкторіях транспортних засобів за допомогою алгоритму Isolation Forest (на валідаційному наборі даних)

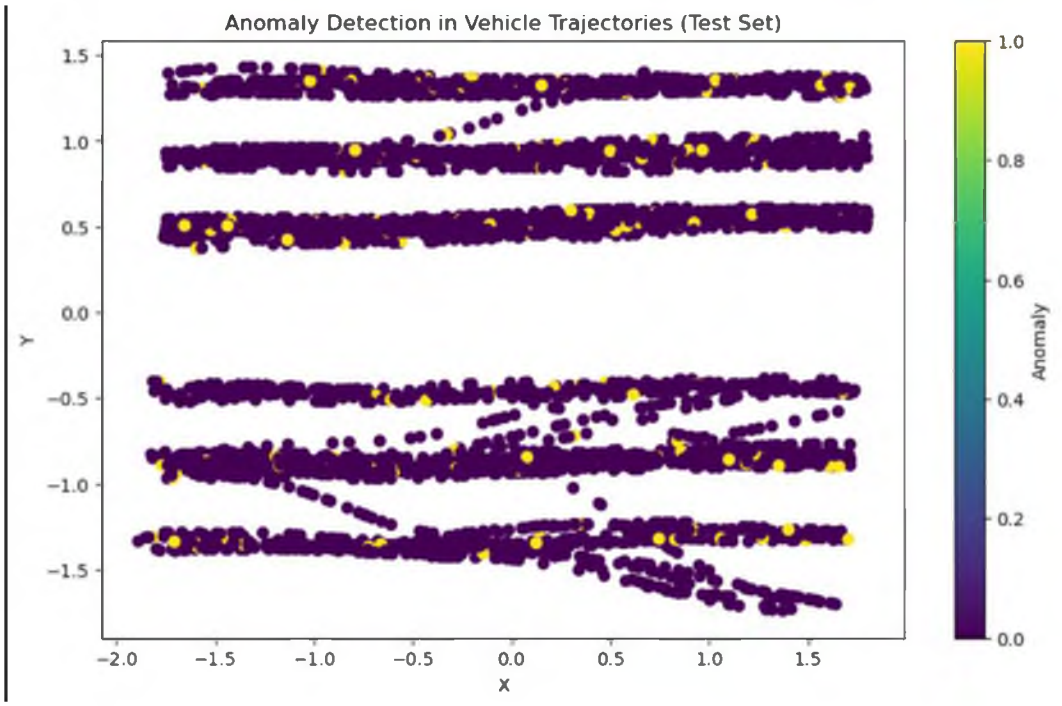

Рис. 13. Виявлення аномальних точок у траєкторіях транспортних засобів за допомогою алгоритму Isolation Forest (на тестовому наборі даних)

3.4. Аналіз результатів та оцінка ефективності запропонованих алгоритмів

У цьому підрозділі буде проведено оцінку ефективності запропонованих алгоритмів.

3.4.1. Оцінка ефективності алгоритму LSTM:

Для оцінки ефективності алгоритму LSTM було використано наступні метрики: validation loss, validation accuracy, test loss, test accuracy, precision, recall, F-1 score.

Отримані результати (для прикладу взято дані з запису номер 58):

- Validation Loss: 0.10469966381788254;
- Validation Accuracy: 0.955390214920044;
- Test Loss: 0.10325556993484497;
- Test Accuracy: 0.9579886794090271.

Модель досягла високої точності на тестовому наборі даних, що свідчить про добру узагальнюючу здатність моделі на нових даних.

Також були обчислені метрики для класифікації класу "машина":

- Точність (Precision) для класу "машина": 0.9776;
- Повнота (Recall) для класу "машина": 0.9669;
- F1-міра (F1-score) для класу "машина": 0.9722.

Ці метрики показують, що розроблена модель має високу здатність правильно класифікувати машини серед усіх передбачених як машини (precision), а також серед усіх реальних машин (recall). F1-score, який є гармонійним середнім між precision та recall, свідчить про загальну ефективність моделі.

3.4.2. Оцінка ефективності алгоритму Isolation Forest:

Результати роботи алгоритму були візуалізовані у вигляді двох графіків для валідаційного та тестового наборів даних (Рис. 12).

- кількість точок, визначених як аномальні у валідаційному наборі: 1786;
- кількість точок, визначених як аномальні у тестовому наборі: 1763.

Важливо зазначити, що кількість виявлених точок, визначених як аномальні, завжди є різною при тренуванні алгоритму Isolation Forest, адже цей алгоритм використовує випадкове підвибірку даних для побудови ізоляційних дерев. Цей процес включає випадкові вибірки і випадковий вибір ознак, що призводить до невеликої варіативності у результатах при кожному запуску алгоритму. Це є природною властивістю алгоритму і може бути враховано шляхом проведення кількох запусків і усереднення результатів для досягнення більш стабільних оцінок.

Головний принцип роботи розробленого алгоритму LSTM : алгоритм кваліфікує аномалії як випадки, коли очікуваний клас транспортного засобу не співпадав з передбаченим моделлю класом. Тобто факт неправильної класифікації класу транспортного засобу був взятий як індикатор потенційної аномалії.

Головний принцип роботи розробленого алгоритму Isolation Forest : алгоритм знаходить точки, які відрізняються від нормальної поведінки транспортних засобів на основі таких характеристик траєкторій, як просторово-швидкісні параметри та відстані до інших транспортних засобів.

Підсумовуючи, обидва алгоритми, LSTM і Isolation Forest, продемонстрували здатність ефективно виявляти аномалії у траєкторіях транспортних засобів, але мають свої переваги і недоліки:

- LSTM: Ефективний алгоритм для пошуку аномалій в траєкторіях транспортних засобів, оскільки він здатний враховувати часові залежності і допомагає виявляти патерни у послідовних даних. Проте, підготовка та попереднє закодовування даних в траєкторії вимагає велику кількість часових та обчислювальних ресурсів
- Isolation Forest: Завдяки здатності ізолювати потенційно аномальні точки на основі багатовимірних ознак, алгоритм Isolation Forest доводить свою ефективність у виявленні підозрілих точок, що відрізняються від нормальних патернів за своїми просторово-швидкісними параметрами.

Вибір алгоритму повинен базуватися на пріоритетних показниках для виконання завдання: швидкість, легкість масштабування, точність та простота у використанні. Залежно від конкретних вимог, один алгоритм може мати переваги над іншим, забезпечуючи оптимальний баланс між продуктивністю та ефективністю.

ВИСНОВКИ

У науковій роботі було детально розглянуто актуальність задачі пошуку аномалій у траєкторіях руху транспортних засобів. Особливу увагу приділено статистичним методам пошуку аномалій та сучасним методам машинного навчання, що застосовуються для розв'язання цієї проблеми. Значна увага була приділена вибору та розгляду наборів даних, які можуть бути використані в контексті цієї наукової задачі. Було детально проаналізовано важливість створення нових атрибутів, таких як загальна швидкість і прискорення, для покращення навчання моделей. У практичній частині дослідження було застосовано дві технології машинного навчання для виявлення аномалій у траєкторіях руху транспортних засобів: Long Short-Term Memory (LSTM) та Isolation Forest. Ці алгоритми демонструють особливості використання двох різних підходів до розв'язання задачі: контрольованого та неконтрольованого машинного навчання.

Результати цього дослідження мають важливе значення для подальшого розвитку методів аналізу траєкторій, що, своєю чергою, сприятиме покращенню безпеки та ефективності транспортних систем. Подальший аналіз запропонованих алгоритмів сприятиме створенню передових інтелектуальних систем керування транспортом, здатних автономно виявляти та реагувати на аномалії в реальному часі. Це значно підвищить безпеку дорожнього руху та ефективність транспортної інфраструктури.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Jiao, Ruochen, Juyang Bai, Xiangguo Liu, Takami Sato, Xiaowei Yuan, Qi Alfred Chen, and Qi Zhu. "Learning Representation for Anomaly Detection of Vehicle Trajectories." Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2023, pp. 9699-9706. DOI: 10.1109/IROS55552.2023.10342070.
2. Chandola, Varun, Arindam Banerjee, and Vipin Kumar. "Anomaly Detection: A Survey." ACM Computing Surveys, vol. 41, no. 3, article 15, July 2009, pp. 1-58. DOI: 10.1145/1541880.1541882.
3. Bou Nassif, Ali, et al. "Machine Learning for Anomaly Detection: A Systematic Review." IEEE Access, vol. 9, 2021, pp. 78658-78700. DOI: 10.1109/ACCESS.2021.3083060
4. Laxhammar, Rikard, and Göran Falkman. "Online Learning and Sequential Anomaly Detection in Trajectories." IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 36, no. 6, 2014, pp. 1158-1173. DOI: 10.1109/TPAMI.2013.2297918.
5. Giacomo, V., and A. Capasso. *Hands-On Industrial Internet of Things: Create a Powerful Industrial IoT Infrastructure Using Industry 4.0*. Packt Publishing, 2018. https://www.oreilly.com/library/view/hands-on-industrial-internet/9781789537222/efbdcbd9-8914-43db-8045-54f80221e5fc.xhtml
6. Algan, Gokhan, and Ilkay Ulusoy. "Time Series Anomaly Detection with Adjusted-LSTM GAN." arXiv preprint arXiv:2009.02040, 2020.
7. Al Farizi, W. S., I. Hidayah, and M. N. Rizal. "Isolation Forest Based Anomaly Detection: A Systematic Literature Review." 2021 8th International Conference on Information Technology, Computer and Electrical

Engineering (ICITACEE), Semarang, Indonesia, 2021, pp. 118-122. DOI: 10.1109/ICITACEE53184.2021.9617498.

8. Mane, Deepak, Sunil Sangve, Gopal Upadhye, Sahil Kandhare, Sanket Sonar, and Satej Tupare. "Detection of Anomaly Using Machine Learning: A Comprehensive Survey." International Journal of Emerging Technology and Advanced Engineering, vol. 12, 2022, pp. 134-152. DOI: 10.46338/ijetae1122_15.
9. Wu, Yang, Junhua Fang, Wei Chen, Pengpeng Zhao, and Lei Zhao. "Safety: A Spatial and Feature Mixed Outlier Detection Method for Big Trajectory Data." Information Processing & Management, vol. 61, 2024, p. 103679. DOI: 10.1016/j.ipm.2024.103679.
10. ApolloScape Dataset. ApolloScape: Large-Scale, Diverse, High-Definition Dataset for Autonomous Driving. ApolloScape, https://apolloscape.auto/ . Accessed 2023-12-10 .
11. ZenTraffic Dataset. ZenTraffic: A Comprehensive Traffic Dataset for Research and Development. ZenTraffic, https://zen-traffic-data.net/english/ . Accessed 2023-12-10.
12. Krajewski, R., J. Bock, L. Kloeker, and L. Eckstein. "The highD Dataset: A Drone Dataset of Naturalistic Vehicle Trajectories on German Highways for Validation of Highly Automated Driving Systems." *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, 2018, pp. 2118-2125. IEEE, DOI:10.1109/ITSC.2018.8569552.
13. U.S. Department of Transportation Federal Highway Administration. (2016). Next Generation Simulation (NGSIM) Vehicle Trajectories and Supporting Data. [Dataset]. Provided by ITS DataHub through Data.transportation.gov. Accessed 2023-12-10 from http://doi.org/10.21949/1504477

14. Zhang, Q., S. Hu, J. Sun, Q. A. Chen, and Z. Morley. "On Adversarial Robustness of Trajectory Prediction for Autonomous Vehicles." *arXiv preprint*, arXiv:2201.05057, 2022.
15. Pazho, A. D., V. Katariya, G. A. Noghre, and H. Tabkhi. "VT-Former: A Transformer-Based Vehicle Trajectory Prediction Approach for Intelligent Highway Transportation Systems." *arXiv preprint*, arXiv:2311.06623, 2023.
16. Lumen Learning. "Addition of Velocities." Physics, Lumen Learning. Accessed March 2, 2024. https://courses.lumenlearning.com/suny-physics/chapter/3-5-addition-of-velocities/.

ІНТЕРНЕТ РЕСУРСИ

1. Статистика ДТП в Україні за 2023 рік - дані від Патрульної Поліції
2. Z-Score: Meaning and Formula
3. Statistical Methods for Anomaly Detection using Python: A Comprehensive Guide
4. ApolloScape Trajectory Data
5. Zen-Traffic-Data
6. Next Generation Simulation (NGSIM) Open Data
7. Introduction to Long Short-Term Memory