RBDPM: Risk-Based Differential Privacy Model for Trajectory Data

Alofe Olasunkanmi Matthew

Department of Computer Science and Engineering
University of Derby

This dissertation is submitted for the degree of
Doctor of Philosophy

March 2025

This thesis is dedicated to my loving parents and siblings, who have always been my biggest supporters and have encouraged me to pursue my dreams. Their unwavering faith in me and their selfless sacrifices have been my source of strength and motivation throughout this journey. I am also grateful to my dear friends and colleagues who have been with me every step of the way, offering me invaluable advice and sharing both my struggles and successes. Your companionship and support have made this journey worthwhile and enjoyable. Lastly, I dedicate this work to all those who believe in the power of knowledge and the pursuit of truth. To all those who have contributed to my work, both in ways big and small, I offer my heartfelt thanks. May this work contribute to our collective understanding and inspire those who embark on this challenging, yet rewarding, path of research. This doctoral thesis is dedicated to you all with gratitude and appreciation.

Declaration

I hereby declare that, except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and acknowledgements. This dissertation contains fewer than 65,000 words including appendices, bibliography, footnotes, tables, and equations, and has fewer than 150 figures. Furthermore, I acknowledge that any assistance I have received in my research work and the preparation of this thesis has been explicitly acknowledged.

Alofe Olasunkanmi Matthew
March 2025

Acknowledgements

Firstly, I would like to express my deepest appreciation to my supervisors, Dr. Kaniz Fatema, Dr. Maria Papadaki, Dr. Muhammad Ajmal Azad, and Prof. Fatih Kurugollu. Their invaluable advice, enthusiasm, support, and encouragement have played a pivotal role throughout my research journey. Their insightful feedback and patient guidance, coupled with their vast knowledge and experience, have not only propelled my academic research forward but have also been a source of inspiration in my daily life. Special thanks are due to Dr. Hany Atlam, Dr. Ovidiu Bagdasar, and Dr. John Pannerselvam, whose technical expertise has been instrumental in advancing my study. Your unwavering support and valuable assistance have been indispensable to my work. I extend my heartfelt gratitude to the Department of Electronics, Computing, and Mathematics at the University of Derby. Their continual support and invaluable input have played a crucial role throughout my research. To all the members of the University of Derby, I express my sincere thanks. It is through their kindness, help, and support that my study and life in the United Kingdom have been such a remarkable journey. They have truly made my time here an enriching and unforgettable experience.

Last but certainly not least, I would like to express my profound gratitude to my parents, siblings friends, lab mates, and colleagues. Their unwavering understanding and encouragement over the past few years have been the cornerstone of my success. Without their unfailing support and faith in me, completing this study would have been an insurmountable task. In conclusion, I am eternally grateful to everyone who has contributed to my journey for their part in shaping this rewarding and fulfilling experience. Your contributions will be forever remembered and appreciated.

Abbreviations

ACPP Adaptive Changing of Periodical Pseudonyms

APT Advanced Persistent Threats

ARM Advanced Reduced Instruction Set Computer Machine

ADF Augmented Dickey-Fuller

AR AutoRegressive

ARIMA AutoRegressive Integrated Moving Average

ARMA AutoRegressive Moving Average

ACF Autocorrelation Function

bi-LSTM Bidirectional Long Short-Term Memory

C3 Carto-Car Cooperation

CEL Common Exposure Library

CVSS Common Vulnerability Scoring System

COBIT Control Objectives for Information and Related Technologies

CEM Convolutional Embedding Model

CNNs Convolutional Neural Networks

CMIX Cryptographic Mix-Zone

CRRA Cyber Risk Remediation Analysis

CTSA Cyber Threat Susceptibility Analysis

DLNLP Deep Learning Approach for Next Location Prediction

DoS Denial of Service

DP Differential Privacy

DES Discrete-Event Simulation

DM Downtown Model

DMLP Dynamic Mix-Zone for Location Privacy

EKF Extended Kalman Filter

FM Flow Model

GPR Gaussian Process Regression

GPL General Public License

GPS Global Positioning System

HMM Hidden Markov Model

IRAM 2 Information Risk Assessment Methodology 2

ISF Information Security Forum

ISACA Information Systems Audit and Control Association

ITS Intelligent Transport System

IoT Internet of Things

KPSS Kwiatkowski, Phillips, Schmidt, and Shin

LBS Location Based Services

LPPMs Location Privacy-Preserving Mechanisms

LSTM Long Short-Term Memory

MM Manhattan Model

MAE Mean Absolute Error

MSE Mean Squared Error

MOL Method and Objective Library

MMM Mixed Markov-Chain Model

MPE Mobility Pattern Embedding

MOVE Mobility Model Generator for Vehicle Networks

MA Moving Average

MTD Moving Target Defense

MOP Mutually Obfuscating Paths

NALUs Neural Arithmetic Logic Units

NIST National Institute of Standards and Technology

OCTAVE Operationally Critical Threat, Asset, and Vulnerability Evaluation

OLS Ordinary Least Squares

PSC Pseudonyms Synchronously Change

PII Personally Identifiable Information

POI Places of Interest

PACF Partial Autocorrelation Function

RM Random Model

RGRASPSemTS Reactive Greedy Randomized Adaptive Search Procedure for Semantic Semi-Supervised Trajectory Segmentation

RSU Road-Side Unit

RMSE Root Mean Squared Error

SARIMA Seasonal ARIMA

SMS Short Message Service

SM Simple Model

SUMO Simulation of Urban Mobility

SEI Software Engineering Institute

STF-RNN Space-Time Features-Based Recurrent Neural Network

SP Special Publication

STRAW STreet Random Waypoint

SAS/ETS Statistical Analysis System Econometrics and Time Series Analysis

SVR Support Vector Regression

SeqPT Sequential Prefix Tree

TTC Time-To-Collision

TTP Tactics, Techniques, and Procedures

TIDE Technology Insertion, Demonstration, and Evaluation

TAL Threat Agent Library

TARA Threat Assessment & Remediation Analysis

UKF Unscented Kalman Filter

VANET Vehicular Ad-Hoc Network

V2I Vehicle-to-Infrastructure

V2V Vehicle-to-Vehicle

VM Virtual Machine

WHO World Health Organisation

Abstract

Personal safety applications enable users to communicate emergency situations to relevant third parties and local authorities. Location-Based Services play a crucial role in the capture and exchange of data, including location and personal identifiable information, to better inform emergency response efforts. Maximising the effectiveness of these safety applications requires the data to be accurate and informative yet prevent the exposure of sensitive user information. Current solutions often fail to adequately protect this sensitive data in the attempt to maintain accurate and useful information for emergency response. Therefore, personal safety solution safety applications should be able to protect the privacy of individuals without compromising the overall utility and accuracy of the data. This thesis presents a Risk-Based Differential Privacy Model for Location Data that is designed to assess safety-critical factors and attributes associated with users and scenarios to provide a dynamic balance for trajectory data utility and privacy trade-off. The model assesses the safety-critical factors facing the user from the data and quantifies the risk in the Hazard Assessment Module. The quantified risk informs the level of privacy parameters in the Privacy Preservation Module, which will determine the levels of noise to be added to the dataset in the Noise Application Module to ensure that lower risk levels can afford maximum privacy, whereas high-risk scenarios will result in reduced privacy without losing data utility. The resulting noise-injected trajectory dataset is processed using the Linear Regression model to validate this concept and evaluate the impact of data utility and privacy trade-off in the dataset during processing. The performance of the dataset to retain utility while ensuring privacy during processing is analysed using evaluation criteria metrics that explore the efficiency, generalisation, and robustness of the dataset. The metrics outcome show that the noise-injected dataset can maintain good data utility while safeguarding the privacy of the user when processed. The outcome emphasises the importance of exploring factors and attributes associated with safety-critical data by the user and the dataset to dynamically find the optimal balance for the data utility and privacy trade-off.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Personal safety has become important with consistent cases of assault and the disappearance of individuals being reported. Some of these cases occur on lonely roads, quiet environments, and areas with low lighting, while others occur in moving vehicles (Lewis 2016; Qureshi 2015; "AsiaOne" 2017; *BBC News* 2012; McSorley 2018; Utehs 2018; *BBC News* 2018; Longnecker 2019). The circumstances leading to these assaults typically involve the vehicle starting from a stationary position or maintaining a consistent movement pattern, which then transitions into a random movement pattern during the assault. The assailant during the assault attempts to flee his current location diverting attention away from the immediate surroundings to a remote location. In trying to change location, they frequently adopt a random movement pattern characterised by sudden changes in speed, acceleration, and direction. This endangers all road users and can lead to safety situations unavoidable for other road users. The safety of victims in such situations can be preserved by immediate implementation of rapid response measures that mitigate the impact of the attack on the victim and prevent subsequent damage to the road infrastructure and the users, as evidenced in various instances (McSorley 2018; Utehs 2018; *BBC News* 2018; Longnecker 2019).

Personal safety solutions function is a critical application that leverages the deployment of an Internet module and various other components to empower road users with the ability to communicate with third parties and local authorities during emergencies. Thus, facilitating the dissemination of information with accurate emergency reports and alerting authorities necessary to provide immediate assistance to victims (Rohilla, Deshwal, and Balasubramanian 2019). There is infrastructure that is instrumental in the acquisition and exchange of essential information related to transportation, such as location-based data and services, traffic-related information and accident-related data. These communication channels provide information on users' places of interest, road behaviour, movement patterns, and the shortest travel routes to ultimately improve the overall road experience for users (Yang and Hua 2019;

Lin 2017). The information derived from these communication channels is used to make informed decisions about traffic conditions and to prevent traffic injuries, which according to the World Health Organisation (WHO), are the eighth leading cause of death in all age groups and the main cause of death in children and young adults aged 5 to 29 years (W.H.O 2018). This information is applicable in safety and non-safety related functions, including vehicle safety, automated toll payment, traffic management, enhanced navigation, and other location-based services (Feng et al. 2015).

Location-Based Services (LBS) represent a significant component that facilitates the acquisition and exchange of information such as traffic updates, weather forecasts, and recommendations for nearby shops or restaurants among network components. Location information by LBS contains more than coordinates or point of interests information, they also include data such as user's identity, spatial information, and temporal information. These services play an essential role in traffic management and Intelligent Transport Systems (ITS) (Sun and Kim 2021). The information used by LBS encompasses sensitive data related to Personally Identifiable Information (PII), traffic updates in real time, personal interests, shopping preferences, tourist routes, and recommendations for nearby Points of Interest (POI) intended to enhance the daily lives of individuals. This information contains inferable details that shed light on user lifestyle patterns, religious affiliations, and health conditions, making it susceptible to attacks from adversaries. Information leakage or disclosure represents a common risk associated with the exchange of data between nodes and LBS servers, potentially allowing attackers to intercept information transmitted within LBS and gain access to sensitive network data. This risk, among others, underscores the importance of safeguarding transmitted information within the system while preserving user privacy (Chan and Lars 2003; Kolvoord, Keranen, and Rittenhouse 2017; Gupta and Sutar 2014).

The protection of sensitive information within LBS is imperative and must be maintained consistently. Location privacy is a subcategory of data privacy, which revolves around an individual's expectation of moving through public spaces without their location information being systematically recorded. Key privacy concerns is preventing unauthorised disclosure, leakage, or exposure of an individual's past or present location and personal information. The rapid evolution of information technology and the constant growth in the volume of information incorporated into daily life require frequent adjustments in privacy expectations (Liu et al. 2018). The deployment of privacy techniques is vital to safeguarding user identity information, driving routes, and/or location data of users within the network by validating the legitimacy of each user's identity. Common attacks targeting location information include information forgery, information manipulation and alteration, replay attacks, message delay, and privacy breaches (Ferrag et al. 2018). The information acquired by LBS is vast and

subject to constant changes over time, exhibiting variable levels of importance and sensitivity, often linked to highly temporally correlated coordinates. These characteristics require ongoing preservation of privacy, with the application of varying degrees of protection to strike a balance between data security and utility (Liu et al. 2018; Ying, Makrakis, and Mouftah 2013; Julien et al. 2007; Buttyán et al. 2009; Li et al. 2019).

When an individual's safety is compromised, the level of safeguarding assigned to the information must be adjusted downward to enhance the information utility and facilitate the dispatch of rapid responses. Determining the appropriate level of protection during an event requires evaluating location data to determine the level of sensitivity that is necessary to trade off data utility and information security. This evaluation involves assessing the risks associated with the information and anticipating the challenges that may arise when attempting to enhance the utility of the information. For this evaluation, a robust risk management approach is employed, taking into account the dynamic characteristics and enhancements of the components in personal safety applications (Houmer and Hasnaoui 2020; Bayad, Rziza, and Oumsis 2016; Ren, Du, and Zhu 2011).

The primary challenge in determining the balance between data privacy and utility trade-off lies in selecting the appropriate equilibrium that maximises data utility without compromising information security and vice versa. LBS-collected information is dynamic and highly sensitive, necessitating the implementation of privacy mechanisms to prevent data disclosure or leakage.

Different traffic situations necessitate varying approaches to data utility in ensuring personal safety. In normal traffic conditions, where road users maintain stable speeds, predictable behavior, and safe following distances, the likelihood of accidents is significantly reduced. These conditions promote a safer driving environment by allowing adequate reaction time for braking, lane changes, and intersection navigation. However, in collision-prone traffic scenarios, which is characterised by sudden stops, rapid deceleration, and erratic lane changes and this escalates the risk to personal safety due to the unpredictability of road user behavior. These abrupt movements reduce reaction time, increase the likelihood of accidents, and present challenges in modeling real-world traffic behaviors (Tayal and Triphathi 2012; Fiore et al. 2007). Simulators are widely used to study traffic patterns and predict risk factors, yet they struggle to capture the full complexity of dynamic road conditions, aggressive driving behaviors, and congestion-induced hazards (Härri et al. 2006; Fiore et al. 2007; Kaisser, Gransart, and Berbineau 2012; Lim et al. 2017). The reliance on real-time location data for safety systems raises concerns regarding the trade-off balance between data utility and privacy. Striking a balance between leveraging location data for public safety and preserving individual privacy remains a significant challenge. Adaptive

privacy-preserving mechanisms, such as anonymization techniques and Differential Privacy, are essential to ensure that personal data are protected while still maintaining data utility (Wu et al. 2018; Madi and Al-Qamzi 2013; Boucetta, Guichi, and Johanyák 2021; Tayal and Triphathi 2012; Fiore et al. 2007).

The role of personal safety applications is to ensure that the safety of road users is maintained by providing components and features that improve road safety (Mohamed, Ahmed, and Sadek 2021). These applications rely on safety-critical information to reliably assess road conditions, allowing road users to make informed decisions about the necessary precautions for their safety. This application encompasses components such as collision detection and avoidance systems, as well as navigation and traffic awareness aids that contribute to a safe travel experience.

1.1 Motivation

The rapid development of connected devices has created a growing need for more robust personal safety measures and applications. Personal safety solutions aim to enhance user protection by offering an additional layer of information that helps minimize potential hazards. For these applications to be truly effective, the information they provide must remain reliable and accurate. The safety capabilities of personal safety applications can be significantly enhanced by incorporating a privacy-preserving mechanism that evaluates context-specific attributes (such as user conditions and situational variables). This mechanism determines the most appropriate level of privacy for each scenario, to dynamically set the optimal balance between data utility and privacy. Because different users and circumstances call for different privacy requirements, using an unsuitable privacy level can expose data to leaks and compromise personal safety. In contrast, the application of excessive privacy measures can reduce the overall usefulness of the information provided.

Personal safety applications commonly handle sensitive, safety-critical information that must be protected once the individual is no longer in immediate danger. When an individual's safety is at risk, rapid response is crucial to minimize the impact of any hazard and restore secure conditions (Yang and Hua 2019; Lin 2017). During periods when the person's well-being is secured, data privacy takes precedes and sacrificed over utility. Conversely, in emergencies where safety is compromised, privacy protection may be temporarily relaxed to enhance data utility and expedite assistance. Even under these conditions, it remains essential to preserve the confidentiality of critical user data, ensuring that privacy is not unduly compromised.

When an individual faces hazard during road travel, they often adopt manoeuvrers to avoid the incident that changes immediately changes the dynamics of their travel and jeopardise the safety of other road user. Such as turning into oncoming traffic which changes the structured safe travel constraints of oncoming users and disrupting their movement mechanics such as velocity, and stopping distance (Tayal and Triphathi 2012; Song et al. 2017; Tian et al. 2019; Xin et al. 2018). This chaotic traffic condition endangers road users and this situation requires favouring data utility over privacy to process data for assisting users ensure safe conditions are restored (Navidi, Camp, and Bauer 2004; Akhtar, Ergen, and Ozkasap 2014). The decrease and increase in privacy level alter the utility of the data captured by the system, thereby necessitating careful consideration of the impact that altering the privacy level would have on the information critical to the safety of the user.

Changes in environmental conditions influence the likelihood of a road incident and as the likelihood of incident occurring changes, the trade-off balance between data utility and privacy changes. This trade-off can be dynamically balanced by assessing the incident likelihood and depending on the safety condition inferred from the data, the trade-off balance shifts to prioritise either utility over privacy or privacy over utility without significantly compromising of both features.

1.2 Problem Definition

Personal safety solutions are designed to protect individuals from accidents, facilitate immediate assistance during emergencies, and improve overall road safety. These tools handle sensitive and time-critical data, such as a user's location and identity, making robust privacy-preserving mechanisms essential. Research in this field continues to expand, focusing on real-time monitoring of vehicles, traffic conditions, road hazards, and user status to provide timely support when unexpected situations arise. Given the sensitive nature of these data, safeguarding user information is crucial for maintaining both security and peace of mind.

These solutions enable individuals to request assistance and maintain safe road conditions while in motion. It is crucial that every component of these systems is secured to maintain user safety whenever the emergency response functionality is active. Given that such solutions rely on sensitive, real-time information, implementing a robust privacy-preserving mechanism is vital for safeguarding user data and while maintaining overall system efficiency during emergencies (Gharaibeh et al. 2017; Li et al. 2020).

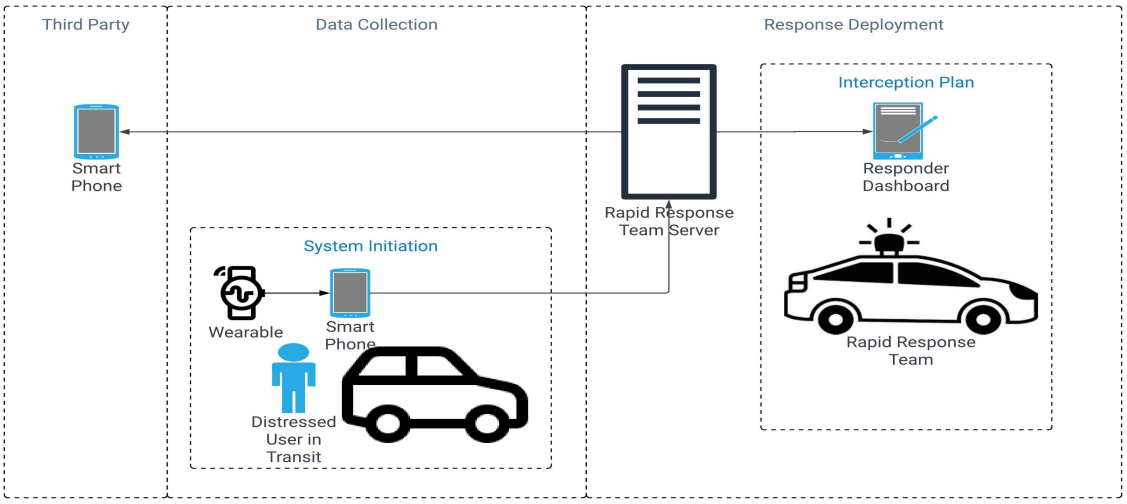

Figure 1.2.1: Emergency System Operational Scenario in Personal Safety Solution Concept

Various personal safety solutions have been proposed to support individuals during emergencies. Many of these tools can track and monitor a user's location, alert and contact third parties about an incident, capture image or video evidence, offer self-defense features such as tasers, and emit loud alarms. While such features are effective for users in stationary settings, those who are on the move require additional capabilities to account for their mobility situations. Efficient data processing within personal safety solutions can greatly enhance functionality by providing real-time insights into user context and attributes. However, to fully leverage these benefits, it is crucial to strike the right trade-off balance between data utility and user privacy. Robust data protection measures such as Differential Privacy can help ensure sensitive information remains secure, while maintaining sufficient data utility for rapid response teams to deliver timely assistance.

The preservation of privacy in these solutions and dynamically balancing the utility and privacy level trade-off based on data requirement are the main focus of this research, driving the need for innovative privacy-preserving mechanisms to protect sensitive location and trajectory data. This thesis adopts the operational model of the Raspberry Pi-based personal safety solution proposed by Sogi et al. (2018), illustrated in Figure 1.2.1, to address situations in which an individual experiences distress while on the move. In this model, the user relies on an interconnected emergency system: activation occurs when a wearable device's panic button is pressed, triggering a signal to a GPS-enabled mobile phone. The phone tracks and records the user's movements, sending continuous alerts, updates, and notifications to designated contacts and emergency responders. These real-time location and trajectory

updates enable third parties to identify the user's path, positioning themselves in an optimal location to provide assistance if a rapid response is needed.

During safety-critical situations such as collisions, there is high demand for data utility while that changes when safe condition resumes. Balancing data utility and user privacy trade-off is crucial for an effective emergency response system during incidents to dynamically shift in the direction needed for the scenario. High risk situations requires low privacy but high utility information while during low risk events, the requirements changes to high privacy and low utility. While high data utility about user location and movement can significantly improve rapid intervention, handling this information responsibly is paramount. Privacy preservation in these solutions has been neglected to maximize data utility, The preservation of privacy in these solutions has been quite non-existent due to the prioritisation of data utility, which enables rapid emergency intervention to restore safe conditions for road users.

The RBDPM presented in this thesis addresses the identified gap by integrating privacy preservation into the operational framework of solution. This integration ensures that user information is protected without significantly degrading data utility, thus enabling prompt and effective interventions. Moreover, the privacy preservation mechanism is dynamically adjusted to prioritize user safety, balancing data utility and privacy in accordance with the prevailing safety conditions.

1.3 Aim and Objectives

The aim of this thesis is to design and develop a novel Risk-Based Differential Privacy Model that dynamically balances the trade-off between preserving the privacy of location/trajectory data and maintaining the utility based on safety-critical information within the data. This model is intended to strike an optimal balance between data utility and user privacy trade-off by assessing distinct attributes within location/trajectory data. The assessments would determine the precise degree of noise required to dynamically meet the changing privacy needs of users and ensure their safety in diverse, safety-critical scenarios.

The aim is achieved by meeting the following objectives:

- **Objective 1:** Conduct a critical review of state-of-the-art personal safety solutions and privacy preservation methodologies for trajectory data, evaluating their strengths, limitations, and relevance to safeguarding sensitive location information in emergency response contexts.

- **Objective 2:** Implement a validation methodology through the location prediction mechanism, such as Linear Regression, to validate the performance of the novel privacy preservation framework proposed for trajectory data in personal safety applications.
- **Objective 3:** Design and develop a novel Risk-Based Differential Privacy Model that evaluates distinct attributes within trajectory data to produce a transformed dataset that can achieve an optimal balance between data utility and privacy trade-off for different risk levels and safety-critical scenarios during processing.
- **Objective 4:** Conduct a comprehensive experimental validation of the proposed Risk-Based Differential Privacy Model for trajectory data, processing the data using the implemented prediction model to demonstrate how assessing distinct data attributes influences the data privacy and utility trade-off hypothesis.

1.4 Contributions

The novel contributions of this thesis are as follows:

- An innovative Risk-Based Differential Privacy Model for trajectory data perturbation. This model prioritises safety-critical information to provide privacy preservation in safeguarding personal information in location datasets.
- A privacy preservation concept that can dynamically tune data privacy based on hazard thresholds to fine-tune trajectory data output. This adaptive solution that enhances operational efficiency and addresses privacy needs in unpredictable situations.
- A dynamic framework designed to balance the trade-off between data utility and privacy during processing, ensuring that sensitive information is protected without compromising utility. This approach is experimentally validated using Linear Regression predictive model, which evaluates framework effectiveness by measuring the ability to maintain data utility while applying privacy-preserving techniques.

These contributions advance the field of privacy preservation in personal safety applications by offering refined methods for managing the trade-off between privacy and data utility. Through the development and validation of the proposed Risk-Based Differential Privacy Model, this work lays a strong foundation for accurate, privacy-focused decision-making in both normal and high-risk traffic conditions, setting the stage for further research and innovation in this domain.

1.5 Thesis Outline

The structure of the remainder of this thesis is as follows:

Chapter 2 explores background information regarding the components that are involved in the Risk-Based Differential Privacy Model. The chapter evaluates location-based services, location prediction schemes for proof-of-concept validation, and the Differential Privacy mechanism. These components are crucial in creating an effective and secure privacy-preserving scheme.

Chapter 3 offers a comprehensive overview of current state-of-the-art personal safety applications. This chapter explores these applications, privacy preservation mechanisms, and the challenges facing these safety applications.

Chapter 4 explores the efficiency of the prediction models used for the validation of RBDPM on trajectory data. This chapter shows the implementation of the prediction model and the performance evaluation.

Chapter 5 delves into the novel concepts of the Risk-Based Differential Privacy Model. This chapter discusses the conceptual foundation of the model, the roles and interplay of the various modules, and their collective contribution to the model's architecture.

Chapter 6 details the proof-of-concept validation for the novel Risk-Based Differential Privacy Model. The chapter evaluates the performance of the noise-injected private data set during processing and the impact of assessing distinct attributes within the data to determine the data utility and privacy trade-off balance. This contributes to the dynamic balancing of data utility and privacy trade-off of trajectory data to meet the changing privacy requirements of scenarios and users.

Chapter 7 provides an overview of the research, examining the limitations of the research and future research directions for the research. It emphasises the promise of the model in augmenting the efficiency of safety and privacy measures in both VANET and trajectory data.

Chapter 2

Background on Risk-Based Differential Privacy Model

This chapter provides the essential background for understanding the development of the Risk-Based Differential Privacy Model (RBDPM). Central to the development of the RBDPM are Location-Based Tracking Services (LBS), which generate trajectory data critical for applications such as traffic management and emergency response. Location prediction schemes serves as the data processing framework for the validation of the model. Differential Privacy as the privacy preservation mechanism that protects data sensitivity. The evaluation metrics criteria for assessing the data utility and privacy trade-off balance during processing. The sections in this chapter provide the essential background for the components that make up the RBDPM and how it meets the privacy demands of safety-critical scenarios.

2.1 Location-Based Tracking Service (LBS)

The LBS plays a crucial role in the acquisition and processing of location data, using technologies such as GPS, wireless communication, and cloud computing. This service medium collects data about a user in a coordinated and systematic manner in exchange for personalised semantic information related to the current geographical position of the user (Das and Sadhukhan 2014; Shin et al. 2012). The response is personalised and provides versatile value added services ranging from navigation assistance services, intelligent tour guides, and PoI closest to the current geographical location of the user (Das and Sadhukhan 2014; Shin et al. 2012). LBS harnesses the operating capabilities of the Internet service, mobile service and GPS services of the requesting device to acquire data and exchange these data for better personalised road experiences such as places of interest, local amenities, and

traffic situations (Khan et al. 2015; Aasha, Monica, and Brumancia 2015; Kolvoord, Keranen, and Rittenhouse 2017; Lai et al. 2013).

The information transmitted by this service is sensitive and contains personal details attributed to the user. Information is not protected against leakage and attacks, making it vulnerable to threats. The utility of the information transmitted by the service is crucial and safeguarding the sensitivity of transmitted information within the service is imperative. This thesis works on balancing the data utility and privacy trade-off that can optimise the data utility and privacy of transmitted data within LBS.

The components that ensure the efficient operation of LBS within VANET are categorised into technological and data components. The technological components manage the representation and quality of the information acquired within the network. The main technologies used by LBS are as follows:

Position technology: This is responsible for the acquisition of data and the accuracy of the acquired location information sent in the query (Pontikakos et al. 2006; Khan et al. 2015).

Application technology: This is responsible for presenting the response data received from the server that serves as a response to the user request query (Das and Sadhukhan 2014; Gupta and Sutar 2014).

Data components handle information transmission between nodes within the network and manage the type of information the nodes access and present. The central data components operating within VANET are:

Geographic Data: These data manage geographic information, such as road structure and infrastructure of the geographical location, and the possible PoI within the vicinity of the location requested by the device (Das and Sadhukhan 2014).

Communication Data: This data component manages the information exchange between the LBS control centre and the requesting device. They are essential in maintaining high-quality communication and service for nodes within the network (Gupta and Sutar 2014).

2.2 Location Prediction Schemes

A comprehensive understanding of location prediction methodologies is fundamental for the experimental validation of the novel RBDPM in this thesis. The focus of this section is to analyse the implementation of location prediction models on location data. The proposed model will leverage location data collected through LBS during mobility tracking, to anticipate the user's location in real time. The evaluation of this model will be based on simulations, and the findings will be examined in detail.

The landscape of location prediction has seen a variety of methodologies aiming to achieve high accuracy. These range from traditional statistical methods to more advanced machine learning and deep learning models. Most of these studies have used historical location data and mobility patterns for predictions. Others have incorporated auxiliary factors, such as travel time and meteorological conditions. The location prediction literature classifies predictions into two categories: long-term and short-term predictions. Long-term predictions aim to anticipate the location of a vehicle over an extended period, such as hours or days, with the precision of this prediction classification being less critical. While short-term predictions aim to determine the vehicle's location within a few seconds, the accuracy of the predictions in this classification type is crucial, as they are required to provide a highly accurate forecast.

2.2.1 ARIMA Prediction Model

Time series analysis leverages historical data to forecast future events, particularly in systems exhibiting trends and seasonality. The AutoRegressive Integrated Moving Average (ARIMA) model combines three components to identify, estimate, and suggest the most suitable ARIMA notation for the data (Kumar and Anand 2014; Chen, Yuan, and Shu 2008; Islam and Raza 2020; Alofe et al. 2019; Ye, Szeto, and Wong 2012):

- **Auto-Regressive (AR):** Uses past values to predict future observations.
- **Integrated (I):** Applies differencing to achieve stationarity.
- **Moving Average (MA):** Models forecast errors as a linear combination of past error terms.

The implementation of the ARIMA model for the prediction for location data has been deployed to forecast traffic flow (Lin 2016; Kumar and Vanajakshi 2015; Ghosh, Basu, and O'Mahony 2005; Williams and Hoel 2003; Dhingra, Mujumdar, and Gajjar 1993), traffic volume (Tong and Xue 2008; Ding et al. 2011; Wang et al. 2017), traffic speed (Song et al. 2019), and traffic road congestion (Alghamdi et al. 2019).

An ARIMA model is denoted as ARIMA(p,d,q) (Kumar and Anand 2014; Chen, Yuan, and Shu 2008; Islam and Raza 2020; Alzyout and Alsmirat 2020; Lai and Dzombak 2020), where:

- p: Number of lag observations in the AR component.
- d: Degree of differencing in the I component required for stationarity.

- q: Order of the MA component.

The general form of the ARIMA model (Agrawal and Adhikari 2013; Liu et al. 2016; Yuanhui et al. 2022) is expressed in Equation 2.2.1:

$$Y_t = \alpha + \sum_{i=1}^{p} \beta_i Y_{t-i} + \varepsilon_t + \sum_{j=1}^{q} \phi_j \varepsilon_{t-j}, \tag{2.2.1}$$

where α is a constant, β_i and ϕ_j are the coefficients for the AR and MA terms respectively, and ε_t is the error term.

Model Implementation Steps

The essential steps in implementing an ARIMA model (Panneerselvam, Liu, and Antonopoulos 2018; Ariyo, Adewumi, and Ayo 2014; Liu et al. 2016; Mohamed 2020)are:

1. **Data Exploration and Preparation:** Plot the time series and perform statistical tests (e.g., the Augmented Dickey-Fuller test) to assess stationarity. Apply transformations or differencing to achieve stationarity.

2. **Model Identification:** Use autocorrelation (ACF) and partial autocorrelation (PACF) plots to determine suitable values for p and q.

3. **Model Fitting:** Estimate parameters using methods such as maximum likelihood estimation.

4. **Model Diagnostics:** Evaluate residuals for randomness, stationarity, and normality.

5. **Forecasting and Evaluation:** Generate forecasts using the fitted model. Evaluate performance using metrics such as Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE).

6. **Model Selection:** Compare models using information criteria like the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC):

$$AIC(p) = n\ln(\sigma_e^2/n) + 2p, \tag{2.2.2}$$
$$BIC(p) = n\ln(\sigma_e^2/n) + p\ln(n), \tag{2.2.3}$$

where n is the number of observations and σ_e^2 is the error variance.

The model performance is validated by determining the model's accuracy and identify any potential overfitting through comparing the predicted values to the actual values (Alzyout and Alsmirat 2020; Ingdal, Johnsen, and Harrington 2019; Thiruchelvam et al. 2021). . A modified version, such as Seasonal ARIMA (SARIMA), extends the ARIMA model to capture periodic effects and has been effectively applied in domains like traffic forecasting in vehicular networks (Alzyout and Alsmirat 2020; Lippi, Bertini, and Frasconi 2013).

2.2.2 Regression Tree Ensemble

A Regression Tree Ensemble combines multiple decision trees to predict a continuous target variable. Each tree is trained on different subsets of the data and features, and the final prediction is typically obtained by averaging the outputs of all trees. This ensemble approach reduces variance and overfitting, often leading to improved accuracy compared to a single decision tree.The suggestion that regression models perform admirably compared to neural networks by Wang et al. (2019) and the implementation of the regression model to predict location has prompted the use of a regression tree ensemble on the data set to evaluate performance and possibly provide improved accuracy relative to the ARIMA model (Goli, Far, and Fapojuwo 2018; Zhao et al. 2020b).

Lu et al. (2012) used a regression ensemble learning model to predict the next place of the Nokia Mobile Data Challenge 2012 spatial-temporal location information dataset. The prediction models employed during the challenge used spatial-temporal information within the data to make predictions (Etter, Kafsi, and Kazemi 2012; Wang and Prabhala 2012; Gao, Tang, and Liu 2012; Lu et al. 2012). Two common ensemble techniques are:

- **Bagging (Bootstrap Aggregating):** The model combines Bootstrapping and Aggregation into one model which works on improving unstable estimation or classification schemes. Builds multiple trees on bootstrapped (randomly sampled with replacement) subsets of the training data. The predictions are averaged to reduce variance and improve stability. Bagging is a variance and Mean Squared Error (MSE) reduction technique that is effective in improving the predictive performance of regression or classification trees (Zhao et al. 2020b; *MathWorks* n.d.[a]; Anagnostopoulos et al. 2009).

- **Boosting:** Trains trees sequentially, where each subsequent tree focuses on correcting the errors of the previous one. This method reduces bias by giving more weight to misclassified observations (Zhao et al. 2020b; *MathWorks* n.d.[a]; Anagnostopoulos et al. 2009)..

Choosing between the two depends on the specific problem and the trade-off between computational cost and model performance (*MathWorks* n.d.[b]; Zhao et al. 2020b).

Ensemble Regression Tree Model Implementation

The implementation process involves the following key steps:

1. **Data Preparation:** This involves cleaning, transforming, and normalizing the data.
2. **Feature Selection:** Identify the most relevant features affecting the target variable.
3. **Base Model Training:** Train individual decision tree models on different subsets of the data.
4. **Ensemble Construction:** Combine the base models using either bagging or boosting techniques.
5. **Hyperparameter Tuning:** Optimize model parameters using techniques such as grid search or random search.
6. **Model Evaluation:** Validate performance through cross-validation or hold-out testing.
7. **Model Deployment:** Deploy the best-performing model for predictions on new data.

2.2.3 Linear Regression

Linear Regression is a widely used statistical method for modeling the relationship between a dependent variable and one or more independent variables using a linear equation. The model is typically trained via Ordinary Least Squares (OLS), which minimizes the sum of squared errors (Maulud and Abdulazeez 2020; Khuri 2009). The basic form is represented in equation (2.2.1) below:

$$y = \beta_0 + \beta_1 X + \varepsilon, \tag{2.2.1}$$

where:y is the dependent variable, X is the independent variable, β_0 and β_1 are the intercept and slope, and ε is the error term.

In applications such as privacy-preserving location prediction, Linear Regression is favored due to the computational efficiency, ease of interpretation, and compatibility with Differential Privacy. Differential Privacy mechanisms inject noise into the data to obscure individual information while retaining overall data trends. Other reasons aside from the fact that Linear Regression models are faster to train and perform predictions faster, unlike the ARIMA and Regression Tree Ensemble models, which are computationally intensive for

large datasets, include the accurate estimation provided by the Linear Regression model when the independent and dependent variables within the dataset are linear. The model is capable of incorporating the operation of a privacy preservation mechanism with minimal impact on the performance of the model.

Linear Regression Prediction Model Implementation

The Linear Regression process uses suitable packages and libraries such as the sci-kit library in the Python programming language to perform the operation.

Figure 2.2.1 illustrates the overall Linear Regression prediction scheme and the essential implementation steps are discussed below:.

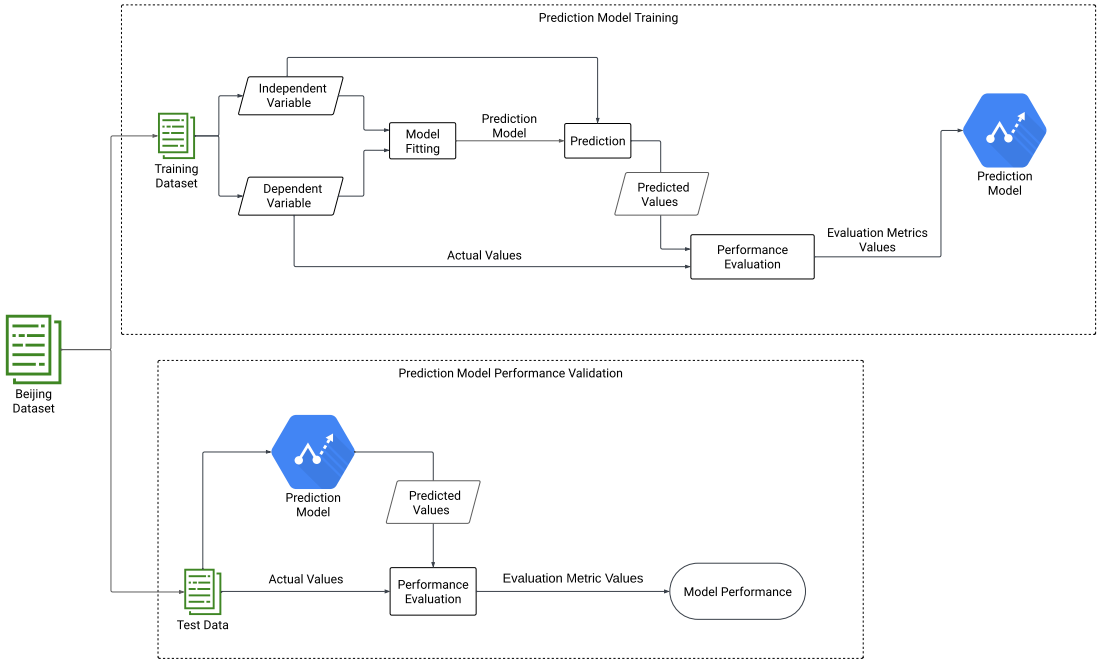

Figure 2.2.1: Linear Regression Location Prediction Scheme Operation

1. **Data Acquisition and Preparation:** Splitting the dataset into training and test sets and perform necessary cleaning and normalization.

2. **Model Fitting:** Separate the data into independent and dependent variables. Fit the Linear Regression model using tools such as Python's scikit-learn library.

3. **Prediction and Evaluation:** Generate predictions on the test dataset. Evaluate performance using metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE).

4. **Privacy Integration:** Implement Differential Privacy by adding controlled noise to the data, ensuring individual privacy with minimal impact on model accuracy.

2.3 Differential Privacy

Differential privacy (DP), first introduced by Dwork (2008), provides a mathematical framework for protecting individual privacy in statistical datasets. The primary goal is to ensure that the output of any data analysis does not reveal sensitive information about any single individual, regardless of an adversary's background knowledge or computational resources. This is achieved by incorporating carefully calibrated random noise into query results, where the level of noise is controlled by a non-negative privacy parameter ε. A higher privacy parameter ε means more noise and thus greater privacy, at the expense of data utility (Alda and Rubinstein 2017; Sarwate and Chaudhuri 2013).

Formally, DP shown in equation (2.3.1) guarantees that for any two datasets D_1 and D_2 differing by a single record, and for any set of possible outputs S, a randomized function K satisfies

$$Pr[K(D_1) \in S] \leq \exp(\varepsilon) \times Pr[K(D_2) \in S]. \tag{2.3.1}$$

A central concept in this framework is the sensitivity of a function Δf, which quantifies the maximum change in f's output when one record in the dataset is altered. Sensitivity is defined as

$$\Delta f = \max_{D_1, D_2} \frac{\|f(D_1) - f(D_2)\|}{\|D_1 - D_2\|}. \tag{2.3.2}$$

DP has been applied for protecting trajectory data and location-based services, where traditional methods often aggregate or cloak data, complicating trajectory analysis. For instance, Chen et al. (2012) proposed releasing large amounts of sequential data using a hybrid granular prefix tree (SeqPT) with Laplace noise to obscure node counts, though this approach struggled with computational complexity in high-dimensional spatio-temporal datasets. Improvements were later introduced by Al-Hussaeni et al. (2018), who proposed the SafePath algorithm—a noisy prefix tree model that reduced empty node generation but required higher privacy budgets for complex datasets. Similarly, Zhao, Dong, and Pi (2019) proposed an SR-tree structure in the Cons-SRT algorithm to mitigate non-location sensitive information attacks, though query efficiency was affected by the increased tree complexity.

In practice, the implementation of DP typically relies on one of several noise-addition mechanisms. The Laplace mechanism is the most widely used (Gursoy et al. 2018; Hua, Gao, and Zhong 2015; Galdames, Gutierrez-Soto, and Curiel 2019; Jang et al. 2012).; it adds Laplace-distributed noise to numerical query outputs according to equation (2.3.3)

$$F(x) = f(x) + Lap\left(\frac{s}{\varepsilon}\right), \tag{2.3.3}$$

where s represents the sensitivity of the function f and $Lap(s/\varepsilon)$ denotes a random draw from the Laplace distribution centered at zero with scale s/ε. This mechanism is favored for the simplicity and the ability to minimize mean-squared error for identity queries (Holohan et al. 2020; Koufogiannis, Han, and Pappas 2015).

Alternatively, the Gaussian mechanism introduces independent Gaussian noise with variance calibrated based on the function's sensitivity and the privacy parameters, thereby ensuring approximate differential privacy. It is characterized in equation (2.3.4) by the condition

$$\sigma \geq \sqrt{2\log(1.25/\delta)}\,\frac{\Delta_2 f}{\varepsilon}, \tag{2.3.4}$$

where $\Delta_2 f$ is the L_2 sensitivity and δ is an additional parameter that quantifies the probability of the privacy guarantee being slightly exceeded.

For non-numeric outputs or when a broader range of utility functions is required, the Exponential mechanism is employed. This mechanism selects outputs by sampling from a probability distribution with input x defined by a utility function υ and sensitivity $\Delta\upsilon$. The probability of choosing an output o is given in equation (2.3.5) by

$$Pr[o] = \frac{\exp\left(\frac{\varepsilon\,\upsilon(x,o)}{2\Delta\upsilon}\right)}{\sum_{o'}\exp\left(\frac{\varepsilon\,\upsilon(x,o')}{2\Delta\upsilon}\right)}. \tag{2.3.5}$$

In summary, DP offers a robust and quantifiable approach to preserving individual privacy by ensuring that the outcome of any data analysis is minimally affected by the presence or absence of any single record. By introducing noise through mechanisms such as the Laplace, Gaussian, or Exponential methods, DP enables the secure analysis of sensitive datasets across diverse applications in statistical analysis, machine learning, and data mining.

2.3.1 Differential Privacy and Risk Relationship

Dandekar, Basu, and Bressan (2021) explored the relationship between risk and privacy with a primary focus on the Laplace noise mechanism. This thesis focuses on the analytical

relationship that links the privacy level ε and risk Υ for the Laplace mechanism, which is calibrated by the sensitivity Δ_f and the privacy level ε_0 as specified in Eqn (2.3.1). The use of the formula makes the correlation between privacy level ε and risk Υ possible to achieve the goals of the Laplace mechanism. This equation demonstrates the risk Υ_1 that is bounded between 0 and 1 for the Laplace mechanism of $Lap(\frac{s}{e})$ for a numerical query denoted by $f : D \rightarrow R^k$ and satisfies the condition where the privacy level ε is greater than 0 (Dandekar, Basu, and Bressan 2021)

$$\Upsilon_1 = \frac{P(T \leq \varepsilon)}{P(T \leq \varepsilon_0)} \tag{2.3.1}$$

Here T is a random variable that follows a distribution with a density function that can be mathematically expressed as a certain equation.

The analytical formula to represent risk Υ_1 is subjective and Eqn (2.3.2) demonstrates the methodology used to calculate the level of privacy ε_0 based on the associated privacy at risk Υ_1 (Dandekar, Basu, and Bressan 2021).

$$\varepsilon = In\left(\frac{1}{1-\Upsilon_1(1-e^{-\varepsilon_0})}\right) \tag{2.3.2}$$

The relationship between the risk of impact and the amount of privacy noise added to the data is inverse. As the risk of impact increases, the amount of privacy noise that must be added to the data decreases. This inverse relationship is a result of the privacy goals specified for the data and the measures put in place to achieve them. To maintain a high level of privacy, a larger amount of noise must be added to the data when the risk of impact is low. However, if the risk of impact is high, the amount of noise that must be added to the data can be reduced. In this way, privacy noise acts as a protective mechanism for the data and the privacy goals, with a specific value depending on the perceived risk of impact.

2.3.2 Evaluation Criteria

Evaluation criteria are vital tools for assessing the performance and effectiveness of prediction models, particularly in terms of predictive capacity, generalisability, and overall operational quality. In the context of DP, these criteria help determine how well a model preserves data utility while protecting sensitive information. The concept of utility loss quantifies the amount of data that must be altered or removed to maintain an acceptable trade-off between utility and privacy in DP-based regression models (Eom, Lee, and Leung 2018).

Saleem et al. (2021) and Yang et al. (2018) has used these metrics to privacy preservation in their work to understand the relationship between data privacy, utility, and accuracy that

impacts the data utility-privacy trade-off balance. Saleem et al. (2021) work shows that data utility is impacted by the introduction of noise as the values of the privacy parameter ε increase. The comparison of the performance of the DP model to a non-private one assists in the assessment of the trade-off between privacy and utility. A small difference between the metrics of both models indicates that utility is provided while privacy is guaranteed (Zhao et al. 2020a).

In regression analysis, the accuracy of predictions is often evaluated by measuring the residual spread, which is the difference between the actual and predicted values using metrics such as the Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE). A model whose MAE, MSE, and RMSE values are closer to zero indicates that the predictions are closely aligned with actual outcomes, even in the presence of noise introduced by DP (Fan et al. 2020; Saleem et al. 2021).

The *Mean Absolute Error (MAE)* is defined in equation (2.3.1) as

$$\mathrm{MAE}(y,\hat{y}) = \frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}-1} |y_i - \hat{y}_i|, \tag{2.3.1}$$

where n_{samples} is the number of data points, y_i is the actual value, and $\hat{y}_i$ is the predicted value. Since MAE focuses on absolute differences, it is robust to outliers and remains unaffected by the direction of errors. In a DP framework, MAE quantifies how injected noise affects prediction accuracy and shows the efficiency of the model in balancing the utility and privacy trade-off; a low MAE suggests that the Regression-Based Differential Privacy Model (RBDPM) retains high utility despite privacy constraints (Gupta et al. 2021; Yan et al. 2023; Hao, Wu, and Wan 2023; Jiang et al. 2021b).

The *Mean Squared Error (MSE)* is given in equation (2.3.1) by

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2, \tag{2.3.2}$$

where n representing the total number of data points and y_i the actual value of the i-th data point. The $\hat{y}_i$ is the predicted value of the i-th data point and $\sum_{i=1}^{n}$ is the summation symbol that indicates that we are summing the squared differences for all data points from $i = 1$ to $i = n$. $\frac{1}{n}$ is the average of the squared differences by dividing the sum by the total number of data points. It penalises larger errors due to the squaring operation. In a DP setting, a small MSE indicates that, despite the noise, the model effectively captures the underlying data patterns and maintains an acceptable level of accuracy (Hao, Wu, and Wan 2023; Chicco, Warrens, and Jurman 2021; Hodson, Over, and Foks 2021).

The *Root Mean Squared Error (RMSE)* is the square root of the MSE and shown in equation (2.3.3):

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}. \tag{2.3.3}$$

where RMSE represents the Root Mean Squared Error, and n is the total number of data points. The y_i is the actual value of the i-th data point and $\hat{y}_i$ is the predicted value of the i-th data point. $\sum_{i=1}^{n}$ represents the sum of the squared differences for all data points from $i = 1$ to $i = n$. $\frac{1}{n}$ represents the average of the squared differences by dividing the sum by the total number of data points.

The Root Mean Squared Error (RMSE) is measured in the same units as the target variable, providing an interpretable metric that allows the assessment of the model's predictive accuracy in a meaningful and intuitive manner. A lower RMSE reflects closer agreement between predicted and true values, implying that noise perturbation has only a limited adverse effect on data utility (Zhang et al. 2022; Yan et al. 2023; Neera et al. 2021; Jiang et al. 2021b).

In summary, these evaluation metrics play a pivotal role in model selection and tuning by providing quantitative insights into the trade-off between privacy and prediction performance. By comparing DP-enabled models against non-private baselines, model configurations that deliver an optimal balance between data utility and privacy protection can be identified.

Chapter 3

State-of-the-Art Personal Safety Solution

Personal safety solutions aim to enhance the security of individuals, particularly road users, by minimizing accident risks, promoting safe behavior, and enabling rapid emergency responses. These systems leverage technologies such as microcontrollers, smartphones, and wearables to monitor user conditions, track locations, and alert third parties during distress. However, the collection and transmission of sensitive data for example, location traces, vital signs, and event footage introduce significant privacy risks, often overlooked in existing designs. Differential privacy (DP) offers a mathematically grounded framework to protect individual data while preserving utility, yet the standard form applies uniform privacy guarantees, which may not suit the variable risk profiles in personal safety contexts. This systematic review assesses current personal safety solutions, identifies shortcomings, and explores the potential of risk-based DP to address these gaps, laying the groundwork for a novel model.

This systematic review adopts a narrative synthesis with a systematic search approach, a hybrid methodology that combines a structured, reproducible search strategy with a qualitative synthesis of findings to explore the structured approach to identify and analyze relevant studies, aligning with guidelines from the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) framework adapted for a qualitative focus (Helbach et al. 2023; Parums 2021).

Literature was sourced from academic databases including IEEE Xplore, ACM Digital Library, and ResearchGate. Search terms included combinations and variations of “Personal safety device”, “Personal emergency alert system”, “Location-based safety application”, “IoT personal safety”, “Women safety application”, “Location Privacy”, “Location tracking safety apps”. Studies published between 2010 and 2022 were prioritised, although earlier foundational works were included if they were frequently cited or demonstrated clear relevance. Inclusion criteria encompassed Peer-reviewed journal articles, conference proceedings, and high-quality technical reports that focused on women’s safety, personal

safety, and emergency alert systems that covers variety of user groups as shown in Table 3.1.1. Solutions that include location tracking, alert mechanisms, or preventive measures that presented evidence, performance evaluations, technical operations, or distress detection for solution were reviewed. Exclusion criteria filtered out studies that are non-technological interventions, news articles, literature that does not present results or peer-reviewed, and when full text is not available or is of insufficient quality.

The analysis categorised solutions by evaluating their functionality, mobility, and data protection measures. These categories are discussed below

3.1 Personal Safety Solutions Applications

Personal safety applications that rely on location or trajectory data are designed to provide rapid emergency response by tracking users and alerting third parties when distress is detected. However, while many solutions offer effective tracking and alerting mechanisms, yet many current solutions transmit sensitive data without robust safeguards, thereby compromising user privacy.

Sharma et al. (2017) presented an Advanced Reduced Instruction Set Computer Machine 7 (ARM7) processor-based safety device that monitors real-time location and issues loud alerts when activated. Although the device can notify nearby individuals and transmit the current location to third parties, it suffers from several drawbacks. The bulky design prevents continuous tracking, and the lack of privacy-preserving mechanisms means that the transmitted location trail is exposed. Similarly, Bhavale et al. (2016) proposed a portable safety device integrated with a bus tracking system. While this solution can capture and transmit both location information and images, it requires pre-installation on vehicles, which limits portability, and does not implement any robust safeguards to protect the sensitive data it collects.

In response to the mobility limitations of microcontroller-based systems, Pawar et al. (2018) proposed a wearable safety device that incorporates sensors to monitor vital signs, track movement, and capture images connected to a micro-controller. Despite the advanced features, the device faces challenges related to computational overhead and rapid battery drain, which limit the continuous operation. it does not incorporate any dynamic privacy-preserving measures to protect the sensitive location data it processes. Monisha et al. (2016) developed another solution using an ARM controller with GSM, GPS, Bluetooth, and RF modules that communicates with an Android application. Although it reliably sends SOS messages with location data even under low connectivity, the approach suffers from mobility challenges and a lack of data protection measures.

Choudhary et al. (2017) presented an automated safety device that combines multiple sensing units—heartbeat and temperature sensors along with a panic button—connected to an ATmega8L microcontroller and interfaced with GPS and GSM modules. The system continuously monitors physiological parameters and, upon detecting deviations from set thresholds, activates an alert by sending the user's location to a third party. While this approach enhances responsiveness through sensor fusion, it is prone to false positives due to normal variations in sensor readings, leading to unnecessary activations and resource wastage. Additionally, the absence of privacy-preserving techniques means that sensitive location data is directly transmitted, raising significant privacy concerns.

The category evaluated above has highlighted limited mobility issues, prompting the exploration of smartphone-based personal safety solutions that leverage the inherent mobility and connectivity of modern smartphones to track users' location, send alerts to third parties, and notify nearby individuals during emergencies. These solutions offer enhanced scalability and real-time data utility; however, they often fall short in preserving the privacy of sensitive location and trajectory data—a gap that poses significant risks in personal safety applications.

Shinde et al. (2012) presented an Android-based personal safety solution that alerts third parties when danger is detected. Upon activation, the system acquires the user's current location and sends it via SMS or email. While effective for immediate alerting, the solution does not monitor the user's subsequent movement, making it unsuitable for dynamic scenarios. Moreover, the use of HTTP/SOAP protocols for data transmission provides inadequate safeguards, exposing the location data to potential interception.

Vithu is another smartphone-based application that initiates the cycle of operation with a double-press activation, subsequently capturing and transmitting the user's location at two-minute intervals until the operation is terminated. Although Vithu monitors the trajectory of the user's path, there are significant concerns regarding the safeguarding of the transmitted information; without robust privacy measures, sensitive data may be vulnerable to unauthorized access or leakage (Harikiran, Menasinkai, and Shirol 2016; Thavil, Durdhawale, and Elake 2017).

BSafe is designed to continuously track the user's location and send alerts to third parties with a single tap, while triggering an alarm to notify nearby individuals. However, the constant transmission of location updates increases the risk of data disclosure, as the system lacks the necessary safeguarding measures to protect sensitive information, thereby compromising user privacy (Azman et al. 2018).

Smartphone-based personal safety solutions have increasingly leveraged the mobility and connectivity of modern devices to provide scalable functionality through continuous tracking of location or trajectory data. During events, access to devices may be deterred, and contact

of the third party may be hampered, thereby leading to the preventive safety solutions. These systems are designed not only to alert third parties during emergencies but also enable users to make informed decisions about their routes and the areas they visit.

Safetipin is a smartphone application that aims to empower users by providing real-time safety assessments. The app monitors the user's location and computes a safety score for an area based on user-reported disturbances and risk factors. Additionally, it displays alternate routes to the destination and allows users to invite a third party to track their movements. Although these features enhance situational awareness, the underlying methodology for calculating safety scores has been shown to be vulnerable. Reports indicate that the score can be manipulated, where unsafe areas might be assigned artificially high safety ratings, and thereby potentially misdirecting users into hazardous zones (Viswanath and Basu 2015; Kartik, Jose, and MK 2017; Manazir, Govind, and Rubina 2019). This vulnerability highlights a significant gap in the integration of robust privacy-preserving and data integrity mechanisms within the system.

Similarly, Street Smart proposed by Chaudhari et al. (2018) offers users detailed contextual information about locations through articles, reviews, and safety-level recommendations, enhanced by augmented reality (AR) and sentiment analysis. While this approach provides a comprehensive overview to aid in decision-making, it is accompanied by a lack of information security measures. The absence of proper safeguards means that sensitive location data and user interactions are at risk of exposure, thereby undermining the privacy of the individuals relying on the app.

Furthermore, to the preventive solutions, there is a category of alarm-based systems that serve to alert nearby individuals and third parties when a distress event occurs. These devices typically function by emitting a loud ringing sound and transmitting the user's approximate location. While the immediate notification function is crucial for rapid response, these systems often lack continuous tracking capability and employ simplistic data transmission methods.

StreetSafe, as proposed by Yarrabothu and Thota (2015), is a smartphone-based safety application that integrates multiple distress response features such as activating an alarm, sharing a user's location on social media, sending an SMS to preselected contacts, and placing a call to a third party. Although this multifaceted approach aims to maximize data utility by disseminating location information for rapid emergency response, a critical concern lies in the method of broadcasting sensitive location data into the public domain. This practice exposes users to significant privacy risks, as the publicly accessible location data can be intercepted or misused by malicious actors. Consequently, while StreetSafe achieves high

immediate accessibility, it fails to adequately safeguard the trajectory data that is integral to personal safety applications, thus presenting a substantial privacy-utility trade-off.

In a related approach, Srikrishna and Veena (2017) proposed a mechanism that leverages network provider information to improve emergency communication. Upon activation, the system retrieves the user's network provider and clusters nearby users within the same provider, then broadcasts the location information of the user in distress to this confined group. Although this proximity-based clustering is an innovative attempt to limit the exposure of sensitive data, it still involves the transmission of location information without robust safeguards. The risk of privacy breaches remains high if the data is intercepted within the network cluster. Moreover, the mechanism does not incorporate continuous tracking of location changes, which diminishes the efficacy in dynamic emergency scenarios where real-time trajectory data is crucial.

Consequently, they are susceptible to information leakage, as the sensitive location data is not protected by encryption or other privacy-preserving techniques. This led to wearable personal safety solutions category reviewed below that is designed to complement smartphone-based systems by providing additional, immediate layers of protection through integrated sensors and communication modules. These devices capture and transmit location or trajectory data during distress events, yet significant gaps remain regarding the preservation of sensitive data and the balance between data utility and privacy.

Patel and Hasan (2018) designed a smart bracelet equipped with various sensors to detect assaults and analyze sensor data using machine learning on an Arduino controller. The bracelet communicates with a smartphone app via Bluetooth, transmitting the user's location to a third party when distress is detected. Although this approach enables rapid alerts, the reliance on sensor readings can result in false positives—normal safe postures might be misinterpreted as threats—leading to unnecessary activations. Moreover, the system does not support continuous location tracking during transit, thereby limiting the usefulness of the trajectory data for ongoing situational awareness. Critically, no privacy-preserving measures are implemented, leaving the sensitive location data vulnerable during transmission.

A similar challenge is observed in the smart shoe approach proposed by Viswanath, Pakyala, and Muneeswari (2016). This system uses sensors connected to a microcontroller, which communicates with the user's smartphone via Bluetooth. Activation is achieved by a specific foot tap gesture, and the system then sends an SMS with the user's location. While this method offers an intuitive trigger based on user movement, the operation is contingent upon the user being in motion; when the user is stationary, even during distress, the system may fail to trigger. Additionally, the approach does not incorporate robust privacy safeguards, thereby exposing sensitive location data to potential interception.

The concept of electronic jackets, as proposed by Gadhave et al. (2017), Shaikh and PB (2008), Priya et al. (2021), and Bhadula, Benjamin, and Kakkar (2021), further extends the wearable safety paradigm by integrating GSM, GPS, microcontrollers, cameras, and buzzers into a garment. These jackets can emit loud alerts, deliver electric shocks in self-defense, and continuously transmit the wearer's location to a designated contact. However, the bulky design and limited adaptability to varying climates impede user mobility and comfort. More importantly, continuous transmission of location and trajectory data without proper encryption or adaptive privacy controls raises substantial privacy concerns.

Additional smartphone-based solutions, such as FightBack (Miriyala et al. 2016; Yarrabothu and Thota 2015), Vanitha Alert (Hariharan et al. 2021; Walkunde, Shinde, and Pandhare 2022), Raksha-Women Safety Alert (Saranya et al. 2021; Prashanth, Patel, and Bharathi 2017), and Glympse (Reddy et al. 2021; Aminuddin et al. 2019), offer similar functionalities by sending SMS alerts with location data. While these applications improve scalability and the immediacy of emergency response, they also share the common shortcoming of transmitting sensitive data without adequate protection, thereby risking unauthorized tracking or data breaches.

Personal safety solutions that leverage location and trajectory data incorporate a diverse array of features designed to rapidly alert third parties and provide immediate assistance during emergencies. These solutions, spanning both smartphone-based and wearable devices, offer significant improvements in mobility and scalability by harnessing real-time location tracking. They enable users to communicate distress through multiple channels such as SMS, email, social media, or direct phone calls thus enhancing the likelihood of prompt intervention.

Despite these advances, a critical gap emerges in the safeguarding of sensitive location information. Many current systems prioritize high data utility and accurate, continuous tracking to ensure timely emergency response. However, they often do so at the expense of robust privacy protection. Sensitive data is frequently transmitted without adequate encryption or privacy-preserving measures, thereby exposing users to risks such as unauthorized tracking, data breaches, and potential misuse of personal information.

In conclusion, while personal safety solutions based on location and trajectory data have substantially improved emergency response capabilities, they remain vulnerable due to insufficient privacy safeguards. This research is focusing on introducing privacy preservation to safeguard data transmitted by these solutions. Furthermore, the research will be extended to dynamically balance data privacy levels while preserving data utility, based on the user's safety conditions.

Table 3.1.1: Systematic Review Study Criteria and Key Features

Reference	Detect Distress	Alert Third Parties	Location Updates	Physical Defense	Privacy Protection	Type	Included
Sharma et al. (2017)				X	X	Microcontroller	Yes
Bhavale et al. (2016)				X	X	Microcontroller	Yes
Pawar et al. (2018)					X	Microcontroller	Yes
Monisha et al. (2016)				X	X	Microcontroller	Yes
Choudhary et al. (2017)					X	Microcontroller	Yes
Shinde et al. (2012)				X	X	Smartphone	Yes
Harikiran, Menasinkai, and Shirol (2016)				X	X	Smartphone	Yes
Thavil, Durdhawale, and Elake (2017)				X	X	Smartphone	Yes
Azman et al. (2018)				X	X	Smartphone	Yes
Yarrabothu and Thota (2015)				X	X	Smartphone	Yes
Walkunde, Shinde, and Pandhare (2022)				X	X	Smartphone	Yes
Muralidhar and Bharathi (n.d.)				X	X	Smartphone	Yes
Chand et al. (2015)				X	X	Smartphone	Yes
Kanagaraj, Arjun, and Shahina (2013)				X	X	Smartphone	Yes
Rengaraj and Bijlani (2016)				X	X	Smartphone	Yes
Viswanath and Basu (2015)	X	X		X	X	Preventive	Yes
Kartik, Jose, and MK (2017)	X	X		X	X	Preventive	Yes
Manazir, Govind, and Rubina (2019)	X	X		X	X	Preventive	Yes
Chaudhari et al. (2018)	X	X		X	X	Preventive	Yes
Srikrishna and Veena (2017)				X	X	Preventive	Yes
Patel and Hasan (2018)				X	X	Wearable	Yes
Viswanath, Pakyala, and Muneeswari (2016)				X	X	Wearable	Yes
Gadhave et al. (2017)					X	Wearable	Yes
Shaikh and PB (2008)					X	Wearable	Yes
Priya et al. (2021)					X	Wearable	Yes
Bhadula, Benjamin, and Kakkar (2021)					X	Wearable	Yes
Miriyala et al. (2016)				X	X	Smartphone	Yes
Hariharan et al. (2021)				X	X	Smartphone	Yes
Saranya et al. (2021)				X	X	Smartphone	Yes
Prashanth, Patel, and Bharathi (2017)				X	X	Smartphone	Yes
Reddy et al. (2021)				X	X	Location Sharing	Yes
Aminuddin et al. (2019)				X	X	Location Sharing	Yes

3.2 Privacy Preserving Mechanism

Location-based services involve transmitting sensitive personal and location information, which may be exploited to deduce an individual's identity or behavior (Kalaiarasy, Sreenath, and Amuthan 2019). The protection of user privacy have led to the proposal of several mechanisms. The most notable ones include:

Pseudonyms : Pseudonyms replace true identities with temporary identifiers to obscure a user's identity. Approaches include: Synchronous pseudonym changes that randomly exchange pseudonyms and vehicle status information (Liao and Li 2009). Cooperative pseudonym-changing processes based on the number of neighboring vehicles (Pan and Li 2013). Despite their effectiveness, message content may still leak identifying information (Khacheba et al. 2017).

K-Anonymity : K-anonymity ensures that a user's location is indistinguishable from at least $k-1$ other users, thus limiting the success probability of linking attacks below $\frac{1}{k}$ (Kido, Yanagisawa, and Satoh 2005; Masoumzadeh and Joshi 2011). Enhanced methods use cloaking strategies: **Data-dependent cloaking** creates anonymity regions based on the spatial distribution of users. **Space-dependent cloaking** forms regions covering the entire anonymizer area (Gedik and Liu 2007; Kang and Meng 2012; Shokri et al. 2010; Zuberi, Lall, and Ahmad 2012; Bettini, Mascetti, and Wang 2008). A key challenge is balancing privacy with service quality trade-off.

Group Signatures : Group signatures allow any member of a group to sign messages on behalf of the group without revealing individual identities (Chaum and Van Heyst 1991; Yue et al. 2019). Hybrid approaches offer conditional anonymity (Rajput et al. 2017), though they often incur significant computational overhead.

Mix-Zones: Mix-zones are defined spatial regions where users change their pseudonyms, making it hard for adversaries to link old and new identities (Beresford and Stajano 2004). For a mix-zone to be effective, it should: Ensure a minimum (k) number of participants. Have randomized entry and exit points and impose unpredictable dwell times.

Obfuscation : Obfuscation techniques reduce the precision of location data by adding intentional errors (perturbation) or by generalizing the data (Tyagi and Sreenath 2015; Maharaj and Hosein 2016). While effective in masking exact locations, they may also diminish the quality of LBS.

Silent Period A silent period temporarily halts the transmission of location information, thereby breaking the link between old and new pseudonyms (Kasori and Sato 2015). Although this improves privacy, it can reduce service quality if timely location data are required.

Dummy Node The dummy node approach transmits both the true and one or more dummy locations during communication with LBS (Kasori and Sato 2015). This masks the actual location but may result in service degradation if too many dummy nodes are used.

Cloaking Region Cloaking regions blur the exact location by mixing a user's position with those of $k-1$ other users, effectively creating an anonymity region (Kasori and Sato 2015). While this increases anonymity, it typically comes with higher computational overhead and reduced accuracy.

Table 3.2.1 summarizes the key features of these privacy techniques, where "X" indicates the presence of a feature and "-" the absence.

Table 3.2.1: Evaluation Features of Privacy Techniques

Feature	Cloaking Region	Differential Privacy	Dummy Node	Group Signature	K-Anonymity	Mix-Zones	Obfuscation	Pseudonyms	Silent Period
Anonymity	X	X	-	X	X	-	-	-	-
Blurring	X	-	-	-	-	-	-	-	-
Cloaking	X	-	-	-	X	-	-	-	-
Data privacy-utility balance	-	X	-	-	X	-	X	-	-
Masking	-	-	X	X	X	-	X	X	-
Multiple nodes	-	-	X	X	-	-	-	-	-
Multiple datasets	-	X	X	-	-	-	-	-	-
Perturbation	-	X	-	-	-	-	X	-	-
Post-processing	-	X	-	-	-	-	-	-	-
Pseudonym exchange	X	-	X	-	-	X	-	X	X
Silent period	-	-	-	-	-	X	-	-	X
Time limit	-	-	-	-	-	X	-	-	X
Unlinkability	-	X	-	-	-	-	X	X	X
Vehicle Threshold	-	-	-	-	-	X	-	X	-
Verification	-	-	-	X	-	-	-	-	-

Having reviewed these techniques, the next section will focus on location privacy preservation mechanisms used to establish privacy by implementing various strategies. These mechanisms aim to protect users' sensitive location data from different attacks while maintaining the functionality and usability of location-based services.

3.3 Location Privacy Preservation Mechanism

The study by Zhong et al. (2022) introduced a sensitivity-based pseudonym change mechanism that leverages the regularity of a vehicle's movement patterns to provide personalized location privacy. This approach tailors privacy measures to individual preferences, potentially increasing user trust. However, it relies heavily on frequent visits to the same locations, and frequent pseudonym changes could interrupt service continuity, compromising safety-critical applications like real-time traffic updates.

Hou et al. (2021) developed two neural network-based vehicle tracking methods to assess the effectiveness of the Mix-Zone scheme in preventing vehicle tracking. These methods quantify the protection level achieved but face challenges due to the complexity of the neural network models. Specifically, the inclusion of numerous repetitive parameters can degrade the training process, while the fully connected Backpropagation Neural Network (BPNN) risks converging to a local optimum, reducing the reliability.

Nisha, Natgunanathan, and Xiang (2022) proposed a dummy location scattering scheme to safeguard user location privacy. This method generates dummy locations to obscure real position data from untrusted entities, supplemented by pseudonym-based mechanisms and time-delay techniques to enhance privacy further. However, introducing dummy locations may compromise data utility, illustrating a persistent challenge in privacy-preserving schemes: achieving an optimal balance between privacy and practical usability.

Hayat et al. (2023) designed a location privacy preservation strategy tailored to sparse traffic areas to counter colluding attacks. In dense areas, pseudonym changes occur within mix-context zones, whereas in sparse areas, the scheme shifts to differential privacy, adding noise to raw beacon message attributes via Local Differential Privacy to produce multiple perturbed messages. This confuses adversaries but focuses solely on collusion attacks, neglecting safety-critical scenarios such as collision avoidance for road users.

Ren et al. (2023) introduced a privacy protection model for Location-Based Services (LBSs) that allows the LBS server to access valid location distributions while ensuring robust user location privacy. The model generates perturbed locations meeting strict privacy criteria and includes a retrieval radius determination method to balance query accuracy with privacy. A key limitation is that preserving distribution without adequately addressing the

utility-privacy trade-off can diminish the data utility of LBS outputs, potentially affecting safety-critical applications.

Hara et al. (2016) proposed a dummy-based user anonymization scheme suitable for real-world settings. This method anonymizes user locations by generating dummy positions and addresses traceability issues, noting that the relative positioning of users and dummies could enable Location Service Providers (LSPs) to identify real users. The scheme's effectiveness hinges on careful dummy placement, but this may still expose users if not dynamically adjusted.

Zhang et al. (2018) developed a privacy-enhancing substructure for LBSs using a uniform grid framework, integrating order-preserving encryption and k-anonymity techniques. A semi-trusted third-party anonymizer handles caching and encrypted coordinate matching. While well-suited for continuous queries, the consistent use of a single encryption key and the need for users to share location-related data with others to form cloaking regions weaken the privacy assurances.

Zhang et al. (2020) presented a trajectory privacy-preserving mechanism for continuous LBSs based on a dual-K approach. Multiple anonymizers sit between the user and the LSP, distributing K query locations to achieve k-anonymity. Dynamic pseudonyms and location selection mechanisms further bolster trajectory privacy. However, the accuracy of predicted locations inversely affects privacy: selecting more predicted values reduces data utility, undermining service quality.

The common shortcomings exhibited by the reviewed privacy preservation mechanisms from the literatures shows that schemes such as Nisha, Natgunanathan, and Xiang (2022), Ren et al. (2023), and Zhang et al. (2020) struggle to balance privacy with data utility. Techniques like dummy locations, perturbations, or noise addition often degrade the quality of location-based services, impacting real-time or safety-critical applications. While some Zhong et al. (2022) and Hayat et al. (2023) rely on specific conditions such as frequent location visits or sparse traffic, limiting generalizability across diverse environments. Dummy-based (Hara et al. 2016) and pseudonym-based (Zhang et al. 2018) methods risk exposure if adversaries exploit patterns such as relative positions or static keys, undermining privacy. Several studies such as Hayat et al. (2023) and Ren et al. (2023) prioritize privacy over safety-critical use cases, such as collision avoidance or emergency response, which are vital in emergency situations.

3.4 Challenges Facing Existing Frameworks

Despite the advancements in effectiveness and user acceptance of personal safety solutions, there are several challenges identified. Moreover, usability and operational constraints pose significant challenge. Microcontroller-based solutions suffer from limited mobility, bulky hardware designs, and issues related to battery drain. Wearable solutions, such as smart bracelets and smart shoes offers portability, yet it relies on sensor data that can generate false positives. These false activations not only waste emergency resources but risk desensitising users and responders to genuine threats. Furthermore, systems that depend on specific activation gestures or require continuous movement to trigger alerts may fail in scenarios where the user is stationary or unable to perform the required actions.

In addition, the integration and scalability of these systems remain a concern. Many smartphone-based solutions rely on legacy communication protocols like HTTP/SOAP or SMS for data transmission, which are not inherently secure. Some applications attempt to enhance functionality through features such as augmented reality, social media integration, or contextual safety scoring, these enhancements often come with increased computational overhead and do not adequately address the critical need for secure data handling.

A primary challenge is the lack of robust privacy-preserving mechanisms. Many frameworks prioritise the rapid transmission of accurate location data to ensure timely emergency response. However, this comes at the expense of data protection, as sensitive information is transmitted using unsecured protocols or made publicly accessible.

Another challenge is the trade-off between data utility and privacy preservation. High data utility demands continuous, accurate tracking of a user's location or trajectory, yet frequent data collection, increases the risk of compromising user privacy. Many of the reviewed solutions do not offer dynamic privacy adaptation; they fail to modulate the level of data protection in real time based on contextual needs. Consequently, this gap makes it difficult to strike a balance between ensuring data utility and safeguarding sensitive user data.

3.5 Conclusion

The review of existing personal safety solutions reveals that despite significant progress in developing systems capable of using location and trajectory data to offer rapid emergency response, substantial challenges persist. The primary issues include the insufficient safeguarding of sensitive location information, the delicate balance between achieving high data utility and ensuring robust privacy preservation, and practical concerns related to device mobility, false positives, and system scalability.

The state of personal safety applications requires future research focus on integrating adaptive, privacy-preserving frameworks directly into these systems. Such integration would help maintain the data utility necessary for timely interventions while ensuring that sensitive user information remains protected. Addressing these challenges is crucial for developing next-generation personal safety solutions that can offer both high data utility and strong privacy guarantees, ultimately enhancing user trust and overall system effectiveness.

Chapter 4

Trajectory Data Prediction Validation Methodologies

4.1 Introduction

The analysis of existing approaches in the previous chapter has shown the challenges of personal safety solutions. The privacy preservation model proposed by this research to tackle one of the challenges is validated using prediction model for data processing. This is an essential step to establish the experimental performance for the methodological implementation. This chapter centers on the validation framework of the RBDPM, focusing on the application of three prediction models (ARIMA, Regression Tree Ensemble, and Linear Regression) to assess the model's effectiveness in achieving privacy preservation for personal safety solutions while maintaining data utility. These models, identified in Section 2 shows ability to handle trajectory data processing. This chapter examines the model's ability to effectively process trajectory data, offering preliminary insights into its strengths and limitations

4.2 ARIMA Model Performance Evaluation

The data exploration phase of the ARIMA prediction model is where stationarity test is performed to determine if the data are stationary. This is a common practise in time series analysis to explore the dataset, obtaining the mean, standard deviation, minimum, and maximum value of observations (Panneerselvam, Liu, and Antonopoulos 2018; Ariyo, Adewumi, and Ayo 2014) and the mean procedure test used to perform this operation. The outcome of this test for the dataset is shown in Table 4.2.1 with the standard deviation identified as the crucial factor in determining the stationarity of the dataset. The standard

deviation value for this dataset is less than 0.05 for both latitude and longitude and is insignificantly close to zero, which shows that the mean of the time series is constant and is interpreted as the dataset being stationary. The result of the test means that differencing (d) is not required for the dataset, as the dataset shows stationarity.

Table 4.2.1: The MEANS Procedure

Variable	Label	N	Mean	Std Dev	Minimum	Maximum
lat	lat	120	52.9252574	0.0043273	52.9196800	52.9295710
lon	lon	120	-1.4913512	0.0077818	-1.4990940	-1.4816560
elevation	elevation	70	109.2222736	15.4281721	43.2273404	131.0873011
accuracy	accuracy	120	19.8872000	3.6176629	6.0000000	32.7580000
bearing	bearing	8	181.629754	73.5573774	72.7008400	301.7622000

The model identification phase of the ARIMA prediction model involves the evaluation of the autocorrelation (ACF) and partial autocorrelation (PACF) plots of the transformed time series to determine the number of AutoRegressive (p) and Moving Average (q) terms required by the dataset. ACF and PACF graphs are common tools used in ARIMA modelling to evaluate the dataset and identify the model best suited to implement on the dataset (Ariyo, Adewumi, and Ayo 2014; Panneerselvam, Liu, and Antonopoulos 2018; Liu et al. 2016). The ACF plot is used to visualise the correlation between time series and lagged versions of the dataset and Fig. 4.2.1 shows the ACF for the dataset where a gradual decrease in the lag indicates that the correlation between time series and lagged versions of itself is statistically significant. This means that AR should be included for the dataset when identifying the model to apply to the dataset. The Partial Autocorrelation Function is a statistical tool used to measure the correlation between a time series and a lagged version while controlling for the correlation of all lower-order lags and visualising the partial autocorrelation between the time series and lagged versions. The PACF of this dataset shown in Figure 4.2.1 shows a sharp drop after a few lags and a gradual increase across the lag that indicates a statistically significant correlation, which implies the need for the MA model to be included in the model.

The model selection phase that involves selecting the best set of parameters for an ARIMA model is performed using information metric criteria such as AIC and BIC to evaluate the relative quality of different ARIMA models. The best model to use for prediction is the model with the lowest value of AIC and BIC and this approach aims to select the model with the lowest AIC or BIC value. The AIC metric tends to favour models with more parameters, while the BIC metric tends to favour models with fewer parameters (Ingdal, Johnsen, and Harrington 2019; Thiruchelvam et al. 2021). Table 4.2.2 shows the AIC and BIC criteria metrics values for different models tried on the dataset, and the metrics as seen in the table are negative and show that they are all suitable for the research. The best model to use for

prediction is the model with the lowest value of AIC and BIC, which according to the table is ARIMA (2,0,1) where 2 represents the number of lags, 0 is the degree of differencing and 1 is the order of moving average. ARIMA (2,0,1) provided the lowest AIC and BIC values, which are 1735.68 and 1721.74 for latitude and 1478.09 and 1464.15 for longitude, respectively.

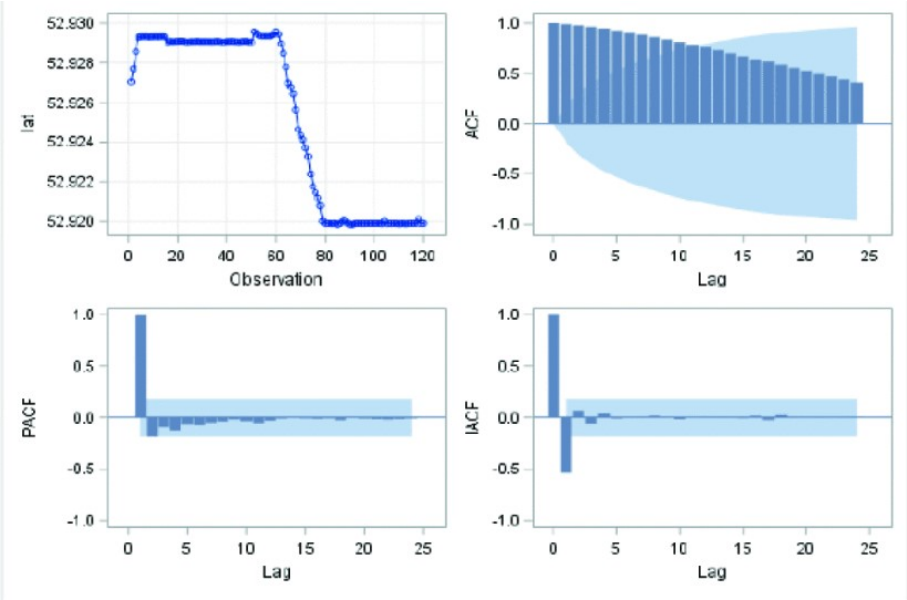

Figure 4.2.1: Correlation Analysis of training data

Table 4.2.2: AIC and SBC Result of ARIMA Model

ARIMA Model	Latitude		Longitude	
	AIC (-)	SBC (-)	AIC (-)	SBC (-)
ARIMA(1,0,0)	1627.95	1622.38	1419.60	1414.02
ARIMA(0,0,1)	1122.44	1116.87	981.49	975.92
ARIMA(1,0,1)	1696.50	1688.14	1445.14	1436.78
ARIMA(1,1,0)	1627.06	1621.49	1417.11	1411.53
ARIMA(0,1,1)	1122.44	1116.87	981.49	975.92
ARIMA(1,1,1)	1696.50	1688.14	1445.14	1436.78
ARIMA(1,1,3)	1712.58	1698.65	1455.00	1441.07
ARIMA(2,0,0)	1724.07	1715.70	1454.37	1446.01
ARIMA(2,1,3)	1459.45	1442.72	1459.45	1442.72
ARIMA(2,0,1)	**1735.68**	**1721.74**	**1478.09**	**1464.15**

The final step is the model validation process, which aims to assess the model's performance on unseen data. This is a critical step, as it determines whether the model generalises well to new data and how accurately it compares predicted values to actual values. Figures 4.2.2 and 4.2.4 illustrate the trends of the latitude and longitude training data, respectively, which are used to train the ARIMA model and identify patterns in the location data. Figures 4.2.3 and 4.2.5 show the difference between the test data and the predicted data for the latitude and longitude data, respectively. They show statistically insignificant differences in a single direction, suggesting that the ARIMA model is able to identify a pattern in the training data. The results of the predictions are relatively consistent in a single direction that does not reflect the capacity for use for a high-speed moving object.

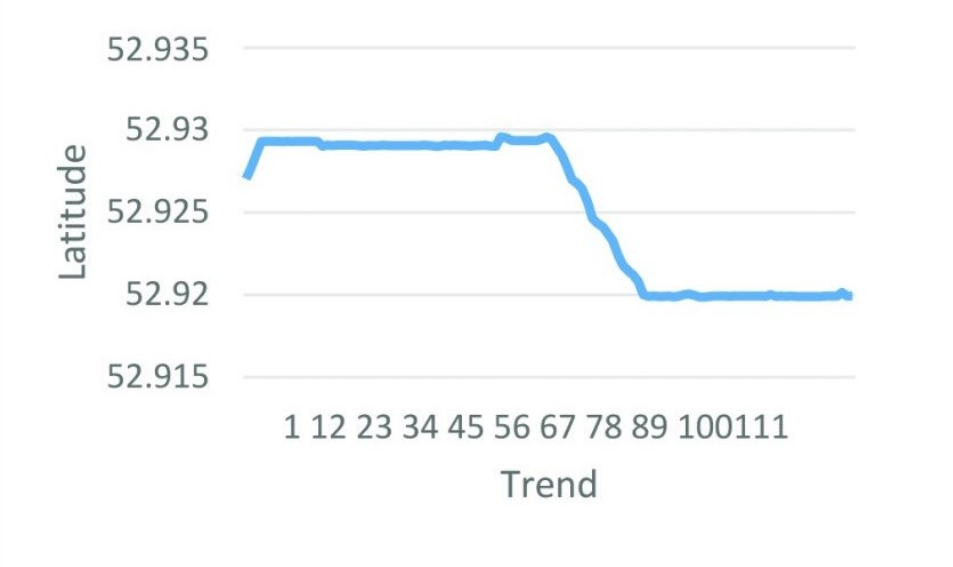

Figure 4.2.2: Latitude Training Data

4.3 Ensemble Regression Tree Model Performance Evaluation

The second model implemented is the Ensemble Regression Tree model which uses a weighted combination of multiple regression trees to construct a linear combination of models that enhances the predictive performance of the ensemble model. The bagged tree ensemble method involves training multiple decision tree models on different subsets of data and then combining the predictions of these models to make a final prediction (*MathWorks* n.d.[b]). Theoretically, this ensemble model should improve prediction accuracy compared

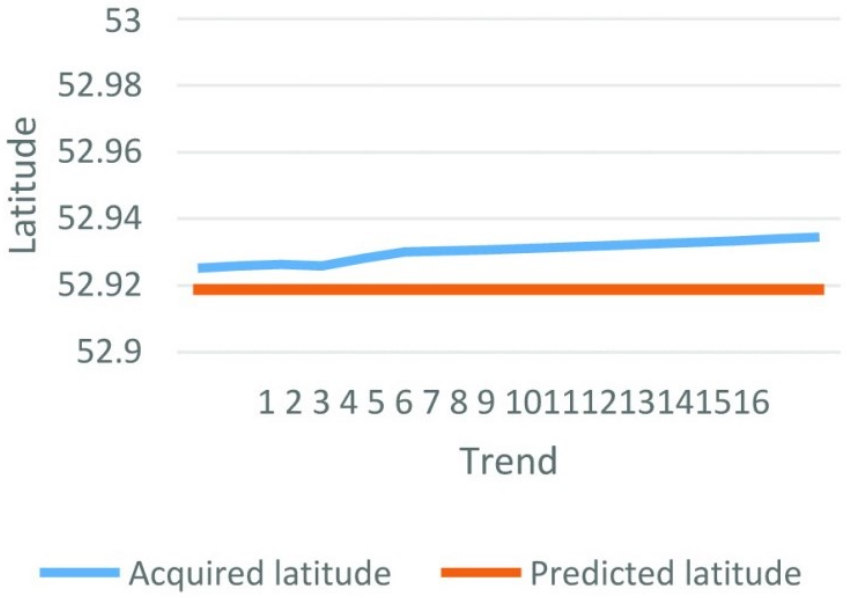

Figure 4.2.3: Trend Difference for Latitude Test Data and Predicted Data

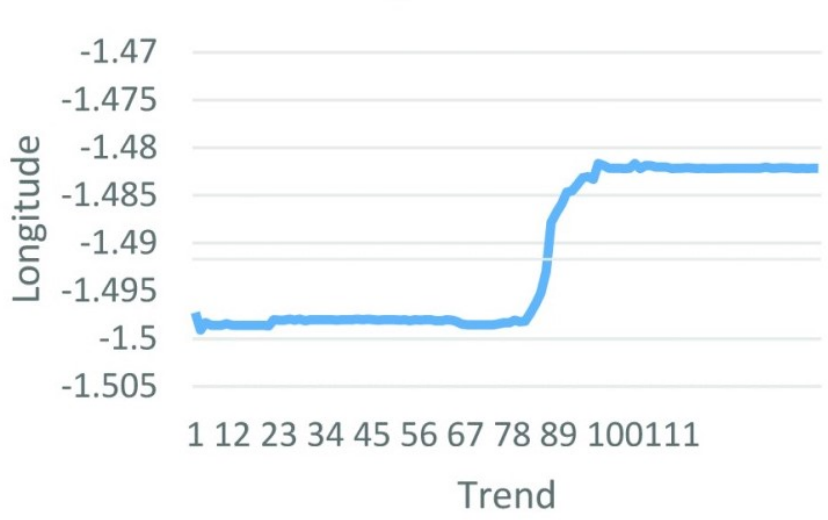

Figure 4.2.4: Longitude Training Data

to using a single decision tree model due to the reduction of overfitting and an increase in generalisation. The results of the model would depend on the quality and size of the training data, and the specific implementation of the ensemble method such as the number of trees used, the method for selecting subsets of the data. Based on privacy and prediction requirements needed for emergency response in personal safety solutions, a high level of

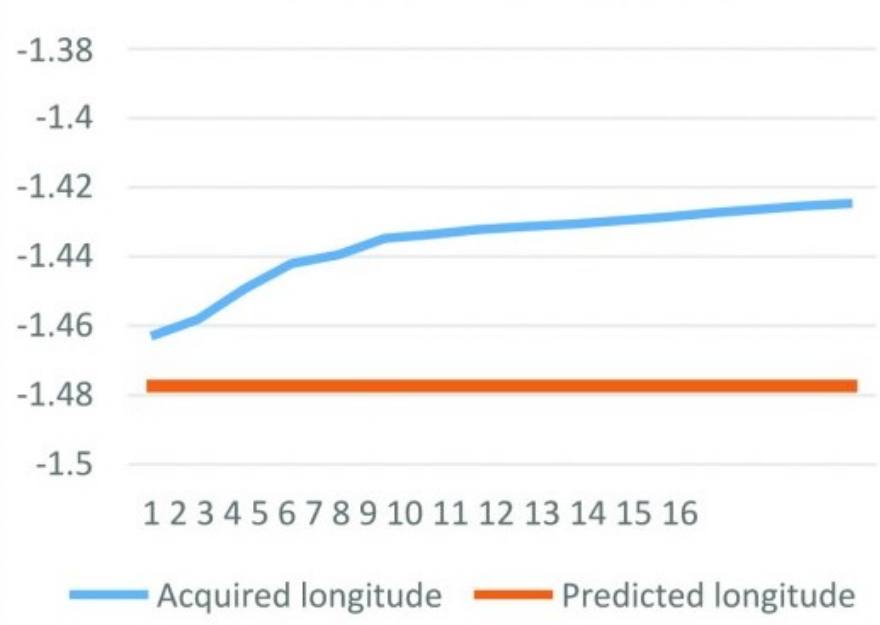

Figure 4.2.5: Trend Difference for Longitude Test Data and Predicted Data

accuracy is critical for a rapid and accurate identification of intercept points for the rapid response team to intervene and offer assistance to the distressed individual.

The trend of the training data shown in Figures 4.2.2 and 4.2.4 is compared with the predictions in Figures 4.3.1 and 4.3.2 for latitude and longitude, respectively. While the trends of the predicted values tend to follow a trend similar to the training data, there are noticeable differences in the trend from the test data. This suggests that, while the model may be efficient in capturing the general trend of the training data, it may have limitations in accurately predicting values for vehicle in motion. The trend exhibited by the predicted values for the latitude data has an upward trend similar to that for the test data with slight inconsistencies in the slope. However, the longitude data show a greater disparity with the test data and an upward slope that is different from the test data's consistent downward slope. These observations suggest that despite the bagged tree ensemble model that captures some of the underlying patterns in the data, it may not accurately predict the location of the event.

The boosted tree model uses a sequential weight adjustment process and is based on the fitting of the successive algorithm to the previous one. Sequential fitting can be observed in the consistent intervals shown in Figures 4.3.3 and 4.3.4, and provides a graphical representation of the latitude and longitude values for the predicted value and the test data. The figures infer that the boosted tree ensemble model is capable of capturing the general trend of the training data, and the predicted values follow a pattern similar to the actual test data. Despite the general alignment of the trends, there are still significant differences between the test data

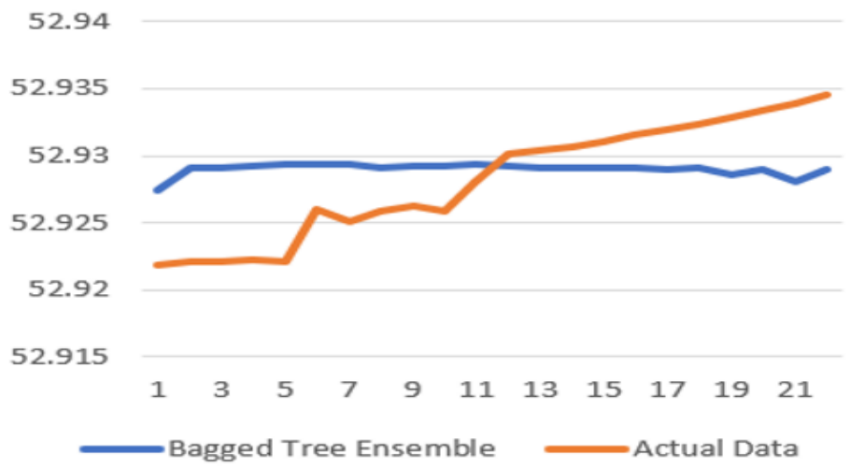

Figure 4.3.1: Bagged Model Trend Difference for Latitude Test Data and Predicted Data

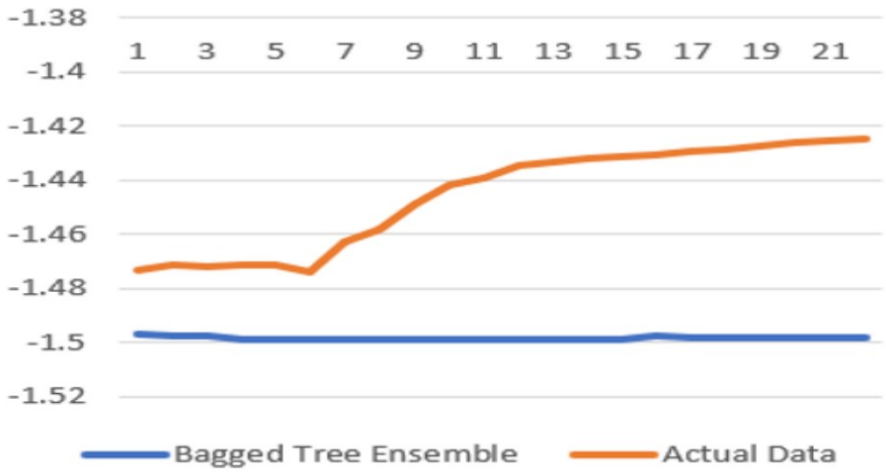

Figure 4.3.2: Bagged Model Trend Difference for Longitude Test Data and Predicted Data

and the predicted data. These are particularly noticeable in the longitude values, and, moving along the trend, the trend of the predicted data deviates from the trend of the test data.

This analysis suggests that the boosted tree ensemble model can capture some of the underlying patterns in the data, but it fails to accurately predict location data. This indicates a limitation in the predictive accuracy of the model for the purpose of processing data for personal safety solutions. These limitations could be due to various factors, such as the complexity of the data, the choice of hyperparameters, or the inherent limitations of the boosting algorithm. Despite the promise of the model's ability to identify general trends, extensive tuning and testing would be required to improve the predictive accuracy for specific values.

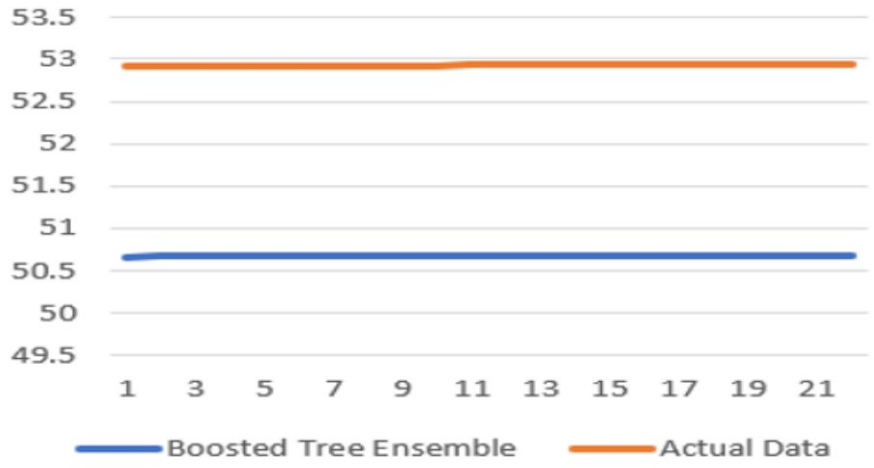

Figure 4.3.3: Boosted Model Trend Difference for Latitude Test Data and Predicted Data

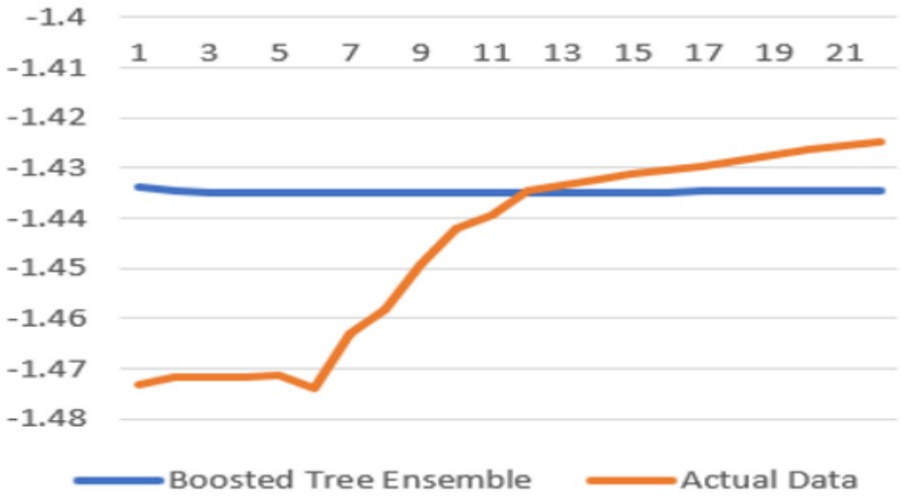

Figure 4.3.4: Boosted Model Trend Difference for Longitude Test Data and Predicted Data

4.4 Linear Regression Prediction Model Performance Evaluation

During the training phase of the Linear Regression prediction model, once the model is fitted, the subsequent step involves using the independent variables to predict the dependent variable. In this phase, the model generates predicted outputs that are directly compared with the actual observed values to assess the performance. The evaluation process is critical as it provides insights into how well the model captures the underlying relationship in the data and indicates whether adjustments or refinements are necessary.

The discrepancies between the predicted and actual values are quantified using several evaluation metrics, as discussed in Section 2.3.2, include the Root Mean Squared Error (RMSE), Mean Squared Error (MSE), and Mean Absolute Error (MAE). These metrics are fundamental in assessing the quality of the predictions, each providing a different perspective on the error characteristics:

The table below presents the computed values for these metrics during the training process, along with the recorded training time:

Table 4.4.1: Evaluation Metrics for the Model Training Process

Metric	**Value**
MAE	3.887×10^{-14}
MSE	1.193×10^{-26}
RMSE	1.092×10^{-13}
Training time	0.596 seconds

The values shown in Table 4.4.1 are close to zero, indicating that the model's predictions are closely identical to the actual observations. Such minimal error values suggest that the linear regression model has effectively captured the relationship between the independent and dependent variables, yielding a high degree of accuracy.

In addition to assessing the numerical error values, the evaluation process helps in identifying potential issues such as overfitting or underfitting. When error metrics are near zero on the training data, it is a good indicator that the model has learned the underlying patterns. However, it is essential to validate the model on unseen data to ensure that this performance is maintained outside the training set. The consistency of these metrics across different datasets would further strengthen the confidence in the model's generalizability.

Moreover, the inclusion of the training time metric serves as an indicator of the computational efficiency of the model. A training time of 0.596 seconds demonstrates that the model not only performs accurately but also does so in a computationally efficient manner, which is particularly important for real-time applications or scenarios where the model needs frequent updating.

In summary, the evaluation phase confirms that the linear regression model exhibits excellent performance based on the minimal discrepancies between predicted and actual values. The comprehensive analysis using RMSE, MSE, and MAE, coupled with a fast training time, validates that the model meets the thesis performance expectations. This robust performance is critical for applications such as trajectory prediction, where even minor inaccuracies can lead to significant deviations in practical scenarios during processing.

4.5 Discussion

The application of this concept to a personal safety solution using the capabilities of the existing framework (Sogi et al. 2018) that tracks the movement of people and collects trajectory data to improve the system by predicting possible intersection points to provide assistance to distressed individuals in transit. This should amplify the efficiency of the emergency system, facilitating a rapid response without exhausting resources or compromising the safety of others.

This study provides a insightful exploration of the implementation of the prediction model and potential utility in the emergency response system of personal safety solutions in a VANET. This chapter evaluates the implementation of the ARIMA, ensemble regression tree and Linear Regression prediction model on a real-world trajectory as a commendable step towards understanding time-series prediction models and efficiency in obtaining accurate predictions.

In our exploration of predictive models, one of the predictive models reviewed for implementation is the ARIMA model. This model is simple to use, has light computational costs, and can capture seasonal trends, but falls short when it comes to non-linear patterns and non-aggregated data (Petropoulos et al. 2022). This model uses the ARIMA(p,d,q) notation to indicate the best-suited model for prediction based on the degrees of autoregressive, differencing, and moving average. The dataset of this study required the model with the ARIMA(2,0,1) notation that indicates the need for two degrees of lag (AR), no differencing (I), and a forecast error lag of one degree (MA). The prediction outcome based on this notation shows constant gaps between each forecast value, and the trend followed by the predicted values is similar to the trend observed by the model from the training data. This indicates the dependence of this algorithm on trend and pattern observation from the training data.

The implementation of the ARIMA model on the trajectory dataset began with a stationarity test using the mean procedure test to verify the stationarity state of the data and whether differencing would be required when implementing the model on the dataset. The outcome of this test showed an insignificant standard deviation value that is less than 0.05, which means that no differencing (d) was required for this dataset. The evaluation of the Autocorrelation (ACF) and Partial Autocorrelation (PACF) plots in the model identification phase proved valuable for determining the appropriate number of AutoRegressive (p) and Moving Average (q) terms; it shows that AR and MA are required by the dataset. The model selection phase, which is guided primarily by the AIC and BIC values, revealed that the ARIMA (2,0,1) model provided the lowest AIC and BIC values. The model validation phase

illustrates the model's performance on unseen data which shows a satisfactory performance of the model on unseen trajectory data.

Another predictive model that was implemented is the Regression Tree ensemble model, which uses a weighted combination of multiple regression trees. This model has shown proficiency in trajectory data prediction and uses BAGGing and Boosting methods to decrease the variance of the model prediction, leading to more accurate results. The Bagged Tree approach employs the most efficient predictor of the aggregated decision for prediction, echoing the pattern observed with the ARIMA model. The outcome of this approach follows the trend of the training model as observed in the ARIMA model that differs from the test data. The boosted tree technique, on the other hand, uses a sequential process of weight adjustment and is built on the fitting of the successive algorithm to the previous algorithm. This approach produced an outcome pattern that differed from the other two models, with significant deviation from the test data, and the trend differs from that of the training data. This deviation was caused by the dependence of the approach on the outcome of the previous algorithm during the decision-making process.

The implementation of the Ensemble Regression Tree model aims to improve predictive performance through a weighted combination of multiple regression trees, making it a good model to predict trajectory data. The bagging technique employed trains several decision tree models on varied data subsets, subsequently aggregating these models' predictions to produce a final one. The bagged tree method demonstrates the model's capability to capture inherent data patterns, with predicted values largely conforming to the training data trend. Notable discrepancies with the test data indicate potential limitations in predicting specific values accurately. The boosted tree technique, which iteratively adjusts the weights and builds on the fitting of successive algorithms, does not seem to address these shortcomings. Although the model seems to capture the general trend of the training data and the predicted values align to some extent with the actual test data, significant disparities persist, especially in the longitude values.

Using ensemble learning models, such as the regression tree ensemble, can lead to enhanced accuracy in location prediction tasks compared to conventional models such as ARIMA. Ensemble models amalgamate multiple predictor models to surpass the performance of individual models. The ensemble learning model is adept at managing intricate and non-linear relationships in data, where even neural networks, despite their strength, might falter (Wang et al. 2019). Ensemble models are more interpretable than neural networks since each individual model's decision-making process can be comprehended and scrutinised. They consist of several models, thereby reducing the risk of overfitting. This leverages the strengths of various models while counteracting the weaknesses of others. This attribute

makes ensemble models especially beneficial for datasets with a limited and specific sample size.

Both ARIMA and Regression Tree ensemble models display potential. Each model boasts certain benefits: ARIMA excels in discerning data patterns and trends, while Regression Tree ensemble models offer superior accuracy in location prediction tasks, effectively handle complex and non-linear data relationships, and deliver easily interpreted results. Despite their strengths, both models exhibit limitations and differ in test data performance, underscoring the intricacy of location prediction tasks and the imperative of judicious model selection and fine-tuning.

Although ARIMA and Regression Tree Ensemble models showed promising capabilities for emergency response location prediction, they come with constraints. Proper interpretation of results and potential adjustments based on the nature of the data are essential. The ARIMA model, reliant on recent past data, has limited efficacy for longer-term forecasts, and susceptibility to noise might distort the model's parameters. Implementing the Regression Tree Ensemble models can be intricate, necessitating meticulous tuning of parameters like tree count, tree depth; furthermore, overfitting becomes problematic without optimal parameter configuration. These drawbacks led to the choice of the Linear Regression model for validating the privacy model.

4.6 Conclusion

This chapter has provided analysis of different prediction models that are used for processing trajectory data, assessing the conceptual viability in efficiently processing trajectory data. Three distinct prediction models: ARIMA, Regression Tree Ensemble, and Linear Regression were explored and applied to real-world trajectory data.

While ARIMA and Regression Tree Ensemble highlighted challenges such as overfitting to temporal patterns or computational complexity, Linear Regression emerged as a suitable model for trajectory data offering a robust foundation for integrating risk-driven differential privacy without the overfitting or complexity burdens of ARIMA and Regression Tree Ensemble.

Chapter 5

Risk-Based Differential Privacy Model Concept

5.1 Introduction

The validation of the predictive models in Chapter 4 has established a foundation for evaluating the performance of the RBDPM with Linear Regression emerging as a the most suited for processing trajectory data. this chapter shifts focus to the conceptualization and implementation of the RBDPM, a framework proposed in this thesis to enhance privacy preservation in personal safety solutions.

This chapter presents a detailed exploration of the RBDPM, elaborating on the design and the modular structure that supports the implementation. It introduces three integral modules (Hazard Assessment Module, Privacy Preservation Module, and Noise Application Module). Together, these modules analyze distinct data characteristics, provide contextual insights into risk and privacy needs, and culminate in the RBDPM's operational framework.

5.2 Privacy Preservation in Personal Safety Solution

The RBDPM proposed in this thesis categorizes location data records by their risk levels, enabling tailored noise injection to preserve privacy while maintaining data utility of location data. The location data used by this model would be collected during the collection phase from travelling vehicles. The initial phase evaluates the relative velocity and distance between adjacent vehicles to determine the time to collision, which is a critical indicator of immediate risk of collision. The time-to-collision is used for the classification of risk on a scale from 1 lowest to 3 highest levels. The lower the TTC value, the higher the risk of collision and

vice versa. The privacy mechanism uses these risk score to adaptively compute a privacy parameter. This parameter is derived using the Differential Privacy and risk relationship equation by Dandekar, Basu, and Bressan (2021) that shape how privacy levels respond to risk. Having established the privacy parameter, the model proceeds to inject Laplace

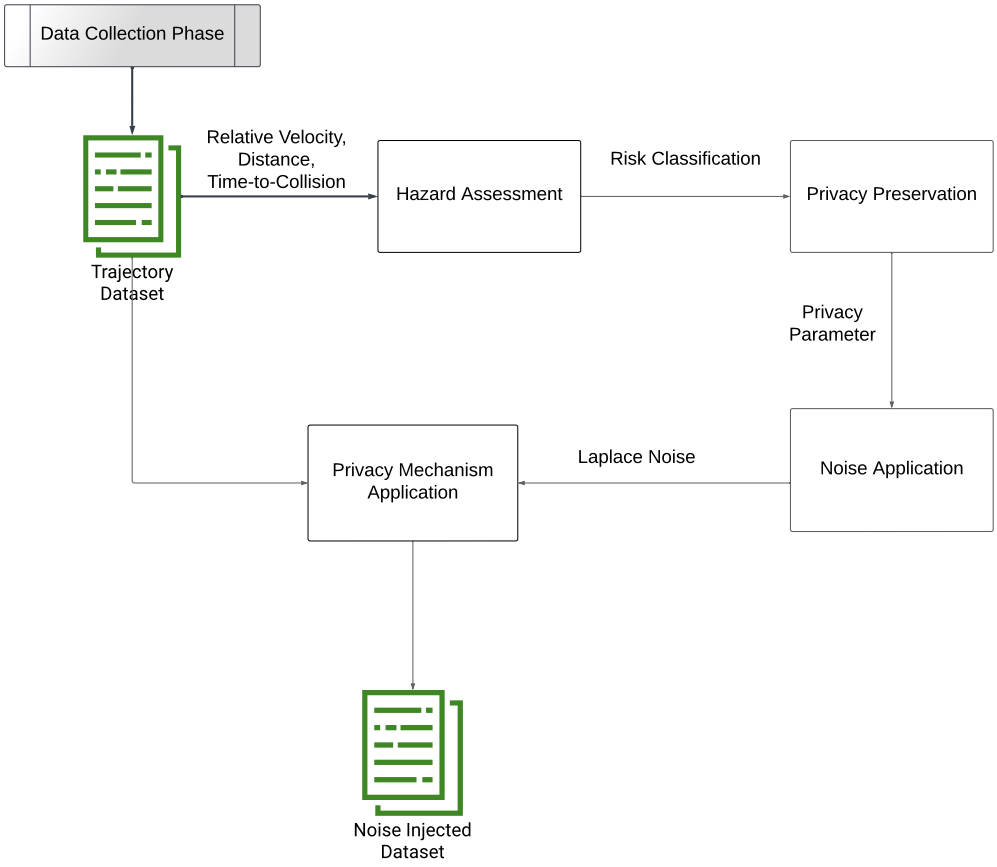

Figure 5.2.1: Design Concept for Risk-Based Differential Privacy Model

noise into the location dataset in preparation for processing. This ensures that the published data does not reveal exact user trajectories. The scale of the noise depends on the privacy parameter ε with large privacy parameter receiving milder noise to ensure good data utility. Unlike low privacy parameter during incident times that receives higher noise injection reflecting the user's reduced need for precise collision warnings in these situations.

5.3 Risk-Based Differential Privacy Model Concept Design

The RBDPM consists of three core modules: Hazard Assessment, Privacy Preservation and Noise Application Module. This model is based on extending privacy preservation into the functionality of personal safety solution such as the solution proposed by Sogi et al. (2018) discussed in section 1.2. The data collection phase provides the trajectory location information of a moving vehicle, and this serves as the input data for the model.

Hazard Assessment Module

The Hazard Assessment module illustrated in Algorithm 1 processes the dataset to obtain kinematic features such as coordinate, velocity, distance and nearby vehicle information. The input dataset consists of location data points, each identified by an id, timestamp, latitude, and longitude. The dataset is sorted chronologically to feed the *CalculateDistance* function to determine the distance travelled between consecutive timestamps. This distance is divided by the time taken to travel the distance between the coordinates (output of the *TimeDifference* function) to derive the velocity.

The next step is to identify the nearby vehicles for a given vehicle at a specific timestamp by checking for vehicles location at the specific timestamp. The distance between the current vehicle and each nearby vehicle, as well as their relative velocity is calculated using the *ComputeRelativeVelocity* function. If the relative velocity is not zero, the time to collision is computed as the distance divided by the relative velocity (Yan et al. 2010). The TTC value for each vehicle at each timestamp is determined, indicating the shortest time to potential collision with another vehicle. This process allows for dynamic risk assessment based on real-time movement data.

Privacy Preservation Module

Building on the output from the Hazard Assessment module, the privacy preservation module as shown in Algorithm 2 applies a risk-based assessment to the data. Specifically, it calculates Time-to-Collision (TTC) values with the *ComputeTTC* function and converts these continuous metrics into categorical risk labels (e.g., Low, Medium, or High) using the *CategorizeRiskScore* function to assign a risk score based on TTC thresholds.

Algorithm 1 Hazard Assessment Module

Require: A dataset X, where each entry contains:

id, timestamp, latitude, longitude,

$X \leftarrow \text{Sort}(X, \text{ by} = [\text{id, timestamp}])$
for $i \leftarrow 1$ to $\text{len}(X)$ **do**
    if $X_i['id'] = X_j['id']$ **then**
        $d \leftarrow \text{CalculateDistance}(X_i['lat'], X_i['lon'], X_j['lat'], X_j['lon'])$
        $\Delta t \leftarrow \text{TimeDifference}(X_i['timestamp'], X_j['timestamp'])$
        if $\Delta t \neq 0$ **then**
            $v \leftarrow d/\Delta t$
        else
            $v \leftarrow 0.0$
        end if
    end if
end for
$min\_ttc \leftarrow \infty$
$nearby\_vehicles \leftarrow \text{Filter}(X, \text{ by} = [\text{id, timestamp}])$
for each vehicle v in $nearby\_vehicles$ **do**
    $d \leftarrow \text{CalculateDistance}(v_1, v_2)$
    $v\_rel \leftarrow \text{ComputeRelativeVelocity}(v_1['\text{velocity}'], v_2['\text{velocity}'])$
    if $v\_rel \neq 0$ **then**
        $ttc \leftarrow d/v\_rel$
        if $ttc < min\_ttc$ **then**
            $min\_ttc \leftarrow ttc$
        end if
    end if
end for
$X['\text{TTC}'] \leftarrow min\_ttc$ **if** $min\_ttc \neq \infty$ **else** NULL
return X

Algorithm 2 Privacy Preservation Module

function PRIVACYPARAMETER(TTC)
    df['Risk_Score'] $\leftarrow$ df['ttc'].apply(categorize_risk_score)
    df['Privacy'] $\leftarrow$ df['Risk_Score_Category'].apply(calculate_privacy(X))
    if $ttc\_value < 7$ **then**
        $risk\_score \leftarrow 3$
    else if $4 \leq ttc\_value \leq 7$ **then**
        $risk\_score \leftarrow 2$
    else
        $risk\_score \leftarrow 1$
    end if
    $privacy\_value \leftarrow \frac{1}{1-risk\_score\_category \times (1-3 \times e^{-\varepsilon\_factor})}$
    return $privacy\_value$
end function

Close interaction with short TTC usually under 4 seconds is deemed high risk situations and the risk score set to 3 (high risk). While those with larger separation receive lower scores, if TTC value falls between 4 and 7 seconds (inclusive), the category is set to 2 (moderate risk), and if it is greater than 7 seconds, the category is 1 (low risk). Once the risk score has been computed, the next step is the computation of the privacy parameter level using the *ComputePrivacy* function. This function uses the Differential Privacy and risk relationship equation by Dandekar, Basu, and Bressan (2021) that relates risk and privacy to compute the privacy value from the risk score.

Noise Application Module

The privacy parameter ε value determine previously would be used by this Noise Application module to dictates the amount of noise to be injected into dataset. This module shown in Algorithm 3 uses sensitivity parameter, which determines the maximum impact a single data point can have on the dataset. Using the Laplace noise mechanism, along with the sensitivity and privacy level, noise is added to the dataset. This mechanism perturbs each location coordinate by adding noise sampled from a Laplace distribution with scale $(\varepsilon, \text{sensitivity})$, where ε (epsilon) is a privacy parameter controlling the level of noise applied. After applying the noise, a modified dataset is returned as the output that would be used for data processing. This process ensures that location data remains differentially private by adding controlled randomness, preventing adversaries from accurately pinpointing individual locations while still preserving statistical properties useful for analysis.

Algorithm 3 Noisy Application Module

1: **function** NOISYDATASET(X)
2: **Set** sensitivity
3: X_noisy[X] $\leftarrow$ AddLaplaceNoise($X, \varepsilon, sensitivity$)
4: **return** X_noisy
5: **end function**

5.4 Risk-Based Differential Privacy Model Implementation

The initial trajectory dataset acquired from the data collection phase is replaced with a generalized and randomized noise-injected trajectory dataset that is larger and exhibits a decrease in linearity while maintaining sufficient data utility to enable high-accuracy predictions. The purpose of the noise-injected trajectory dataset is to limit the amount of information that an attacker can obtain by intercepting the information transmitted by the user in motion.

The concept of dynamically balancing the data utility and privacy trade-off is to preserve the accuracy of the processed information while ensuring data sensitivity protection. This dynamism is achieved by identifying the mobility pattern through the mobility classification module, determining the ε level via the respective modules, and adjusting the noise level applied to the trajectory dataset to balance the trade-off between data utility and privacy, regardless of the user's mobility pattern. The full execution flow of RBDPM is shown in Algorithm 4 encompassing the various modules.

The data processing methodology adopted for the experimental validation of the model is based on a predictive modeling approach. The prediction model, as described in Section 2.2.3, is pre-trained using the training sub-dataset of the Beijing taxi dataset, and the performance is validated on a new, unseen dataset using the test sub-dataset.

The noise-injected dataset is utilized for prediction using the resultant prediction model. The model's performance is evaluated by measuring the error/difference between the predicted values and the actual values of the input data. The errors are assessed using the evaluation metrics: Root Mean Squared Error (RMSE), Mean Squared Error (MSE), and Mean Absolute Error (MAE).

The evaluation criteria indicate the effectiveness of the model in capturing the underlying data patterns and the generalization capability. When the evaluation metric values approach zero, it suggests that the model performs well, as the predicted values closely align with the actual values in the i^{th} term. Conversely, the farther the value is from zero, the higher

the ineffectiveness of the model for the intended purpose. In such cases, the model requires improvements, such as parameter hyper-tuning and retraining with optimized parameters.

Algorithm 4 Main Execution Pipeline for Generating a Noisy Dataset

```
function MAINEXECUTION(dataset)
    df ← dataset
    df ← CALCULATEVELOCITY(df)
    df (TTC) ← COMPUTETTC(df)
    df (RiskScore) ← CATEGORIZERISKSCORE("ttc")
    df (Privacy) ← COMPUTEPRIVACY("RiskScore")
    noisyDF ← NOISYDATASET(df)
    return noisyDF
end function
```

Dataset Description

The Beijing T-Drive dataset is a dataset encompassing information useful for spatio-temporal analysis and urban mobility. This Microsoft Research Asia dataset captures extensive GPS trajectories from taxi fleets operating in Beijing and has been widely adopted by researchers for studying traffic patterns, routing algorithms, and map-matching techniques, among other applications.

The dataset includes the trajectories of approximately 10,357 taxis over a one-week period (February 2-8, 2008). This version consists of millions of GPS points detailing the paths taken by taxis across Beijing's road network, thereby providing granular location data that reflects real-world traffic dynamics.

This dataset as shown in Table 5.4.1 contains Taxi ID (anonymized), Timestamp, Latitude, and Longitude, which allows for comprehensive analysis of urban traffic flow and driver behavior, identifying patterns and decision-making strategies. Although taxi IDs are anonymized, the fine-grained nature of the data requires careful handling to preserve privacy while enabling detailed analysis. The dataset offers spatial coverage reflecting real-world traffic dynamics such as the complexities of urban mobility, driver behavior, and actual travel time that is suitable for training a predictive model for the validation of the trade-off balance hypothesis for data utility and privacy in this thesis.

The dataset does not entail features related to dangerous activities such as accidents or near-miss incidents to inform the investigations of this research. This absence of safety critical data means that the dataset alone cannot support analyses aimed at collision-related

Table 5.4.1: Dataset sample with timestamp, latitude, longitude, and ID.

timestamp	lat	lon	id
2008-02-02 13:38:03	39.9071	116.415	10
2008-02-02 13:38:09	39.907	116.49	366
2008-02-02 13:38:13	39.907097	116.416431	10
2008-02-02 13:38:13	39.907	116.488845	366
2008-02-02 13:38:23	39.907093	116.41782	10
2008-02-02 13:38:23	39.907	116.48769	366
2008-02-02 13:38:29	39.907	116.419293	10
2008-02-02 13:38:29	39.907086	116.482534	366
2008-02-02 13:38:43	39.907	116.4757	10

occurrences. For modelling and investigation of high-risk scenarios, it is essential to generate synthetic data that mimics near-miss and accident events. Such generated data can capture the dynamic and unpredictable nature of dangerous driving situations while preventing danger to user safety. These simulated scenarios serve as a critical supplement to the existing T-Drive dataset by providing the hazardous event information necessary for the analysis of dangerous events.

The SimPy simulation environment offers a comprehensive set of tools for creating event-driven realistic simulations, especially for these dangerous event situations. The versatility and robustness of SimPy make it a valuable tool for simulating mobility models and validating analytical results in various applications. Driver behavior, stopping distance, speed, acceleration, deceleration patterns, and vehicle dynamics all contribute toward modelling dangerous events. In this thesis, the simulation will track features like time-to-collision, thereby enabling the generation of near-miss and accident events.

The generated collision dataset is focused on two users on a two-lane road with both vehicles traveling at moderate speeds and maintaining a safe distance from other vehicles. User A and User B are approaching from opposite directions. User B sees the vehicle ahead brake suddenly and makes a rapid decision to swerve away from their lane to avoid a potential rear-end collision. However, this evasive maneuver puts User B directly in the path of User A at a distance greater than the stopping distance of both vehicles when braking. This leads to a head-on collision between the two users that jeopardizes their safety, thereby calling for a drop in privacy level for assistance to be rendered with high data utility. Similarly to the T-Drive dataset, the generated dataset contains ID, Timestamp, Latitude, and Longitude and the features embodied in the dataset include time-to-collision (TTC) and stopping distance for analyzing the situation.

5.5 Conclusion

This chapter outlines the core modules essential to the operational efficacy of the RBDPM: the Hazard Assessment Module, Privacy Preservation Module, and Noise Application Module. These modules integrate distinct functionalities to produce a noisy dataset for processing.

The Hazard Assessment Module analyses trajectory data attributes, such as TTC, to identify and categorize risk levels relevant to personal safety scenarios. The Privacy Preservation Module then determines the privacy parameter, ε, based on the assessed risk. Next, the Noise Application Module applies the DP mechanism, injecting Laplace noise calibrated to the specified ε value, yielding a privacy-protected trajectory dataset tailored to the risk profile.

This chapter details the cohesive integration of these modules, establishing the RBDPM's operational framework. This foundation sets the stage for Chapter 6, which evaluates the dataset's performance through experimental validation.

Chapter 6

Risk-Based Differential Privacy Model Proof-of-Concept Validation

Chapter 5 detailed the design and implementation of RBDPM through the Hazard Assessment, Privacy Preservation, and Noise Application Modules. This chapter shifts focus to the empirical evaluation of the RBDPM, presenting the results of the experimental validation providing comprehensive analysis of the RBDPM's effectiveness in preserving privacy for personal safety solutions while maintaining the data utility essential for operational success. The model's performance is assessed using a normal traffic and collision traffic trajectory dataset to leverage Linear Regression for data processing. The evaluation results explore key metrics to quantify how RBDPM modules perform. The findings was discussed and compared with other research showing the strength and weaknesses.

6.1 Privacy Parameter Determination

The hazard assessment evaluates the safety-critical information within the location dataset using relative velocity, stopping/braking distance and distance between vehicles to estimate the TTC between the two users.

The calculation of TTC provides significant insights into the safety dynamics of vehicle interactions. The TTC between nearby vehicles is calculated to determine the closeness between these vehicles and the potential for collision. The minimum TTC value, which indicates the closest point to a potential collision within the dataset is used to determine the general measure of safety across the dataset. The underlying concept is that a shorter TTC indicates a higher risk of collision. This value helps in understanding the immediate safety landscape of vehicle interactions, which is crucial for hazard assessment where lower TTC

values signal higher-risk scenario (Wandtner et al. 2018; Hosseini et al. 2016). TTC could not be calculated using the normal traffic dataset as the vehicles maintain safe distances as illustrated in Figure 6.1.2. However, in the collision dataset, TTC values of 3.415243, 2.276829, and 1.138415 were observed as illustrated in Figure 6.1.1 with TTC values printed at key points and 1.138415 the minimum TTC before collision. The figure shows the decrease in TTC until collision with the value changing from 3.415243 to 2.276829 to 1.138415 till the collision occurs.

Figure 6.1.1: Map showing vehicle paths and TTC values for the collision dataset

Based on the computed TTC, a risk score is assigned according to predefined thresholds. A TTC of less than 4 seconds indicates an instance of high risk, and is assigned a risk score of 3; an instance of medium risk, with a risk score of 2, is assigned for TTC between 4 and 7 seconds; while low risk, with an assigned risk score of 1, applies if TTC exceeds 7 seconds. This categorisation directly impacts the level of privacy protection needed, with higher-risk scenarios requiring more noise to safeguard sensitive data, thus influencing how privacy and data utility are balanced within the system.

Higher risk calls for stronger privacy protection. Therefore, if the risk score is high, indicating a greater chance of collision and a more sensitive situation, the privacy parameter mandates a low degree of noise to preserve data utility. Conversely, if the risk score is low, more noise is needed, thus preserving data privacy. The privacy parameter, ε, is derived from the risk score using the Differential Privacy and Risk Relationship function by Dandekar, Basu, and Bressan (2021) described in Section 2. This value directly modulates the level of Laplace noise added to the trajectory data, ensuring that data with higher collision risk

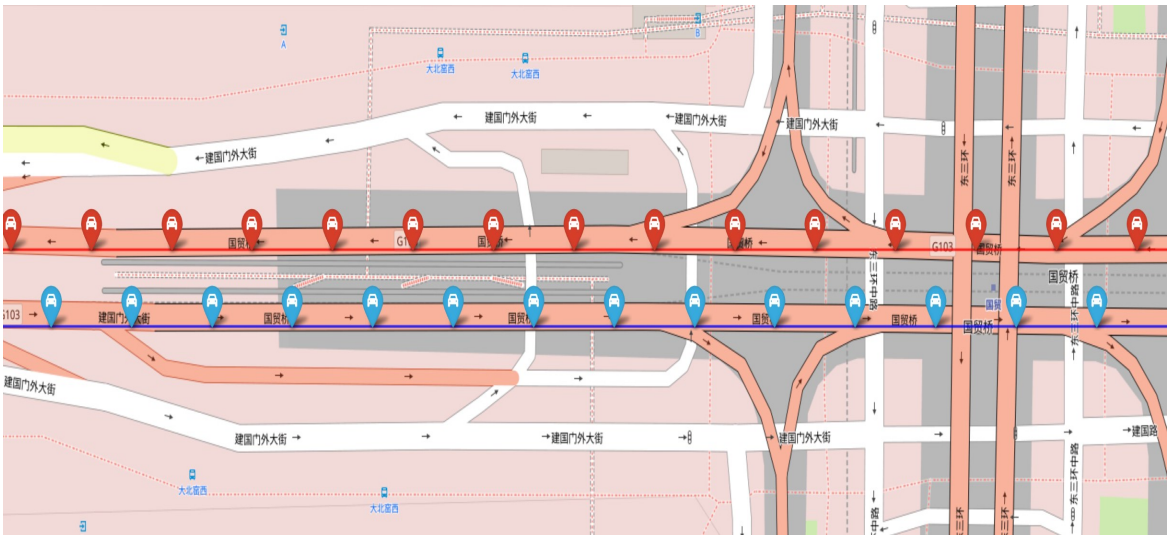

Figure 6.1.2: Map showing vehicle paths for normal traffic

receive a noise level that adequately protects sensitive location information and is tailored to the risk profile of the data. The privacy parameter levels, as shown in Table 6.1.1, equate to 0.5 for a risk score of 3, 0.6 for a risk score of 2, and 0.7 for a risk score of 1.

Table 6.1.1: Risk and ε Level Based on Time-to-Collision (TTC)

TTC	**Risk Level**	**ε Level**
TTC $<$ 4 secs	High (3)	0.5
4 $\leq$ TTC $\leq$ 7	Medium (2)	0.6
TTC $>$ 7	Low (1)	0.7

This process ensures that the level of privacy protection is dynamically tailored to the risk associated with the data. By linking TTC and risk score to the privacy parameter, the framework effectively balances the trade-off between maintaining data utility and ensuring robust privacy protection in critical scenarios. This approach allows the system to adapt to varying safety-critical situations, ensuring that sensitive information is protected without unduly compromising the utility of the data used for processing.

6.2 Differential Privacy Preservation Performance Evaluation

The initial implementation of the prediction model focused on integrating a DP scheme to safeguard data while maintaining predictive accuracy. A pre-trained Linear Regression

model as discussed in Section 2.2.3 was adopted, and the performance on new data was evaluated using various levels of Laplace noise, controlled by the privacy parameter ε. The performance was assessed using MSE, MAE, and RMSE metrics explained in Section 2.3.2.

For the baseline dataset, the test sub-dataset produced MAE, MSE, and RMSE values of 3.89×10^{-14}, 1.10×10^{-13}, and 1.20×10^{-26} respectively. This outcome shows values close to zero for all evaluation metrics and represents an efficient prediction model that works well with a new unseen dataset.

Table 6.2.1: Privacy Preservation Prediction Scheme Performance

Epsilon Level	Mean Absolute Error	Root Mean Squared Error	Mean Squared Error	Time taken (seconds)
0	3.89×10^{-14}	1.10×10^{-13}	1.20×10^{-26}	0.59
0.1	2.69×10^{-13}	4.87×10^{-13}	2.37×10^{-25}	1.43
0.2	3.12×10^{-13}	5.41×10^{-13}	2.93×10^{-25}	1.42
0.3	3.29×10^{-13}	5.64×10^{-13}	3.18×10^{-25}	0.96
0.4	3.38×10^{-13}	5.76×10^{-13}	3.32×10^{-25}	0.94
0.5	3.43×10^{-13}	5.84×10^{-13}	3.41×10^{-25}	1.10
0.6	3.47×10^{-13}	5.89×10^{-13}	3.47×10^{-25}	1.29
0.7	3.50×10^{-13}	5.93×10^{-13}	3.52×10^{-25}	1.02
0.8	3.52×10^{-13}	5.96×10^{-13}	3.55×10^{-25}	0.92
0.9	3.54×10^{-13}	5.99×10^{-13}	3.58×10^{-25}	0.92
1.0	3.55×10^{-13}	6.00×10^{-13}	3.61×10^{-25}	0.92

Subsequently, Laplace noise was systematically injected into the dataset using ε values ranging from 0.1 to 1.0. Table 6.2 presents the performance metrics across these privacy levels. The outcome indicated that as the value of ε increased, a minor yet consistent rise in the error metrics was observed. This trend suggests that while the introduction of noise to enhance privacy slightly reduces data utility, the model performance remains robust. The slight increments in MAE, MSE, and RMSE demonstrate that the model's predictions remain very close to the true values even when privacy-preserving noise is applied.

Table 6.2.2 shows the differences between the evaluation metrics of the noise-injected dataset and the baseline dataset. The baseline dataset produced a MAE of 3.89×10^{-14}, which confirms that, the absolute difference between the predicted and actual values is nearly zero. Similarly, the MSE value of 1.10×10^{-13} indicates that large prediction errors are extremely rare due to the squaring of errors, resulting in a minimal overall error. Finally, the RMSE of 1.20×10^{-26} shows that when the error measure is converted back to the original data units, it remains negligible.

Table 6.2.2: Differential Privacy-Applied Prediction vs Baseline Prediction Comparative Evaluation

Epsilon Level	Δ Mean Absolute Error	Δ Root Mean Squared Error	Δ Mean Squared Error	Δ Time taken (seconds)
0.1	2.30×10^{-13}	3.77×10^{-13}	2.25×10^{-25}	0.84
0.2	2.73×10^{-13}	4.32×10^{-13}	2.81×10^{-25}	0.83
0.3	2.90×10^{-13}	4.54×10^{-13}	3.06×10^{-25}	0.37
0.4	2.99×10^{-13}	4.67×10^{-13}	3.20×10^{-25}	0.35
0.5	3.04×10^{-13}	4.74×10^{-13}	3.29×10^{-25}	0.51
0.6	3.08×10^{-13}	4.79×10^{-13}	3.35×10^{-25}	0.70
0.7	3.11×10^{-13}	4.83×10^{-13}	3.40×10^{-25}	0.43
0.8	3.13×10^{-13}	4.86×10^{-13}	3.43×10^{-25}	0.33
0.9	3.15×10^{-13}	4.89×10^{-13}	3.46×10^{-25}	0.33
1.0	3.17×10^{-13}	4.91×10^{-13}	3.49×10^{-25}	0.33

The best-performing models are dependent on the lowest values of the evaluation metrics. Based on the MAE metric obtained for these data, the baseline dataset shows the best performance with a low MAE value. This is expected as the baseline dataset maintains high data utility without any added noise to the dataset (Dwork and Roth 2014). The same observation can be made of other evaluation metrics such as MSE and RMSE. Evaluation metrics assessment for the models shows that the increase in Laplace noise level introduced into the dataset leads to a slight decrease in data utility, and the noise-injected dataset processing is capable of providing an accurate result similar to the result from processing the initial trajectory dataset.

The differences in MAE, MSE, and RMSE remained close to zero, which implies that the DP mechanism introduces minimal degradation in data utility. Figure 6.2.1 further illustrates the relationship between the privacy parameter ε and the evaluation metrics, highlighting that data utility gradually decreases as privacy protection is strengthened. This confirms the critical role of the privacy parameter in trade-off balancing of data utility and privacy, as the error increases slightly when more noise is added (Harder, Bauer, and Park 2020).

Data Utility vs Data Privacy Trade-off

The assessment of the privacy guarantee required to balance the trade-off between data utility and privacy is depicted in Fig. 6.2.1. The graphs plot the levels of privacy parameters against the values for both the data utility and the privacy utility evaluation metric. This indicates the degree of data utility and privacy required by each privacy parameter to offer a balanced data utility-privacy trade-off (Harder, Bauer, and Park 2020). The graphs suggest similar

trends for the utility and privacy features of the data as the level of ε increases, with slight incremental differences between these values and the values of ε. This trend indicates that the utility of data decreases as privacy preservation increases, highlighting the significant role of properly adjusting the value of the privacy parameter in moderating the trade-off between the utility of data and privacy to meet the privacy requirements of the dataset.

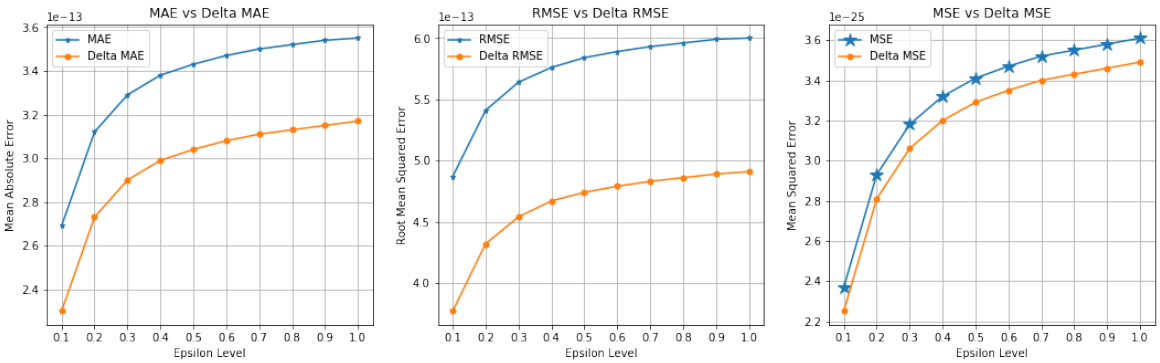

Figure 6.2.1: Data Utility vs Privacy Utility Analysis

The irregular differences exhibited by the data utility and privacy utility output during the analysis show that there is no linear relationship between the degradation of the data utility and the increased preservation of privacy. The graph shows that there may be certain levels of privacy where the loss in data utility is close to zero, and other levels show a small increase in privacy results with a significant drop in data utility. The challenge in this involves identifying an optimal value of ε that provides an acceptable level of privacy protection while minimising loss in data utility. This optimal value could vary depending on the context, the type of data involved and the sensitivity of the information. Thus, balancing the trade-off between data utility and privacy is crucial, and this can be based on various factors in the dataset, such as data type and information sensitivity. The evaluation of factors within the dataset considers attributes and additional information associated with the dataset that would assist the determination of the trade-off point between the utility of the data and privacy.

6.3 Prediction Performance Evaluation of Risk Assessment-Driven Privacy-Preservation Scheme Noise-Injected Data

The degree to which the trade-off between data utility and privacy must be offset under varying privacy requirements can be estimated by assessing intrinsic dataset factors. Determining the hazard level with TTC, which informs the situational risk provide the degree of offset needed by the data to balance the trade-off balance between data utility and privacy. Evaluating the performance of the noise-injected dataset from the Privacy Preservation Module with the evaluation criteria metrics using the Linear Regression prediction model will provide insight into the impact of balancing the data utility and privacy trade-off and the degree of offset that is needed to balance the trade-off for the dataset.

Performance evaluation would assess data utility and privacy for different datasets that meet the requirements of normal and hazard scenarios after the dataset has passed through the three modules. In the second module, the risk score was used to define the privacy parameter, and the designed Laplace noise distribution driven DP mechanism is applied to the datasets individually to obtain the noise-injected dataset that was evaluated by applying the prediction model to the dataset. The outcome of the performance evaluation criteria contributes to understanding the impact of the risk-based privacy preservation concept on balancing the trade-off between data utility and privacy.

6.3.1 Data Utility Performance Evaluation

The performance evaluation outcome of applying the Linear Regression model on the normal traffic and collision dataset without noise injection into the dataset serves as the baseline output for the analysis. The baseline evaluation metrics, where no noise is added to the dataset, show the data utility provided by the dataset for processing. The metric values for both datasets are shown in Table 6.3.1 and Table 6.3.2 , with values close to zero, indicating a model that performs well on the dataset. The low error values of 3.89×10^{-14} for MAE, 1.10×10^{-13} for RMSE, and 1.20×10^{-26} for MSE after processing the normal traffic dataset indicate the high data utility in providing accurate predictions with minimal deviation from actual values. When more noise is added, the errors gradually rise, yet the degradation in performance remains relatively small. The increase in noise levels for MSE, RMSE, and MAE leads to a slight increase in errors and suggests that the model retains substantial processing utility, making it suitable for dynamically balancing the data utility and privacy trade-off.

Metric	Privacy Parameter			
	Baseline	**0.5**	**0.6**	**0.7**
Mean Absolute Error	3.89×10^{-14}	3.43×10^{-13}	3.47×10^{-13}	3.50×10^{-13}
Root Mean Squared Error	1.10×10^{-13}	5.84×10^{-13}	5.89×10^{-13}	5.93×10^{-13}
Mean Squared Error	1.20×10^{-26}	3.41×10^{-25}	3.47×10^{-25}	3.52×10^{-25}

Table 6.3.1: Evaluation Metrics For Normal Traffic Dataset

The Collision Dataset exhibits significantly higher sensitivity to noise addition, leading to a more pronounced decline in predictive accuracy. The baseline result has higher errors, with an MAE value of 2.17×10^{-13}, RMSE of 1.85×10^{-12}, and MSE of 2.58×10^{-25}. When noise is introduced at level 0.5, MAE increases sharply to 1.31×10^{-11}. This trend continues at noise levels 0.6 and 0.7, where MAE further rises to 1.46×10^{-11} and 1.52×10^{-11}, along with a similar increase in RMSE and MSE. This suggests that the size of the errors made during processing increases because the model is unable to extract significant insight from the data.

Metric	Privacy Parameter			
	Baseline	**0.5**	**0.6**	**0.7**
Mean Absolute Error	2.17×10^{-13}	1.31×10^{-11}	1.46×10^{-11}	1.52×10^{-11}
Root Mean Squared Error	1.85×10^{-12}	2.15×10^{-12}	2.24×10^{-12}	2.37×10^{-12}
Mean Squared Error	2.58×10^{-25}	1.61×10^{-24}	1.81×10^{-24}	1.90×10^{-24}

Table 6.3.2: Evaluation Metrics For Collision Dataset

The Collision Dataset is far more affected by noise than the Normal Traffic Dataset, which suggests that the collision events already have high variability, and adding noise further distorts the dataset, leading to a drop in data utility. This results in the model showing high efficiency during processing for the normal traffic dataset in comparison to the collision dataset, which contains more imbalance due to the accident event.

6.3.2 Data Privacy Performance Evaluation

The introduction of noise into the dataset is expected to impact the data utility during processing. This thesis focuses on the dynamic balancing of data utility and the privacy trade-off to meet the privacy requirements of various scenarios. Data privacy is determined by calculating the loss of utility, measured as the increase in error from the baseline (with no noise) to the error at a given noise level. Table 6.3.2, which presents the privacy level

based on the loss of utility, shows that the increases in MAE, RMSE, and MSE remain small as noise levels increase. This suggests that the dataset retains high data utility despite the addition of privacy-preserving noise.

Noise Level	Δ MAE	Δ RMSE	Δ MSE
0.5	3.04×10^{-13}	4.74×10^{-13}	3.29×10^{-25}
0.6	3.08×10^{-13}	4.79×10^{-13}	3.35×10^{-25}
0.7	3.11×10^{-13}	4.83×10^{-13}	3.40×10^{-25}

Table 6.3.1: Privacy Evaluation for the Normal Traffic Dataset

Given the minimal impact on error metrics, the normal traffic dataset can accommodate privacy-preserving noise without significant loss of utility. Processing of this dataset has shown that noise-adding, privacy-preserving mechanisms may lead to slight inaccuracies after data processing; however, these inaccuracies do not significantly impact overall analyses.

The privacy evaluation of the collision dataset is presented in Table 6.3.2. As with the data utility evaluation, the MAE for the collision dataset is much larger than that for the normal traffic dataset. This is also true for MSE and RMSE. This shows that the dataset is highly sensitive to noise, with each increase in privacy protection leading to a severe degradation in data utility. Collision events are less frequent and are influenced by factors such as weather conditions, road anomalies, and driver behavior, making them highly sensitive to disruptions. The addition of an inappropriate level of noise can obscure and derail critical details required during processing.

Noise Level	Δ MAE	Δ RMSE	Δ MSE
0.5	1.29×10^{-11}	3.00×10^{-13}	1.35×10^{-24}
0.6	1.44×10^{-11}	3.90×10^{-13}	1.55×10^{-24}
0.7	1.50×10^{-11}	5.20×10^{-13}	1.64×10^{-24}

Table 6.3.2: Privacy Evaluation for the Collision Dataset

The MAE increase for the collision dataset is nearly two orders of magnitude larger than that for the normal traffic dataset across all noise levels. This suggests that collision-related predictions are much more sensitive to the introduction of noise, implying a significant loss of data utility when privacy measures are applied. In contrast, the normal traffic dataset maintains a steady structure despite the noise.

The findings, as seen in Figure 6.3.1, have shown the need for dynamic, privacy-preserving methods to balance the trade-off between data utility and privacy during processing. The high sensitivity of collision data requires delicate noise introduction to maintain data utility while preserving privacy. Given the differences in dataset behavior, a one-size-fits-all

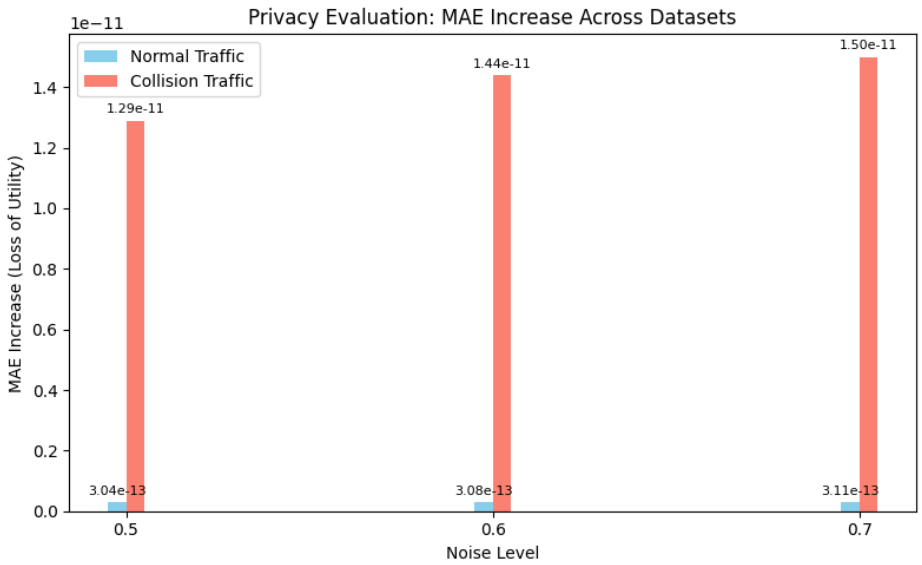

Figure 6.3.1: Placeholder for privacy evaluation visualization.

approach to privacy preservation is insufficient. A dynamic trade-off balancing approach that adapts the privacy level based on requirements and data characteristics is crucial in this case. In situations where data is less critical, such as low-safety-critical scenarios, a higher level of privacy can be applied; however, when situations become safety-critical, the approach is reversed, and data utility becomes paramount while still maintaining data privacy. This dynamic approach ensures that data utility is preserved where it matters most while still providing adequate privacy protection.

The metric values of the collision dataset have shown the importance of carefully controlling the privacy level to avoid significant utility degradation. During safety-critical situations, precise location data is essential for planning targeted interventions, as errors in location coordinates can misplace a collision incident, leading to incorrect assessments of high-risk zones and delaying emergency responses.

6.4 Discussion

The experimental validation outcome for the proof-of-concept design has demonstrated the feasibility of assessing the risk associated with the intrinsic factors within the trajectory dataset to specify the level of privacy parameter ε. The ε level defines the degree of the Laplace noise mechanism of DP to protect sensitive information while maintaining good utility capabilities. The evaluation metrics, which are MAE, MSE, and RMSE, used to gauge the model's performance, exhibit admirable performance when applied to a new dataset

despite the introduction of Laplace noise for privacy. However, the increasing level of noise applied to the dataset results in noticeable error values for these metrics. This indicates that increasing privacy levels dampen data utility, and the privacy-preservation prediction model in this thesis has been shown to exhibit a high level of resilience, efficiently managing the added complexity and providing commendable performance compared to actual values (Dwork 2008).

A key observation was that as the level of privacy increased (through the amplification of the amount of noise added to the data), there was a decline in model performance (through an increase in the evaluation metric errors). The complexity of balancing the trade-off between data utility and privacy is a challenge in the application of DP. The relationship between data utility and privacy is not linear, necessitating adjustment of the privacy requirement according to the level of privacy parameters (Abadi et al. 2016). This decrease in model performance can be attributed to the reduction in the information available for use by the model due to the introduction of noise by the data privacy preservation technique. This reduces the ability of the model to explore the data and compromises prediction efficiency (Sarathy and Muralidhar 2010). This highlights that the pursuit of high levels of privacy may require some compromise in model performance and data utility. Other privacy-preserving schemes for trajectory data, such as location generalisation and location perturbation, demonstrate similar trends where an increase in privacy level leads to a decrease in model performance. This in turn compromises data utility and impacts the accuracy of forecast outcomes (Shokri et al. 2011).

One of the most significant contributions of the framework is the successful integration of the three modules when handling location trajectory data. This achievement carries a high degree of complexity due to the unique nature of location data, which often exhibits temporal correlations. The successful application of DP based on hazard assessment marks a critical milestone in protecting privacy, demonstrating its ability to manage high-dimensional data characterised by temporal correlations (Shokri et al. 2011; Dwork and Roth 2014). This study builds on previous work such as (Zhang et al. 2022; Hao, Wu, and Wan 2023) that applied differential privacy to location data in data publishing or data aggregation settings. However, those works mostly protected user-level privacy, whereas this study needed to enforce protection on the location information of a single user, considering the temporal correlations between a user's locations to enhance information credibility, improve the security of sensitive information, and maintain accurate results during processing.

The evaluation criteria metric output for the collision dataset demonstrated commendable performance. This indicates that the framework is robust and adaptable, effectively handling collision trajectory data. The different evaluation metrics exhibited show that a one-size-

fits-all approach to implementing privacy preservation mechanisms for different trajectory situation is not optimal.

The validation of this thesis concept uses the baseline dataset that does not have any level of noise introduced. The normal traffic dataset exhibited the lowest error values across all evaluated metrics, while the collision dataset showed significantly higher errors. This indicates that the model achieves greater data utility with the normal traffic data, whereas collision-related predictions are less accurate. The baseline performance across these two datasets demonstrates the model's adaptability in handling diverse data types, which is crucial in real-world scenarios where data can originate from varied sources with distinct properties. However, the pronounced differences in accuracy underscore the need for tailored parameter tuning and feature selection for each dataset to optimise model performance for specific applications.

The baseline error metrics for the collision dataset are higher due to the infrequent nature of collision events, and the values increase further when noise is injected into the dataset. The error value increase is noticeable in MAE, MSE, and RMSE, and this demonstrates that enhancing privacy compromises utility, making the trade-off between utility and privacy even more pronounced. In collision events, minor noise-induced perturbations can obscure critical details, such as location information essential for mapping accident hotspots and planning timely interventions. This suggests that rare trajectory events increase the data's sensitivity to noise, thereby necessitating a careful balance between utility and privacy. The importance of tuning the privacy parameters to optimally balance the trade-off between data utility and privacy cannot be overstated.

The privacy-utility evaluation for the collision dataset underscores the critical need to balance data privacy and utility when processing trajectory data. In the context of collision events, which are inherently infrequent and complex, evaluating the risks associated with this complexity is essential for determining the appropriate level of DP noise to inject into the dataset. The introduction of noise based on the risk score for collision events demonstrates that the framework can effectively adapt to various scenarios, dynamically balancing the data utility and privacy trade-off. The high sensitivity of the collision dataset to noise injection shows that a lack of careful calibration of the privacy parameter would lead to a significant drop in data utility. When the privacy parameter is appropriately tuned, effective balancing of data utility and privacy is achieved.

The privacy-utility evaluation shows the importance of balancing the data privacy-utility trade-off when processing trajectory data. Evaluating and quantifying the risk associated with collision events enables the specification of the optimal privacy parameter for DP noise injection, thereby safeguarding sensitive information while maintaining data utility. The

introduction of noise in the collision dataset, based on the risk score, demonstrates that the framework can effectively adapt to scenarios where trajectory data is acquired, successfully balancing the trade-off between data privacy and utility.

The findings of this study align with previous research that emphasises the importance of a customised approach to the preservation of privacy in diverse data scenarios and support the data utility and privacy trade-off hypothesis (Thakurta and Smith 2013). The adaptability of the model, as demonstrated by the performance across different datasets, supports the assertion of Johnson and Shmatikov (2013) that effective data models must be able to handle diverse data sources. The observation that the introduction of noise to preserve privacy can impact the outcome of data processing is consistent with the findings of Dwork et al. (2016), who noted a trade-off between data privacy and utility in their seminal work on Differential Privacy. The results of the privacy-utility evaluation show the importance of balancing the data privacy-utility trade-off, a concept that has been extensively discussed in the literature of Sweeney (2002) and Narayanan and Shmatikov (2019).

The ability of RBDPM to maintain this balance when noise is introduced at calculated privacy levels suggests that it is effective in various data scenarios and aligns with the work of Machanavajjhala et al. (2007). The performance of the model on different datasets and privacy parameters demonstrates its robustness, adaptability, and sensitivity to privacy levels. These findings contribute to the growing body of literature on privacy preservation in data models and provide valuable insights for future research and application in this field.

6.5 Model Comparative Analysis

The investigation by Jiang et al. (2021b) into the data-driven framework focused on protecting the location privacy of users by processing the trajectories of users that can be applied to multiple types of destination prediction methods. The central idea of their framework is the incorporation of a unique differential privacy design to construct a privacy-preservation, data-driven model, which uses Multiple Linear Regression to formulate the relationship between the injected noise and privacy preservation. They focused on providing an optimised quantifiable framework that combines a Recurrent Neural Network and Multi-hill Climbing for adding fine-grained noise to obtain the trade-off between privacy preservation and the utility of the predicted results.

The first sample had noise added to various degrees in the trajectory data to generate multiple samples from the trajectory dataset. Then, the raw and noisy samples were fed into the prediction model as input to obtain the multigroup prediction results used to construct a data-driven model. Thereafter, Multiple Linear Regression is used to fit the relationship

between the noise scale and privacy protection. Assessing the utility of the predicted results, they optimised the noise scales at each location along the sub-trajectories to obtain the trade-off between privacy preservation and the utility of the predicted results. Then, Neural Arithmetic Logic Units (NALUs) were used to formulate a neural network model and Multi-hill Climbing was utilised to find a sub-optimal setting because of the huge overhead of fine-density traversal. The performance of the model is validated on the T-Drive sample dataset. The extensive results validate that their framework can be applied to different prediction methods, provide quantifiable preservation of location privacy, and guarantee the utility of the predicted results simultaneously.

This result proves that the framework performs similarly for both equally distributed privacy requirements and randomly distributed privacy requirements. This is because the effects of protecting privacy for each user are independent of each other, and different users select different privacy requirements. Therefore, their privacy preservation frameworks can provide effective quantitative protection (low MAE) and robust (low RMSE) performance to satisfy users' privacy requirements. Although they discovered that optimisation does harm the performance of privacy preservation, the optimised framework greatly retains the utility of the predicted results.

The contribution of their work includes providing proven guarantees for both the users' location privacy and the utility of the predicted results, and it can be applied to several destination prediction methods. Constructing an optimisation framework by employing NALUs and Multi-hill Climbing to obtain the trade-off between privacy preservation and the utility of the predicted results. This framework considers the trade-off for sub-trajectories rather than the full trajectory dataset. However, this does not consider the intrinsic factors of the trajectories or sub-trajectories to obtain the optimal trade-off between privacy preservation and processing results utility. The usage of Multiple Linear Regression for the determination of the trade-off was not described as a means to efficiently balance the trade-off, and the results for equally distributed privacy requirements and randomly distributed privacy requirements appear to be similar.

The work of Jiang et al. (2021b) uses similar privacy preservation and linear regression components as in this thesis to achieve their purpose of quantifiable privacy preservation for destination prediction in LBS. This is comparable to the approach taken in this thesis. Their result shows that the effects of privacy preservation for each user are independent of each other, and the low MAE and RMSE values obtained during the evaluation show that the privacy preservation framework provides effective and robust quantitative protection that satisfies users' privacy requirements despite optimisation reducing utility. However, the framework still provides a comparable privacy preservation effect compared to the general

framework. They did not consider the reduction of the utility of the predicted results. The thesis obtains low MAE and RMSE values that indicate that exploring the attributes within the dataset and using them to specify the noise level to inject into the data will provide an effective and robust privacy protection system for processing information.

The use of optimisation of the noise scales at each location along the sub-trajectories to obtain the trade-off between privacy preservation and the utility of the predicted results using NALUs and Multi-hill Climbing to find a sub-optimal setting produces a huge overhead of fine-density traversal, unlike the use of a hazard assessment methodology in this thesis that requires minimal human input or computational overhead to determine the trade-off between privacy and the utility of the predicted results. The adoption of varying mobility patterns by an individual in transit shows that privacy requirement would differ as different mobility patterns are adopted, and using NALUs and Multi-hill Climbing to balance the trade-off between privacy preservation and the utility of the predicted results would rack up huge computational overhead, but the usage of context-based hazard assessment methodology reduces computational overhead and is adaptable to quickly changing mobility patterns that are interpreted to provide suitable quantified privacy levels that balance privacy and utility trade-off for the model.

The work of Jiang et al. (2021b) focuses on destination prediction in LBSs rather than trajectory prediction as seen in this thesis. The privacy requirements differ for each scenario as the requirement for Jiang et al. (2021b) is based on predicting a final destination point and ensuring the privacy budget ε is close to the privacy requirement, unlike in this thesis which requires the generation of multiple predictive location points that focus on highly accurate results while preserving privacy. The paper uses the Beijing T-Drive dataset that is used in this thesis. The low-evaluation-criterion metric outcome obtained shows good performance for the framework and models. This indicates that the model is effective, generalised and robust to provide accurate predictions while preserving privacy.

The use of Differential Privacy to preserve privacy for vehicular trajectories was shown in the work of Arif et al. (2021). The sensitivity of information within the trajectory is a big responsibility that can lead to identity theft, which can disrupt a person's life when it is leaked. Users want to avoid the relational connection between the information within their data and do not want to worry about disclosing multiple locations. This led to the work by Arif et al. (2021) that implemented different variants of differential privacy in four real-time datasets, including the T-Drive sample dataset, with MAE being one of the six evaluation metrics used to validate the efficiency of their proposed method. They found that the MAE values provided for the different variants of Differential Privacy are low, with the highest MAE value being 0.12 over the period of their work. The introduction of an anonymisation mechanism on the

trajectory data increases the accuracy and information accessibility during data publication. Their proposed framework shows the lowest MAE value compared to the other models in the paper and the framework. The proposed framework consumes the security spending plan during development adjustments, and the number of advancement adjustments is not exactly equivalent to the total number of queries. The complexity of the proposed technique for the process requires long calculation periods, leading to computational overhead.

The paper by Arif et al. (2021) focuses on privacy preservation for the publication of vehicular trajectory data with differential privacy using continuous static informational datasets. This is different from this thesis, which focuses on maintaining data utility while preserving data privacy. The thesis and the paper of Arif et al. (2021) aim to protect the privacy of the data involved in the publication of vehicular trajectories. The main limitation of the method of Arif et al. (2021) is the use of continuous static informational datasets, and this technique may not be useful for dynamic datasets. The paper mentioned, in terms of future work, the need to apply differential privacy on other moving objects to guarantee the privacy of the moving objects. This thesis focuses on applying differential privacy to moving objects. This thesis evaluates safety-critical situations such as collision events during transit and uses that to determine the suitable privacy level to balance the trade-off between data utility and privacy during processing.

The work of Cheng et al. (2022) focuses on establishing the probabilistic mobility model of trajectories and clustering the locations to achieve semantic location matching between different trajectories. Based on the semantic similarity, they identify the trajectory and propose a privacy level allocation method based on stay points and frequent sub-trajectories. Then, according to the location matching results, they automatically identify the privacy level of all locations. Combined with the optimal location-differential privacy mechanism, they perturb the location points on the user's trajectory before publishing, where different location privacy levels correspond to different privacy budgets.

The paper assessed the introduction of a personalised DP mechanism to balance privacy protection and data utility in trajectory privacy protection. Personalised DP allows for the adjustment of the privacy level of each record based on the privacy needs of the data/user. The paper proposes an optimal personalised trajectory DP mechanism for trajectory data publication to avoid manual designation of privacy level by exploiting the relationship between location features and privacy requirements. They proposed a location matching method based on semantic similarity derived from a cluster of locations on different trajectories and built a probabilistic mobility model of trajectories. This aims to identify the privacy level of a location by automatically learning beyond the feature extraction step.

The use of the GPS trajectory dataset from the T-Drive Data with AdvError and QualityLoss as their evaluation metrics shows that their model can improve the data utility of published trajectory data compared to other algorithms. The degree of privacy protection gradually decreases, which indicates that the degree of privacy protection for sensitive locations is higher than that for non-sensitive locations and concludes that their model can achieve the balance between privacy protection and data utility.

This thesis focuses on combining privacy preservation mechanisms to apply privacy in safety-critical situations through hazard assessment, risk quantification, and privacy parameter determination that help to determine the trade-off balance for data privacy and utility. The thesis has shown that the combination of these components enhances the preservation of privacy while maintaining utility to produce accurate results during processing that are similar or better than the outputs of existing state-of-the-art models. The capacity of the model discussed in the thesis provides lower MAE and RMSE values than the values given by the paper by Arif et al. (2021). Compared to Jiang et al. (2021b), the thesis model provides low evaluation metric values and is capable of doing more than predicting a single destination point by predicting multiple location points that can form a trajectory path or a moving direction that helps with route planning, which is an extension of the future research work stated by Arif et al. (2021). Similarly to the trajectory model concept used by Cheng et al. (2022), this model focuses on the inherent information provided by the location dataset by evaluating hazard to classify risk that is translated to a privacy parameter to balance the trade-off between the utility of data and the preservation of privacy to be deployed during processing. Semantic matching requires a higher computational overhead compared to the use of a hazard assessment mechanism for risk quantification that is translated into the privacy level.

The novelty of the thesis evolves around the combination of the Hazard Assessment, Privacy Preservation, and Noise Application Modules, which is the first of its kind to provide a personalised privacy parameter based on the safety-critical situation associated with a user's location information to balance the trade-off between data utility and privacy. Some research in areas such as data publishing (Dwivedi 2017; Liu and Liu 2021) and parking recommendation (Saleem et al. 2021) has combined some of the components such as DP in this thesis to different capabilities with good results. The thesis has shown that the combination of these components would provide noise-injected trajectory data that balance data utility and privacy trade-offs for data processing. The use of a prediction model to validate the concept shows that the modality proposed in this thesis for the injection of noise into trajectory data provides a generalised, robust, and effective means of personalising privacy preservation levels based on personnel or environmental requirements.

6.6 Conclusion

The experimental validation for the RBDPM using the Linear Regression prediction model was carried out to substantiate the robust performance of the model during the processing of location data across diverse travel patterns. This framework consists of three integral modules: the Hazard Assessment, Privacy Preservation, and Noise Application Modules, which collectively operate to ensure that user privacy is preserved while maintaining data utility. The framework processes trajectory data produced by the user and transforms it into a Laplace noise-injected trajectory dataset, which is then processed with the prediction model to validate the concept (Mohr, Zhang, and Schueller 2017; Schwarting, Alonso-Mora, and Rus 2018; Siegel, Erb, and Sarma 2017).

The Hazard Assessment Module evaluates hazards associated with mobility patterns by analysing the attributes of trajectory data such as stopping distance, relative distance, and Time-To-Collision which inform the risk associated with the situation. The Privacy Preservation Module focuses on the determination of the privacy parameter level from the risk score and the privacy noise mechanism selection for the scheme. The Noise Application module is where noise is added to the dataset to create a dataset that balances privacy concerns with the need for data utility.

The process involves transforming user trajectory data into a noise-injected dataset, which is then used for validation with the Linear Regression model. This validation demonstrates the successful possibility of dynamically balancing the trade-off between data utility and privacy preservation. It demonstrates that the assessment of trajectory dataset attributes provides key factors that determine the optimal amount of noise required to balance the trade-off between data utility and privacy. The assessment of risk and the application of noise, guided by the specified privacy parameters, are crucial in evaluating the effectiveness of this privacy-preserving approach (Mohr, Zhang, and Schueller 2017; Schwarting, Alonso-Mora, and Rus 2018; Siegel, Erb, and Sarma 2017).

The Linear Regression prediction model was adopted to validate the concept, aiming to overcome the limitations observed in earlier models like ARIMA and Regression Tree Ensemble, which were discussed in Chapter 3. These earlier models, while effective at predicting location coordinates from trajectory data, exhibited a critical flaw where their predictions too closely mimicked the patterns seen in the training data. This resulted in forecasts that were almost identical to those of the training set, indicating a potential risk in terms of predictability and privacy (Petropoulos et al. 2022). In contrast, Linear Regression introduces variability by deviating from these patterns, maintaining prediction accuracy and showing enhanced performance (Petropoulos et al. 2022; Jiang et al. 2021a; Yuan et al. 2021).

Differential Privacy is a privacy preservation technique that allows for the analysis of datasets while safeguarding individual privacy. The necessity for robust privacy mechanisms for location information is increasingly acknowledged, particularly when deploying predictive models. DP enables a balance between data utility and privacy, ensuring that the outputs vary sufficiently when a single data point changes, thus protecting individual data without losing the overall informative value of the dataset. This chapter explores the potential of integrating Differential Privacy as a privacy preservation strategy in the context of location data processing (Dwork 2008; Dwork, Kohli, and Mulligan 2019).

The concept of optimal utility-privacy trade-off balance has become a topic for research aimed at balancing the trade-off between dataset utility and privacy. This concept serves as a privacy metric applicable to DP systems for assessing both utility and privacy. The privacy metric can be influenced by inherent information within location data and the research is still in infancy as shown by the work of Cheng et al. (2022). This thesis explores the impact of these inherent features in determining the privacy parameter level necessary for balancing the trade-off between privacy and utility, thereby protecting privacy during processing while maintaining accurate results.

This thesis has examined the application of the Laplace Noise mechanism of DP to inject noise into trajectory dataset to protect sensitive information while striving to achieve the best possible results during dataset processing. The hazard assessment approach ensures that the specified privacy parameter levels are tailored to the safety-critical information within the data. The level of noise introduced is modulated by the privacy parameter ε, and the performance of the prediction model used for concept validation is evaluated based on the evaluation metric criteria. This approach seeks to tailor privacy measures to the dynamic nature of mobility data, providing insights into how privacy can be maintained without significantly undermining the utility of the data during processing.

The hazard assessment concept that evaluates the risk associated with collision events provides a means to assess the varying risk profiles that such events might possess to offer a customised approach to privacy protection that helps overcome the struggle with the delicate balance between maintaining data privacy and ensuring utility (Saleem et al. 2021; Rassouli and Gündüz 2019; Loukides and Shao 2008; Alvim et al. 2018; Erdogdu and Fawaz 2015). The necessity for effective mechanisms to study collision data, which can offer insights into road user behaviours and trajectories is emphasised by Ghaleb, Razzaque, and Isnin (2013). Time-to-Collision (TTC) has been used as a metric for crash analysis for location-based services (LBS) data (Barhoumi, Zaki, and Tahar 2024). The implementation of DP brings forth considerable challenges in managing the trade-off between privacy and utility,

prompting research into defining an optimal balance for this trade-off (Yao, Zhou, and Ma 2016; Kalantari, Sankar, and Sarwate 2018).

The integration of privacy preservation with hazard assessment represents an advancement over current personal safety solutions (Sogi et al. 2018; Viswanath and Basu 2015; Kartik, Jose, and MK 2017; Manazir, Govind, and Rubina 2019; Chaudhari et al. 2018; Srikrishna and Veena 2017; Patel and Hasan 2018; Miriyala et al. 2016; Yarrabothu and Thota 2015; Hariharan et al. 2021; Walkunde, Shinde, and Pandhare 2022; Saranya et al. 2021; Prashanth, Patel, and Bharathi 2017; Reddy et al. 2021; Aminuddin et al. 2019; Gadhave et al. 2017; Shaikh and PB 2008; Priya et al. 2021; Bhadula, Benjamin, and Kakkar 2021). These systems typically activate emergency responses and track user movements but fail to consider the privacy implications and potential risks associated with the data collected during the assistance process. This thesis thus introduces a framework that enhances personal safety by incorporating privacy preservation using inherent data information to manage utility and privacy trade-off balance.

Chapter 7

Conclusion

7.1 Conclusion

This thesis is a significant contribution to the advancement of privacy preservation for location data, striving to improve the safety, efficiency, and privacy of road users and location-based services. The adoption of predictive model for the proof-of-concept validation experimentation follows research that used predictive model for the improvement of traffic management, early hazard detection, and enhancement of personal safety applications (Yang and Hua 2019; Lin 2017). LBS generated data contains sensitive information that poses substantial privacy risks if not adequately protected. Balancing the trade-off between preserving data utility for operational effectiveness and ensuring robust privacy protection remains a central challenge in this domain. This thesis addresses this challenge through the development of the Risk-Based Differential Privacy Model (RBDPM), a novel framework designed to dynamically tailor privacy preservation to the safety-critical contexts of location data, thereby advancing the field of privacy protection and personal safety solutions. The feasibility of the proposed model to enhance a range of applications, including emergency response, traffic management, and location-based services has been demonstrated. This is a valuable addition to the field of privacy preservation and safety applications.

Despite the limitations recognised in the current research, this thesis lays a robust groundwork for future exploration within the field of privacy preservation of location data and the dynamic balancing of data utility and privacy trade-off. The suggested directions for further research will allow for expansion of the existing knowledge base, devising of innovative solutions, and addressing of the multifaceted challenges associated with personal safety, emergency response, and privacy protection in intelligent transportation systems. This concluding chapter reflects on the contributions to privacy preservation in safety-critical situations, acknowledges the study's limitations and outlines future research directions.

7.2 Contributions to the Field: Enhancing Personal Safety and Privacy Preservation

This thesis focuses mainly on advancing privacy preservation for location data. This led to the design and development of the RBDPM, a framework comprising three integral modules: the Hazard Assessment Module, the Privacy Preservation Module, and the Noise Application Module. These components combine to assess collision risks, quantify associated privacy needs, and apply differential privacy mechanisms to trajectory data, ensuring a tailored trade-off balance between data privacy and utility. The experimental validation, conducted using a Linear Regression prediction model, substantiated the RBDPM's robust performance in processing location data across diverse collision scenarios, outperforming earlier models like ARIMA and Regression Tree Ensemble, which struggled with overfitting to training data patterns (Petropoulos et al. 2022).

Key contributions include:

- **Integration of Component-Driven Hazard Assessment and Privacy Preservation:** This thesis uses TTC as a key hazard metric, calculated to determine the proximity of potential vehicle collisions and reflects the typical safety margin across interactions. These TTC values informed the assignment of risk scores and the risk categorisation directly influenced the privacy parameter epsilon. This shows how features inherent to data can be used to balance the trade-off between data utility and privacy.

- **Dynamic Privacy-Utility Balance:** The linking of risk scores to noise levels forms the basis for the dynamic balance that protects sensitive trajectory data without unduly compromising the utility for applications like emergency response and traffic management. This balance is evidenced by the tailored epsilon adjustments and their measurable impact on prediction accuracy.

- **Data Attribute Assessment on Privacy Preservation:** The analysis of distinct data attributes, such as TTC, stopping distance, and relative distance in this case shapes privacy needs. This facilitates a personalised privacy-preservation strategy that aligns with the specific requirements of the scenario, contributing to the development of more effective privacy-preserving mechanisms.

These contributions position the RBDPM as a transformative tool for state-of-the-art personal safety solutions that often neglect privacy implications (Sogi et al. 2018; Viswanath and Basu 2015).

7.3 Limitations

While the RBDPM represents a significant step forward, several limitations warrant consideration, providing insights for refinement and future exploration:

- **Implementation Complexity:** The validation is based on a specific collision dataset, which may not fully represent the complexity and scale of real-world VANETs with thousands of vehicles interacting simultaneously. The computational overhead of calculating Time-to-Collision (TTC), assigning risk scores, and applying Laplace noise for each vehicle could become prohibitive as the data size increases.

- **Accuracy of TTC in Diverse Conditions:** The TTC metric, central to the hazard assessment, relies on accurate velocity and position data. In real-world scenarios, factors such as sensor noise, signal interference (e.g., in tunnels or urban canyons), or adverse weather conditions could degrade the precision of these inputs, leading to unreliable TTC values and subsequent risk scores.

- **DP Parameter Tuning:** Fine-tuning epsilon dynamically and adaptively to capture nuanced risk variations, rather than relying on static thresholds, is complex and requires further exploration to avoid over- or under-protecting data.

- **Real-Time Processing Constraints:** The multi-step process of TTC calculation, risk scoring, epsilon derivation, and noise application introduces latency that may not align with the real-time requirements of personal safety applications, such as instantaneous hazard warnings. The Linear Regression model's validation, while effective, assumes batch processing rather than streaming data, further complicating real-time deployment.

- **Privacy Evaluation Metrics and Benchmarking:** Evaluating the performance of privacy-preserving mechanisms is inherently challenging due to the lack of standardised metrics for both privacy and utility. Establishing a comprehensive and fair benchmarking framework to compare the proposed model with existing solutions remains a challenge, potentially affecting the perceived effectiveness of the approach.

- **Data Variability:** The framework relies on trajectory data, which may vary significantly in quality, density, and noise characteristics across different real-world scenarios. This heterogeneity could impact the model's performance and limit the generalisability of the findings across different datasets and mobility patterns.

7.4 Future Direction

These limitations highlight important areas for future research and refinement to ensure the practical applicability and scalability of the proposed model in real-world scenarios. Future research would explore:

- **Expansion to Multiple Mobility Patterns:** Expand the RBDPM to incorporate a broader range of mobility patterns beyond collision-focused data, such as routine traffic flows, pedestrian movements, or mixed vehicle types (e.g., cars, trucks, bicycles). This could involve testing the model with diverse datasets, including traffic mobility patterns, and behavioral patterns, to assess the adaptability and effectiveness across varied scenarios.

- **Dynamic Thresholds:** Dynamic threshold mechanism for Time-to-Collision (TTC) and risk scoring that adjusts in real-time based on contextual factors, such as road conditions, traffic density, or weather. Machine learning techniques, such as reinforcement learning, could be employed to refine thresholds and optimize epsilon derivation for nuanced risk profiles.

- **Integration of Additional Contextual Factors:** Enhance the hazard assessment process by incorporating supplementary contextual information (e.g., environmental conditions, road infrastructure, weather data) to refine risk quantification and improve the determination of optimal privacy levels.

The identified future directions would enhance RBDPM and empower the privacy preservation model with the ability to safeguard trajectory datasets with multiple mobility patterns and offer an enhanced dynamic system that balances data utility and privacy trade-off to meet the privacy requirements of the user or scenario (Martinez et al. 2016).

7.5 Practical Implications: Applications of RBDPM

The existing effort to enhance data privacy associated with trajectory or location information is expanding with this research offering a unique approach to balancing privacy and utility trade-off while maintaining good data utility performance. This model addresses several challenges associated with safety-critical data and usage in the preservation of privacy. The following points highlight the relevance of the research findings. This section highlights how this research can be applied in real-world contexts to enhance safety, efficiency, and privacy of personal safety solutions and intelligent transport systems. These include:

- **Enhanced Emergency Response Systems:** The RBDPM can be integrated into Location-based emergency response systems to protect sensitive location data of distressed users while enabling rapid and efficient deployment of assistance. By applying tailored noise levels based on for example TTC-derived risk scores, the model ensures that responders receive accurate routing information without compromising individual privacy. This enhances the reliability of emergency services, reducing response times and improving outcomes in critical situations like accidents, while adhering to privacy regulations (Nekovee and Bogason 2007; Papadimitratos et al. 2008).

- **Traffic Management:** Traffic management authorities can deploy the model to analyze collision-prone trajectory data, optimizing traffic flow and reducing congestion while preserving driver privacy. The model's ability to balance utility and privacy allows for effective signal timing and route planning without exposing individual movement patterns. This improves road safety and efficiency in urban environments, supporting congestion mitigation efforts while addressing privacy concerns in large-scale traffic monitoring (Wei et al. 2020; Naumov and Gross 2007).

- **Privacy-Conscious Smart City Infrastructure:** In smart city infrastructure, the model can underpin intelligent transportation systems by providing privacy-protected trajectory data for urban planning and public safety management. The risk-based noise application ensures that sensitive collision data is safeguarded, while still supporting data-driven decisions like infrastructure upgrades or emergency preparedness. This facilitates secure, scalable urban development, enabling cities to leverage location data for sustainability and safety initiatives without risking citizen privacy (Lecuyer et al. 2019).

- **Improved Vehicle Safety Alerts:** The model can enhance vehicle-to-vehicle (V2V) communication systems by embedding privacy-preserving collision risk alerts. Using TTC metrics, the model can trigger warnings for drivers in high-risk scenarios while anonymizing shared location data, preventing potential collisions without exposing precise vehicle positions. This strengthens real-time safety features in connected vehicles, reducing accident rates and fostering trust in VANET technologies by ensuring privacy compliance.

In general, research on RBDPM can significantly improve various aspects of emergency response, traffic management, location-based services, and smart cities, including faster response times, improved traffic flow, personalised services, data-driven decision-making, and improved safety and security, while also protecting the privacy of individuals.

7.6 Final Remarks

This thesis is a significant contribution to the advancement of privacy preservation for personal safety solutions, striving to improve the safety, efficiency, and privacy of transportation networks and location-based services. The central accomplishment of this thesis is the development of the RBDPM that focuses on addressing the complex challenge of balancing data privacy and utility trade-offs, minimizing the impact of privacy preservation on the utility of trajectory data during processing. The model's successful development and validation, anchored by the integration of TTC-based hazard assessment, and tailored Laplace noise application to demonstrate the feasibility and adaptability across diverse real-world applications, including emergency response, traffic management, and smart city systems.

An integral part of this study lies in the departure from static, one-size-fits-all privacy approaches. By leveraging TTC to quantify collision risks and assigning risk scores, the RBDPM dynamically adjusts the privacy parameter epsilon to suit specific safety-critical contexts. This adaptability, validated through Linear Regression with evaluation metrics, underscores the model's capacity to protect sensitive data while preserving data utility during processing (Petropoulos et al. 2022).

The practical implications of the RBDPM are profound, promising significant enhancements across multiple domains. In emergency response, it safeguards distressed users' data while optimizing rapid assistance deployment; in traffic management, it improves flow and reduces congestion with privacy-conscious analytics; and in smart cities, it supports secure, data-driven urban planning (Nekovee and Bogason 2007; Wei et al. 2020; Lecuyer et al. 2019). These applications not only validate the model's utility, it also highlight the potential to transform location operations by improving response times, traffic efficiency, personalized services, and overall safety and security, all while upholding stringent privacy standards.

Despite the contributions, this research acknowledges limitations, such as implementation complexity, epsilon calibration challenges, and a focus on limited data scenario, that temper the immediate scalability and generalisability. The suggested directions for further research will allow the scientific community to expand the existing knowledge base, integrate multimodal mobility patterns, optimize real-time processing, devise innovative solutions, and address the multifaceted challenges associated with personal safety, emergency response, and privacy protection in the realm of intelligent transportation systems.

The findings of this thesis provide a substantial contribution to the understanding and evolution of privacy preservation solutions within VANET. They pave the way towards the realisation of more efficient, secure, and privacy-conscious transportation networks and location-based services.

Finally, this thesis meets the objectives as outlined in Chapters 2 through 6, by delivering a novel RBDPM that enhances personal safety solution privacy preservation by dynamically balancing the data utility and privacy trade-off. The model lays emphasis on customizing data parameters to dataset characteristics and risk profiles, reinforcing the model's adaptability and robustness. The findings pave the way for more efficient, secure, and privacy-conscious transportation networks, contributing substantially to the understanding and evolution of privacy preservation in personal safety solutions. This research not only advances theoretical knowledge but also sets a scheme for practical, impactful innovations, placing the RBDPM as a fundamental component for future explorations in privacy-aware intelligent systems.

Reference

Aasha, J. O., Monica, S, and Brumancia, E 2015. “A tracking system with high accuracy using location prediction and dynamic threshold for minimizing SMS delivery”. In: pp. 1–6.

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L. 2016. “Deep learning with differential privacy”. In: *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pp. 308–318.

Agrawal, R. and Adhikari, R. 2013. “An introductory study on time series modeling and forecasting”. In: *Nova York: CoRR*.

Akhtar, N., Ergen, S. C., and Ozkasap, O. 2014. “Vehicle mobility and communication channel models for realistic and efficient highway VANET simulation”. In: *IEEE Transactions on Vehicular Technology* 64.1, pp. 248–262.

Al-Hussaeni, K., Fung, B. C., Iqbal, F., Dagher, G. G., and Park, E. G. 2018. “SafePath: Differentially-private publishing of passenger trajectories in transportation systems”. In: *Computer Networks* 143, pp. 126–139.

Alda, F. and Rubinstein, B. I. 2017. “The bernstein mechanism: Function release under differential privacy”. In: *Thirty-First AAAI Conference on Artificial Intelligence*.

Alghamdi, T., Elgazzar, K., Bayoumi, M., Sharaf, T., and Shah, S. 2019. “Forecasting traffic congestion using ARIMA modeling”. In: *2019 15th international wireless communications & mobile computing conference (IWCMC)*. IEEE, pp. 1227–1232.

Alofe, O. M., Fatema, K., Panneerselvam, J., and Kurugollu, F. 2019. “Saving Victims in Moving Vehicles: an IoT based prediction model aided solution”. In: *2019 2nd IEEE Middle East and North Africa COMMunications Conference (MENACOMM)*. IEEE, pp. 1–6.

Alvim, M., Chatzikokolakis, K., Palamidessi, C., and Pazii, A. 2018. “Local differential privacy on metric spaces: optimizing the trade-off with utility”. In: *2018 IEEE 31st Computer Security Foundations Symposium (CSF)*. IEEE, pp. 262–267.

Alzyout, M. S. and Alsmirat, M. A. 2020. “Performance of design options of automated ARIMA model construction for dynamic vehicle GPS location prediction”. In: *Simulation Modelling Practice and Theory* 104, p. 102148.

Aminuddin, M. A. I. M., Osman, M. A., Zainon, W. M. N. W., and Talib, A. Z. 2019. “Location Tracking and Location Prediction Techniques for Smart Traveler Apps”. In: *Proceedings of SAI Intelligent Systems Conference*. Springer, pp. 83–96.

Anagnostopoulos, T., Anagnostopoulos, C., Hadjiefthymiades, S., Kyriakakos, M., and Kalousis, A. 2009. “Predicting the location of mobile users: a machine learning approach”. In: *Proceedings of the 2009 international conference on Pervasive services*, pp. 65–72.

Arif, M., Chen, J., Wang, G., Geman, O., and Balas, V. E. 2021. “Privacy preserving and data publication for vehicular trajectories with differential privacy”. In: *Measurement* 173, p. 108675.

Ariyo, A. A., Adewumi, A. O., and Ayo, C. K. 2014. “Stock price prediction using the ARIMA model”. In: *2014 UKSim-AMSS 16th international conference on computer modelling and simulation*. IEEE, pp. 106–112.

“AsiaOne” Aug. 2017. In: *AsiaOne*. URL: https://www.asiaone.com/asia/bangladeshi-law-student-killed-after-five-men-gang-raped-her-bus.

Azman, F., Suraya, Q., Rahim, F. A., Mohd, M. S., and Ariffin, N. A. M. 2018. “My guardian: A personal safety mobile application”. In: *2018 IEEE Conference on Open Systems (ICOS)*. IEEE, pp. 37–41.

Barhoumi, O., Zaki, M. H., and Tahar, S. 2024. “A Formal Approach to Road Safety Assessment Using Traffic Conflict Techniques”. In: *IEEE Open Journal of Vehicular Technology* 5, pp. 606–619. DOI: 10.1109/OJVT.2024.3387414.

Bayad, K., Rziza, M., and Oumsis, M. 2016. “Information Security Risk Analysis of Vehicular Ad Hoc Networks”. In: *International Conference on Mobile Networks and Management*. Springer, pp. 192–205.

BBC News Dec. 2012. URL: http://www.bbc.co.uk/news/world-asia-india-20765320.

BBC News Dec. 2018. URL: https://www.bbc.co.uk/news/uk-england-oxfordshire-46554714.

Beresford, A. R. and Stajano, F. 2004. “Mix zones: User privacy in location-aware services”. In: *IEEE Annual conference on pervasive computing and communications workshops, 2004. Proceedings of the Second*. IEEE, pp. 127–131.

Bettini, C., Mascetti, S., and Wang, X. S. 2008. “Privacy protection through anonymity in location-based services”. In: *Handbook of Database Security*. Springer, pp. 509–530.

Bhadula, G., Benjamin, A., and Kakkar, P. 2021. “Stree Aatmanirbharta Jacket–An IOT based Women Safety System”. In: *2021 Fourth International Conference on Computational Intelligence and Communication Technologies (CCICT)*. IEEE, pp. 350–354.

Bhavale, M. D. M., Bhawale, M. P. S., Sasane, M. T., and Bhawale, M. A. S. 2016. "IoT based unified approach for women and children security using wireless and GPS". In: *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume* 5.

Boucetta, S. I., Guichi, Y., and Johanyák, Z. C. 2021. "Review of Mobility Scenarios Generators for Vehicular Ad-Hoc Networks Simulators". In: *Journal of Physics: Conference Series*. Vol. 1935. 1. IOP Publishing, p. 012006.

Buttyán, L., Holczer, T., Weimerskirch, A., and Whyte, W. 2009. "Slow: A practical pseudonym changing scheme for location privacy in vanets". In: *2009 IEEE vehicular networking conference (VNC)*. IEEE, pp. 1–8.

Chan, N. and Lars, H 2003. "Introduction to location-based services". In: *Lund University GIS Centre*, pp. 1–12.

Chand, D., Nayak, S., Bhat, K. S., Parikh, S., Singh, Y., and Kamath, A. A. 2015. "A mobile application for Women's Safety: WoSApp". In: *TENCON 2015-2015 IEEE Region 10 Conference*. IEEE, pp. 1–5.

Chaudhari, P., Kamte, R., Kunder, K., Jose, A., and Machado, S. 2018. "'Street smart': Safe street app for women using augmented reality". In: *2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)*. IEEE, pp. 1–6.

Chaum, D. and Van Heyst, E. 1991. "Group signatures". In: *Workshop on the Theory and Application of of Cryptographic Techniques*. Springer, pp. 257–265.

Chen, P., Yuan, H., and Shu, X. 2008. "Forecasting crime using the arima model". In: *2008 fifth international conference on fuzzy systems and knowledge discovery*. Vol. 5. IEEE, pp. 627–630.

Chen, R., Fung, B. C., Desai, B. C., and Sossou, N. M. 2012. "Differentially private transit data publication: a case study on the montreal transportation system". In: *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 213–221.

Cheng, W., Wen, R., Huang, H., Miao, W., and Wang, C. 2022. "OPTDP: Towards optimal personalized trajectory differential privacy for trajectory data publishing". In: *Neurocomputing* 472, pp. 201–211.

Chicco, D., Warrens, M. J., and Jurman, G. 2021. "The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation". In: *PeerJ Computer Science* 7, e623.

Choudhary, Y., Upadhyay, S., Jain, R., and Chakraborty, A. May 2017. "WOMEN SAFETY DEVICE (SAFETY USING GPS, GSM, SHOCK, SIREN AND LED)". In: *nternational Journal of Advance Research in Science and Engineering* 6 5, 413–421.

Dandekar, A., Basu, D., and Bressan, S. 2021. "Differential privacy at risk: Bridging randomness and privacy budget". In: *Proceedings on Privacy Enhancing Technologies* 2021.1, pp. 64–84.

Das, S. and Sadhukhan, P. 2014. "Performance evaluation of a LBS system delivering location-based services using wireless local area network". In: *2014 Applications and Innovations in Mobile Computing (AIMoC)*. IEEE, pp. 85–90.

Dhingra, S., Mujumdar, P., and Gajjar, R. H. 1993. "Application of time series techniques for forecasting truck traffic attracted by the Bombay metropolitan region". In: *Journal of advanced transportation* 27.3, pp. 227–249.

Ding, Q. Y., Wang, X. F., Zhang, X. Y., and Sun, Z. Q. 2011. "Forecasting traffic volume with space-time ARIMA model". In: *Advanced Materials Research*. Vol. 156. Trans Tech Publ, pp. 979–983.

Dwivedi, I. 2017. "Privacy-Preserving Trajectory Data Publishing via Differential Privacy". In.

Dwork, C. 2008. "Differential privacy: A survey of results". In: *Theory and Applications of Models of Computation: 5th International Conference, TAMC 2008, Xi'an, China, April 25-29, 2008. Proceedings 5*. Springer, pp. 1–19.

Dwork, C., Kohli, N., and Mulligan, D. 2019. "Differential privacy in practice: Expose your epsilons!" In: *Journal of Privacy and Confidentiality* 9.2.

Dwork, C., McSherry, F., Nissim, K., and Smith, A. 2016. "Calibrating noise to sensitivity in private data analysis". In: *Journal of Privacy and Confidentiality* 7.3, pp. 17–51.

Dwork, C. and Roth, A. 2014. "The algorithmic foundations of differential privacy." In: *Found. Trends Theor. Comput. Sci.* 9.3-4, pp. 211–407.

Eom, C. S.-H., Lee, W., and Leung, C. K.-S. 2018. "STDP: secure privacy-preserving trajectory data publishing". In: *2018 IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData)*, pp. 892–899.

Erdogdu, M. A. and Fawaz, N. 2015. "Privacy-utility trade-off under continual observation". In: *2015 IEEE International Symposium on Information Theory (ISIT)*. IEEE, pp. 1801–1805.

Etter, V., Kafsi, M., and Kazemi, E. 2012. "Been there, done that: What your mobility traces reveal about your behavior". In: *Mobile data challenge by nokia workshop, in conjunction with int. conf. on pervasive computing*. CONF.

Fan, W., He, J., Guo, M., Li, P., Han, Z., and Wang, R. 2020. "Privacy preserving classification on local differential privacy in data centers". In: *Journal of Parallel and Distributed Computing* 135, pp. 70–82.

Feng, H., Liu, C., Shu, Y., and Yang, O. W. 2015. "Location prediction of vehicles in VANETs using a Kalman filter". In: *Wireless personal communications* 80.2, pp. 543–559.

Ferrag, M. A., Derhab, A., Maglaras, L., Mukherjee, M., and Janicke, H. 2018. "Privacy-preserving schemes for fog-based IoT applications: Threat models, solutions, and challenges". In: *2018 International Conference on Smart Communications in Network Technologies (SaCoNeT)*. IEEE, pp. 37–42.

Fiore, M., Harri, J., Filali, F., and Bonnet, C. 2007. "Vehicular mobility simulation for VANETs". In: *40th Annual Simulation Symposium (ANSS'07)*. IEEE, pp. 301–309.

Gadhave, S. N., Kale, S. D., Shinde, S. N., and Bhosale, A. C. 2017. "Electronic jacket for women safety". In: *IRJET*.

Galdames, P., Gutierrez-Soto, C., and Curiel, A. 2019. "Batching location cloaking techniques for location privacy and safety protection". In: *Mobile Information Systems* 2019.

Gao, H., Tang, J., and Liu, H. 2012. "Mobile location prediction in spatio-temporal context". In: 41.2, pp. 1–4.

Gedik, B. and Liu, L. 2007. "Protecting location privacy with personalized k-anonymity: Architecture and algorithms". In: *IEEE Transactions on Mobile Computing* 7.1, pp. 1–18.

Ghaleb, F. A., Razzaque, M., and Isnin, I. F. 2013. "Security and privacy enhancement in VANETs using mobility pattern". In: *2013 Fifth International Conference on Ubiquitous and Future Networks (ICUFN)*. IEEE, pp. 184–189.

Gharaibeh, A., Salahuddin, M. A., Hussini, S. J., Khreishah, A., Khalil, I., Guizani, M., and Al-Fuqaha, A. 2017. "Smart cities: A survey on data management, security, and enabling technologies". In: *IEEE Communications Surveys & Tutorials* 19.4, pp. 2456–2501.

Ghosh, B., Basu, B., and O'Mahony, M. 2005. "Time-series modelling for forecasting vehicular traffic flow in Dublin". In: *84th Annual Meeting of the Transportation Research Board, Washington, DC*.

Goli, S. A., Far, B. H., and Fapojuwo, A. O. 2018. "Vehicle Trajectory Prediction with Gaussian Process Regression in Connected Vehicle Environment \star". In: *2018 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, pp. 550–555.

Gupta, K. K., Kalita, K., Ghadai, R. K., Ramachandran, M., and Gao, X.-Z. 2021. "Machine Learning-Based Predictive Modelling of Biodiesel Production—A Comparative Perspective". In: *Energies* 14.4, p. 1122.

Gupta, P. and Sutar, S. 2014. "Study of various location tracking techniques for centralized location, monitoring & control system". In: *IOSR Journal of Engineering* 4.03, pp. 27–30.

Gursoy, M. E., Liu, L., Truex, S., and Yu, L. 2018. "Differentially private and utility preserving publication of trajectory data". In: *IEEE Transactions on Mobile Computing* 18.10, pp. 2315–2329.

Hao, M., Wu, W., and Wan, Y. 2023. "Hierarchical Aggregation for Numerical Data under Local Differential Privacy". In: *Sensors* 23.3, p. 1115.

Hara, T., Suzuki, A., Iwata, M., Arase, Y., and Xie, X. 2016. "Dummy-based user location anonymization under real-world constraints". In: *Ieee Access* 4, pp. 673–687.

Harder, F., Bauer, M., and Park, M. 2020. "Interpretable and differentially private predictions". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 04, pp. 4083–4090.

Hariharan, K., Jain, R. R., Prasad, A., Sharma, M., Yadav, P., Poorna, S., and Anuraj, K 2021. "A Comprehensive Study Toward Women Safety Using Machine Learning Along with Android App Development". In: *Sustainable Communication Networks and Application*. Springer, pp. 321–330.

Harikiran, G., Menasinkai, K., and Shirol, S. 2016. "Smart security solution for women based on Internet Of Things (IOT)". In: *2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)*. IEEE, pp. 3551–3554.

Härri, J., Filali, F., Bonnet, C., and Fiore, M. 2006. "VanetMobiSim: generating realistic mobility patterns for VANETs". In: *Proceedings of the 3rd international workshop on Vehicular ad hoc networks*, pp. 96–97.

Hayat, A., Iftikhar, Z., Khan, M. I., Mehbodniya, A., Webber, J. L., and Hanif, S. 2023. "A Novel Pseudonym Changing Scheme for Location Privacy Preservation in Sparse Traffic Areas". In: *IEEE Access* 11, pp. 89974–89985. DOI: 10.1109/ACCESS.2023.3303846.

Helbach, J., Hoffmann, F., Pieper, D., and Allers, K. 2023. "Reporting according to the preferred reporting items for systematic reviews and meta-analyses for abstracts (PRISMA-A) depends on abstract length". In: *Journal of Clinical Epidemiology* 154, pp. 167–177.

Hodson, T. O., Over, T. M., and Foks, S. S. 2021. "Mean squared error, deconstructed". In: *Journal of Advances in Modeling Earth Systems* 13.12, e2021MS002681.

Holohan, N., Antonatos, S., Braghin, S., and Mac Aonghusa, P. 2020. "The Bounded Laplace Mechanism in Differential Privacy". In: *Journal of Privacy and Confidentiality* 10, p. 1.

Hosseini, S., Murgovski, N., Campos, G. R. de, and Sjöberg, J. 2016. "Adaptive forward collision warning algorithm for automotive applications". In: *2016 American Control Conference (ACC)*, pp. 5982–5987. DOI: 10.1109/ACC.2016.7526608.

Hou, L., Yao, N., Lu, Z., Zhan, F., and Liu, Z. 2021. "Tracking Based Mix-Zone Location Privacy Evaluation in VANET". In: *IEEE Transactions on Vehicular Technology* 70.10, pp. 10957–10969. DOI: 10.1109/TVT.2021.3109065.

Houmer, M. and Hasnaoui, M. L. 2020. "A risk and security assessment of VANET availability using attack tree concept." In: *International Journal of Electrical & Computer Engineering (2088-8708)* 10.6.

Hua, J., Gao, Y., and Zhong, S. 2015. "Differentially private publication of general time-serial trajectory data". In: *2015 IEEE Conference on Computer Communications (INFOCOM)*. IEEE, pp. 549–557.

Ingdal, M., Johnsen, R., and Harrington, D. A. 2019. "The Akaike information criterion in weighted regression of immittance data". In: *Electrochimica Acta* 317, pp. 648–653.

Islam, K. and Raza, A. 2020. "Forecasting crime using ARIMA model". In: *arXiv e-prints*, arXiv–2003.

Jang, M., Yoon, M., Kim, H.-i., and Chang, J.-W. 2012. "A privacy-aware location cloaking technique reducing bandwidth consumption in location-based services". In: *Proceedings of the Third ACM SIGSPATIAL International Workshop on Querying and Mining Uncertain Spatio-Temporal Data*, pp. 2–9.

Jiang, B., Li, J., Yue, G., and Song, H. 2021a. "Differential privacy for industrial internet of things: Opportunities, applications, and challenges". In: *IEEE Internet of Things Journal* 8.13, pp. 10430–10451.

Jiang, H., Wang, M., Zhao, P., Xiao, Z., and Dustdar, S. 2021b. "A utility-aware general framework with quantifiable privacy preservation for destination prediction in LBSs". In: *IEEE/ACM Transactions on Networking* 29.5, pp. 2228–2241.

Johnson, A. and Shmatikov, V. 2013. "Privacy-preserving data exploration in genome-wide association studies". In: *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 1079–1087.

Julien, F, Raya, M, Felegyhazi, M, and Papadimitratos, P 2007. "Mixzones for location privacy in vehicular networks". In: *Association for Computing Machinery (ACM) Workshop on Wireless Networking for Intelligent Transportation Systems (WiN-ITS)*.

Kaisser, F., Gransart, C., and Berbineau, M. 2012. "Simulations of vanet scenarios with opnet and sumo". In: *International Workshop on Communication Technologies for Vehicles*. Springer, pp. 103–112.

Kalaiarasy, C, Sreenath, N, and Amuthan, A 2019. "Location Privacy Preservation in VANET using Mix Zones–A survey". In: *2019 International Conference on Computer Communication and Informatics (ICCCI)*. IEEE, pp. 1–5.

Kalantari, K., Sankar, L., and Sarwate, A. D. 2018. "Robust privacy-utility tradeoffs under differential privacy and hamming distortion". In: *IEEE Transactions on Information Forensics and Security* 13.11, pp. 2816–2830.

Kanagaraj, S. A., Arjun, G, and Shahina, A 2013. "Cheeka: A mobile application for personal safety". In: *9th IEEE International Conference On Collaborative Computing: Networking, Applications and Worksharing*. IEEE, pp. 289–294.

Kang, H and Meng, W 2012. "Protecting location privacy with personalized k-anonymity". In: *Journal of Nanjing University of Posts and Telecommunications (Natural Science)* 6, p. 014.

Kartik, P., Jose, S., and MK, G. K. 2017. "Safetipin: A Mobile Application Towards Women Safety". In: *Rajagiri Journal of Social Development* 9.1, pp. 5–12.

Kasori, K. and Sato, F. 2015. "Location Privacy Protection Considering the Location Safety". In: *2015 18th International Conference on Network-Based Information Systems*, pp. 140–145. DOI: 10.1109/NBiS.2015.24.

Khacheba, I., Yagoubi, M. B., Lagraa, N., and Lakas, A. 2017. "Location privacy scheme for VANETs". In: *2017 International Conference on Selected Topics in Mobile and Wireless Networking (MoWNeT)*. IEEE, pp. 1–6.

Khan, S., Ahmad, W., Ali, R., and Saleem, S. 2015. "A Research on Mobile Applications for Location Tracking through Web Server and Short Messages Services (SMS)". In: *VFAST Transactions on Software Engineering* 7.2, pp. 12–17.

Khuri, A. I. 2009. *Linear model methodology*.

Kido, H., Yanagisawa, Y., and Satoh, T. 2005. "An anonymous communication technique using dummies for location-based services". In: *ICPS'05. Proceedings. International Conference on Pervasive Services, 2005.* IEEE, pp. 88–97.

Kolvoord, R., Keranen, K., and Rittenhouse, P. 2017. "Applications of location-based services and mobile technologies in K-12 classrooms". In: *ISPRS International Journal of Geo-Information* 6.7, p. 209.

Koufogiannis, F., Han, S., and Pappas, G. J. 2015. "Optimality of the laplace mechanism in differential privacy". In.

Kumar, M. and Anand, M. 2014. "An application of time series ARIMA forecasting model for predicting sugarcane production in India". In: *Studies in Business and Economics* 9 1, pp. 81–94.

Kumar, S. V. and Vanajakshi, L. 2015. "Short-term traffic flow prediction using seasonal ARIMA model with limited input data". In: *European Transport Research Review* 7.3, pp. 1–9.

Lai, Y.-C., Lin, J.-W., Yeh, Y.-H., Lai, C.-N., and Weng, H.-C. 2013. "A tracking system using location prediction and dynamic threshold for minimizing SMS delivery". In: *Journal of Communications and Networks* 15 1, pp. 54–60.

Lai, Y. and Dzombak, D. A. 2020. "Use of the autoregressive integrated moving average (ARIMA) model to forecast near-term regional temperature and precipitation". In: *Weather and Forecasting* 35.3, pp. 959–976.

Lecuyer, M., Atlidakis, V., Geambasu, R., Hsu, D., and Jana, S. 2019. "Certified robustness to adversarial examples with differential privacy". In: *2019 IEEE Symposium on Security and Privacy (SP)*. IEEE, pp. 656–672.

Lewis, K. 2016. *Mother gang-raped on bus as two-week old baby dies in attack*. Publication Title: The Independent. Independent Digital News and Media. URL: https://www.independent.co.uk/news/world/asia/india-gang-rape-women-violence-bus-bareilly-daughter-baby-killed-a6925371.html.

Li, B., Liang, R., Zhu, D., Chen, W., and Lin, Q. 2020. "Blockchain-based trust management model for location privacy preserving in VANET". In: *IEEE Transactions on Intelligent Transportation Systems* 22.6, pp. 3765–3775.

Li, Q., Wu, H., Wu, X., and Dong, L. 2019. "Multi-Level Location Privacy Protection Based on Differential Privacy Strategy in VANETs". In: *2019 IEEE 89th Vehicular Technology Conference (VTC2019-Spring)*. IEEE, pp. 1–5.

Liao, J. and Li, J. 2009. "Effectively changing pseudonyms for privacy protection in vanets". In: *2009 10th International Symposium on Pervasive Systems, Algorithms, and Networks*. IEEE, pp. 648–652.

Lim, K. G., Lee, C. H., Chin, R. K. Y., Yeo, K. B., and Teo, K. T. K. 2017. "SUMO enhancement for vehicular ad hoc network (VANET) simulation". In: *2017 IEEE 2nd international conference on automatic control and intelligent systems (I2CACIS)*. IEEE, pp. 86–91.

Lin, J. 2016. "Study on the Prediction of Urban Traffic Flow Based on ARIMA Model". In: *DEStech Transactions on Engineering and Technology Research* ICETA.

Lin, X. 2017. "Vehicular Networking". In: *IEEE Communications Standards Magazine* 1.2, pp. 68–68. DOI: 10.1109/MCOMSTD.2017.7992932.

Lippi, M., Bertini, M., and Frasconi, P. 2013. "Short-term traffic flow forecasting: An experimental comparison of time-series analysis and supervised learning". In: *IEEE Transactions on Intelligent Transportation Systems* 14.2, pp. 871–882.

Liu, B., Zhou, W., Zhu, T., Gao, L., and Xiang, Y. 2018. "Location privacy and its applications: A systematic study". In: *IEEE access* 6, pp. 17606–17624.

Liu, C., Hoi, S. C., Zhao, P., and Sun, J. 2016. "Online arima algorithms for time series prediction". In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 30. 1.

Liu, X. and Liu, H. 2021. "Data Publication Based On Differential Privacy In V2G Network". In: *Int. J. of Electronics Engineering and Applications* 9.2, pp. 34–44.

Longnecker, E. 2019. *12-year-old calls 911 during frightening ride in stolen car*. Publication Title: 13 WTHR Indianapolis. URL: https://www.wthr.com/article/12-year-old-calls-911-during-frightening-ride-stolen-car.

Loukides, G. and Shao, J. 2008. "Data utility and privacy protection trade-off in k-anonymisation". In: *Proceedings of the 2008 international workshop on Privacy and anonymity in information society*, pp. 36–45.

Lu, Z., Zhu, Y., Zheng, V. W., and Yang, Q. 2012. "Next place prediction by learning with multiple models". In: *Proceedings of the Mobile Data Challenge Workshop*.

Machanavajjhala, A., Kifer, D., Gehrke, J., and Venkitasubramaniam, M. 2007. "l-diversity: Privacy beyond k-anonymity". In: *ACM Transactions on Knowledge Discovery from Data (TKDD)* 1.1, 3–es.

Madi, S. and Al-Qamzi, H. 2013. "A survey on realistic mobility models for vehicular ad hoc networks (VANETs)". In: *2013 10th IEEE International Conference on Networking, Sensing and Control (ICNSC)*. IEEE, pp. 333–339.

Maharaj, K. and Hosein, P. 2016. "Location obfuscation using smart meter readings". In: *2016 International Conference on Advances in Computing and Communication Engineering (ICACCE)*, pp. 449–453. DOI: 10.1109/ICACCE.2016.8073790.

Manazir, S. H., Govind, M., and Rubina 2019. "My Safetipin Mobile Phone Application: Case Study of E-Participation Platform for Women Safety in India." In: *J. Sci. Res.* 8.1, pp. 47–53.

Martinez, C. M., Hu, X., Cao, D., Velenis, E., Gao, B., and Wellers, M. 2016. "Energy management in plug-in hybrid electric vehicles: Recent progress and a connected vehicles perspective". In: *IEEE Transactions on Vehicular Technology* 66.6, pp. 4534–4549.

Masoumzadeh, A. and Joshi, J. 2011. "An alternative approach to k-anonymity for location-based services". In: *Procedia Computer Science* 5, pp. 522–530.

MathWorks n.d.(a). URL: https://www.mathworks.com/help/stats/ensemble-algorithms.html.

MathWorks n.d.(b). URL: https://www.mathworks.com/help/stats/regression-tree-ensembles.html.

Maulud, D. and Abdulazeez, A. M. 2020. "A review on linear regression comprehensive in machine learning". In: *Journal of Applied Science and Technology Trends* 1.4, pp. 140–147.

McSorley, A. 2018. *Who is Jastine Valdez? Wicklow woman abducted by Dublin dad-of-two Mark Hennessy*. Publication Title: dublinlive. URL: https://www.dublinlive.ie/news/dublin-news/jastine-valdez-mark-hennessy-14690587.

Miriyala, G. P., Sunil, P., Yadlapalli, R. S., Pasam, V. R. L., Kondapalli, T., and Miriyala, A. 2016. "Smart intelligent security system for women". In: *International Journal of Electronics and Communication Engineering & Technology (IJECET)* 7.2, pp. 41–46.

Mohamed, J. 2020. "Time series modeling and forecasting of Somaliland consumer price index: a comparison of ARIMA and regression with ARIMA errors". In: *American Journal of Theoretical and Applied Statistics* 9.4, pp. 143–53.

Mohamed, T. M., Ahmed, I. Z., and Sadek, R. A. 2021. "Efficient VANET safety message delivery and authenticity with privacy preservation". In: *PeerJ Computer Science* 7, e519.

Mohr, D. C., Zhang, M., and Schueller, S. M. 2017. "Personal sensing: understanding mental health using ubiquitous sensors and machine learning". In: *Annual review of clinical psychology* 13, pp. 23–47.

Monisha, D. G., Monisha, M, Pavithra, G, and Subhashini, R 2016. "Women safety device and application-FEMME". In: *Indian Journal of Science and Technology* 9 10.

Muralidhar, K and Bharathi, N. P. n.d. "A "Simple-to-Use" Personal Safety System through Smart phones". In:

Narayanan, A. and Shmatikov, V. 2019. "Robust de-anonymization of large sparse datasets: a decade later". In: *May* 21, p. 2019.

Naumov, V. and Gross, T. R. 2007. "Connectivity-aware routing (CAR) in vehicular ad-hoc networks". In: *IEEE INFOCOM 2007-26th IEEE International Conference on Computer Communications*. IEEE, pp. 1919–1927.

Navidi, W., Camp, T., and Bauer, N. 2004. "Improving the accuracy of random waypoint simulations through steady-state initialization". In: *Proceedings of the 15th International Conference on Modeling and Simulation*, pp. 319–326.

Neera, J., Chen, X., Aslam, N., Wang, K., and Shu, Z. 2021. "Private and utility enhanced recommendations with local differential privacy and gaussian mixture model". In: *IEEE Transactions on Knowledge and Data Engineering*.

Nekovee, M. and Bogason, B. B. 2007. "Reliable and effcient information dissemination in intermittently connected vehicular adhoc networks". In: *2007 IEEE 65th Vehicular Technology Conference-VTC2007-Spring*. IEEE, pp. 2486–2490.

Nisha, N., Natgunanathan, I., and Xiang, Y. 2022. "An Enhanced Location Scattering Based Privacy Protection Scheme". In: *IEEE Access* 10, pp. 21250–21263. DOI: 10.1109/ACCESS.2022.3152770.

Pan, Y. and Li, J. 2013. "Cooperative pseudonym change scheme based on the number of neighbors in VANETs". In: *Journal of Network and Computer Applications* 36.6, pp. 1599–1609.

Panneerselvam, J., Liu, L., and Antonopoulos, N. 2018. "InOt-RePCoN: Forecasting user behavioural trend in large-scale cloud environments". In: *Future Generation Computer Systems* 80, pp. 322–341.

Papadimitratos, P., Poturalski, M., Schaller, P., Lafourcade, P., Basin, D., Capkun, S., and Hubaux, J.-P. 2008. "Secure neighborhood discovery: a fundamental element for mobile ad hoc networking". In: *IEEE Communications Magazine* 46.2, pp. 132–139.

Parums, D. V. 2021. "Review articles, systematic reviews, meta-analysis, and the updated preferred reporting items for systematic reviews and meta-analyses (PRISMA) 2020 guidelines". In: *Medical science monitor: international medical journal of experimental and clinical research* 27, e934475–1.

Patel, J. and Hasan, R. 2018. "Smart bracelets: Towards automating personal safety using wearable smart jewelry". In: *2018 15th IEEE Annual Consumer Communications & Networking Conference (CCNC)*. IEEE, pp. 1–2.

Pawar, R, Kulabkar, M, Pawar, K, Tambe, A, and Khairnar, P. S. 2018. "Smart Shield for Women Safety". In: *International Research Journal of Engineering and Technology (IRJET) e-ISSN* 5.4, pp. 56–2395.

Petropoulos, F., Apiletti, D., Assimakopoulos, V., Babai, M. Z., Barrow, D. K., Taieb, S. B., Bergmeir, C., Bessa, R. J., Bijak, J., and Boylan, J. E. 2022. "Forecasting: theory and practice". In: *International Journal of Forecasting*.

Pontikakos, C, Sambrakos, M, Glezakos, T, and Tsiligiridis, T 2006. "Location-based services: A framework for an architecture design". In: *Neural Parallel and Scientific Computatiions* 14.2/3, p. 273.

Prashanth, D. S., Patel, G., and Bharathi, B 2017. "Research and development of a mobile based women safety application with real-time database and data-stream network". In: *2017 International Conference on Circuit, Power and Computing Technologies (ICCPCT)*. IEEE, pp. 1–5.

Priya, M. K., Venkateswaran, M. N. P., Narayanan, M. M. S., Brabakarraj, M. G., and Srinath, M. J. 2021. "Electronic jacket for women safety". In: *INTERNATIONAL JOURNAL* 6.6.

Qureshi, I. 2015. *India woman raped in moving bus in Karnataka*. Publication Title: BBC News. BBC. URL: http://www.bbc.co.uk/news/world-asia-india-34742833.

Rajput, U., Abbas, F., Eun, H., and Oh, H. 2017. "A hybrid approach for efficient privacy-preserving authentication in VANET". In: *IEEE Access* 5, pp. 12014–12030.

Rassouli, B. and Gündüz, D. 2019. “Optimal utility-privacy trade-off with total variation distance as a privacy measure”. In: *IEEE Transactions on Information Forensics and Security* 15, pp. 594–603.

Reddy, M. V., SnehithRaju, B, Nurja, S., and Reddy, M. S. K. 2021. “GPS BASED FEMININE RESCUE SYSTEM”. In: 8 14.

Ren, D., Du, S., and Zhu, H. 2011. “A novel attack tree based risk assessment approach for location privacy preservation in the VANETs”. In: *2011 IEEE International Conference on Communications (ICC)*. IEEE, pp. 1–5.

Ren, Y., Li, X., Miao, Y., Deng, R. H., Weng, J., Ma, S., and Ma, J. 2023. “DistPreserv: Maintaining User Distribution for Privacy-Preserving Location-Based Services”. In: *IEEE Transactions on Mobile Computing* 22.6, pp. 3287–3302. DOI: 10.1109/TMC.2022.3141398.

Rengaraj, V. and Bijlani, K. 2016. “A study and implementation of smart ID card with M-learning and child security”. In: *2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT)*. IEEE, pp. 305–311.

Rohilla, S., Deshwal, A., and Balasubramanian, L. 2019. “Intelligent Vehicular Networks”. In: *2019 International Conference on Advances in Computing and Communication Engineering (ICACCE)*, pp. 1–4. DOI: 10.1109/ICACCE46606.2019.9079957.

Saleem, Y., Rehmani, M. H., Crespi, N., and Minerva, R. 2021. “Parking recommender system privacy preservation through anonymization and differential privacy”. In: *Engineering Reports* 3.2, e12297.

Saranya, K, Nandhini, S, Adish, C., and Manikandan, A 2021. “E-DEFENCE WOMEN SAFETY APPLICATION”. In: *International Journal of Advanced Engineering Science and Information Technology* 4.4.

Sarathy, R. and Muralidhar, K. 2010. “Some additional insights on applying differential privacy for numeric data”. In: *Privacy in Statistical Databases: UNESCO Chair in Data Privacy, International Conference, PSD 2010, Corfu, Greece, September 22-24, 2010. Proceedings*. Springer, pp. 210–219.

Sarwate, A. D. and Chaudhuri, K. 2013. “Signal processing and machine learning with differential privacy: Algorithms and challenges for continuous data”. In: *IEEE signal processing magazine* 30.5, pp. 86–94.

Schwarting, W., Alonso-Mora, J., and Rus, D. 2018. “Planning and decision-making for autonomous vehicles”. In: *Annual Review of Control, Robotics, and Autonomous Systems* 1, pp. 187–210.

Shaikh, M. A. M. and PB, M. D. 2008. “Women’s Safety Jacket”. In: *International Journal on Integrated Education* 1.1, pp. 77–81.

Sharma, S., Ayaz, F., Sharma, R., Jain, D., and Student, B. 2017. "IoT based women safety device using ARM7". In: *IJESC* 7.5, pp. 11465–11466.

Shin, K. G., Ju, X., Chen, Z., and Hu, X. 2012. "Privacy protection for users of location-based services". In: *IEEE Wireless Communications* 19.1, pp. 30–39.

Shinde, P., Taware, P., Thorat, S., Waghmare, T., and Kadam, A 2012. "Emergency Panic Button". In: *International Journal of Scientific & Engineering Research* 3.3.

Shokri, R., Theodorakopoulos, G., Le Boudec, J.-Y., and Hubaux, J.-P. 2011. "Quantifying location privacy". In: *2011 IEEE symposium on security and privacy*. IEEE, pp. 247–262.

Shokri, R., Troncoso, C., Diaz, C., Freudiger, J., and Hubaux, J.-P. 2010. "Unraveling an old cloak: k-anonymity for location privacy". In: *Proceedings of the 9th annual ACM workshop on Privacy in the electronic society*, pp. 115–118.

Siegel, J. E., Erb, D. C., and Sarma, S. E. 2017. "A survey of the connected vehicle landscape—Architectures, enabling technologies, applications, and development areas". In: *IEEE Transactions on Intelligent Transportation Systems* 19.8, pp. 2391–2406.

Sogi, N. R., Chatterjee, P., Nethra, U, and Suma, V 2018. "SMARISA: a raspberry pi based smart ring for women safety using IoT". In: *2018 International Conference on Inventive Research in Computing Applications (ICIRCA)*. IEEE, pp. 451–454.

Song, D., Tharmarasa, R., Kirubarajan, T., and Fernando, X. N. 2017. "Multi-vehicle tracking with road maps and car-following models". In: *IEEE transactions on intelligent transportation systems* 19.5, pp. 1375–1386.

Song, Z., Guo, Y., Wu, Y., and Ma, J. 2019. "Short-term traffic speed prediction under different data collection time intervals using a SARIMA-SDGM hybrid prediction model". In: *PloS one* 14.6, e0218626.

Srikrishna, G. and Veena, G. 2017. "Uncovering the nearest neighbours for personal safety". In: *2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS)*. IEEE, pp. 3469–3473.

Sun, J. and Kim, J. 2021. "Joint prediction of next location and travel time from urban vehicle trajectories using long short-term memory neural networks". In: *Transportation Research Part C: Emerging Technologies* 128, p. 103114.

Sweeney, L. 2002. "k-anonymity: A model for protecting privacy". In: *International journal of uncertainty, fuzziness and knowledge-based systems* 10.05, pp. 557–570.

Tayal, S. and Triphathi, M. R. 2012. "VANET-challenges in selection of vehicular mobility model". In: *2012 Second International Conference on Advanced Computing & Communication Technologies*. IEEE, pp. 231–235.

Thakurta, A. G. and Smith, A. 2013. “Differentially private feature selection via stability arguments, and the robustness of the lasso”. In: *Conference on Learning Theory*. PMLR, pp. 819–850.

Thavil, J., Durdhawale, V., and Elake, P. 2017. “Study on smart security technology for women based on IoT”. In: *International Research Journal of Engineering and Technology (IRJET)* 4.02.

Thiruchelvam, L., Dass, S. C., Asirvadam, V. S., Daud, H., and Gill, B. S. 2021. “Determine neighboring region spatial effect on dengue cases using ensemble ARIMA models”. In: *Scientific Reports* 11.1, pp. 1–9.

Tian, J., Zhang, H., Treiber, M., Jiang, R., Gao, Z.-Y., and Jia, B. 2019. “On the role of speed adaptation and spacing indifference in traffic instability: Evidence from car-following experiments and its stochastic model”. In: *Transportation research part B: methodological* 129, pp. 334–350.

Tong, M. and Xue, H. 2008. “Highway traffic volume forecasting based on seasonal ARIMA model”. In: *Journal of Highway and Transportation Research and Development (English Edition)* 3.2, pp. 109–112.

Tyagi, A. K. and Sreenath, N 2015. “Location privacy preserving techniques for location based services over road networks”. In: *2015 International Conference on Communications and Signal Processing (ICCSP)*. IEEE, pp. 1319–1326.

Utehs, K. 2018. *ABC7 NEWS EXCLUSIVE: 5-year-old girl kidnapped during San Jose car theft, reunited with parents*. Publication Title: ABC7 San Francisco. URL: https://abc7news.com/abc7-news-exclusive-5-year-old-girl-kidnapped-during-san-jose-car-theft-reunited-with-parents/4527498/.

Viswanath, K. and Basu, A. 2015. “SafetiPin: an innovative mobile app to collect data on women’s safety in Indian cities”. In: *Gender & Development* 23.1, pp. 45–60.

Viswanath, N., Pakyala, N. V., and Muneeswari, G 2016. “Smart foot device for women safety”. In: *2016 IEEE region 10 symposium (TENSYMP)*. IEEE, pp. 130–134.

Walkunde, M. K. M., Shinde, M. B. G., and Pandhare, S. D. 2022. “An Android Application for Women Safety”. In: *Journal homepage: www. ijrpr. com ISSN* 3.6, pp. 3194–3199.

Wandtner, B., Schmidt, G., Schoemig, N., and Kunde, W. 2018. “Non-driving related tasks in highly automated driving - Effects of task modalities and cognitive workload on take-over performance”. In: *AmE 2018 - Automotive meets Electronics; 9th GMM-Symposium*, pp. 1–6.

Wang, C., Ma, L., Li, R., Durrani, T. S., and Zhang, H. 2019. “Exploring trajectory prediction through machine learning methods”. In: *IEEE Access* 7, pp. 101441–101452.

Wang, J. and Prabhala, B. 2012. "Periodicity based next place prediction". In: *Nokia Mobile Data Challenge 2012 Workshop. p. Dedicated task*. Vol. 2. 2.

Wang, Z.-H., Lu, C.-Y., Pu, B., Li, G.-W., and Guo, Z.-J. 2017. "Short-term forecast model of vehicles volume based on ARIMA seasonal model and holt-winters". In: *ITM Web of Conferences*. Vol. 12. EDP Sciences, p. 04028.

Wei, K., Li, J., Ding, M., Ma, C., Yang, H. H., Farokhi, F., Jin, S., Quek, T. Q., and Poor, H. V. 2020. "Federated learning with differential privacy: Algorithms and performance analysis". In: *IEEE Transactions on Information Forensics and Security* 15, pp. 3454–3469.

W.H.O 2018. *Global status report on road safety 2018: Summary*. Tech. rep. World Health Organization.

Williams, B. M. and Hoel, L. A. 2003. "Modeling and forecasting vehicular traffic flow as a seasonal ARIMA process: Theoretical basis and empirical results". In: *Journal of transportation engineering* 129.6, pp. 664–672.

Wu, R., Luo, G., Shao, J., Tian, L., and Peng, C. 2018. "Location prediction on trajectory data: A review". In: *Big data mining and analytics* 1.2, pp. 108–127.

Xin, Q., Yang, N., Fu, R., Yu, S., and Shi, Z. 2018. "Impacts analysis of car following models considering variable vehicular gap policies". In: *Physica A: Statistical Mechanics and its Applications* 501, pp. 338–355.

Yan, G., Yang, W., Weigle, M. C., Olariu, S., and Rawat, D. 2010. "Cooperative Collision Warning through mobility and probability prediction". In: *2010 IEEE Intelligent Vehicles Symposium*, pp. 1172–1177. DOI: 10.1109/IVS.2010.5547990.

Yan, L., Li, L., Mu, X., Wang, H., Chen, X., and Shin, H. 2023. "Differential Privacy Preservation for Location Semantics". In: *Sensors* 23.4, p. 2121.

Yang, M., Zhu, T., Liu, B., Xiang, Y., and Zhou, W. 2018. "Machine learning differential privacy with multifunctional aggregation in a fog computing architecture". In: *IEEE Access* 6, pp. 17119–17129.

Yang, Y. and Hua, K. 2019. "Emerging Technologies for 5G-Enabled Vehicular Networks". In: *IEEE Access* 7, pp. 181117–181141. DOI: 10.1109/ACCESS.2019.2954466.

Yao, X., Zhou, X., and Ma, J. 2016. "Differential privacy of big data: an overview". In: *2016 IEEE 2nd International Conference on Big Data Security on Cloud (BigDataSecurity), IEEE International Conference on High Performance and Smart Computing (HPSC), and IEEE International Conference on Intelligent Data and Security (IDS)*. IEEE, pp. 7–12.

Yarrabothu, R. S. and Thota, B. 2015. "Abhaya: An Android App for the safety of women". In: *2015 Annual IEEE India Conference (INDICON)*. IEEE, pp. 1–4.

Ye, Q., Szeto, W. Y., and Wong, S. C. 2012. "Short-term traffic speed forecasting based on data recorded at irregular intervals". In: *IEEE Transactions on Intelligent Transportation Systems* 13.4, pp. 1727–1737.

Ying, B., Makrakis, D., and Mouftah, H. T. 2013. "Dynamic mix-zone for location privacy in vehicular networks". In: *IEEE Communications Letters* 17.8, pp. 1524–1527.

Yuan, R., Wang, X., Xu, J., and Meng, S. 2021. "A Differential-Privacy-based hybrid collaborative recommendation method with factorization and regression". In: *2021 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech)*. IEEE, pp. 389–396.

Yuanhui, L., Wen, Z., Haiyun, H., Zhipeng, O., and Xia, L. 2022. "Comparison of ARIMA Model and GM (1, 1) Model in Passenger Flow Prediction of Sanya Airport". In: *2022 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC)*. IEEE, pp. 318–323.

Yue, X., Xu, J., Chen, B., and He, Y. 2019. "A practical group signatures for providing privacy-preserving authentication with revocation". In: *International Conference on Security and Privacy in New Computing Environments*. Springer, pp. 226–245.

Zhang, M., Zhou, J., Zhang, G., Cui, L., Gao, T., and Yu, S. 2022. "APDP: Attribute-Based Personalized Differential Privacy Data Publishing Scheme for Social Networks". In: *IEEE Transactions on Network Science and Engineering*.

Zhang, S., Choo, K.-K. R., Liu, Q., and Wang, G. 2018. "Enhancing privacy through uniform grid and caching in location-based services". In: *Future Generation Computer Systems* 86, pp. 881–892.

Zhang, S., Mao, X., Choo, K.-K. R., Peng, T., and Wang, G. 2020. "A trajectory privacy-preserving scheme based on a dual-K mechanism for continuous location-based services". In: *Information Sciences* 527, pp. 406–419.

Zhao, J., Mei, J., Matwin, S., Su, Y., and Yang, Y. 2020a. "Risk-aware individual trajectory data publishing with differential privacy". In: *IEEE Access* 9, pp. 7421–7438.

Zhao, X., Dong, Y., and Pi, D. 2019. "Novel trajectory data publishing method under differential privacy". In: *Expert Systems with Applications* 138, p. 112791.

Zhao, Z., Karimzadeh, M., Gerber, F., and Braun, T. 2020b. "Mobile crowd location prediction with hybrid features using ensemble learning". In: *Future Generation Computer Systems* 110, pp. 556–571.

Zhong, H., Ni, J., Cui, J., Zhang, J., and Liu, L. 2022. "Personalized Location Privacy Protection Based on Vehicle Movement Regularity in Vehicular Networks". In: *IEEE Systems Journal* 16.1, pp. 755–766. DOI: 10.1109/JSYST.2020.3047397.

Zuberi, R. S., Lall, B., and Ahmad, S. N. 2012. "Privacy protection through k. anonymity in location. based services". In: *IETE Technical Review* 29.3, pp. 196–201.

Appendix A

Saving Victims in Moving Vehicles: an IoT-based Prediction Model-aided Solution

Content removed due to copyright restrictions

Content removed due to copyright restrictions

Content removed due to copyright restrictions

Content removed due to copyright restrictions

Content removed due to copyright restrictions

Content removed due to copyright restrictions

Appendix B

Persation: an IoT-Based Personal Safety Prediction Model-Aided Solution

International Journal of Computing and Digital Systems
ISSN (2210-142X)
Int. J. Com. Dig. Sys. 9, No.6 (Nov-2020)

http://dx.doi.org/10.12785/ijcds/090602

Persation: an IoT Based Personal Safety Prediction Model Aided Solution

Olasunkanmi Matthew Alofe[1], Kaniz Fatema [2], Muhammad Ajmal Azad [3] and Fatih Kurugollu [4]

[1,3,4] Department of Electronics, Computing and Mathematics, University of Derby, Derby, UK
[2]Department of Computer Science, Aston University, Birmingham, UK

Received 30 May 2020, Revised 21 Oct. 2020, Accepted 24 Oct. 2020, Published 1 Nov. 2020

Abstract: The number of attacks on innocent victims in moving vehicles, and abduction of individuals in their vehicles has risen alarmingly in the past few years. One common scenario evident from the modus operandi of this kind of attack is the random motion of these vehicles, due to the driver's unpredictable behaviours. To save the victims in such kinds of assault, it is essential to offer help promptly. An effective strategy to save victims is to predict the future location of the vehicles so that the rescue mission can be actioned at the earliest possibility. We have done a comprehensive survey of the state-of-the-art personal safety solutions and location prediction technologies and proposes an Internet of Things (IoT) based personal safety model, encompassing a prediction framework to anticipate the future vehicle locations by exploiting complex analytics of current and past data variables including the speed, direction and geolocation of the vehicles. Experiments conducted based on real-world datasets demonstrate the feasibility of our proposed framework in accurately predicting future vehicle locations. In this paper, we have a risk assessment of our safety solution model based on the OCTAVE ALLEGRO model and the implementation of our prediction model.

Keywords: IoT, Mobile Application, Vehicle Location Identification, GPS, Location Prediction

1. INTRODUCTION

Assaulting females have been frequently witnessed in the past few years around the world, with most of such incidents causing serious consequences to the victims [1, 2, 3, 4]. In most incidents, abnormal behaviour of the drivers is a commonly witnessed pattern such as diverted routes than normal, and the vehicle has been on the move while the assaults were taking place. For example, in January 2018, an abducted student forcefully driven away in her car from a car park in Atlanta, Georgia was able to prevent a more sinister ending to her ordeal by sharing her location with a third party and exchanged messages before the attacker seized the phone [5]. Furthermore, a quick response from a rescue team in San Jose, California helped to reunite a 5-year-old, who was in her father's vehicle when snatched, with her family in October 2018 [6]. In December 2018, a victim managed to escape assault in Berkshire, after she was forced into the boot of her car [7]. In February 2019, a 12-year-old girl suffered the same fate while waiting for her mother in the parking lot of a mall in Indiana [8]. Despite saving the lives of some of the victims, some had a more sinister ending, or death on some occasions. One of the survivors was saved because she was able to share her location with a third party, but not before some level of damage was done. The damage done to these victims would have been prevented earlier if there were any means to track and predict the location of the moving vehicle and inform the responder immediately. These scenarios do not share identical factors. In some cases, the speed, bearing, and location of the vehicle change rapidly and continuously while some changes were consistent. For either scenario, the delay in administering help would be damaging. Therefore, the solution should cater to the following requirements:

The helping device should be easily accessible when the attack occurs instead of attempting to unlock her phone, which might be taken away by the attacker. Therefore, the solution should cater to the following requirements.

- The device has to be lightweight for easy mobility.
- The solution should provide a way to detect and predict the location of the victim or the moving vehicle.
- It needs to inform the third party to offer the quickest help.

Fulfilling these requirements would ensure prompt actions taken. The solution needs to provide an avenue to

E-mail: o.alofe1@unimail.derby.ac.uk, k.fatema@aston.ac.uk, m.azad@derby.ac.uk, f.kurugollu @derby.ac.uk

obtain location information from the user, predict location from the obtained information, inform the third party with the predicted location, and create an intersection path for help to be offered by the third party. The contributions of this paper are:

- Proposing an IoT based personal safety solution model for ensuring prompt help for victims attacked in moving vehicle
- Reviewing solutions available for location tracking and personal safety
- Implementing the 1st and 2nd version of our location prediction model as a step towards the implementation of the safety solution model.
- Risk assessment of the safety solution model based on OCTAVE ALLEGRO.

The rest of the paper is organized as follows: Section 2 which is related works addresses the review of location tracking and personal safety solution reveals the related work and Section 3 explores the proposed personal safety solution. Section 4 is the review of the location tracking and prediction algorithm. Section 5 explores various risk assessment approach. Section 6 provides discussion about OCTAVE ALLEGRO risk assessment methodology deployed. Section 7 presents the Implementation of the location prediction model and Section 8 concludes with the results and discussion.

2. RELATED WORK

In this section, we are reviewing solutions available from the literature on providing help for victims attacked in moving vehicles and personal safety solutions for individuals. Shinde et al., [9] have presented an IoT based solution to notify some pre-saved numbers via text message in case of an accident. Although the text message provides the current location of the accident, it does not detect the location of a moving vehicle. Sharma et al., in [10] have presented an ARM7 processor-based safety device, which activates GPS location tracking and sends text messages to the responder with a single button press. Although it activates GPS tracking, it does not predict the location of a victim in a moving vehicle and does not find the nearest police station to take prompt action to save the victim. Furthermore, the device is heavy to carry and may not be suitable to carry all the time. The safety solution presented by Bhavale et al., [11] alerts pre-registered phones with the captured images. However, in a panicking attack, the biggest concern would be to send alerts in the quickest possible way. The bus- monitoring unit used will need to be pre-installed in the bus to activate tracking which may not be a practical expectation.

A wearable device was proposed by Pawar et al., [12] consisting of a microcontroller, Raspberry pi, GPS and Global System for Mobile communications (GSM) module. Readings are continually taken from the sensors and compared against assigned threshold values. Computational overhead is excessively consumed with continuous tracking and comparison of readings from sensors.

Monisha et al., [13] proposed an ARM controller incorporated with GSM, GPS, Bluetooth, and RF detector and powered by 12V for the controller. The device gets activated by pressing the emergency button and sends out messages containing location to a pre-set number. The proposed method uses a hardware device that is too heavy to carry by the individual; furthermore, access to the device is required for activation, which is disastrous in situations where the mobile is not accessible.

Choudhary et al., proposed a safety device [14] which consists of a variety of sensing units such as heartbeat sensor, temperature sensor, and a push button. It also uses ATmega8L, GPS and GSM modules, flashlight, and a taser. The device fetches heartbeat and temperature reading and compares the readings against a set threshold. If there is a variation, the device would be activated, and a message containing the location would be sent to the police alongside known personals with the help of the GSM module. The proposed method may result in false positives as the heartbeat and temperature readings may spike due to other reasons, which are not detrimental to the individual. A false message is sent to the police and known personals to initiate emergency protocols, which leads to the waste of resources.

In the bid to offer easily accessible and less cumbersome personal safety solutions, smartphone applications were developed that are capable of harnessing the modules present in the smartphone to track user's location and send alerts to third party during distress. Safetipin is a smartphone personal safety application that helps users make informed decisions about visiting an area and location tracking of the user. The app operates by providing a safety score for the intended area of visit based on disturbance and risk within the area and alternate routes are displayed to the destination. Tracking of the user can be done when the user invites a third party to track their location. Street smart is another smartphone personal safety application that allows users to make informed decisions about an area before visiting by providing articles and reviews about the safety level of an area of interest by holding the camera at the location. The safety level is determined from posted reviews and articles as positive, negative or neutral using sentimental analysis [15].

Life360 serves as a smartphone family locator application and personal safety application. The application is used for tracking the location of the user and provide the location for wellbeing centres that could be required during distress. The concern remains privacy issues regarding the location information of the user [16].

Vithu is a smartphone personal safety application that alerts selected third party when the user is in distress.

When the cycle of operation is initiated by the double press of the activation button, an SMS message with the location information of the user is sent to the third party every two minutes to track the trajectory of the user [17].

BSafe is a smartphone personal safety application that alerts selected guardians to inform them that the user is in distress. The alert contains location information of the user at the point of distress and phone call is placed to one of the selected guardians with tracking of the user possible. Guard works similarly with placing calls with the name of the user, present location and alert of help required by the user to the third party. The requirement for the use of the app is cumbersome and rigorous [18].

Streetsafe measures up as a smartphone personal safety application that operates by sending alerts for users in distress based on four features. These features are high volume alarm is initiated, the current location is uploaded to the user's Facebook account, SMS sent to preferred associates in the area and lastly call is placed to the user's chosen emergency number. Fightback is a smartphone personal safety application that is similar to Streetsafe. It allows user to send alerts for users in distress making use of features such as e-mail, GPS, SMS and GSM and track location of user on map [19].

The reviewed solutions show their ability to track the present location of devices, send alert to third parties but do not predict the future location. The proposed personal safety solution model, presented in the following section, is enriched with the capacity to predict location using acquired location information.

3. PROPOSED PERSONAL SAFETY SOLUTION (PERSATION)

This paper provides a possible solution for helping victims in violent attacks in moving vehicles, where prompt response is crucially needed for saving the victim. In the implementation of the solution, as seen in Fig. 1, the problem can be separated into two parts. The first part of the solution involves finding a wearable that can work as a panic button, when pressed it can communicate with the gateway, which could be a GPS enabled mobile device held by the victim that communicates through the internet to a cloud-based application server to send help alert and the location information to third party and nearest responder. Once the button is activated, the second part of the solution will be activated, which involves using a suitable solution for tracking and predicting the location of moving vehicles and informing the nearest first responder. To accomplish tracking and prediction, implementation of state-of-the-art location tracking and prediction technologies is required and empowered with geographical data delivery service to aid the identification of the nearest police station. This way prompt action can be taken and the victim might be saved timely. The novelty of our proposed model lies in the usage of prediction capacity for providing timely help in a crucial situation.

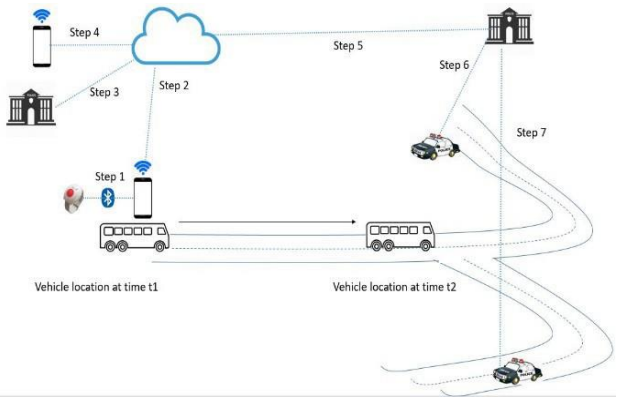

Figure 1. Personal safety solution scenario

Therefore, as a step towards the implementation of our proposed model we have first reviewed location tracking systems and location prediction algorithms in Section IV. We then have implemented the 1st version of our location prediction system by using a simple method for gathering location data and using prediction algorithm for location prediction.

4. REVIEW OF LOCATION TRACKING SERVICE AND PREDICTION MODEL

A. Selecting Location Tracking Service

Location based tracking service (LBS) offers a medium where data about a user can be collected in a coordinated and systematic manner and provides the user with the capability to find their bearing, and find other locations using semantic information about the present location and immediate environment. This service combines mobile services, location awareness, internet and GPS in the collection of a new layer of client's data and authentic data [20, 21, 22, 23, 24] Various technologies are involved in the operation of LBS with positioning technology, which handles the accurate location of the user regarded as the most important. The other technologies required by LBS are application technology that consists of two elements and deals with the presentation of information to the user. Geographic data that renders structures such as road network and manage data of the point of interest, and communication data for the transmission of user's location to the control centre for the provision of necessary service [25].

The delivery service of LBS can be categorized mainly as time- based delivery service, and distance-based delivery service. Time-based delivery service updates location information periodically to maintain high location tracking accuracy while Distance-based delivery service updates location information based on distance [26].

B. Review of Prediction Model

Time series analysis is a method used to gain insight into time series data about statistical patterns in the data and develop a suitable model for forecasting events. The most widely used linear time series models are AutoRegressive (AR) and Moving Average (MA). Autoregression is a model that relies on the dependent relationship between past period values and some degree of lagged observations to predict future values. Moving Average uses dependency within the dataset to provide output that depends linearly on the current and past values of stochastic distribution.

1) AutoRegressive Integrated Moving Average (ARIMA)

Autoregressive Moving Average (ARMA), which is suitable for univariate time series combines both AR and MA. The forecasted value by the model is a linear combination of past observations and random error with a constant term. The model is suitable for stationary time series data but comes short with data exhibiting non-stationary trends and seasonal patterns. AutoRegressive Integrated Moving Average (ARIMA) implements an integrated model that uses differencing to account for the establishment of stationarity. The suitability of the ARIMA model for the dataset is based on the exploration of trends and seasonality features in the dataset. ARIMA uses the ARIMA(p,d,q) notation based on the three models incorporated in the model (AR, integrated and MA). the p stated in the notation represents the AR that indicates the lag present in the stationarised series, the d stands for the integrated model that indicates the differencing required to attain stationarity and the q stands for the MA that indicates the lagged forecast errors in the series.

ARIMA time series model was introduced by Box and Jenkins. The model uses sets of activities to identify, estimate and diagnose the ARIMA algorithm suitable for time series data. ARIMA model forecast time series data by accounting for growth/decline pattern in the data with the Auto-Regressive part, rate of change of growth/decline with the Integrated part and the moving average to account for the noise between consecutive points in the data [27, 28]. Time series is the non-deterministic model for sequential observation of data in relation to a trend or seasonality. Time series applies a model to historic facts from data and forecast futures value of the series making use of movement along with the data over a long period of time (Trend), fluctuations available in the data over a particular period of time (Seasonality) and autocorrelation to distinguish time series operation from other types of statistical analysis. Autocorrelation (ACF), partial autocorrelation (PACF), inverse correlation and cross-correlation are used to identify and specify the form of time series model [29, 30, 31].

The appropriate model for the series is identified by initially determining the degree of differencing required to remove gross features of seasonality and non-stationarity of the series. After differencing, the next step involves checking for autocorrelated errors using the ACF to determine the order of AR required and the lagged error to determine the order of MA. The ACF displays the correlation between past values and helps in determining the term to use for the time series. Positive autocorrelation (PACF) at the first lag indicates that AR model can be used and Moving Average (MA) model indicates the random jumps for calculating error in subsequent periods within the plot.

ARIMA (p, d, q) is the standard notation used to indicate a specific model used by ARIMA where p is a number of lagged observations to be taken in, d is the degree of differencing and q is the size of the moving average window [25, 26]. The ARIMA equation after combining AR and MA becomes

$$Y_t = \alpha + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + .. + \beta_p Y_{t-p} \varepsilon_t + \phi_1 \varepsilon_{t-1} + \phi_1 \varepsilon_{t-1} + .. + \phi_q \varepsilon_{t-q} \quad (1)$$

where it can be translated in words as: Predicted output Y_t is the addition of the Constant α with Linear combination Lags of input Y up to p (number of observations) lags and Linear Combination of Lagged forecast errors up to q (moving average) lags.

2) Regression Tree Ensemble

This is a predictive model composed of a weighted combination of multiple regression trees. Ensemble methods combine several base model decision trees classifier to provide an optimal predictive model and increase the accuracy of the model. The model aims at constructing a linear combination of various models and attempt using the combinations for the improvement of the predictive performance of a model fitting technique. There are two approaches for the model bagging or bootstrap aggregating and boosting.

a) BAGGing, or Bootstrap AGGregating

The model combines Bootstrapping and Aggregation into one model which works on improving unstable estimation or classification schemes. Bagging is a variance and Mean Squared Error (MSE) reduction technique that is effective for the improvement of the predictive performance of regression or classification trees. Given a sample of data, multiple bootstrapped subsamples are pulled. For each of the bootstrapped subsamples, a decision tree is formed. After Decision Trees have been formed for all the subsamples, an algorithm is used for the aggregation of the decision trees for the most efficient predictor.

b) Boosting

Boosting is a sequential process that adjusts the weight of an observation based on the last classification. The first algorithm is trained on the entire dataset and subsequent algorithms are built by fitting the residual of the previous algorithm. The principle of the model is to decrease bias

error to build strong predictive models. The prediction uses a weighted majority vote (classification) or a weighted sum (regression) to produce the final prediction.

5. RISK ASSESSMENT APPROACH

The compounding of a framework, models and methodologies knowledge for the assessment of security and privacy for the Internet of things device is crucial in integrating and examining the cyber risk standards and governance to understand the risk faced by devices and IoT networks. The adaptation of traditional cybersecurity standards indicates the need for the identification of specific IoT cyber risk vectors and integrated into a holistic cyber risk impact assessment model. To reduce cyber risk in cloud technologies, proper design of cloud architecture is maintained between the cloud services and devices connected to it [32]. This scenario involves the interaction between humans and technology in providing real-time feedback that demands the security of data in transit.

The majority of the established framework applies the quantitative approach to measuring cyber risk. One widely accepted framework is the Operationally Critical Threat, Asset, and Vulnerability Evaluation (OCTAVE) methodology. The goal of OCTAVE methodology is to help organisations with operational and strategic mediums to perform information security risk assessment. OCTAVE works to connect organisation and their operational point of view activities of information security risk management. There are three publicly available OCTAVE methodologies. The first methodology introduced is the OCTAVE-consistent methodology that is defined by the implementation guide and training. Series of workshops facilitated by an interdisciplinary analysis team from various units that are in the organisation connect the organisation and the operational point of view activities of information security risk management [33]. The method is designed for large organisations with multi layered hierarchy, independence in performing vulnerability evaluation and interpret the results when maintaining the organisation computing infrastructure. The method can be used to tailor the approach to suit the distinct environment they operate. This method is performed in three phases, the first phase is the identification of assets and protection strategies presently been implemented, the second phase evaluates the infrastructure to supplement the analysis performed in the first phase. Risk mitigation plans for critical assets are developed after performing risk identification activities. In providing OCTAVE methodologies for small organisations, Technology Insertion, Demonstration, and Evaluation (TIDE) developed OCTAVE-S. The criteria are similar to OCTAVE and operate in with the same three phases except OCTAVE-S does not depend on formal knowledge workshop to obtain information. OCTAVE-S does not require extensive examination into the organisation infrastructure and helps practitioner address a wide range of risk which they have no familiarity about them. The risk identification, assessment, and mitigation processes are developed based on the collaborative aspect from an interdisciplinary perspective. With the collaboration strengthening the quality of risk assessment and mitigation, there are limitations in the interdisciplinary collaboration such as varying levels of expertise in threat evaluation, disparity in communication channels, practices, and intended efforts [34].

With the landscape of the information security risk changing coupled with the above limitations and the change in the required capability to manage the risks, the development of a new approach was inevitable to accommodate these changes. OCTAVE ALLEGRO adopts a different approach to organisation information technology environment and information assets than the other OCTAVE methodologies [35]. This method maps information assets to all containers where they are stored, transported or processed. Unlike the other two methodologies, OCTAVE ALLEGRO streamlines and optimizes the process of assessing the security risk of an organization and eliminates the use of vulnerability tools for threat identification by introducing the concept of an information risk environment map. The map help user defines all places information has been stored, transported or processed. The map serves as baseline documentation of the risk environment for the asset and helps establish boundaries of the threat environment and scope of risk assessment. The method uses a value known as relative risk score derived from the evaluation of qualitative description of risk probability combined with the prioritization of the organizational impact of risk in terms of the organization's risk measurement criteria. Mitigation guidelines, and specific strategies are considered for each container where the asset resides [34].

The methodology has four distinct activity areas carried out in various steps. The methodology establishes quantitative measures that are used as criteria for the evaluation of risk effect and serves as the foundation of information risk assessment, identification of the location of the asset and possible situations that threatens the asset, identification of threats and risks that could impact the asset, analysing the discovered risk and selecting the mitigation approach. The methodology does not provide details of methodology implementation and does not adequately address the impact of risk on assets. The methodology can serve as the starting point for risk assessment [34].

Threat Assessment & Remediation Analysis (TARA) is another system level quantitative methodology for the identification, prioritization and response based on three activities. Cyber Threat Susceptibility Analysis (CTSA) to assess the susceptibility of the asset to threats, Cyber Risk Remediation Analysis (CRRA) to determine best-suited countermeasures, and data and tools development to deliver recommendations for informed decisions. The

methodology targets the most crucial risk and offers a complementary form of protection [36].

Common Vulnerability Scoring System (CVSS) is a combination of the qualitative and quantitative framework for providing metric groups for assigning metric values to vulnerabilities and allocating cyber risk into levels and calculate an overall risk level. The methodology applies numerical values ranging from 0 – 10 to indicate the severity of vulnerability along with 3 color-coded levels to differentiate among the actual system. During result simplification, different vulnerabilities can produce the same level and similar score values that can increase the number of the colors in the color-code to enhance the visibility of different score values. The worthiness of the score can be faulted based on the basic mathematical formalism [37].

The Capability Maturity Model Integrated (CMMI) focus on the enterprise risk and development life cycle risk by integrating five levels of the Capability Maturity Model (CMM). The methodology identifies vulnerabilities and does not indicate ways to address the identified vulnerability [38].

The National Institute of Standard and Technology (NIST) provides the combination of risk assessment and risk management with a collection of standards and guidelines when combined with automated tools, aims to improve the security infrastructure [39].

The Factor Analysis of Information Risk Institute (FAIR) approach is focused on impact assessment and complementary with existing risk frameworks. The methodology address weaknesses of ISO standardized approach and creates a standardization reference for compliance. The FAIR model enhanced the deployment of RiskLens. This is a software-as-a-Service (SaaS) platform for the management and quantification of cyber-risks. The methodology is a quantitative model for cybersecurity and technology risk with integrated advanced quantitative risk analytics, best-practice risk assessment and reporting workflows. Another quantitative approach is the Cyber VaR (CyVaR), which shares complexity similarity with RiskLens but allows for the addition of a new type of risk [40, 41, 42].

6. RISK ASSESSMENT RESULT

This section aims to collect security threats discovered from information security risk assessments with OCTAVE Allegro methodology. OCTAVE ALLEGRO focus on information asset in various contexts such as how the assets were used; their exposure to threats, vulnerability and disruptions; where they are stored, transported and processed. The approach implores the user to explicitly consider the implication of risk consequence on security requirements and risk mitigation. The requirement of the approach is to allow focus on assets by ensuring they are selected through a systematic and consistent review process. The approach streamlines and improves threat identification and risk mitigation process without extensive risk assessment knowledge required. OCTAVE ALLEGRO development minimizes certain features which contribute to the ease of use requirement, minimal resource commitment and approach usability through fewer and more focused activities directed towards risk management. The adoption of scenario questionnaires by the approach instead of threat trees used by the earlier version of OCTAVE further help user in the identification of threats associated with information assets. OCTAVE ALLEGRO uses an information asset risk worksheet to capture the relevant information regarding specific risk for an information asset. The worksheet reduces documentation, organization, and data manipulation required to perform the risk assessment and help in producing a concise view of risk. A simple quantitative analysis of risk introduces a relative risk score that is computed on the worksheet using threat and impact information associated with risks captured. The introduced relative risk score is used to compare the significance of individual risks and computed from the combination of the risk probability qualitative description and the prioritization of risk impact based on the organization's risk measurement criteria.

The worksheet is used to compute relative risk scores for the assets are based on the component of the identified assets. Primary usage of the worksheet includes the identification of assets; determining the area of concerns accompanying the identified assets; threats and risks associated with the assets. From the relative risk score shown in table 2, identity theft and privacy violation possess higher risk to the framework as indicated by their high score compared to the other risks. Based on the relative high score of these two risks, security and privacy of user's information remains of utmost importance to maintain efficiency of the framework. Privacy violation could lead to tracking and monitoring of user activities and lead to a replay of the user's previous activities when the user is in distress, the event can be masqueraded by replaying user's activities instead of alerts been sent to the third party.

Based on the identified assets and area of concern related to the assets, associated threats are identified. Table 2 shows the threats associated with the assets and area of concern. The threats associated with these assets undermines the efficiency of the framework as user impersonation could lead to prevention of mechanism activation when the user is in need, device spoofing and data spoofing could result in the system receiving fake data or device information leading to waste of resources and inability to provide help to victims. Location information plays a crucial role in the framework forming the basis required for the prediction. In providing help for individuals in distress, accurate location information is essential. Threats jeopardizing the precision of location information reduces the chances of offering help that ensures the health and safety of the user.

TABLE I. INFORMATION ASSETS AND AREA OF CONCERN

Asset ID	Asset	Area of Concern
1	Personal Information	User's privacy
		Personal Identifiable Information
2	Device Information	Device configuration
		Data stored on device
		Network Topology
3	Location information	User's behavioural pattern
		Area of interest of the user
		User's movement
4	Log Information	Device operation
		User and device activities

TABLE II. INFORMATION ASSETS AND SECURITY THREATS

Asset ID	Asset	Threats
1	Personal Information	Data disclosure
		User impersonation
2	Device Information	Device spoofing
		Device breach/Theft
		Data spoofing
3	Location information	Tracking of user
		Disruption of the user's activity
		Monitoring of user
4	Log Information	Clearing of attack traces
		Map activities of service and application in the device

From the identified assets, reviewed area of concern and threats associated with the assets, risks are computed for the scenario. The relative risk score computed from OCTAVE ALLEGRO worksheet is based on the impact of the risk on the reputation and customer confidence, financial, productivity, safety and health, fines and legal penalties, user-defined impact area. Risks related with Asset ID 2 which includes data manipulation, data leakage and service denial/starvation could lead to disruption in delivering help during distress with inaccurate data been sent or service denial to prevent alerts been sent to the third party for help and poses a high risk to the productivity of the framework and safety and health of the user. Risks related to asset ID 1 and asset ID 3 pose a high risk to the safety and health of the user while risk related to asset ID 4 poses a high risk to productivity. The relative risk score associated with risk linked with asset ID 1 is based on their impact on the user. The health and safety component has high-risk value among the other components with financial and productivity components exhibiting medium risk impact on the asset.

TABLE III. RISKS AND COMPUTED RELATIVE RISK SCORES

Asset ID	Risks	Score
1	Privacy violation	25
	Identity Theft	32
2	Denial/Starvation of service	19
	Data leakage	22
	Data manipulation	12
3	Unauthorised app execution	19
	Interruption of activity	21
	Tracking of user	22
4	Exposure to extra service	13
	Loss of information	14

Table 4 shows different mitigation for the threats identified in table 2. For the framework, the major mitigation approach is to ensures the health and safety of the user. The mitigation approach stated in the table ensure that the safety of the user is not jeopardized at any time and especially when the user is in distress. The mitigation techniques improve the efficiency of the framework and provide adequate protection for users calling out for help.

TABLE IV. POSSIBLE MITIGATION APPROACHES

Asset ID	Mitigation Approach
1	Ensuring a good understanding of user privacy concerns
	Encryption of data
2	Device hardening
	Physical security of the device
	Device firmware update
3	Monitoring device security permissions
	Secured means of communication (VPN)
	Permission restriction
4	Secured system configuration
	Device hardening

7. IMPLEMENTATION OF ARIMA PREDICTION MODEL

The tool that is used for the implementation of the first model is SAS/ETS® and the data used was obtained using the location-based tracking services of a mobile device in a moving vehicle. From the review of section IV, the essential variables latitude, longitude and time used for the evaluated algorithm are collected by LBS. The algorithm gets an error when there is a disparity in subsequent time values or an empty value for any of the

variables. For the evaluation, the data was divided into two datasets. The first dataset which is made up of 88% of the original data is the data used for training the model and the second dataset is the test data.

Mean procedure test is the first test carried out. This test result shows the mean, standard deviation, minimum value and maximum value of the acquired data. The relevance of the test is to determine whether differencing is required based on the standard deviation result. If the value for standard deviation is insignificant i.e STD < 0.05, then differencing of the data is not required but when the value is significant, the Augmented Dickey-Fuller (ADF) test is carried out. The ADF test is based on the hypothesis that time-series data is non- stationary [27, 43] or has a significant standard deviation value.

Application of ARIMA procedure for the forecast of the next values. As discussed in section IV., this test is to determine the ARIMA (p,d,q) notation to be used for forecasting. Positive autocorrelation (PACF) at the first lag indicates that AR model can be used and Moving Average (MA) model indicates the random jumps for calculating error in subsequent periods within the plot. If the AR model is required there is a negative autocorrelation at Lag-1, a sharp drop in PACF after few lags and a gradual increase in PACF [44]. Other factors which help determine the most suitable model is the relatively small value provided by (2) for Akaike Information Criterion (AIC), (3) for Schwarz Bayesian Information Criterion (SBC) and the standard error value of regression (S.E. of regression)

$$AIC(p) = n\ln(\sigma^2_e/n) + 2p \quad (2)$$

$$BIC(p) = n\ln(\sigma^2_e/n) + p + p\ln(n) \quad (3)$$

The last test is to verify the efficiency of the result produced by the ARIMA model. The test involves the comparison of the acquired data from the model forecast and the test data obtained through LBS.

8. Implementation of Ensemble Regression Tree Model

The tool used for the implementation of the second implementation model is the regression learner app of MATLAB 2019a. The app can be used to train regression models for prediction of data, perform automated training for determination of the best regression model type, select features, specify validation schemes and assess results. The model types are including linear regression models, regression trees, Gaussian process regression models, support vector machines, and ensembles of regression trees.

After loading the dataset into the app, the validation method is to be selected to examine the predictive accuracy of the fitted models. Validation helps prevent overfitting, estimate the performance of the model and choose the best model.

The first type of validation is cross-validation, this selects the number of divisions to partition the dataset. If k division is selected, then the app:

- Partitions the data into k disjoint divisions
- For each fold, out-of-fold observation is used to train the model and model performance is assessed using in-fold data.
- Calculates the average test error over all divisions.

The method makes efficient use of all data and requires multiple fits which makes it suitable for small datasets.

The second type of validation is the holdout validation. For this validation type, a percentage of the data is reserved and used as the validation set. This type of validation is appropriate for large data sets as it segregates the data efficiently based on percentage.

If no validation is selected, there is no protection against overfitting. All the data are used for training and computing the error rate on the same data. The lack of test data makes the model performance for the estimation of new inaccurate and unrealistic.

Without any test data or validation data, the model can provide unrealistic estimates when used for new data. With the dataset used for the implementation, the cross-validation was selected as it suits the dataset.

9. Results and Discussion

From the result of the means procedure test as shown in Fig 2, the standard deviation for the latitude and longitude is insignificant with a value less than 0.05. The insignificant value of the standard deviation indicates that differencing is not required for the data and the ADF test would not be carried out.

The MEANS Procedure

Variable	Label	N	Mean	Std Dev	Minimum	Maximum
lat	lat	120	52.9252574	0.0043273	52.9198600	52.9295710
lon	lon	120	-1.4913512	0.0077818	-1.4990940	-1.4816560
elevation	elevation	70	109.2222736	15.4281721	43.2273404	131.0873011
accuracy	accuracy	120	19.8872000	3.6176629	6.0000000	32.7580000
bearing	bearing	8	181.6297544	73.5573774	72.7008400	301.7622000

Figure 2. Means procedure

The next test is to determine the specific ARIMA model to be used. The major factors for determining the most suitable model to used are the ACF, PACF graphs, AIC and SBC values.

As shown in Fig. 3, the ACF obtained shows a gradual decrease across the lag which indicates that AR model can be use and the PACF plot showing a sharp drop after a few lags and gradual increase across the lag indicates the MA model can be used. Based on these findings, various models of ARIMA was implemented to determine the AIC and SBC value and determine the best model for forecasting.

The values of AIC and SBC of various models of ARIMA are indicated in Table 5. The models explored have low AIC and SBC which indicates that they are all suitable for the research. The best model to use for prediction is the model with the lowest value of AIC and SBC. From Table 5, the lowest AIC and SBC value are 1735.68 and 1721.74 for latitude and 1478.09 and 1464.15 for longitude respectively, indicating the most suitable model is ARIMA (2,0,1) where 2 represent the number of lags, 0 is the degree of differencing, and 1 is the order of moving average.

TABLE V. AIC AND SBC RESULT OF ARIMA MODEL

ARIMA Model	Latitude		Longitude	
	AIC	SBC	AIC	SBC
ARIMA(1,0,0)	1627.95	1622.38	1419.60	1414.02
ARIMA(0,0,1)	1122.44	1116.87	981.49	975.92
ARIMA(1,0,1)	1696.50	1688.14	1445.14	1436.78
ARIMA(1,1,0)	1627.06	1621.49	1417.11	1411.53
ARIMA(0,1,1)	1122.44	1116.87	981.49	975.92
ARIMA(1,1,1)	1696.50	1688.14	1445.14	1436.78
ARIMA(1,1,3)	1712.58	1698.65	1455.00	1441.07
ARIMA(2,0,0)	1724.07	1715.70	1454.37	1446.01
ARIMA(2,1,3)	1459.45	1442.72	1459.45	1442.72
ARIMA(2,0,1)	1735.68	1721.74	1478.09	1464.15

For the last test, the predicted values from the model are compared with the test data from the acquired data. Fig 4 and Fig 6 shows the trend of the latitude training data for latitude and longitude respectively while Fig 5 and Fig 7 shows the difference between the test data and the predicted data. The pattern of the difference between values of the predicted data seems to be constant and swaying in a single direction, either increasing or declining gradually unlike the test data, which decreases, stagnates and increases over the trend. The model determines a pattern from the training data and uses the data to derive the predicted data.

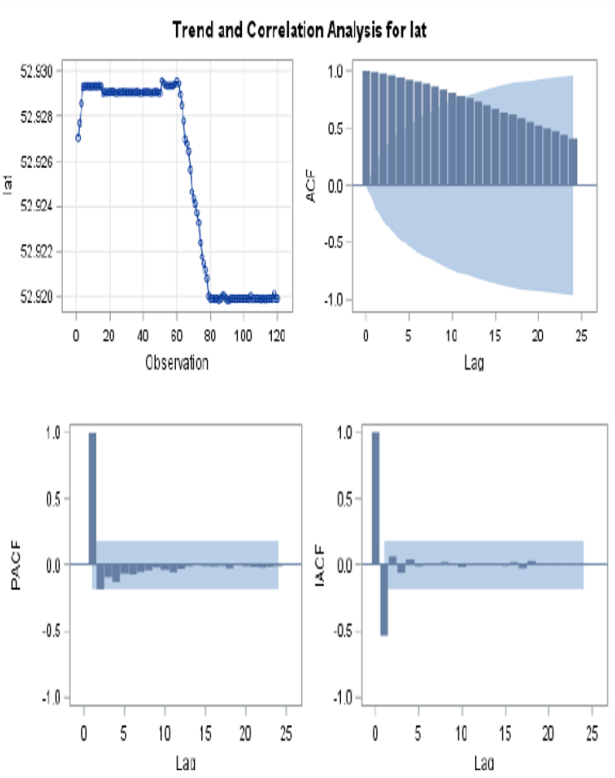

Figure 3. Correlation Analysis of training data

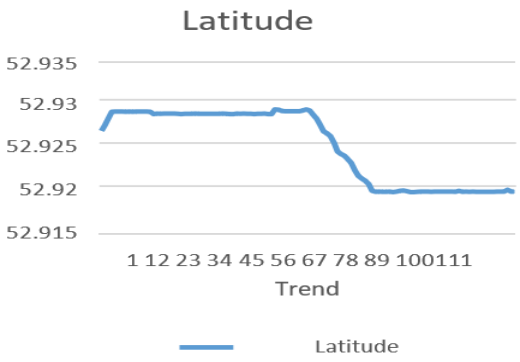

Figure 4. Latitude training data

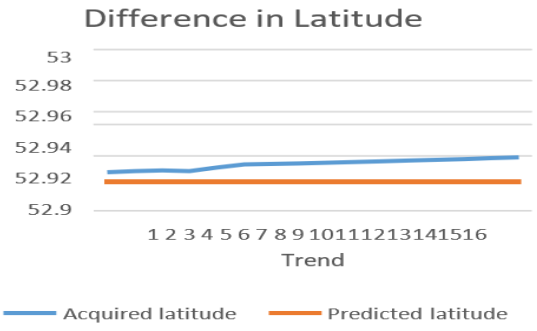

Figure 5. The difference in latitude data

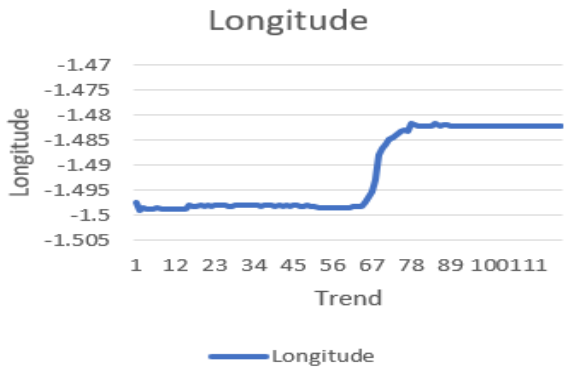

Figure 6. Longitude training data

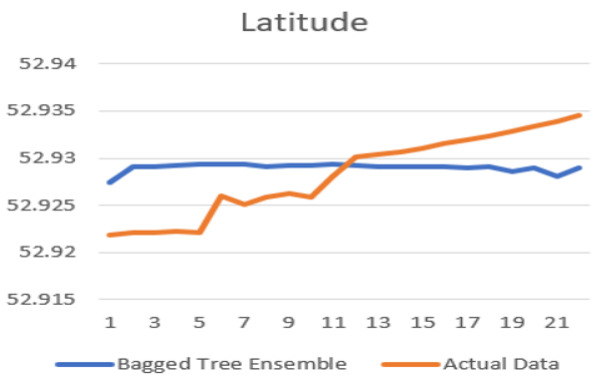

Figure 8. The difference in Bagged tree latitude data

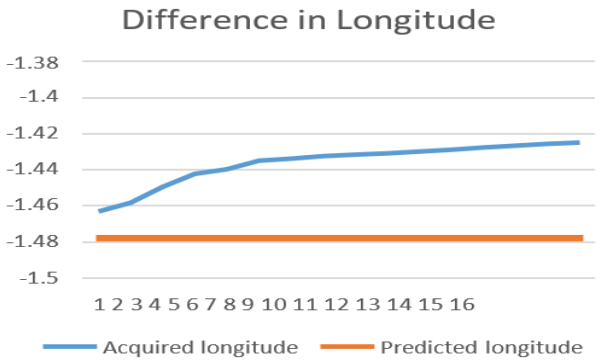

Figure 7. The difference in longitude data

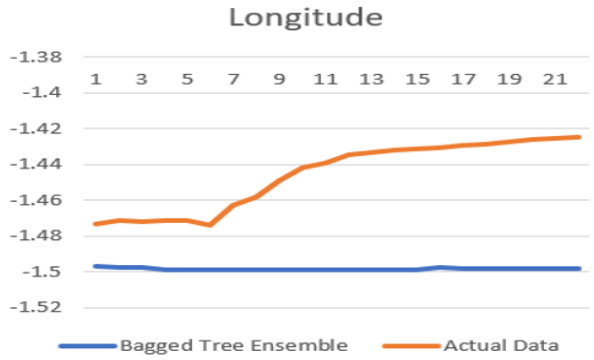

Figure 9. The difference in Bagged tree longitude data

This second implementation model uses a weighted combination of multiple regression trees to construct a linear combination that improves the predictive performance of the model.

The bagged tree ensemble aggregates the decision tree for the most efficient predictor. The figure shows the predicted values use trend which is similar to the trend of the training data. The similarity in trend shows the efficiency of the model. For the latitude in Fig 8, the trend of the values between the verification data and predicted values shows the same initial trend of an upward slope with an inconsistent upward and downward slope along with the trend. The longitude values show a great disparity along with the trend as indicated in Fig 9. Unlike the actual data that has a consistent downward slope, the predicted data displayed a conspicuous upward slope along with the trend.

The boosted tree uses a sequential process of weight adjustment and built on the fitting of the successive algorithm on the previous one. The sequential fitting can be observed in the consistent intervals shown in the figure as the predicted values seem constant across the trend. Fig 10 shows the latitude values for both the predicted value and the test data shows similar movement along with the trend but shows a significant difference between the test data value and the predicted data values. Longitude values show a significant difference in values between test data and predicted data and the movement along the trend shows significant difference along with the trend as shown in Fig 11.

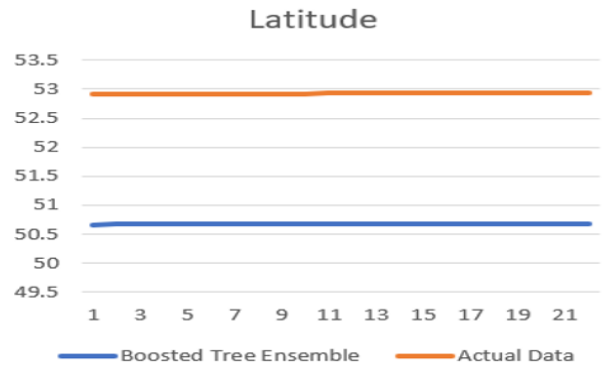

Figure 10. The difference in Boosted latitude data

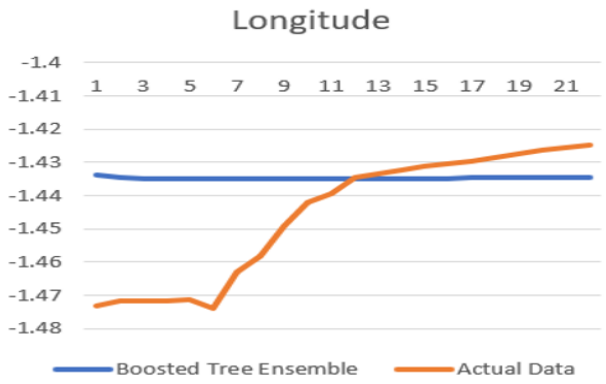

Figure 11. The difference in Boosted longitude data

10. Conclusion

The solution provided for our problem space is believed to help the prevention of violent attacks in high-speed vehicles and provides easily accessible devices to aid the communication of the on the move violent attacks to a third party and the nearest police station to save the lives of victims. This paper takes into consideration the prediction model for the second part of the proposed IoT solution. The first algorithm applied in this paper offers a consequential prediction based on the pattern of historical data. The results from the first implementation are predicted based on the pattern derived from the trend of the training data, the trend displayed from the result indicates the suitability of the model for forecasting slow movement along a straight path rather than random movement.

The result of the second implementation unlike the first implementation does not follow the trend of the data rather it manifested a curve based on the weighted combination of multiple regression trees.

To maximize the efficiency of the framework and prevent sabotage by insider or external vulnerabilities, risk analysis of the framework was performed using OCTAVE ALLEGRO for the identification of various assets available in the scenario, area of concerns regarding these assets, threats and mitigation approach deployable in the framework. The risk assessment looks to prevent undermining the performance of the framework.

As a next step of the research, we would look into adopting deep learning algorithms that best suit the purpose of forecasting the next location of a moving vehicle with random motion.

References

[1] Shock and outrage over India Delhi bus gang rape, BBC, 2012.

[2] Bangladeshi law student killed after five men gang-raped her on bus, 2017.

[3] K. Lewis, Mother gang-raped on bus as two-week old baby dies in attack, Independent Digital News and Media, 2016.

[4] I. Qureshi, India woman raped in moving bus in Karnataka, BBC, 2015.

[5] A. McSorley, Who is Jastine Valdez? Wicklow woman abducted by Dublin dad-of-two Mark Hennessy, 2018.

[6] K. Utehs, ABC7 NEWS EXCLUSIVE: 5-year-old girl kidnapped during San Jose car theft, reunited with parents, 2018.

[7] Woman kidnapped in boot of her own car in Emmer Green, BBC, 2018.

[8] E. Longnecker, 12-year-old calls 911 during frightening ride in stolen car, 2019.

[9] P. Shinde, P. Taware, S. Thorat, T. Waghmare and A. Kadam, "Emergency Panic Button," International Journal of Scientific & Engineering Research, vol. 3, p. 3, 2012.

[10] S. Sharma, F. Ayaz, R. Sharma and D. Jain, "IoT Based Women Safety Device using ARM 7," 2017.

[11] M. D. M. Bhavale, M. P. S. Bhawale, M. T. Sasane and M. A. S. Bhawale, "IOT Based Unified Approach for Women and Children Security Using Wireless and GPS," International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume, vol. 5, 2016.

[12] R. Pawar, M. Kulabkar, K. Pawar, A. Tambe and P. Smita Khairnar, "Smart Shield for Women Safety," International Research Journal of Engineering and Technology (IRJET) e-ISSN, vol. 05, no. 4, pp. 56-2395, 4 2018.

[13] D. G. Monisha, M. Monisha, G. Pavithra and R. Subhashini, "Women safety device and application-FEMME," Indian Journal of Science and Technology, vol. 9, no. 10, 2016.

[14] Y. Choudhary, S. Upadhyay, R. Jain and A. Chakraborty, "Women Safety Device (Safety Using GPS, GSM, Shock, Siren and LED)," International Journal of Advance Research in Science and Engineering, vol. 6, no. 5, p. 413–421, 5 2017.

[15] P. Kartik, S. Jose and G. K. MK, "Safetipin: A Mobile Application Towards Women Safety," Rajagiri Journal of Social Development, vol. 9, no. 1, pp. 5-12, 2017.

[16] Life360 - Feel free, together..

[17] M. Umar, Akash and Naveen, VithU App: A Woman Safety App by Gumrah, 2018.

[18] K. Sharma and A. More, "Android Application for women security system," International Journal of Advanced Research in Computer Engineering & Technology, vol. 5, no. 3, pp. 725-729, 2016.

[19] K. J. M. Baker, The Street Safety App for Proactive and Paranoid Woman, Jezebel, 2013.

[20] S. Khan, W. Ahmad, R. Ali and S. Saleem, "A Research on Mobile Applications for Location Tracking through Web Server and Short Messages Services (SMS)," VFAST Transactions on Software Engineering, vol. 7, no. 2, pp. 12-17, 2 2015.

[21] C. Pontikakos, M. Sambrakos, T. Glezakos and T. Tsiligiridis, "Location-based services: A framework for an architecture

design," Neural Parallel and Scientific Computatiions, vol. 14, no. 2/3, p. 273, 2006.

[22] J. O. Aasha, S. Monica and E. Brumancia, "A tracking system with high accuracy using location prediction and dynamic threshold for minimizing SMS delivery," in 2015 International Conference on Computation of Power, Energy, Information and Communication (ICCPEIC), 2015.

[23] R. Kolvoord, K. Keranen and P. Rittenhouse, "Applications of location-based services and mobile technologies in K-12 classrooms," ISPRS International Journal of Geo-Information, vol. 6, no. 7, p. 209, 2017.

[24] Y.-C. Lai, J.-W. Lin, Y.-H. Yeh, C.-N. Lai and H.-C. Weng, "A tracking system using location prediction and dynamic threshold for minimizing SMS delivery," Journal of Communications and Networks, vol. 15, no. 1, pp. 54-60, 2013.

[25] N. Chan and H. Lars, "Introduction to location-based services," Lund University GIS Centre, p. 1–12, 8 2003.

[26] P. Gupta and S. S. Sutar, "Study of Various Location Tracking Techniques for Centralized Location, Monitoring & Control System," IOSR Journal of Engineering, vol. 4, no. 03, pp. 27-30, 3 2014.

[27] M. Kumar and M. Anand, "An application of time series ARIMA forecasting model for predicting sugarcane production in India," Studies in Business and Economics, vol. 9, no. 1, pp. 81-94, 2014.

[28] P. Chen, H. Yuan and X. Shu, "Forecasting crime using the arima model," in 2008 Fifth International Conference on Fuzzy Systems and Knowledge Discovery, 2008.

[29] A. A. Ariyo, A. O. Adewumi and C. K. Ayo, "Stock price prediction using the ARIMA model," in 2014 UKSim-AMSS 16th International Conference on Computer Modelling and Simulation, 2014.

[30] R. Adhikari and R. K. Agrawal, "An introductory study on time series modeling and forecasting," arXiv preprint arXiv:1302.6613, 2013.

[31] C. Liu, S. C. H. Hoi, P. Zhao and J. Sun, "Online arima algorithms for time series prediction," in Thirtieth AAAI conference on artificial intelligence, 2016.

[32] P. Radanliev, D. C. De Roure, J. R. C. Nurse, R. M. Montalvo, S. Cannady, O. Santos, P. Burnap, C. Maple and others, "Future developments in standardisation of cyber risk in the Internet of Things (IoT)," SN Applied Sciences, vol. 2, no. 2, p. 169, 2020.

[33] C. Woody, J. Coleman, M. Fancher, C. Myers and L. Young, "Applying OCTAVE: practitioners report," 2006.

[34] R. A. Caralli, J. Stevens, L. Young and I. W. R. Wilson, "Introducing OCTAVE Allegro: Improving the Information Security Risk Assessment Process," 2007.

[35] C. J. Alberts and A. Dorofee, Managing information security risks: the OCTAVE approach, Addison-Wesley Longman Publishing Co., Inc., 2002.

[36] J. Wynn, J. Whitmore, G. Upton, L. Spriggs, D. McKinnon, R. McInnes, R. Graubart and L. Clausen, "Threat assessment & remediation analysis (tara): Methodology description version 1.0," 2011.

[37] P. Mell, K. Scarfone and S. Romanosky, "A complete guide to the common vulnerability scoring system version 2.0," in Published by FIRST-forum of incident response and security teams, 2007.

[38] C. W. I. C. M. Model, "Integration (CMMI)®?| CMMI Institute," CMMI Institute, 2017.

[39] C. NIST, "Cybersecurity framework| NIST," NIST Website, 2016.

[40] N. Sanna, How FAIR Can Ensure The Success of COSO Risk Management Programs, 2017.

[41] Risk Analytics Platform | FAIR Platform Management | RiskLens.

[42] R. Shaw, V. Takanti, T. Zullo, M. Director and E. Llc, Best Practices in Cyber Supply Chain Risk Management Boeing and Exostar Cyber Security Supply Chain Risk Management Interviews, 2017.

[43] J. Panneerselvam, L. Liu and N. Antonopoulos, "InOt-RePCoN: Forecasting user behavioural trend in large-scale cloud environments," Future Generation Computer Systems, vol. 80, pp. 322-341, 2018.

[44] Sangarshanan, Time series Forecasting - ARIMA models, Towards Data Science, 2018.

[45] ISO - ISO 31000 — Risk management.

Alofe Olasunkanmi Matthew presently working towards PhD in Cybersecurity at the University of Derby having received his M.Sc in cybersecurity from the same university. His current research is focused on relating cybersecurity, Internet of Things and machine learning. His research interest includes cyber security, machine learning, Internet of Things, cryptography and encryption.

Dr. Kaniz Fatema is working as a Lecturer at Aston University, UK. Previously she worked as a Senior Lecture at the University of Derby, UK. She has many years of experience of working as a Research Fellow at Trinity College Dublin and University College Cork, Ireland. She completed her PhD in Computer Science (Information Security) from the University of Kent, UK and MSc in Data Communications from the University of Sheffield, UK. She has almost a decade of research experience in the domains of Information and Cyber Security, such as, access control, data protection, compliance assurance for data protection regulation and Cloud Computing.

Muhammad Ajmal Azad received the Ph.D. (2016) degree in Electrical and Computer Engineering from the University of Porto, Portugal, and MS (2008) in Electronics Engineering from the International Islamic University Pakistan. He is a lecturer in Cyber Security at the University of Derby UK, before joining The University of Derby, he was research fellow (an equivalence of lecturer in the UK) in the department of computer science at The University of Warwick and research associate at Newcastle University. He also spent more than 5 years in the telecommunication company. His research interests include privacy-aware collaboration, reputation aggregation, privacy protection, privacy-aware outsourcing of network logs, and spam detection in a telecommunication network.

Fatih Kurugollu (Senior Member, IEEE) received the B.Sc., M.Sc., and Ph.D. degrees in computer engineering from Istanbul Technical University, Istanbul, Turkey, in 1989, 1994, and 2000, respectively. From 1991 to 2000, he was a Research Fellow with Marmara Research Centre, Kocaeli, Turkey. In 2000, he joined the School of Computer Science, Queen's University Belfast, Belfast, U.K., as a Postdoctoral Research Assistant. He was appointed as a Lecturer with Queen's University Belfast in 2003 and was promoted to a Senior Lecturer in 2011. He is currently a Professor of cyber security and the Head of the Cyber Security Research Group, University of Derby, Derby, U.K. His research interests include cyber security, multimedia security, big data analysis and AI, image and video processing applications, biometrics, and hardware architectures for image and video applications.

Appendix C

Risk Assessment Based Privacy-Preserving Scheme Source Code

```
#
   ------------------------------------------------------------

#                Thesis Code
#
   ------------------------------------------------------------


import pandas as pd
from geopy.distance import geodesic
from datetime import datetime
import math
import numpy as np
import time
import matplotlib.pyplot as plt

from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_absolute_error,
   mean_squared_error, r2_score

def calculate_distance(lat1, lon1, lat2, lon2):
```

```
    return geodesic((lat1, lon1), (lat2, lon2)).meters

def calculate_velocity(df):
    df = df.sort_values(by=['id', 'timestamp'])

    # Ensure timestamps are in datetime format
    df['timestamp'] = pd.to_datetime(df['timestamp']).dt.
       floor('s')

    df['distance'] = float('nan')
    df['time_diff'] = float('nan')
    df['velocity'] = 0.0

    for i in range(1, len(df)):
        if df.iloc[i]['id'] == df.iloc[i-1]['id']:
            distance = calculate_distance(
                df.iloc[i-1]['lat'], df.iloc[i-1]['lon'],
                df.iloc[i]['lat'], df.iloc[i]['lon'])
            df.at[i, 'distance'] = distance

            # Convert timestamps explicitly
            t1 = pd.Timestamp(df.iloc[i]['timestamp'])
            t0 = pd.Timestamp(df.iloc[i-1]['timestamp'])

            time_diff = (t1 - t0).total_seconds()
            df.at[i, 'time_diff'] = time_diff
            df.at[i, 'velocity'] = distance / time_diff if
               time_diff != 0 else 0.0

    return df

def calculate_relative_velocity(vel1, vel2):
    return (vel1**2 + vel2**2)**0.5

def calculate_ttc(row, df):
```

```
lat1, lon1, vel1 = row['lat'], row['lon'], row['
   velocity']
min_ttc = float('inf')
timestamp = row['timestamp']

nearby_vehicles = df[(df['timestamp'] == timestamp) & (
   df['id'] != row['id'])]

#print(f"Timestamp: {timestamp}, ID: {row['id']},
   Velocity: {vel1:.2f} m/s")
for _, other_row in nearby_vehicles.iterrows():
    lat2, lon2, vel2 = other_row['lat'], other_row['lon
       '], other_row['velocity']
    distance = calculate_distance(lat1, lon1, lat2,
       lon2)

    relative_distance = ((vel1 + vel2)* 7)
    #print (f"This is relative_distance {
       relative_distance:.5f}")

    if distance <= relative_distance:  # Check if the
       distance is within the car's velocity value
        #print(f"Nearby Vehicle ID: {other_row['id']},
           Distance: {distance:.2f} meters, Other
           Velocity: {vel2:.2f} m/s")

        relative_velocity = calculate_relative_velocity
           (vel1, vel2)
        if relative_velocity != 0:
            ttc = distance / relative_velocity
            if ttc < min_ttc:
                min_ttc = ttc

            # Debug
```

```
                #print(f"Relative Velocity: {
                    relative_velocity:.2f} m/s, TTC: {ttc:.2
                    f} seconds")

    return min_ttc if min_ttc != float('inf') else None

#Privacy Presevation Module
#Probabilitic Quantification
def epsilon_derivation(Upsilon_1, epsilon_0):
    return (1/(1 - Upsilon_1 * (1 - 3 * (math.exp(-
        epsilon_0)))))


#Add Selected Laplace Noise Mechanism of Differential
   Privacy
def add_laplace_noise(data, epsilon, sensitivity):
    """
    Add Laplace noise to the data
    :param data: Data to be perturbed
    :param epsilon: Privacy parameter
    :param sensitivity: Sensitivity of the data
    :return: Perturbed data
    """
    laplace_noise = np.random.laplace(loc=0, scale=
        sensitivity/epsilon, size=data.shape)
    return data + laplace_noise


def categorize_risk_score(ttc_value):
    """
    Categorizes risk score based on TTC (Time to Collision)
        values.

    Parameters:
        ttc_value (float): The TTC value to be categorized.
```

```
    Returns:
        int: The risk score category (1, 2, or 3)
    """
    if ttc_value < 4:
        return 3
    elif 4 <= ttc_value <= 7:
        return 2
    else:
        return 1

def calculate_privacy(risk_score_category, epsilon_factor
   =0.8):
    """
    Calculates privacy value based on the risk score
       category using epsilon_derivation.

    Parameters:
        risk_score_category (int): The categorized risk
           score.
        epsilon_factor (float): The epsilon factor for
           privacy derivation. Default is 0.8.

    Returns:
        float: The calculated privacy value, rounded to one
            decimal place.
    """
    return round(epsilon_derivation(risk_score_category,
       epsilon_factor), 1)


def prepare_data(df):
    """
    Prepares the dataset by handling missing values,
       creating duplicate columns,
    and dropping unnecessary features.
    """
```

```
    df["Latitude"] = df["lat"]
    df["Longitude"] = df["lon"]
    df.dropna(inplace=True)
    features_to_drop = ["Latitude", "Longitude", "timestamp
       "]
    return df, features_to_drop

def split_data(df, features_to_drop):
    """
    Splits the dataset into training and testing sets.
    """
    train_data = df.sample(frac=0.8, random_state=0)
    test_data = df.drop(train_data.index)
    X_train = train_data.drop(columns=features_to_drop)
    y_train = train_data[["Latitude", "Longitude"]]
    X_test = test_data.drop(columns=features_to_drop)
    y_test = test_data[["Latitude", "Longitude"]]
    return X_train, y_train, X_test, y_test

def train_model(X_train, y_train):
    """
    Trains a linear regression model.
    """
    model = LinearRegression().fit(X_train, y_train)
    return model

def evaluate_model(model, X, y):
    """
    Evaluates the model using MAE, MSE, RMSE, and training
       time.
    """
    start_time = time.time()
    y_pred = model.predict(X)
    end_time = time.time()
    mae = mean_absolute_error(y, y_pred)
    mse = mean_squared_error(y, y_pred)
```

```
    rmse = np.sqrt(mse)
    time_taken = end_time - start_time
    print("Baseline Metrics (   = 0):")
    print("Mean Absolute Error (Baseline):", mae)
    print("Mean Squared Error (Baseline):", mse)
    print("Root Mean Squared Error (Baseline):", rmse)
    print("Time taken for baseline process: %s seconds" % (
       time.time() - start_time))
    print
       ("-------------------------------------------------------------

    return mae, mse, rmse, time_taken

def test_privacy_preserved_data(model, X_test, y_test,
   sensitivity, mae, mse, rmse, training_time):
    """
    Tests the model on privacy-preserved data by adding
       Laplace noise.
    """
    results = []

    for epsilon in np.arange(0.1, 1.1, 0.1):
        start_time = time.time()
        print("Epsilon level is: ", epsilon)
        noisy_X_test = X_test.copy()
        noisy_X_test["lat"] = add_laplace_noise(
           noisy_X_test["lat"], epsilon, sensitivity)
        noisy_X_test["lon"] = add_laplace_noise(
           noisy_X_test["lon"], epsilon, sensitivity)
        new_predictions = model.predict(noisy_X_test)
        new_mae = mean_absolute_error(y_test,
           new_predictions)
        new_mse = mean_squared_error(y_test,
           new_predictions)
        new_rmse = np.sqrt(new_mse)
        new_accuracy = r2_score(y_test, new_predictions)
```

```
        results.append({
            "epsilon": epsilon,
            "noisy_mae": new_mae,
            "baseline_mae": mae,
            "delta_mae": new_mae - mae,
            "noisy_mse": new_mse,
            "baseline_mse": mse,
            "delta_mse": new_mse - mse,
            "noisy_rmse": new_rmse,
            "baseline_rmse": rmse,
            "delta_rmse": new_rmse - rmse,
            "r2_accuracy": new_accuracy,
            "baseline_time": training_time,
            "process_time": ((time.time() - start_time) -
               training_time)
        })
        print("Mean Absolute Error (Noisy Test Data):",
           new_mae)
        print("Mean Squared Error (Noisy Test Data):",
           new_mse)
        print("Root Mean Squared Error (Noisy Test Data):",
            new_rmse)
        print("R^2 Accuracy (Noisy Test Data):",
           new_accuracy)
        print("Time taken for process: %s seconds" % (time.
           time() - start_time))
        print("\n")
    results_df = pd.DataFrame(results)
    print("\nSummary of Outcomes and Delta Metrics:")
    print(results_df)
    return results_df

sensitivity = 0.1  # Sensitivity of the data
# Load the dataset
df = pd.read_excel(r'Dataset\output.xlsx')
df_copy = df.copy()
```

```
df = calculate_velocity(df)
df['ttc'] = df.apply(lambda row: calculate_ttc(row, df),
   axis=1)
print (df)
df['Risk_Score_Category'] = df['ttc'].apply(
   categorize_risk_score)
df['Privacy'] = df['Risk_Score_Category'].apply(lambda x:
   calculate_privacy(x))

df.to_excel(r'Dataset\privacy.xlsx', index=False)


# df = load_your_dataframe_here()
df2, features_to_drop = prepare_data(df_copy)
X_train, y_train, X_test, y_test = split_data(df2,
   features_to_drop)

model = train_model(X_train, y_train)
mae_train, mse_train, rmse_train, training_time =
   evaluate_model(model, X_train, y_train)
print("Training MAE:", mae_train, "Training MSE:",
   mse_train, "Training RMSE:", rmse_train, "Training Time
   :", training_time)
results = test_privacy_preserved_data(model, X_test, y_test
   , sensitivity, mae_train, mse_train, rmse_train,
   training_time)

results_df = pd.DataFrame(results)
output_file = f"Dataset\Error Results.xlsx"

# Save DataFrame to Excel
results_df.to_excel(output_file, index=False)
```

Appendix D

Collision Dataset Generation Source Code

```
#
   ------------------------------------------------------------

#                Generate Collision Dataset
#
   ------------------------------------------------------------



import simpy
import numpy as np
import pandas as pd
from datetime import datetime, timedelta

# --------------------------------
# Simulation Setup and Parameters
# --------------------------------

# Collision point coordinates (the destination of each
   vehicle)
collision_lon = 116.45650
collision_lat = 39.90700

# Vehicle starting points (given as [lat, Longitude])
```

```
vehicle1_start = [39.90710, 116.41500]  # Will be assigned
   id 10
vehicle2_start = [39.90700, 116.49000]  # Will be assigned
   id 366

# Simulation configuration:
n_steps = 30               # Total number of simulation
   steps (rows)
interval_seconds = 10      # Each subsequent timestamp is
   10 seconds apart

# Collision event time (the final timestamp in the
   simulation)
collision_time = datetime(2008, 2, 2, 13, 42, 53)

# Calculate the total duration of the simulation.
total_duration = interval_seconds * (n_steps - 1)

# Determine the simulation start time so that the final
   step occurs at collision_time.
start_time = collision_time - timedelta(seconds=
   total_duration)

# --------------------------------
# Function to Simulate Vehicle Movement using SimPy
# --------------------------------

def vehicle_process(env, vehicle_id, start, collision_point
   , results):
    """
    Simulate vehicle movement over time using SimPy.

    Parameters:
        env (simpy.Environment): The SimPy environment
           handling simulation events.
```

```
        vehicle_id (int): The unique identifier of the
           vehicle.
        start (list): The starting [latitude, longitude] of
            the vehicle.
        collision_point (tuple): The target (latitude,
           longitude) where the vehicle is heading.
        results (list): A shared list to store results (
           timestamp, lat, lon, vehicle_id).
    """
    start_lat, start_lon = start
    collision_lat, collision_lon = collision_point

    # Generate linear interpolation points
    interp_param = np.linspace(0, 1, n_steps)
    lats = np.linspace(start_lat, collision_lat, len(
       interp_param))
    lons = np.linspace(start_lon, collision_lon, len(
       interp_param))

    for i in range(n_steps):
        current_time = start_time + timedelta(seconds=i *
           interval_seconds)
        results.append([vehicle_id , current_time, round(
           lats[i], 6), round(lons[i], 6)])
        yield env.timeout(interval_seconds)  # Simulate
           time passing

# Initialize SimPy environment
env = simpy.Environment()

# List to store simulation results
simulation_results = []

# Start vehicle processes
env.process(vehicle_process(env, 10, vehicle1_start, (
   collision_lat, collision_lon), simulation_results))
```

```
env.process(vehicle_process(env, 366, vehicle2_start, (
    collision_lat, collision_lon), simulation_results))

# Run the simulation
env.run()

# -------------------------------
# Creating DataFrame from Simulation Data
# -------------------------------

df_combined = pd.DataFrame(simulation_results, columns=["id
    ", "timestamp", "lat", "lon"])

# -------------------------------
# Saving Data to an Excel File (Single Sheet)
# -------------------------------

# Specify the file path where you want to save the Excel
    file.
excel_file_path = r"Vehicle_paths_combined.xlsx"

# Save the combined DataFrame to one Excel sheet using
    DataFrame.to_excel.
df_combined.to_excel(excel_file_path, index=False)

print("The combined Excel file has been successfully saved
    .")




#
    ------------------------------------------------------------

#                Collision Dataset on Map
```

```
#
   ------------------------------------------------------------


import numpy as np
import pandas as pd
import folium


# Load the simulation data from Excel file
excel_file_path = r"Vehicle_paths_combined.xlsx"
df_combined = pd.read_excel(excel_file_path)

# Extract unique vehicle IDs
vehicle_ids = df_combined["id"].unique()

# Define colors for each vehicle
vehicle_colors = {10: "blue", 366: "red"}

# Create a new Folium map
collision_lat, collision_lon = 39.90700, 116.45650
m = folium.Map(location=[collision_lat, collision_lon],
   zoom_start=16)

# Plot vehicle paths
for vehicle_id in vehicle_ids:
    vehicle_data = df_combined[df_combined["id"] ==
       vehicle_id]
    path = list(zip(vehicle_data["lat"], vehicle_data["lon
       "]))
    folium.PolyLine(path, color=vehicle_colors.get(
       vehicle_id, "gray"), weight=2.5, opacity=1).add_to(m
       )

    # Add markers for vehicle movements
    for i, (lat, lon) in enumerate(path):
```

```
        folium.Marker(
            location=[lat, lon],
            icon=folium.Icon(color=vehicle_colors.get(
               vehicle_id, "gray"), icon="car", prefix="fa
               "),
            tooltip=f"Vehicle {vehicle_id} - Step {i+1}"
        ).add_to(m)

# Add the collision marker
folium.Marker(
    location=[collision_lat, collision_lon],
    icon=folium.Icon(color="black", icon="exclamation-
       triangle", prefix="fa"),
    tooltip="Collision Point"
).add_to(m)

# Save map to an HTML file
map_output_path = "vehicle_simulation_map.html"
m.save(map_output_path)

print("The map has been successfully saved as an HTML file
   .")
```