NOVEL MACHINE LEARNING TECHNIQUES FOR WEATHER-RELATED CRASH PREDICTION

by

Abimbola Rasheed Ogungbire

A dissertation submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in
Infrastructure & Environmental Systems

Charlotte

2024

Approved by:

Dr. Srinivas S. Pulugurtha

Dr. Suzanne Leland

Dr. Omid Shoghli

Dr. Ming-Chun Lee

ABSTRACT

ABIMBOLA RASHEED OGUNGBIRE. Novel Machine Learning Techniques for Weather-Related Crash Prediction. (Under the guidance of DR. SRINIVAS S. PULUGURTHA)

This dissertation addresses critical aspects of traffic safety, focusing on novel approaches for weather-related crash prediction—a significant concern in the transportation field. It is divided into three interconnected studies: geospatial risk mapping, the treatment of imbalanced data in machine learning, and analytics for crash prediction. In the first study, the dissertation advances a novel approach to hotspot mapping by developing a spatio-temporal cube that incorporates both the spatial and temporal dimensions of crash data, providing a dynamic and comprehensive analysis of crash hotspots. In the second study, the dissertation tackles the challenge of imbalanced data, which can bias model outputs from the machine learning techniques, making them less adept at predicting crash severity. By extending methods such as Synthetic Minority Over-sampling Technique (SMOTE) and Adaptive Synthetic Sampling (ADASYN), the dissertation evaluates the effectiveness of these methods in datasets with a prevalence of nominal predictors, aiming to enhance the predictive accuracy of machine learning techniques for crash severity. Lastly, the dissertation proposes the use of a Long Short-Term Memory (LSTM) algorithm for predicting a weather-related traffic crash. This approach seeks to overcome the limitations of traditional predictive models by leveraging the ability of LSTMs to retain relevant information over extended time frames, despite the stochastic nature of weather and human behavior.

DEDICATION

This dissertation is dedicated to my family: my lovely fiancée, Suliat Alli; my siblings, Adedoyin Mariam, Adedolapo Mojeed, Abiola Khadijat; and my parents, Mujidat Olaitan and Jimoh Ademola of the Ogungbire family.

ACKNOWLEDGEMENT

First, I thank God for seeing me through this phase of my life. To my supervisor, Dr. Srinivas Pulugurtha, thank you for your support, mentorship, and feedback at each step of this Ph.D. journey. It was an honor to have worked with you. Also, I appreciate my committee members, Dr. Suzanne Leland, Dr. Omidreza Shoghli, and Dr. Ming-Chun Lee, for your constructive feedback and insight provided to improve this work.

I am incredibly fortunate to hail from a family that holds education in high regard and has backed me throughout my academic journey- all the way from Osogbo to UNC Greensboro and finally UNC Charlotte. I owe my deepest gratitude to my parents, Jimoh and Mujidat, whose unwavering support and sacrifices have fueled my intellectual curiosity from a young age. Their dedication and encouragement have been the foundation of my journey. To my wonderful siblings, thank you for your unwavering support and constant prayers. To my very special person, Suliat, you are truly incredible. To all my amazing friends and colleagues, thank you for your encouragement and motivation. Most importantly, I want to extend a special thanks to my friends, Ismail Olasege, Joseph Udofia, and Panick Kalambay, for their exceptional support.

Finally, I would like to thank the United States Department of Transportation, Graduate School, Pulugurtha's Lab and the Infrastructure and Environmental Systems program at UNC Charlotte for funding the works that culminated to this dissertation. I am very grateful.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

ADASYN	Adaptive Synthetic
ADASYN-N	Adaptive Synthetic- Nominal
ARIMA	Autoregressive Integrated Moving Average
AUC-ROC	Area under Curve- Receiver Operating Characteristics
BART	Bayesian Additive Regression Trees
ConvLSTM	Convolutional Long Short-Term Memory
DTW	Dynamic Time Warping
DTW-G*	Dynamic Time Warping- Getis*
EB	Empirical Bayes
EPDO	Equivalent Property Damage Only
FHWA	Federal Highway Administration
HSIS	Highway Safety Information System
KDE	Kernel Density Estimation
LR	Linear Regression
LSTM	Long Short-Term Memory
MI	Moderate Injury
MNL	Multinomial Logit
MSE	Mean Square Error
MVDM	Modified Value Difference Metric
NB	Negative Binomial
NC	North Carolina
NNC	Nearest Neighbor Classification
NNH	Nearest Neighbor Hierarchical

PDO	Property Damage Only
PFI	Permutation Feature Importance
RF	Random Forest
RMSE	Root Mean Square Error
RNN	Recurrent Neural Network
ROC	Receiver Operating Characteristics Curve
SHAP	Shapely Additive exPlanation
SI	Severe Injury
SMOTE	Synthetic Minority Over-sampling Technique
SMOTE-N	Synthetic Minority Over-sampling Technique- Nominal
SVM	Support Vector Machine
XGBoost	eXtreme Gradient Boosting

CHAPTER 1: INTRODUCTION

Weather-related crashes stand out as a critical concern among the myriad challenges that necessitate rigorous study. Adverse weather, ranging from rain and snow to fog and icy roads, introduces a complex matrix of challenges that compromise traffic safety (Hambly et al., 2013) and effective traffic management (Dey et al., 2014). For example, according to a Federal Highway Administration (FHWA) report, weather-related crashes accounted for more than 21% of all vehicle crashes between 2007 to 2016 (FHWA, 2023). On the technology advancement front, technologies like weather information system (Saarikko et al., 2020), traction control system (Turner & Austin, 2000), and automatic headlight and wiper (Gaikwad & Markande) have revolutionized the approach to navigating adverse weather conditions while driving. Stakeholders have improved traffic safety while saving supply-side resources using these technologies, coupled with policies, increasing awareness, and education (Pahl-Wostl, 2007). Note that stakeholders are entities such as transportation agencies, insurance companies, healthcare institutions, and local governments, whose concerns revolve around reduction of traffic injuries and fatalities, resource allocation, and economic sustainability (WHO, 2015). For end users, such as drivers, pedestrians, and cyclists, the value of improving weather-related safety would be reduced physical and psychological toll on individuals and families (Musselwhite et al., 2021). For stakeholders, the value can be increased emergency response, effective resource allocation and reduced strain on budget (WHO, 2015; Daniel et al., 2016).

The challenges in enhancing traffic safety, particularly in addressing weather-related crashes, can be broadly categorized into three key areas: data-driven hotspot mapping, efficacy of synthetic data in machine learning, and analytics for weather-related crash prediction (Strong et al., 2010). This dissertation primarily focuses on addressing gaps in each of them assuming the backdrop of favorable policies in place. Firstly, it focuses on developing an advanced hotspot map, utilizing spatial-time cubes to identify and visualize high-crash locations, thereby aiding practitioners in targeted interventions. Secondly, the dissertation explores the effectiveness of imbalance data treatment on different machine learning techniques. This exploration is pivotal in understanding how various machine learning techniques can interpret and use artificially generated data, especially in scenarios where real-world data is rare. Lastly, the dissertation employs the use of machine learning techniques in weather-related crash prediction, an area that traditionally relies heavily on historical data and patterns. Future research, however, could delve deeper into refining predictive models for weather-related crashes, potentially revolutionizing how one prepares for and respond to these incidents.

1.1. Geospatial Risk Mapping of Traffic Crashes

The primary challenge in weather-related traffic crashes is to accurately identify and analyze hotspots where these crashes are most likely to occur (Perrels et al., 2015). Traditionally, hotspot analysis for traffic crashes has been explored using spatial data, focusing on geographical locations where crashes frequently happen (Lakshmi et al., 2019; Soltani & Askari, 2017; Songchitruksa & Zeng, 2010). However, this approach often overlooks the temporal dimension - the specific times when these crashes are more likely to happen. Weather-related crashes are not just spatial phenomena; they are also highly time-dependent, as weather conditions fluctuate over time (Ungar, 1999). Therefore, a more comprehensive analysis that includes both spatial and temporal data is needed to better understand and predict these hotspots.

In addition, there exists a substantial gap in how this geospatial information is currently used for dynamic resource planning and allocation (Robin et al., 2019). Most mapping applications provide a static view that does not account for the time factor in case of crash data. To bridge this gap, a novel approach using a spatiotemporal cube is proposed. This method involves creating a three-dimensional model where two dimensions represent the spatial aspects (latitude and longitude) of the traffic crashes, and the third dimension represents time (Nakaya, 2013). This spatiotemporal cube allows for the analysis of traffic crashes in both space and time simultaneously, providing a more holistic view of the hotspots (Nakaya, 2013; Sha et al., 2023). This dissertation aims to explore and apply a time series machine learning technique to uncover intricate and latent patterns in crash data. The clusters identified through this technique are important as they reveal patterns and trends over time, which might not be apparent when considering spatial data alone (Tavenard et al., 2020).

1.2. Effectiveness of Imbalance Data Treatment

Predicting crash severity is complicated by the imbalanced nature of crash data, where certain types of crashes, such as severe or fatal crashes, occur less frequently than others (Wen et al., 2021). This imbalance can lead to biases in traditional model outputs from machine learning techniques, making them less effective at classifying crash severity. The unpredictability and hazardous nature of weather-related crashes can result in significant challenges in traffic management and emergency response planning.

Traditional methods like under-sampling and over-sampling (Gao et al., 2021; Kim et al., 2021) have their drawbacks, such as the potential loss of important information or overfitting. Recent studies have explored sophisticated methods like the synthetic minority over-sampling technique (SMOTE) (Chawla et al., 2002) and adaptive synthetic (ADASYN) (He et al., 2008). However, these methods rely heavily on having numeric predictors in calculating distances and interpolating between data points within a multidimensional feature space to generate new instances in the minority class. In the absence of numeric predictors, or in datasets predominantly composed of nominal variables, the efficacy of these techniques is significantly constrained.

This dissertation aims to develops a technique to extend the SMOTE and ADASYN techniques and test the effectiveness of these data treatment techniques in enhancing the performance of machine learning techniques, particularly with a focus on handling datasets with predominantly nominal predictors. By assessing the effectiveness of these data treatment methods, the study seeks to provide insights into how they can improve the accuracy of models used for predicting the severity of weather-related traffic crashes.

1.3. Predicting Weather-related Crash

Weather-related traffic crashes are inherently random due to the stochastic nature of both weather events and human behavior. Weather conditions like rain, snow, fog, and ice can vary greatly in intensity, duration, and spatial distribution, making them unpredictable to some extent (Saarikko et al., 2020). This unpredictability is compounded by the random nature of human responses to these conditions (Hamdar et al., 2016; Ahmed et al., 2022). Drivers may react differently to the same weather conditions based on their individual driving skills, experience, vehicle condition, and other factors (Ahmed et al., 2022). This randomness in both weather events and driver behavior leads to a spatial point pattern of traffic crashes that is highly irregular and difficult to predict.

Due to the inherent stochastic characteristics of weather-related traffic crashes, it is challenging to predict future weather-related crash events (Khan et al., 2008). Traditional predictive models often rely on historical data and identifiable trends or patterns for prediction. However, the dynamic and rapidly changing nature of weather conditions makes them inaccurate (Ahmed et al., 2022). As a result, any prediction model would need to account for a high degree of variability and uncertainty, both in terms of the weather conditions and the resulting spatial distribution of traffic crashes. This requires sophisticated modeling techniques that can handle randomness and uncertainty, such as machine learning techniques that are capable of learning from complex and non-linear data patterns (Khan et al., 2008; Abdar et al., 2021).

This dissertation aims to develop a long short-term memory (LSTM) algorithm to retain information of weather-related crashes over extended period. Unlike traditional recurrent neural networks (RNNs), which tend to forget earlier information in a sequence, LSTMs utilize gates to control the flow of information (Abdar et al., 2021). These gates can learn which data in a sequence is important to keep or discard, enabling the model to maintain relevant information throughout the sequence of inputs.

1.4. The Interdisciplinary Perspective

This dissertation embodies an interdisciplinary approach, seamlessly blending the domains of traffic engineering, data science, and public policy. It harnesses advanced spatial-temporal geostatistical methods to unravel the complexities of traffic crash hotspots, bridging the gap between traditional traffic safety analysis and cutting-edge computational techniques. It further addresses data science problems by exploring challenges associated with imbalanced datasets through innovative synthetic data generation methods tailored for nominal predictors. The results will provide actionable insights for resource allocation and emergency response planning, emphasizing the societal impact of improved traffic safety measures viz-a-viz their policy implications. The interdisciplinary nature will not only strengthen the dissertation's foundation but also broadens its applicability and relevance across multiple sectors.

1.5. Dissertation Structure

This dissertation follows a three-paper format and is comprised of six chapters. Chapter 1 is the introductory chapter that provides an overview and overall motivation for the dissertation. Chapter 2 delves into a literature review that synthesizes the three key areas geospatial risk mapping, imbalance data treatment in machine learning, and weather-related crash prediction. Chapter 3 focuses on geospatial risk mapping, discussing the development of advanced hotspot maps using spatial-time cubes to identify high-crash locations. Chapter 4 explores the development and application of imbalance data treatment method to evaluate various data preprocessing strategies to improve the predictive performance of machine learning techniques, with a special emphasis on datasets predominantly composed of nominal or categorical predictors. Chapter 5 presents an in-depth analysis of predicting weather-related traffic crashes using advanced machine learning techniques, such as LSTM algorithms, which address the stochastic nature of these crashes. Finally, Chapter 6 concludes the dissertation, summarizing key findings, contributions to the field, and potential areas for future research.

CHAPTER 2: LITERATURE REVIEW

2.1. Weather-Related Crashes

Research in crash severity prediction, focusing specifically on weather-related traffic crashes, is vital for understanding how various factors influence the severity of such incidents. Accurate crash severity prediction is crucial for emergency response planning, resource allocation, and the development of effective countermeasures to reduce the impact of these crashes. Numerous studies have been conducted in the past to investigate the factors influencing crash severity of weather-related traffic crashes (Al-Mistarehi et al., 2022; Zhao et al., 2019; Wei et al., 2023; Das et al., 2020; Yang et al., 2022). These factors can be broadly categorized into three main groups: driver-related, vehicle-related, and environmental-related factors, with emphasis on weather conditions (Yang et al., 2022; Robin & Fotios, 2020; Hou et al., 2022). Driver-related factors include driver age, gender, impairment (e.g., alcohol or drug use), distraction, and fatigue (Dingus et al., 2016). Vehicle-related factors encompass vehicle type, size, and safety features (Strong et al., 2010; Ahmed et al., 2022). Environmental-related factors consist of road conditions, weather conditions, lighting, and traffic characteristics (Hamdar et al., 2016; Ahmed et al., 2022; Khan et al., 2008; Abdar et al., 2021; Ghahramani, 2015). Understanding the impact of these factors is crucial for developing effective crash severity prediction models.

The selection of predictor variables significantly impacts the accuracy and interpretability of crash severity prediction models (Sattar et al., 2023). Previous research has identified a wide range of potential predictor variables, including weather conditions and their interactions with driver characteristics (e.g., age and gender), roadway attributes (e.g., speed limit and road type), and other environmental conditions (e.g., lighting), and crash-specific variables (e.g., crash type and time of day) (Duddu et al., 2019; Shi et al., 2019; Yuan et al., 2019; Islam & Mannering, 2023). Das et al. (2023) identified other contributing factors such as locality and road-specific features (Das et al., 2023a; Das et al., 2023b). Feature selection techniques such as stepwise regression, principal component analysis, or recursive feature elimination have been employed to identify the most influential variables for crash severity prediction (Duddu et al., 2019; Das et al, 2023a; Das et al., 2023b).

Evaluation metrics play a crucial role in assessing the performance of crash severity prediction models. Commonly used metrics include accuracy, precision, recall, F1 score, and area under the curve - receiver operating characteristic curve (AUC-ROC) (Yuan et al., 2019). Additionally, confusion matrix analysis provides insights into model performance across different severity levels. Studies have compared the performance of different models and techniques, highlighting the strengths and limitations of each approach. Higher accuracy and AUC-ROC values indicate better model performance (Ke et al., 2017).

Accurate crash severity prediction models have practical implications for traffic safety management. For instance, emergency response systems can use predicted severity levels to proactively plan or dispatch appropriate medical personnel and resources during adverse weather conditions. Transportation agencies can prioritize weather-specific traffic safety improvements and allocate funding based on predicted crash severity hotspots (Theofilatoa & Yannis, 2014). Furthermore, crash severity prediction models can aid in the development of intelligent transportation systems and advanced driver assistance systems to prevent or mitigate risks associated with weather-related crashes.

2.2. Impact of Weather Events on Traffic Safety

Weather-related traffic crashes have a profound impact at multiple levels, ranging from individual consequences to broader societal implications (Hambly et al., 2013; Theofilatoa & Yannis, 2014). At the individual level, these crashes often result in physical injuries or fatalities, which can have a lasting impact on the victims and their families (Theofilatoa &

Yannis, 2014; Böcker et al., 2013). Injuries sustained in such crashes can range from minor bruises to severe, life-altering conditions, leading to long-term disability, chronic pain, and psychological trauma (Böcker et al., 2013). The emotional and psychological effects, such as post-traumatic stress disorder (PTSD) and anxiety related to driving or traveling in adverse weather conditions, can persist long after the physical injuries have healed. Furthermore, there are significant financial burdens associated with medical treatments, rehabilitation, and potential loss of income due to an inability to work.

From a societal perspective, weather-related traffic crashes contribute to substantial economic costs. These costs include direct expenses such as emergency response services, healthcare for injured individuals, and legal proceedings, as well as indirect costs like traffic congestion, property damage, and reduced productivity due to injury or death (Theofilatoa & Yannis). For instance, it’s estimated that weather-related crashes in the United States alone accounted for $46 billion, in 2014, in economic losses annually (FHWA, 2023). These incidents also strain public resources, with emergency services and healthcare systems often being stretched to respond effectively (Theofilatoa & Yannis, 2014; Böcker et al., 2013). The economic impact extends beyond immediate costs, affecting insurance premiums, public health services, and local economies.

Successfully addressing the challenges posed by weather-related traffic crashes would have significant implications for public policy and infrastructure planning. Improved predictive models and mitigation strategies would enable policymakers to allocate resources more effectively and design targeted safety campaigns, potentially leading to a substantial reduction in crashes and fatalities. Additionally, the findings from research on weather-related crashes could inform legislation related to driving in poor weather condition, promoting safer driving practices. Table 1 present studies illustrating how various weather conditions impact traffic safety.

Table 1. Impact of weather event on traffic safety

Weather event(s)	Impact on traffic safety	Limitations/gap	References
Cloudy	Cloudy conditions impact light levels which affect driver visibility and perception	Limited direct impact on safety compared to other weather events; however, subtle change in lighting has been overlooked in traffic safety studies	Perrels et al., 2015; Mohammed et al., 2020
Rain	Rain decreases driver visibility, reduces tire traction, and could result in hydroplaning	Difficulty in predicting crash events in raining weather condition	Perrels et al., 2015; Das et al., 2020; Mohammed et al., 2020
Snow	Snow obscures road markings, reduces friction, and reduces roadway capacity	Challenges in driver preparedness and response to snowy conditions	Strong et al., 2010; Ashifur Rahman et al., 2022; Mohammed et al., 2020
Fog, smog, smoke	Drastically reduces visibility, increases collision risk especially on high-speed roads	Prediction and real time communication to drivers are inadequate; lack of visibility impairment mitigation strategies	Perrels et al., 2015; Mohammed et al., 2020
Sleet, hail, freezing rain/drizzle	Slippery road surface; physical damage to vehicle	Weather predicting and traffic management systems are not sufficiently responsive	Mohammed et al., 2020
Severe crosswinds	Affects vehicle visibility; increases risk of overturning; increases likelihood of lane deviation	Driving training on handling crosswinds are lacking; inaccurate prediction of wind events	Mohammed et al., 2020; Sawtelle, 2020
Blowing sand	Reduces visibility, causes road abrasion and mechanical failure in vehicles; can also cover road markings and reduce traction	Impacts are localized and not widely explored; prevention and cleanup are not prioritized in non-desert locations	Mohammed et al., 2020; NRC 2004

2.3. Hotspot Identification for Weather-related crashes

The pursuit of enhancing traffic safety has led to the identification of high-risk locations, known as hotspots, which are typically determined based on specific selection criteria. Various studies aimed to refine this process, thereby bolstering the cost-effectiveness of safety programs (AASHTO, 2010). A widely accepted criterion involves analyzing expected collision frequencies at sites of interest, a method that strives to optimize system-wide safety benefits (Hauer, 1992; Saccomanno et al., 2001; Greibe, 2003; Miranda-Moreno, 2006; Cheng & Washington, 2008). Conversely, some experts advocate for considering the collision rate relative to traffic exposure to address individual road user equity (Tarko & Kanodia, 2004). This dichotomy in approach highlights the ongoing debate between system optimization and individual risk assessment in the realm of traffic safety.

Crash prediction models have traditionally served as the cornerstone for estimating expected crash frequencies. By statistically modeling crashes as a function of road characteristics, traffic volume, and weather conditions, these models partition roads into uniform sections for analysis (Hauer, 1992; Saccomanno et al., 2001; Greibe, 2003; Miranda-Moreno, 2006; Cheng & Washington, 2008). The negative binomial (NB) model, particularly within an Empirical Bayes (EB) framework, stands out for its widespread adoption due to its adeptness at capturing local safety experiences (AASHTO, 2010; Cheng & Washington, 2008). However, the success of such models hinges on the accurate specification of crash count distributions and model parameters. Any misstep in these areas can lead to the misidentification of hotspots, not to mention the substantial data collection and required model calibration efforts (Kuo et al., 2011).

Geostatistical techniques offer an alternative approach by incorporating spatial autocorrelation, recognizing the interconnectedness of crash events across the geographical landscape (Pulugurtha et al., 2007). The Kernel Density Estimation (KDE) method, for instance, has been utilized to discern the spatial patterns of crashes and pinpoint hotspots (Pulugurtha et al., 2007; Kuo et al., 2011). Other methods like K-mean clustering (Kim & Yamashita, 2007) nearest neighborhood hierarchical (NNH) clustering (Levine, 2009), and the use of Moran's I Index and Getis-Ord Gi statistics (Prasannakumar et al., 2011) also offer insights into the spatial dynamics of traffic safety, each with unique considerations of spatial correlations.

Illustratively, Anderson et al. (2009) deployed the KDE method in Turkey to unveil high-risk road sections, particularly at intersections. Keskin et al. (2011) further leveraged KDE to capture the temporal shifts in hotspots, while Khan et al. (2008) explored the specific spatial patterns of weather-related crashes. These explorations underscore the diverse impacts of different weather conditions on traffic safety and the importance of tailoring interventions accordingly. Moreover, the KDE method's relative simplicity and focus on spatial autocorrelation of crashes have made it a popular choice in traffic safety studies (Khan et al., 2008; Keskin et al., 2011).

2.3.1. History of Hotspot Identification

Over the past two decades, the examination of crash data from a spatial perspective has gained traction, highlighting significant correlations and heterogeneity of crash occurrences across different areas (Aguero-Valverde & Jovanis, 2006; Quddus, 2008). Models, such as the multivariate Bayesian hierarchical models, laid the groundwork for incorporating spatial correlations into the prediction of crash frequencies, leading to enhanced model performance (Aguero-Valverde & Jovanis, 2010). The application of the geographically weighted Poisson regression model, like by Xu & Huang (2015), marked a progression in accounting for spatial variability in crash data analysis.

Recently, a shift towards incorporating both spatial and temporal data points has emerged, reflecting an increasing interest in the dynamics of crash frequency and risk (Mannering & Bhat, 2014; Wu et al., 2023). These advanced studies suggest that the influence of nearby locations on a specific crash spot is a critical component often overlooked in traditional models. However, such detailed spatiotemporal analysis, especially on fine-grained data, poses a challenge due to its complexity and time-intensive nature (Cai et al., 2019). To navigate these complexities, contemporary research has pivoted towards leveraging sophisticated machine learning techniques, such as Convolutional Neural

Networks (CNNs), demonstrating their efficacy in discerning spatial correlations within high-resolution datasets (Cai et al., 2019). This reinforces the importance of considering spatial location as sets of monocytic homogenous grids to summarize the crash risk profile of an area, an idea that remains to be thoroughly explored.

Factors influencing weather-related crashes can be broadly categorized into road characteristics, environmental condition, traffic flow, traffic management, and driving behavior. Environmental condition is an important factor that requires more studies, especially looking into how weather conditions influence crash risk. These conditions are systematically categorized in Table 1 as examined in past studies. Yet, there remains a gap in understanding how these factors interplay with the spatial and temporal landscape of crash risk, as most research has focused on isolated locations without considering the broader context.

Historically, the identification of crash hotspots has been refined through various scientific methodologies. The Empirical Bayesian approach was initially heralded as the leading method for this purpose (Guo et al., 2019). It was later eclipsed by Full Bayes hierarchical models, which offered greater accuracy in identifying crash hotspots (Guo et al., 2019). Incorporating equivalent property damage only (EPDO) crashes expanded on this by integrating crash frequency and severity into a cohesive risk assessment (Ma et al., 2016). With advancements in spatial statistics, KDE and emerging hotspot analysis became prominent tools for visualizing and identifying crash risk patterns (Plug et al., 2011; Chainey, 2013). Nevertheless, the limitations of these methods, particularly in terms of high false discovery rates, have been noted as potential sources of inaccuracies in identifying true hotspots (Songchitruksa & Zeng, 2010; Ogungbire et al., 2023).

The surge of application of machine learning techniques in traffic safety represents a pivotal shift towards more precise and interpretable risk prediction. Models employing vehicle trajectory data, spatiotemporal dynamics, and advanced tree-based algorithms have been introduced, providing nuanced insights into crash risk prediction (Bao et al., 2019; Gao et al., 2023). Interpretability frameworks such as SHAP and LIME have transformed these complex models into more transparent systems, enabling a deeper understanding of the influential factors in various crash types (Wen et al., 2021; Amini et al., 2022; Veran et al., 2023).

2.3.2. Towards a Spatiotemporal Analysis

Despite the advancements in geostatistical methods, there remains room for innovation, particularly in incorporating the temporal aspect of crash data. Current methods predominantly focus on spatial analysis, which can overlook the temporal patterns that are equally crucial in understanding crash dynamics. This gap provides the motivation for a novel approach: the spatiotemporal cube (Nakaya, 2013).

The spatiotemporal cube method allows for a comprehensive analysis of traffic crashes by considering both space and time dimensions simultaneously. This approach could offer a more dynamic and detailed understanding of crash hotspots, revealing not only where but also when crashes are most likely to occur. By integrating temporal data with spatial analysis, this method could identify patterns over time, such as seasonal variations or the impact of temporary road conditions, providing a richer context for safety interventions. This novel approach has the potential to revolutionize hotspot identification and significantly enhance traffic safety strategies.

2.4. Handling Data Imbalance in Crash Data

A huge amount of crash data is generated year in, year out, often exhibiting a skewed distribution (Kim et al., 2021). This skewness occurs when one category of crash severity-typically PDO crashes- far outnumbers more severe crashes. In this context, the prevalent minor crashes represent the ‘majority class,’ while the less frequent but more severe crashes are the ‘minority class’. The challenge lies in the fact that most standard predictive models are inclined to favor the majority class due to its larger representation in the data, leading to poor prediction performance for the minority class, yet more critical, severe crash instances (Chawla et al., 2002).

To address this imbalance, various strategies have been developed, categorized into algorithmic modifications, data preprocessing, and feature selection techniques (Yijing et al., 2016; Maldonado & López, 2018; Roy et al., 2018). Data preprocessing is one approach, where the data is manipulated before feeding it into the model. This can involve over-sampling, where synthetic severe crash instances are generated to bolster the minority class (Gao et al., 2021), or under-sampling, where instances of the minor crashes are selectively removed to balance the classes (Kim et al., 2021). Another tactic is to refine or create new algorithms that are sensitive to the costs of misclassification, like cost-sensitive learning, and utilize advanced techniques such as kernel-based learning, like support vector machines (SVMs) (Tao et al., 2019).

2.4.1. Approaches for Addressing Imbalance Problem in Crash Data

The phenomenon of class imbalance has been identified as a significant challenge within the realm of statistical modeling and machine learning techniques. The disproportionate representation of classes often results in the dominance of the majority class, thereby undermining the reliability of the predictions pertaining to the minority class. This accentuates the pivotal role that data plays in these modeling frameworks, given their data-dependent nature. This aspect is particularly salient within the context of traffic crash data analysis. Infrequent crashes, while less common in the dataset, are typically associated with higher severity and concomitant socioeconomic costs (Das et al., 2023a). Consequently, the accurate prediction of these minority class crash events assumes critical importance (Wei et al., 2023).

Iranitalab et al. (2017) undertook a comparative analysis of four distinct statistical and machine learning techniques, namely Multinomial Logit (MNL), nearest neighbor classification (NNC), SVM, and random forest (RF), in the context of traffic crash severity prediction. The findings revealed that machine learning techniques, encompassing NNC, SVM, and RF, generally outperformed the traditional MNL model in terms of prediction accuracy. Nonetheless, a common challenge encountered by all four models pertained to the classification of infrequent severe injury crashes, such as those resulting in severe injury or fatal crashes.

To address the prevalent issue of class imbalance in crash severity analysis, minority over-sampling methods such as the SMOTE and the ADASYN have been developed and explored in the past (Chawla et al., 2002; He et al., 2008). SMOTE generates synthetic instances of the minority class by employing a bootstrapping approach in combination with the k-nearest neighbors' algorithm, which has been widely used in traffic safety analysis. Similarly, ADASYN adopts a density distribution-based measure to determine the required number of samples from the minority class, in contrast to SMOTE's uniform weight assignment. Besides over-sampling the minority class, under-sampling of the majority class has also been incorporated in crash analysis. For instance, Fiorentini & Losa (2020) utilized a random under-sampling of the majority class strategy to develop models predicting crash type, specifically focusing on two levels of collision severity: PDO and fatal + injury crashes. Their findings demonstrated that the random under-sampling of the majority class significantly improved the prediction performance for the minority class compared to the model trained on imbalanced data. Table 2 provides a summary of selected weather-related crash severity studies (Call et al., 2019; Ghasemzadeh & Ahmed, 2019; Mondal et al., 2020;

Zeng et al., 2020; Rahman et al., 2022) that have explored machine learning techniques in the past.

In summary, the issue of class imbalance is a major issue in case of statistical methods and machine learning techniques, which can be addressed by using data treatment methods like SMOTE and ADASYN. However, there is a challenge when the dataset has predominantly nominal predictors with few or no numeric predictors. This study aims to address a critical gap in the existing literature by introducing a novel technique for synthetic data generation specifically tailored for nominal predictors in the context of crash severity analysis.

Table 2. Summarization of previous weather-related crash severity studies

Authors	Independent variables	Model used	Findings
Mondal et al., 2020	Manner of crash, weather condition, route class, hour of the day, type of intersection, light condition, road surface condition, work zone related, day of the week, and school bus	Random Forest (RF) and Bayesian Additive Regression Trees (BART)	RF model was better at predicting crash severity than the BART model; performance of the RF model was found to be very good, with a higher skill score of 0.73 compared to 0.61 for BART
Zeng et al., 2020	Driver type, vehicle type, vehicle registered province, crash time, crash type, response time of emergency medical service, and horizontal curvature and vertical grade of the crash location	Bayesian Spatial Generalized Ordered Logit Model	An increase in precipitation is associated with a decrease in the probability of light and severe crashes, and an increase in the probability of medium crashes
Call et al., 2019	Elevation, slope, location, time, and severity, as well as over twenty-five other fields related to the crash, such as single-vehicle, weather-related, and driving under the influence (DUI)	Logistic Regression	Adverse weather-related crashes were most common in the winter season and correlated with snowfall; Excessive speed was more likely in these crashes, but they were generally less severe; Roadway slope was also a factor, with slight increases increasing the likelihood of crashes
Ghasemzadeh & Ahmed, 2019	Speed, age, land use, crash type, DUI, lighting condition, weather, road type, traffic control devices, vehicle age, vehicle type, construction type, location, and surface condition	Probit–Classification Tree	Presence of traffic control device and lighting conditions are significant interacting variables in the developed complementary crash severity model for work zone weather-related crashes

Authors	Independent variables	Model used	Findings
Rahman et al., 2022	Lightning condition, AADT, roadway type, functional class, area type, roadway alignment, shoulder width, and posted speed limit	Cluster Correspondence Analysis (CCA)	Variety of attributes linked to speed limit, lighting condition, alignment, area type, manner of collision, restraint usage, and alcohol/drug can have an Influence on fatal/severe injury crashes and moderate injury crashes in the State of Louisiana

It is hypothesized that these data treatment techniques will improve the prediction accuracy when trained on machine learning techniques compared to the control/raw dataset. The proposed technique is expected to improve accuracy of weather-related crash severity classifications. The findings from this study serve as valuable insights into the impact of these techniques on the accuracy and robustness of machine learning techniques when applied to crash data, contributing to more reliable and effective crash severity analysis methodologies.

2.5. Weather-Related Crash Prediction

Crashes are a result of multiple factors that can be categorized into behavioral, technological, and environmental influences (Ogungbire et al., 2023). While weather is not the primary cause of road crashes (FHWA, 2023; Ogungbire et al., 2023), its significance cannot be overlooked. Studies, including those by Vickery (1996) and Downs (2000), indicate that most people do not consider poor weather as a deterrent to driving unless conditions severely impede travel. Bergel-Hayat et al. (2013) have established a correlation between weather conditions and road transport, detailing how adverse weather can lead to inconvenience or even compel travelers to cancel their travel plan.

Vehicles, unlike other modes of transport, are generally not designed to operate under extreme weather conditions. The impact of bad weather on traffic safety is complex and cannot be reduced to simple cause and effect. Research by Jackson & Sharif (2016) shows that rain increases crash rates, a situation exacerbated by more people choosing to drive under wet conditions. However, the introduction of technologies such as anti-lock brakes and traction control has changed the dynamics of driving in poor weather, potentially leading to riskier driving behaviors as drivers gain confidence from these features (Smiley & Rudin-Brown, 2020).

Weather-related challenges do not always lead to severe crashes. In some situations, like snow, drivers tend to be more cautious, reducing their speed and thus mitigating risk. Decisions to cancel or postpone travel plans can also decrease the likelihood of crashes during unfavorable weather conditions (Kilpeläinen & Summala, 2007).

2.5.1. Crash Prediction Using Traditional Models

Table 3 summarizes example crash prediction techniques. The evolution of traffic crash predicting using traditional models shows a shift from linear statistical models to more dynamic and complex computational models (Duddu & Pulugurtha, 2017; Pulugurtha et al., 2013; Gajera et al., 2023; Kalambay & Pulugurtha, 2023; Pulugurtha & Mahanthi, 2016; Najaf et al., 2018; Khan et al., 2022; Feng et al., 2020; Iranitalab & Khattak, 2017). The integration of different prediction techniques, such as combining grey models with Markov chains or enhancing autoregressive integrated moving average (ARIMA) with neural network analysis for the non-linear components, exemplifies the interdisciplinary approach towards a more accurate and robust prediction of traffic incidents.

Time-series prediction methods have been foundational in predicting traffic crashes, utilizing historical data to estimate future outcomes (Khan et al., 2022). These methods consider the sequence of data points collected over time, analyzing patterns such as long-term trends, seasonality, and irregular factors. Key approaches within time-series prediction include the exponential smoothing method (Khan et al., 2022; Rabbani et al., 2021), which emphasizes the diminishing significance of older data, and ARIMA, a model that integrates differencing of observations (to remove non-stationarity) with autoregression and moving averages (Rabbani et al., 2021; Khan et al., 2022). These methods are based on the premise that past patterns in traffic crash data can offer insights into future occurrences, with techniques like exponential smoothing and ARIMA being particularly noted for their ability to model and predict traffic crashes (Hassouna & Al-Sahili, 2020; Rabbani et al., 2021; Khan et al., 2022).

Exponential smoothing models, including the simple exponential smoothing and its extensions to account for trend and seasonality (Rabbani et al., 2021), provide a framework for smoothing out time series data to identify underlying trends. The sophistication of these models lies in their statistical rationale, which accommodates various forms of trends and seasonality through state-space models (Hassouna & Al-Sahili, 2020). ARIMA and its precursor, autoregressive moving average (ARMA), further advance the field by addressing the stochastic properties of time series and facilitating model selection based on the stationary characteristics of the data. These models have been applied to correct error terms in traffic crash prediction, combining with other methodologies like regression models to enhance the reliability of predictions. Markov chain models introduce a probabilistic approach to predicting, emphasizing the transition probabilities between discrete states over time (Pei et al., 2011). This method suits scenarios with significant random fluctuations but without clear trends, offering insights into the stochastic nature of traffic crashes.

Table 3. Comparison of crash prediction techniques

Authors	Methodologies	Prediction range	Features used	Major improvement in literature
Duddu & Pulugurtha (2017)	Linear regression + back propagation neural network	Short-term	On network characteristics, and land use characteristics	Link-level crash frequency model was developed.
Loo et al. (2023)	Random Forest (RF) + XGBoost + Naïve bayes Negative Binomial) NB)	Short-term	Pedestrian exposure factors, pedestrian jaywalking, and bus stop crowding	Both XGBoost and RF models generated similar results on feature importance for three sets of models. Also, there are non-linear relationships of many risk factors with bus crashes.
Formosa et al. (2020)	Regional convolutional neural network (R-CNN)	Short-term	Speed, vehicle sensor data (yaw rate, velocity, longitudinal displacement, etc.), headway, and occupancy	Predict traffic conflict by integrating and mining heterodox data.
Rabbani et al. (2021)	Seasonal autoregressive integrated moving average (SARIMA) + ES	Medium-term	Historical crash frequency was used to predict future crash risk	Exponential smoothing model has a better fit on crash data over SARIMA.
Cai & Di (2021)	Autoregressive integrated moving average (ARIMA) + boosting	Short-term	Lane traffic flow, weather information, vehicle speed, and truck to car ratio	Integrating time series with a count data model can capture traffic crash features and account for the temporal autocorrelation.
Ivan (2004)	Bayesian framework	Short-term		Shows how traffic volume can be used in crash rate analysis.
Ladron de Guevara et al. (2004)	NB	Short-term	Population, TAZ area, number of housing units, number of schools, total miles of bus routes, and miles of bike routes	Developed a model to predict crashes for equitable planning. In addition, the model would help state agencies in establishing incentive programs to reduce injuries and fatalities.
Huang et al. (2019)	Deep dynamic fusion	Short-term & long-term	Time, road condition, illegal parking, unsanitary condition, and blocked drive	Improved the ability of deep neural network in modeling heterogeneous conditions in a fully dynamic ways for traffic crash prediction.

2.5.2. Deep Learning Models for Weather-Related Crash Prediction

Deep learning, a subset of machine learning techniques characterized by its use of neural networks with multiple layers, has emerged as a powerful tool in the realm of traffic crash prediction (Formosa et al., 2020; Huang et al., 2020; Bibi et al., 2021; Valcamonico et al., 2022; Loo et al., 2023). The application of deep learning in traffic crash prediction in literature has been restricted to specific models such as CNNs, recurrent neural networks (RNNs), and their variations such as LSTM networks and gated recurrent units (GRUs). These architectures are adept at handling the spatial and temporal data inherent in traffic systems, allowing for the modeling of complex patterns and relationships that traditional models might overlook.

CNNs have proved effective in processing spatial data, making them suitable for analyzing crash data aggregated by geographical units, such as grid maps (Loo et al., 2023). By capturing spatial dependencies through their convolutional filters, CNNs can identify patterns related to traffic flow, road infrastructure, and other spatial factors contributing to crash risks (Huang et al., 2020). On the other hand, RNNs and LSTMs, are designed to handle sequential data, allowing for temporal dynamics of traffic crash occurrences. These models can learn from historical crash data, recognizing patterns over time, such as the cyclic nature of traffic volume and its correlation with crash incidents (Shi et al., 2015). In addition, LSTMs, can remember long-term dependencies, are particularly effective in overcoming the vanishing gradient problem common in traditional RNNs.

2.6. Overview of Prior Research Limitations and Dissertation Contributions

The literature review was conducted to scrutinize the limitations of past research and advance the field of weather-related crash severity analysis. The past research has focused on identifying crash hotspots through traditional crash prediction models and geostatistical techniques (Pulugurtha et al., 2007; Kim & Yamashita, 2007; Khan et al., 2008; Levine, 2009; Anderson, 2009; Kuo et al., 2011; Prasannakumar et al., 2011; Keskin et al., 2011). Nevertheless, these methods did not consider the temporal dimension of crashes and require substantial data collection, which was addressed through innovative spatiotemporal analysis in this dissertation.

Past studies have also identified numerous factors that influence crash severity, including driver, vehicle, and environmental variables, with an emphasis on weather conditions (Zhao et al., 2019; Das et al., 2020; Al-Mistarehi et al., 2022; Yang et al., 2022). However, there is a gap in effectively predicting crash events and addressing their multifaceted impacts. This dissertation presents a robust variable selection technique in crash severity prediction and explores data imbalance treatment techniques to improve crash severity prediction models.

Overall, key gaps were identified to set the stage for contribution to the body of knowledge, which involves enhancing the precision of hotspot identification with spatiotemporal analysis, proposing novel techniques for handling nominal predictors in crash data and developing advanced methodologies for traffic crash prediction.

CHAPTER 3: A SPATIOTEMPORAL RISK MAPPING OF STATEWIDE WEATHER-RELATED TRAFFIC CRASHES: A MACHINE LEARNING TECHNIQUE

3.1. Introduction

The dynamics of statewide transportation safety emphasize the impact of weather-related crashes. According to the Federal Highway Administration (FHWA), between 2007 and 2016, over 21% of all crashes were attributed to adverse weather conditions (FHWA, 2023). These weather-related crashes tend to have a higher severity, most of them resulting in fatal outcomes. In addition to this, weather-related crashes in the United States accounted for $46 billion, in 2014, in economic losses (FHWA, 2023). Thus, state agencies are constantly making efforts to roll out strategic initiatives and policies with the aim of improving traffic safety and ensuring resilient transportation infrastructure (HSIP, 2010; PSC, 2008). Central to these efforts is the accurate identification of high-risk zones and a deep understanding of the factors that influence these areas.

Traffic safety units of state DOTs have recognized the profound value of analyzing the spatial characteristics of crash data. This appreciation stems from the critical role such data plays in shaping statewide safety planning and management efforts (Brown 2016; Huang et al., 2014). Equipped with detailed locational insights, this data fosters research that delves into the spatial trends of weather-related crashes, enabling comprehensive safety evaluations at a granular level (Cai et al., 2019).

Spatial grid methodologies have been embraced in recent traffic safety assessments across the state (Bao et al., 2019; Cai et al., 2019; Wu et al., 2023). By defining these grid dimensions, state planners can flexibly assess weather-related crash patterns across various spatial scales. These grids, with their fine granularity, offer a detailed lens for evaluating traffic safety, while also capturing diverse crash-related attributes, ranging from road characteristics patterns to urban infrastructures.

The challenge of identifying crash-prone zones, especially when mapped across time and space, has gained traction in recent state-led research endeavors. By identifying these zones, state transportation agencies are better positioned to devise and implement targeted interventions, thereby optimizing traffic safety outcomes (Drawve, 2019; Wu et al., 2023). Yet, there exists a gap in understanding the myriad factors contributing to these high-risk zones. Often, the interconnectedness of spatial-temporal attributes of surrounding areas with these zones remains underexplored. This interplay can be best understood through the lens of zonal spillover effects, where external factors at one location can have a ripple effect, impacting both the focal and adjacent areas (Zhang et al., 2022). Such effects arise due to the inherent continuity and interconnectedness observed in parameters like road designs and traffic flow dynamics.

Given the huge number of factors influencing statewide crash risks, traditional statistical methods are often too simple to identify the important factors. However, machine learning techniques offer state DOTs the agility and precision to swiftly identify key risk factors from a high dimensional feature set (Santosh & Gaur, 2022; Yuan et al., 2022). Even though past studies have considered the combination of space and time in geospatial risk mapping theory, and these days, spatial data often include a time component, only a few studies have studied this in the context of weather-related crashes. The problem may be due to the complexity introduced by the heterogeneity of weather-related crash data.

The aim of this study is to shed light on how spatiotemporal point process data using geographic location and times of individual traffic crashes can be used to identify weather-related crash hotspots. Here, weather-related crash data are considered as a realization of spatiotemporal point process that lacks both spatial and temporal homogeneity, i.e., the expected number of crashes in each area units depends on their location and time. This helps identify key spatiotemporal dynamics influencing weather-related crash hotspots at the grid level.

The dissertation introduces a novel model for identifying statewide crash risks, weaving together space-time analytics and cutting-edge machine learning techniques. The nuanced effects of identified key factors on crash-prone zones are carefully explained using interpretable machine learning techniques. Drawing from these findings, state agencies can carve out informed policy recommendations, setting the stage for a safer transportation landscape. The subsequent sections will delve deeper into related literature, the proposed model's intricacies, results interpretation, and implications, and culminate with a study summary and conclusion. The contribution of this step is as summarized next.

a) In a statewide weather-related crash risk identification and prediction, the crash risk is not contained to one spatial region or time but spreads across multiple dimensions (Chen et al., 2016). The challenge lies in understanding how weather conditions influence crash risks over the vastness of a state, considering the data generation point process. The non-homogeneous nature of weather-related crashes that vary across space and time is addressed by advancing the current methodologies which assume that crash data is uniformly distributed across a region. An architecture designed to systematically detect intricate patterns in the vast array of grids using an unsupervised machine learning technique is employed.

b) Previous studies have developed techniques in visualizing crash hotspots without explaining factors responsible for crash hotspot (Lee & Khattak, 2019; Plug et al., 2011). This part of the dissertation builds on existing studies to further examine factors responsible for both hotspot and coldspot using a supervised machine learning technique. In addition, the interaction among these risk zones and weather conditions is examined.

3.2. Study Design & Workflow of Developed Techniques

The workflow depicted in Figure 1 describes a process for analyzing weather-related crash data. Initially, multiple datasets detailing cases of weather-related crashes are compiled into a space-time cube, which organizes the data across both spatial and temporal dimensions. This cube enables time series clustering, a method that groups similar patterns over time, to identify crash hotspots. These hotspots are then categorized and labelled as areas of high and low EPDO per mile using the method of integrating Getis-Ord Gi* statistics and dynamic time warping (DTW) as detailed in Algorithm 1. Subsequently, an advanced machine learning technique known as XGBoost, coupled with SHapley Additive exPlanations (SHAP), is used to integrate, and analyze the clustered data. The SHAP values explains which factors are most influential in predicting areas of high and low crash risk. The final step is the identification of risk factors for both crash-prone and low crash areas.

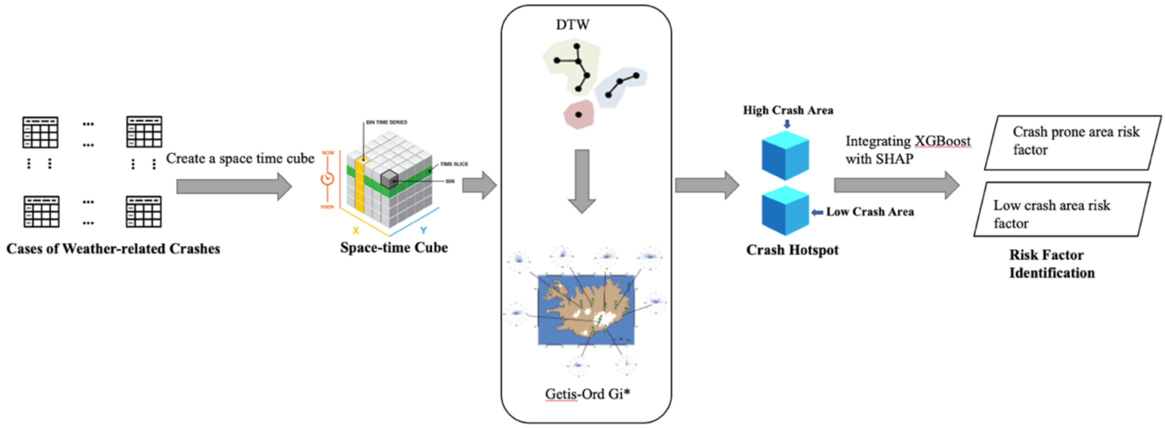

Cases of Weather-related Crashes
Create a space time cube
BIN TIME SERIES
NOW
THEN
TIME SLICE
BIN
X
Y
Space-time Cube
DTW
Getis-Ord Gi*
High Crash Area
Low Crash Area
Crash Hotspot
Integrating XGBoost with SHAP
Crash prone area risk factor
Low crash area risk factor
Risk Factor Identification

Figure 1. Data processing workflow with a multi-layered crash hotspot identification technique

3.2.1. Principle of Time Space Cubes

Consider a random collection of weather-related crash points $X = \{(u_i, t_i)\}_{i=1}^{n}$ observed within a bounded space of plane, $Z \subset \mathbb{R}^2$, where u_i represent the spatial location of the i^{th} crash event and t_i is the time with a positive interval $T \subset \mathbb{R}_+$. Utilizing the crash occurrence time and its geographical coordinates ($W \times T$), the spatiotemporal pattern mining capabilities of the algorithm is used to construct a spatiotemporal cube, capturing both time and spatial nuances, as illustrated in Figure 2. Each cell within this cube corresponds to a distinct spatial (x, y) and temporal (t) coordinate. Traffic crashes within a given cell are assigned a shared location identifier. For the purposes of this research, the spatial base of each cell is configured as a square spanning 5mi × 5mi, with a temporal granularity set at one month. Viewed temporally, this cube can be dissected into multiple time series, differentiated primarily by their spatial coordinates.

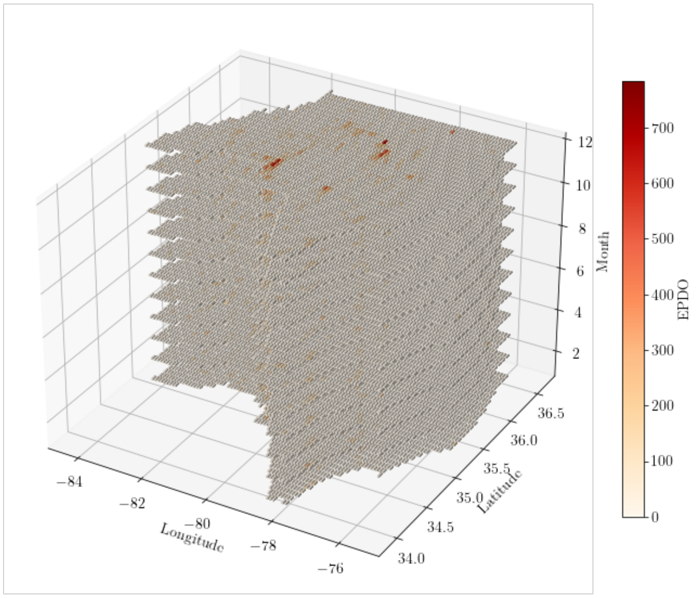


Figure 2. Space-time cube of the case study

3.2.2. Integrating Dynamic Time Warping (DTW) and Getis-Ord Gi* Statistics (DTW-G*) for Crash Risk Labeling

Integrating Getis-Ord G_i^* and DTW may be applied to essentially any crash hotspot learning framework. The exact idea behind the development of the algorithm is explained. The contribution is a simple but effective fix to a problem that will otherwise plague any attempt at identifying hotspot for heterogeneous crash point datasets. The integration of Gestis-Ord G_i^* and DTW is conceptualized as a two-layer model: a) the temporal layer that constitute regions of similar temporal pattern in weather-related crash events, and, b) the spatial layer that identifies which region within the groups of temporally similar regions are spatial hotspots.

The DTW is a method to measure similarity between two temporal sequences. For time series *X* and *Y*, DTW finds the optimal alignment between these sequences to minimize the total distance between them. Assume there are two sequences as shown in Figure 3a, representing the *EPDO* of monthly weather-related crashes, Sequence *X* with length n, $X = [x_1, x_2 \ldots, x_i, \ldots, x_n]$ and sequence *Y* with length *m*, $Y = [y_1, y_2 \ldots, y_j, \ldots, y_m]$, one can create an *m*-by-*n* path matrix where the (i, j)th element of the matrix contain the distance between two points x_n and y_m as shown in Figure 3b. The distance $d(x_i, y_j)$, is calculated using $L_p\ norm, \|\, x_i - y_j \,\|_p$ which measures the difference in EPDO per mile for month *i* in location *X* and month j in location *Y*. The minimum distance path after the alignment of two sequences is recorded as the best match.

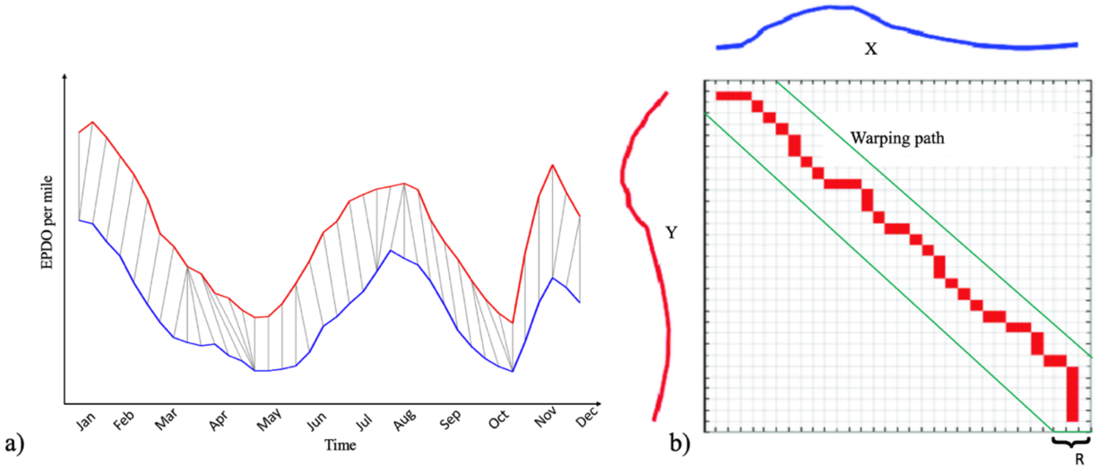


Figure 3. a) Two similar time series that are out of phase, b) a warping matrix and search for optimal warping path (red squares)

Thus, the optimal warping path can be computed by using the recursive formula in Equation 1.

$$DTW(X, Y) = \sqrt{D(i, j)} \qquad (1)$$

where $D(i, j)$ is the cumulative distance as shown in Equation 2.

$$\mathrm{D(i,j) = d(x_i, y_i) + min\{D(i-1, j-1), D(i-1, j), D(i, j-1)\}} \qquad (2)$$

The objective of DTW is to find warping path through this matrix that minimizes the total distance subject to the following constraint:

- **Endpoint Constraint:** The warping path must start at the first month's EPDO for both location ($u_1 = (x_1, y_1)$) and end at the last month's counts ($u_k = (x_k, y_k)$) making sure the entire time span for both sequences is captured in the comparison.
- **Continuity Constraint:** The path advance in one step increment in either sequence, i.e., if at step k the path is (x_i, y_j), then at step $k + 1$, it can only move to (x_{i+1}, y_j), (x_i, y_{j+1}), or (x_{i+1}, y_{j+1}).
- **Monotonicity:** The path must always move forward for both sequences to show the chronological progression of time i.e., if the path is at (x_i, y_j) at step k, then for step $k + 1$, it must move to a point (x_{i+1}, y_{j+1}) where i and j are non-decreasing.

Algorithm 1 is then designed to identify hotspots of weather-related crashes within a 3D grid by calculating a Getis-Ord G_i^* score for each cell *i* within distinct temporal cluster group *k* using equations 3 to 5. This score x_j is based on the values of neighboring cells *j*, assuming each neighboring cell contributes equally (with a weight $w_{i,j}$ of 1) to the score. A cell is considered a neighbor if it touches another cell at any point, meaning a cell can have up to 26 neighbors in the middle of the grid, 11 if it is on the edge, and 7 if it is in a corner. The G_i^* score essentially tells us how unusually high or low a cell's value is compared to its neighbors, using a method that turns these comparisons into a z-score, a statistical measure that indicates the difference from the average in units of standard deviation.

$$G_{k,i}^* = \frac{\sum_{j=1}^{n} w_{i,j}\, x_j - \bar{X} \sum_{j=1}^{n} w_{i,j}}{s\sqrt{\frac{[n \sum_{j=1}^{n} w_{i,j}^2 - (\sum_{j=1}^{n} w_{i,j})^2]}{n-1}}} \qquad (3)$$

where, $\bar{X} = \dfrac{\sum_{j=1}^{n} x_j}{n}$, (4) and $s = \sqrt{\frac{\sum_{j=1}^{n} x_j^{2}}{n} - \bar{X}^2}$ (5)

Algorithm 1: Spatiotemporal Hotspot Identification

```
Input: temporal sequence $\{S_1, S_2 \ldots, S_n\}$, each sequence $S_i$ representing the occurrence of weather-related crash event over time for region i
Output: Group of regions with similar temporal pattern and their hotspots significance scores
Compute temporal similarity using DTW
    Initialize $DTW_{distance}[n][n]$ to store DTW distance
for each unique region $(I, j)$, where $(i \neq j)$ and $(I, j)$ in {1, …, n}
    $DTW_{distance}[i][j] \leftarrow DTW(S_i, S_j)$
    A threshold θ is defined to distinguish between similar and dissimilar temporal patterns
    A list ∀ is initialized to store groups of regions based on their temporal similarity
    for each region $i$ in {1, …, n}
        for each region $j$ in {1, …, n} and $j \neq i$
            if $DTW_{distance}[i][j] < \theta$
                Add $i$ and $j$ to the same group ∀
Initialize list ℶ to store and compare significant hotspots within groups using Getis-Ord $G_i^*$
    for each group $G_k^*$, compute $G^*$ statistics as $G_i^* \leftarrow GetisOrd(S_i)$
        for each region $i$ in $G_k^*$
            if $G_i^*$ is significant within group $G_k^*$
                Add $i$ to ℶ with significant score $G_i^*$
end algorithm
```

3.2.3. Risk Factor Identification Using XGBoost

XGBoost is used to predict the risk pattern at different locations in a study area by predicting the labels as classified in the prior step. This algorithm leverages the ensemble technique of gradient boasting trees (Chen & Guestrin, 2016). Let X denote the input matrix $n \times m$, where n is the number of samples and m is the number of dimensions. The prediction outcome i.e., the labelled risk is given by the sum of the output from K trees, each represented by a function f_k on the input matrix $\mathbf{X}$ as shown in Equation 6.

$$\hat{y} = \phi(\mathrm{X}) = \sum_{k=1}^{K} f_k(\mathbf{X}) \quad (6)$$

where $\hat{y}$ is an n-dimensional vector of the predicted outputs.

The objective function L, combining both the loss and regularization term, can be expressed as Equation 7.

$$L = \sum_{i=1}^{n} l(\hat{y}_i, y_i) + \sum_{i=1}^{n} \Omega f_k \tag{7}$$

For the regularization term Ωf_k, it is applied to each tree and can be seen as a sum of penalties on the complexity of each tree. The process of modifying the model iteratively, incorporating one tree at a time is given by Equation 8.

$$\hat{y}^{(t)} = \hat{y}^{(t-1)} + f_t(\mathbf{X}) \tag{8}$$

where $f_t(\mathbf{X})$ is the combination of the t^{th} tree to the prediction.

For the optimization of the objective function using a second-order Taylor expansion, the objective function $L^{(t)}$ after dropping the constant becomes Equation 9.

$$L^{(t)} = \sum_{i=1}^{n} \left[g_i f_t(\mathbf{X}_i) + \frac{1}{2} h_i f_t(\mathbf{X}_i)^2 \right] + \Omega f_t \tag{9}$$

where g and h are vectors of the first and second derivative statistics of the loss function with respect to the predictions $\hat{y}^{(t-1)}$, and $f_t(\mathbf{X}_i)$ represents the contribution of the t^{th} tree to the prediction for sample *i*.

3.2.4. SHAP for Key Risk Factor Interpretation

The concept of SHapley Additive exPlanations (SHAP) values, introduced by Lundberg & Lee (2017), draws upon the foundational Shapley value estimation method, originally proposed by Shapley (1953). SHAP values provide a mechanism to quantify the importance of each feature by calculating its contribution to the prediction outcome when included in the model. The computation of SHAP values, serving as feature attributions, entails a weighted average of all conceivable differences in predictions, as outlined below in Equation 10 in a more generalized matrix form.

$$\phi_i = \sum_{S \in F_{\{i\}}} \frac{|S|!(|F|-|S|-1)!}{|F|!} [f_{S\cup\{i\}}(\mathbf{X}_{S\cup\{i\}}) - f_S(\mathbf{X}_S)] \tag{10}$$

Here, ϕ_i represents the SHAP value for the i^{th} feature, F denotes the full set of features, S is a subset of features excluding the i^{th} feature, $f_{S\cup\{i\}}$ and f_S signify the model predictions with and without the i^{th} feature, respectively, and $\mathbf{X}_S$ refers to the input values for the features in the set S.

3.3. Case study of North Carolina (NC)

The proposed workflow is demonstrated using the state of North Carolina (NC) as the study area. The traffic crash data used in the study comes from the Highway Safety Information Systems (HSIS) database. This data is the general crash data that occurred in the study area from January 1, 2015, to December 31, 2018. Figure 4 provides a visual analysis of weather-related traffic crashes and traffic volume in NC. The left map is a map detailing the severity of crashes, with different colors indicating the severity level: PDO, minor injuries, and severe injuries. This map is dense with PDO crashes, displayed in blue, and severe crashes displayed in red. The right map displays traffic volume intensity with a color scale ranging from yellow to red, highlighting the areas with the highest traffic density.

Different weights based on societal and economic value of crashes in NC are used to compute the EPDO values as shown in Equation 11 to account for severity level of crashes in the density estimation. Weight of fatal, injury and PDO crashes are set to 249.8, 11.6 and 1 respectively (NCDOT, 2019).

The selected dataset is comprised of 2,045 cubes. The description of features extracted within each spatial cube is as presented in Table 4.

$$EPDO_i = \frac{249.8 \sum_{year=1}^{4}[SI]_i + 11.6 \sum_{year=1}^{4}[MI]_i + \sum_{year=1}^{4}[PDO]_i}{L_i} \tag{11}$$

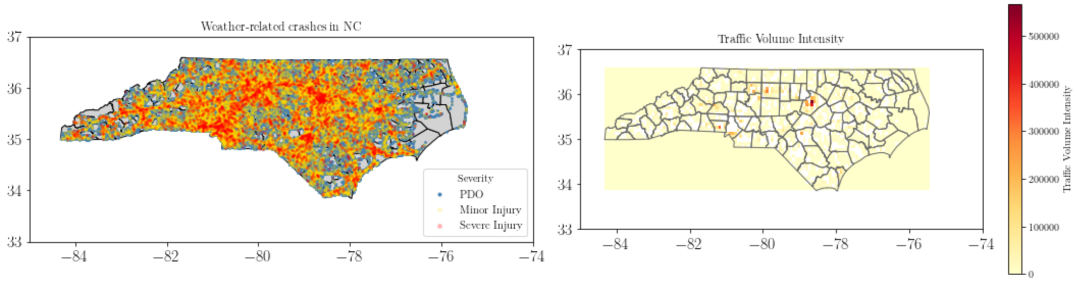


Figure 4. Left: Location of 238,252 weather-related crashes by severity level. Right: NC population density averaged over years 2015-2018

Table 4. Description of features extracted within spatial cubes

Feature Type	Feature Name	Categories	Feature Code	Description
Road	Lanes	-	no_lanes	Average number of lanes in each cube
	Left shoulder width	-	lshldwid	Average left shoulder width in each cube
	Right shoulder type	-	rshldwid	Average right shoulder width in each cube
	Route type	Interstate	rte_type_1_ratio	Proportion of interstate routes within each cube
		US route	rte_type_2_ration	Proportion of US routes within each cube
		NC route	rte_type_3_ratio	Proportion of NC routes within each cube
		Secondary route	rte_type_4_ratio	Proportion of secondary routes within each cube
	Left shoulder type	Bitumen	lft_shldr_Bitum_ratio	Proportion of left shoulder with bitumen within each cube
		Concrete	lft_shldr_Concrete_ratio	Proportion of left shoulder with concrete within each cube
		Curb-concrete	lft_shldr_Curb-Bit_ratio	Proportion of left shoulder with curb-concrete within each cube
		Curb-bituminous	lft_shldr_Curb-Con_ratio	Proportion of left shoulder with curb-bituminous within each cube
		Gravel or stone	lft_shldr_Grass_ratio	Proportion of left shoulder with gravel or stone within each cube
		Grass or sod	lft_shldr_Gravel_ratio	Proportion of left shoulder with grass or sod within each cube
	Right shoulder type	Bitumen	rt_shldr_t_Bitum_ratio	Proportion of right shoulder with bitumen within each cube
		Concrete	rt_shldr_t_Concrete_ratio	Proportion of right shoulder with concrete within each cube
		Curb-concrete	rt_shldr_t_Curb-Bit_ratio	Proportion of right shoulder with curb-concrete within each cube
		Curb-bituminous	rt_shldr_t_Curb-Con_ratio	Proportion of right shoulder with curb-bituminous within each cube
		Gravel or stone	rt_shldr_t_Grass_ratio'	Proportion of right shoulder with gravel or stone within each cube
		Grass or sod	rt_shldr_t_Gravel_ratio	Proportion of right shoulder with grass or sod within each cube
	Speed limit	0-15 mph	spd_limt_1_ratio	Proportion of 0-15 mph speed limit roads within each cube

Feature Type	Feature Name	Categories	Feature Code	Description
		20-40 mph	spd_limt_2_ratio	Proportion of 20-40 mph speed limit roads within each cube
		45-55 mph	spd_limt_3_ratio	Proportion of 45-55 mph speed limit roads within each cube
		>55 mph	spd_limt_4_ratio	Proportion of >55 mph speed limit roads within each cube
	Road surface type	Dry	rdsurf_1.0_ratio	Proportion of dry road surface within each cube
		Wet	rdsurf_2.0_ratio	Proportion of wet road surface within each cube
		Water (standing, moving)	rdsurf_3.0_ratio	Proportion of road surface with water within each cube
		Ice	rdsurf_4.0_ratio	Proportion of road surface with ice within each cube
		Snow	rdsurf_5.0_ratio	Proportion of road surface with snow within each cube
		Slush	rdsurf_6.0_ratio	Proportion of road surface with slush within each cube
		Sand, mud/gravel	rdsurf_7.0_ratio	Proportion of road surface with sand, mud, dirt, or gravel within each cube
		fuel/oil	rdsurf_8.0_ratio	Proportion of road surface with fuel/oil within each cube
		other	rdsurf_9.0_ratio	Proportion of other road surface type within each cube
		Unknown	rdsurf_10.0_ratio	Proportion of unknown road surface type within each cube
Environmental factors	Lighting condition	Daylight	light_1.0_ratio	Proportion of crash during daylight within each cube
		Dusk	light_2.0_ratio	Proportion of crash during dusk within each cube
		Dawn	light_3.0_ratio	Proportion of crash during dawn within each cube
		Dark- lighted roadway	light_4.0_ratio	Proportion of crash during the dark with lighted roadways within each cube
		Dark- roadway not lighted	light_5.0_ratio	Proportion of crash during the dark with roadways not lighted within each cube
		Dark- Unknown lighting	light_6.0_ratio	Proportion of crash during dark with unknown lighting within each cube
		Other	light_7.0_ratio	Proportion of crash with other type of roadway light condition within each cube
		Unknown	light_8.0_ratio	Proportion of crash with unknown type of roadway light condition within each cube
	Weather condition	Unclear, foggy, and cloudy	weather1_1_ratio	Proportion of crash with unclear, foggy, and cloudy weather condition within each cube
		Rain	weather1_2_ratio	Proportion of crash with rainy weather condition within each cube

Feature Type	Feature Name	Categories	Feature Code	Description
		Snow, sleet, hail, freezing rain	weather1_3_ratio	Proportion of crash with snow, sleet, hail, freezing rain/drizzling weather condition within each cube
		Severe crosswinds, blowing sand, dirt	weather1_4_ratio	Proportion of crash with severe crosswinds, blowing sand condition within each cube
Traffic flow/management	Traffic volume	-	aadt	Average traffic volume in each cube
	Traffic control type	No control present	trf_cntl_0.0_rati	Proportion of location with no control device present within each cube
		Stop sign	trf_cntl_1.0_rati	Proportion of location with stop sign present within each cube
		Yield sign	trf_cntl_2.0_rati	Proportion of location with yield sign present within each cube
		Stop and go signal	trf_cntl_3.0_rati	Proportion of location with Stop and Go signal present within each cube
		Flashing signal with stop sign	trf_cntl_4.0_rati	Proportion of location with flashing signal with stop sign present within each cube
		Flashing signal without stop sign	trf_cntl_5.0_rati	Proportion of location with flashing signal without stop sign present within each cube
		RR gate and flasher	trf_cntl_6.0_rati	Proportion of location with railroad gate present within each cube
		RR flasher	trf_cntl_7.0_rati	Proportion of location with railroad flasher present within each cube
		RR crossbucks	trf_cntl_8.0_rati	Proportion of location with railroad crossbucks present within each cube
		Human control	trf_cntl_9.0_rati	Proportion of location with human control present within each cube
		Warning sign	trf_cntl_10.0_rati	Proportion of location with warning sign device present within each cube
		School zone signs	trf_cntl_11.0_rati	Proportion of location with school zone signs present within each cube
		Flashing stop and go signal	trf_cntl_12.0_rati	Proportion of location with flashing stop and go signal present within each cube
		No passing zone	trf_cntl_13.0_rati	Proportion of location with 'no passing' zone present within each cube
		other	trf_cntl_14.0_rati	Proportion of location with other control type present within each cube
Driver characteristics	Driver restraint	None used	drv_rest_ 0_ratio	Proportion of drivers with no restraint used within each cube

Feature Type	Feature Name	Categories	Feature Code	Description
		Lap belt only	drv_rest_ 1_ratio	Proportion of drivers with lap belt only restraint within each cube
		Shoulder and lap belt	drv_rest_ 2_ratio	Proportion of drivers with shoulder and lap belt restraint within each cube
		Child restraint	drv_rest_ 3_ratio	Proportion of drivers with lap belt only restraint within each cube
		Helmet	drv_rest_ 4_ratio	Proportion of drivers with helmet within each cube
		Protective pads	drv_rest_ 5_ratio	Proportion of drivers with protective pad restraint within each cube
		Reflective pads	drv_rest_ 6_ratio	Proportion of drivers with reflective pad restraint within each cube
		Reflective clothing	drv_rest_ 7_ratio	Proportion of drivers with reflective clothing restraint within each cube
		Lighting	drv_rest_ 8_ratio	Proportion of drivers with lighting restraint within each cube
		Other	drv_rest_ 9_ratio	Proportion of drivers with other type of restraint within each cube
		Unable to determine	drv_rest_ 10_ratio	Proportion of drivers with unidentifiable restraint within each cube
	Driver sex	Male	drv_sex_1_ratio	Proportion of male drivers in each cube
		Female	drv_sex_2_ratio	Proportion of female drivers in each cube
		Unknown	drv_sex_3_ratio	Proportion of unknown drivers in each cube
	Driver’s blood alcohol %	-	drv_bac	Average driver blood alcohol % within a cube

Figure 5 shows that the maps exhibit temporal variations in crash occurrences across NC, with the red colored areas indicating high-crash zones and yellow colored areas signifying locations with fewer crashes. Over time, these high-crash hotspots appear to fluctuate in location and concentration, suggesting a dynamic pattern in the incidence of traffic crashes. This fluctuation could be influenced by various factors, including changes in traffic volume, road conditions, seasonal variations, or the impact of traffic safety interventions.

Although past studies have considered the combination of space-time theory (Yoon & Lee, 2021; Azimian et al., 2021), there are no examples in the context of weather-related crash events. Perhaps, the heterogeneity of weather-related crash data is responsible. Here, a multi-layered technique was employed to analyze the complex heterogeneity issues seen in both spatial and temporal settings.

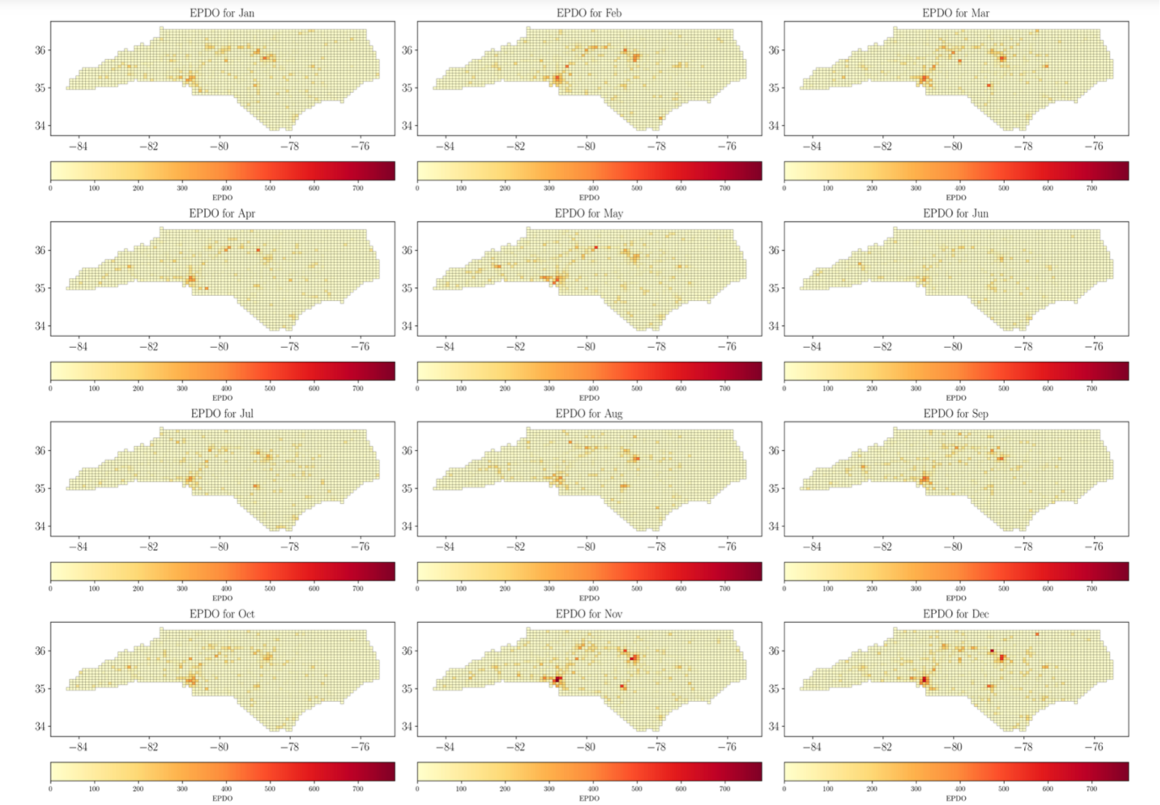


Figure 5. Map showing time series maps for EPDO in NC

3.4. Results

The results as well as the model outputs are discussed in this section. The interactions between crash risk and weather conditions are also discussed.

3.4.1. Distribution of Crashes in Distinct Weather Conditions

Figure 6 is a set of four choropleth maps displaying EPDOs for crashes under different weather conditions: Cloudy/Fog, Rain, Snow/Sleet, and Wind. Each map shows a geographic area (say, NC), with a grid overlay. The intensity of the color on each grid cell indicates the EPDO value for crashes in that specific cell under the given weather condition.

Under the cloudy or foggy conditions, some grid cells have relatively higher EPDO values. These kinds of crashes are more prominent in the Mecklenburg and Wake counties of NC. These conditions significantly reduce visibility, which may lead to more severe crashes due to drivers not being able to see other vehicles or road hazards in time to react (Abdel-Aty et al., 2011). Similar to this is the map showing higher EPDO values for crashes in rainy weather compared to cloudy/fog. Both conditions show similar pattern, what is different is the intensity of the EPDOs. While rain can create slippery road surfaces and reduce tire traction (Perrels et al., 2015), leading to a higher risk of crashes, they also occur because of low visibility (Mohammed et al., 2020), especially with high rain intensity.

The map on the bottom left corner shows generally lower EPDO values for crashes under snow or sleet conditions than for rain. This may be because drivers tend to be more cautious and drive slower under snowy conditions (Strong et al., 2010), which can lead to fewer crashes or less severe crashes with lower property damage. In addition, NC does not experience snow often and the result does not come surprising. However, the spatial footprint of crashes under this weather condition is higher compared to other weather conditions. This finding bolsters the result of a study carried out on personal disaster preparedness in NC that individuals are not prepared to handle the snowy conditions (Foster et al., 2011).

The map in the bottom right corner of Figure 6 shows the lowest EPDO values among the displayed conditions. While wind can affect vehicle stability and control, especially for larger vehicles like trucks, it appears to have less impact on the EPDO for crashes in this region.

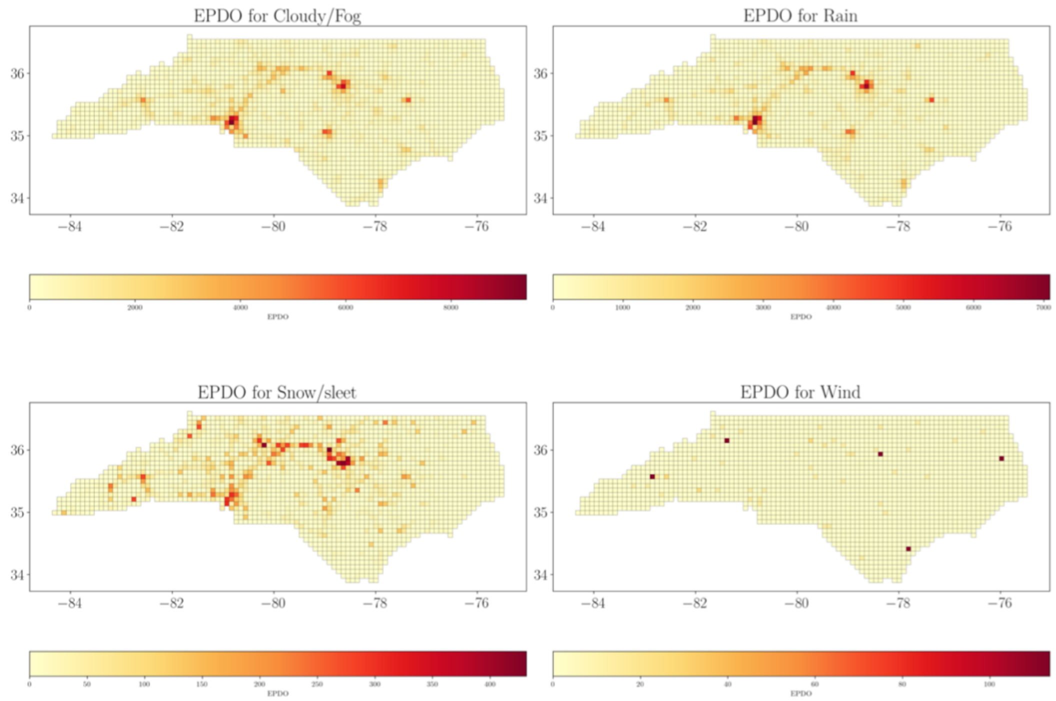


Figure 6. Spatial distribution of crashes in distinct weather condition

3.4.2. Dynamic Time Warping (DTW)

Two statistical methods were employed, the elbow method, and the silhouette score, to evaluate the clustering of crash data. The silhouette score is a measure of how similar data points are within a cluster compared to other clusters, with higher scores indicating better clustering quality, meaning the data points within a cluster are more similar to each other, and the clusters themselves are more distinct from one another. Per Figure 7, a silhouette score analysis suggests that the crash data is most effectively organized into two clusters. This means that the data can be best understood when it is divided into two distinct groups. However, when this finding is compared with results from the elbow method, which is another way to determine the number of clusters that best fits the data, it can be interpreted that three clusters are optimal for the dataset.

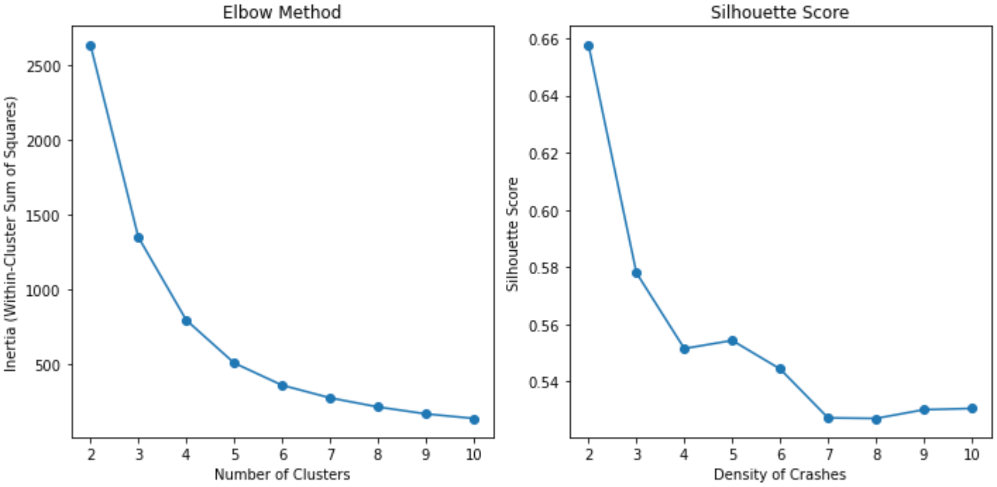


Figure 7. Left: Elbow method and Right: silhouette score technique

The unsupervised clustering analysis provides the homogeneous groups of weather-related crash point process in which each group contains data with similar behavior. This analysis was performed on monthly EPDO of weather-related crashes and the cluster wise spatial distribution is as shown in Figure 8.

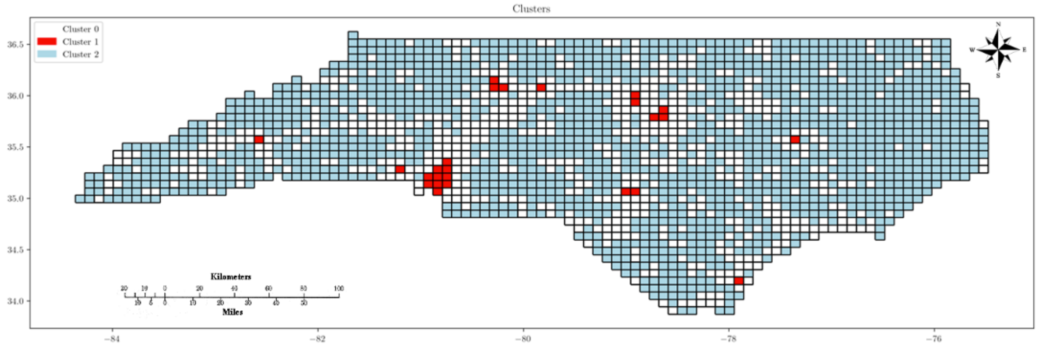


Figure 8. Distribution of locations into DTW clusters

Figure 9 shows the average EPDO across clusters and have two y-axes with different scales since cluster 1 EPDOs are much higher than those for clusters 0 and 2. The left y-axis corresponds to clusters 0 and 2, while the right y-axis corresponds to cluster 1. Given that the EPDO values for cluster 0 range from around 20 to just above 30, and show a gradual increase throughout the year, this cluster can be considered as 'gradually increasing low-EPDO.' Cluster 1 has significantly higher EPDO values, starting off around 200 and growing towards 250 by the end of the year with notable fluctuations. This cluster is referred to as ‘volatile high-EPDO group.’ Since the EPDO values for cluster 2 are consistently the lowest, this cluster is referred to as ‘stable low-EPDO group.’

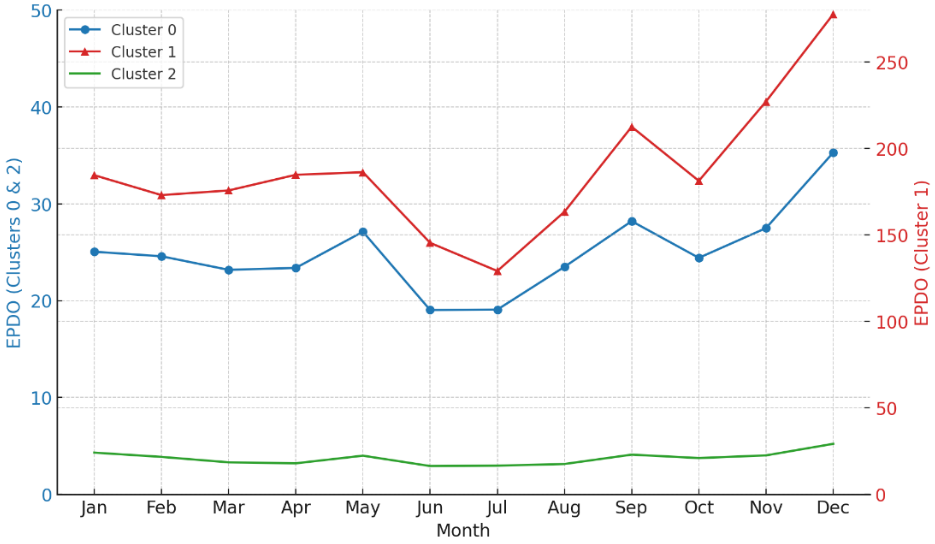


Figure 9. Distribution of average EPDO across DTW clusters

The DTW-G* maps are compared (Figure 10) with the DTW cluster distribution to find patterns and similarities. Cluster 1 is a group of volatile high-EPDOs which is also reflected in the hotspot map as majority of the grids that fall into the cluster 1 groups have high EPDOs. By visual inspection of Figure 5 and Figure 9, it can be stated that the spatial and temporal distribution of weather-related crashes are somewhat similar. Whereas in some cases, certain time and spatial units exhibit different patterns. For example, the positive peak bias between the distribution of crashes for cluster 0 and 1 is during the cold periods of October to December. A good justification for this might be that people are driving more in the colder months, thus increasing vehicle mile traveled (VMT) leading to more crashes. The summer months (June and July) records a dip in EPDOs. This is reasonable as favorable weather is experienced during this period. People want to walk and get involved in recreational activities during this period. The finding obtained from the spatial and temporal distribution of average EPDO across DTW clusters is two-fold. First, the seasonal effect in weather-related crashes may be interacted with other external factors such as traffic volume and VMT. The key risk factors are explained in a later section.

3.4.3. Crash Risk Labeling

The application of the two-layered technique helps segregate high-risk locations from low-risk locations. Figure 10 indicates that the method intelligently classifies significant hotspot and coldspot cubes within their respective clusters. The trends and the intensity of crashes in different clusters can be better observed for the DTW method. The technique shows how these significant cubes distribute across the clusters. The DTW method suggests that cluster 0 is a group of gradually increasing low-EPDOs looking at the average trend line. The two layered technique supports this by showing that the cluster has a mix of hotspots and cold spots grids.

From Table 3, all clusters have more than half of their weather-related crashes under unclear, foggy, and cloudy conditions. Cluster C2 has the highest mean percentage, indicating that such weather is a common factor in crashes, but with the highest variability (±17.1%), meaning that the extent to which this weather contributes to crashes varies the most in this cluster. Cluster C1 has the lowest mean but the least variability, indicating more consistency in the impact of unclear-weather condition on crashes across this cluster. Rain is a significant factor in crashes across all clusters, but it is most prominent in C1, where it is associated with the highest mean percentage of crashes. In addition, a low standard deviation of ± 3.6% indicates that rain-related crashes are quite consistent, meaning that rain is a reliably common condition when crashes occur in C1. C0 and C2 have lower mean percentages but higher standard deviations, suggesting that the contribution of rain to crashes is highly inconsistent- some areas within this clusters may experience many rain-related crashes, while others experience few rain-related crashes.

Snow, sleet, hail, and freezing rain happen in winter weather conditions. These weather conditions have a minor impact on crashes compared to the other weather types, with C1 having the lowest mean percentage. C0 and C2 have similar mean percentages, but C2 has a higher variability, suggesting that in some areas within this cluster, these conditions are more significant factors in crashes. This finding seems reasonable because these weather conditions are rarely experienced in the study area. The high variability is also supported by the spatial distribution shown in Figure 6. The finding suggests that even though these crashes are not frequent within the study area, the outcome ends up more severe.

Severe crosswinds and dusty condition category have the least impact on crashes across all clusters, with very low mean percentages. C1 has the smallest mean percentage and variability, indicating that these conditions are the least significant for crashes. C0 and C2 have marginally higher means, but C2 exhibits a higher variability, again suggesting that the impact of these conditions on crashes varies more across different areas within this cluster. These conditions are expected to happen only for a short period, and it only make sense that Figure 6 shows severe crosswinds related crashes are more prominent along/close to the coast.

Table 5. Interaction of weather and DTW clusters

Weather categories	**Feature Code**	**C0**		**C1**		**C2**	
		Mean %	std. dev.	Mean %	std. dev.	Mean %	std. dev.
Unclear, foggy, and cloudy	weather1_1_ratio	59.6	10.9	56.3	4.2	**61.4**	17.1
Rainy weather	weather1_2_ratio	35.1	10.4	**40.7**	3.6	32.7	16.7
Snow, sleet, hail, freezing rain	weather1_3_ratio	5.2	4.3	2.9	1.4	**5.7**	7.9
Severe crosswinds, blowing sand, dirt	weather1_4_ratio	0.18	0.8	0.07	0.08	**0.2**	1.3

Note: Numbers highlighted in bold are the most significant clusters for various weather conditions.

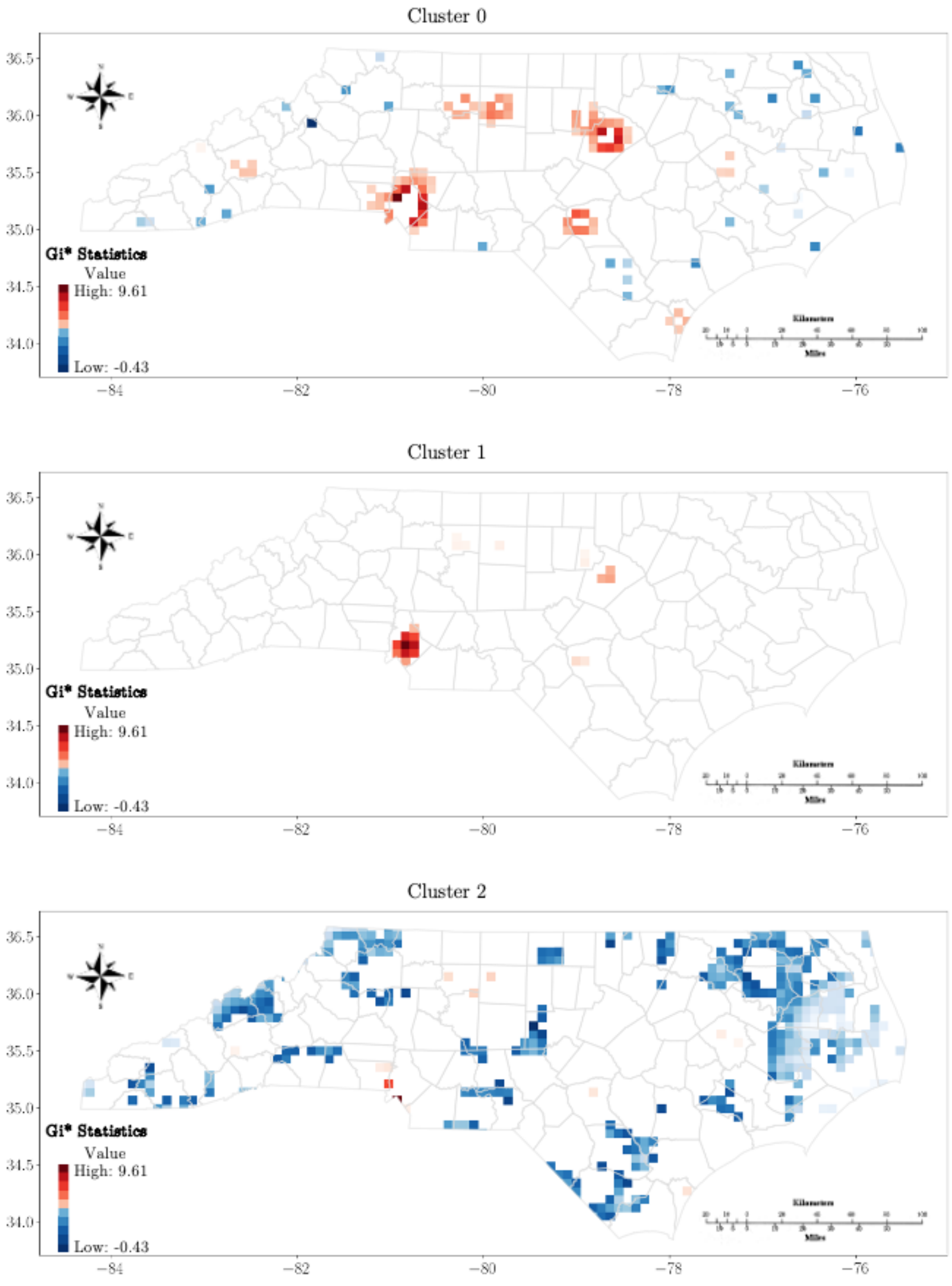


Figure 10. DTW-G* results

3.4.4. Risk Pattern Prediction Performance

Of the 2,045 candidate cubes, 620 belong to cluster 0, 25 belong to cluster 1 and 1400 belong to cluster 2. The training set is formed by drawing 80% samples without replacement from the full dataset and the remaining dataset was used for testing. The computed ROC curve and AUC from the base model are presented in Figure 11a. The plot shows a perfect prediction of cluster 1 which is a cluster of high-risk locations.

It is important to assess how realistic is this result in practice. It is noteworthy that accurately predicting this class would help us put preventive measures in place prior to their occurrence. While ROC curves are not sensitive to changes in outcome class proportion (Carter et al., 2016), they do have considerable impact on the estimation of error rates and AUC. Although the developed model is perfectly able to identify this minority class, the sample size is to be considered while interpreting this curve. Cluster 2 has a higher discriminate capacity compared to cluster 0. While a curve that deviates significantly from the diagonal reflect the performance of a crash risk classification that is much better than chance (Carter et al., 2016), one can improve the performance by tuning and removing redundant features.

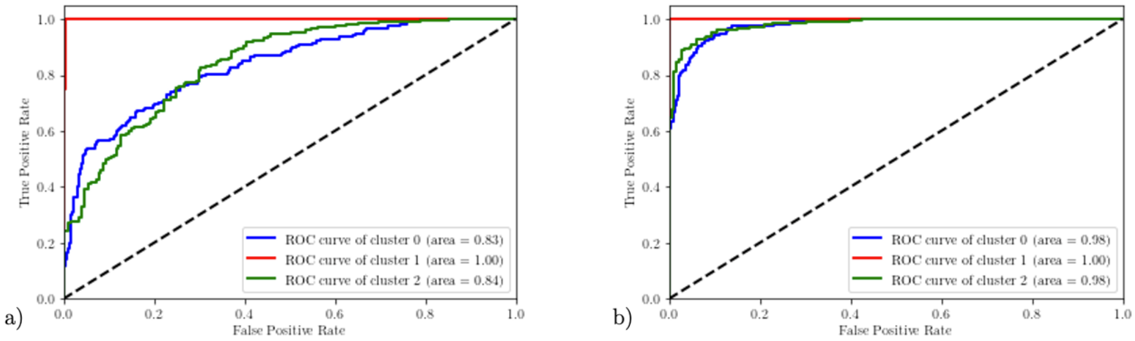


Figure 11. ROC curve of crash risk classification: a) base model; b) tuned model

Figure 12 shows a ranked list of features used to classify the weather-related crash risks. The length of the bar indicates how much influence that feature has on the model's prediction. The most important features are at the top of the plot with the longest bars, and they decrease in importance as you move down to the bottom of the plot. From Figure 12, one can make changes to the dataset eliminating features that do not influence the results while having a better understanding of which features are the most influential.

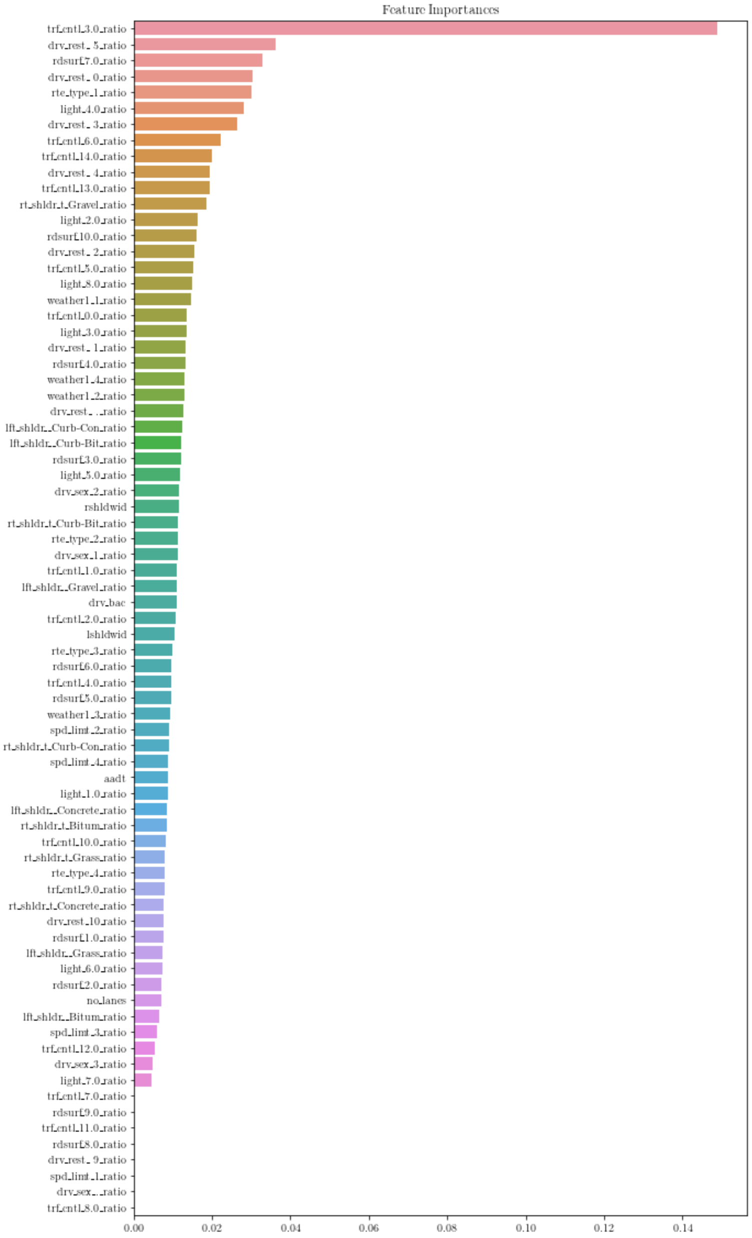


Figure 12. Importance of features described in Table 4 for crash risk prediction

When comparing the ROC curve of the developed tuned model in Figure 11b to that of the base model in Figure 11a, the tuned version shows the ROC curve that lies above the base model's curve, reflecting a superior true positive rate for the same false positive rate across different threshold settings. This improvement suggests that the tuning process has effectively optimized the model's parameters, leading to a more sensitive and specific classification of the crash risk.

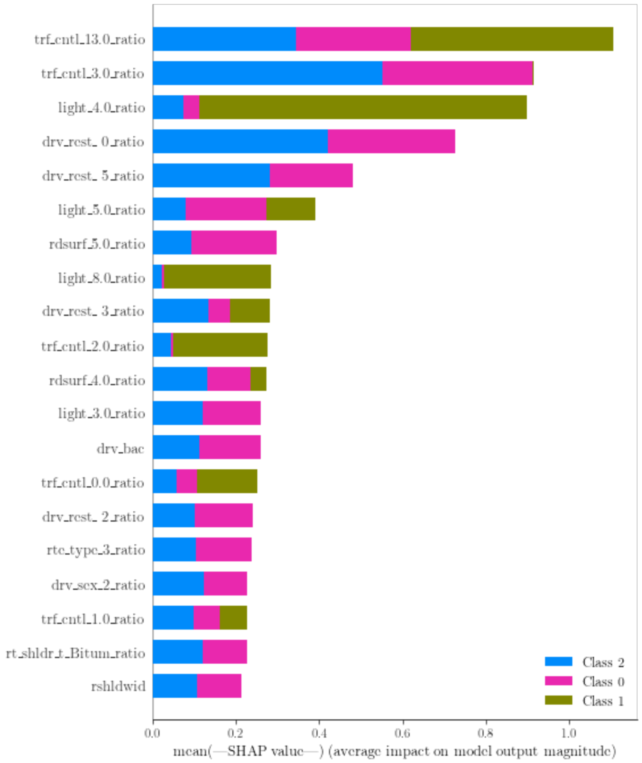


Figure 13. Average impact of selected features described in Table 4 on crash risk

3.4.5. Key Factors for High-Risk Crash Classification

From Figure 13, the contribution of each factor to the crash risk classification model can be examined. Double yellow line or no passing zone have the strongest influence in predicting the volatile high-EPDO crashes or the cluster 1 crashes. This observation posits that two-way roads, exert a significant influence on the model's ability to predict crashes at high-risk locations. The preeminence of double yellow lines or no passing zones in predicting high-EPDO crashes may be attributable to the nature of crashes that occur within such areas. Typically, these zones are instituted on stretches of road where visibility or road conditions make overtaking maneuvers dangerous (Pashkevich et al., 2021). Consequently, violations of these regulations are likely to result in head-on collisions (Das et al., 2018), which are among the most severe types of vehicle crashes due to the high velocity impact and direct contact of the vehicles' front sections. The severity of such crashes, in terms of both human injury and property damage, would naturally elevate their EPDO scores, making the model's sensitivity to this feature as an indicator of high-risk crash location. A study carried out by Ogungbire et al. (2023) confirmed that there is a higher likelihood of weather-related crashes on two-way roads compared to one-way roads. Further, research focusing on road design and traffic regulation effectiveness has illuminated the critical role of no passing zones in mitigating the risk of such crashes by restricting overtaking maneuvers on particularly dangerous road segments (Naheed et al., 2023).

Several other features are found to have huge impact on crash occurrence in high-risk locations. For example, light_4.0_ratio which represent the proportion of crashes during the dark with lighted roadways exert a significant influence on the model's ability to predict crashes at high-risk locations. The importance of lighted roadways in predicting crash occurrences can be interpreted through several lenses. Firstly, while illumination is intended to enhance visibility and safety during nighttime driving, the effectiveness of roadway lighting can vary greatly, influenced by factors such as the intensity of lighting, the spacing of light posts, and the presence of reflective road markers (Bullough, Donnell & Rea, 2013). In scenarios where lighting is insufficient or unevenly distributed, the contrast between light and dark spots can create visual illusions or reduce the reaction time available to drivers, thus escalating the risk of crashes. Moreover, studies examining driver behavior have indicated that perceived safety improvements due to roadway lighting can induce riskier driving behaviors, a phenomenon known as risk compensation (Houston & Richardson, 2007). For statewide transportation planning, these findings reinforce those enhancements to roadway lighting lead to a reduction in roadway safety. Not improving the lightning does reduce the safety as the ratio of crashes during the dark with unlighted roadways having a very high impact on the high-risk cluster prediction. Thus, it becomes important for state DOTs to not only implement and maintain roadway lighting but also ensure that educational campaigns and encouraging cautious driving behavior under all conditions are done.

Amidst other factors that affects locations of high crashes lie locations with yield signs. Yield signs are intended to regulate traffic flow, requiring drivers to slow down or stop to give way to vehicles on the main road, thus avoiding potential conflicts (Rachakonda & Pawar, 2023). In situations where the yield sign's directive is misinterpreted, disregarded, or visibility is compromised, the risk of high-severity crashes increases. Scenarios involving non-visible yield signs due to poor weather conditions tend to result in sideswipe crashes. Several studies have shown that merging into traffic at inappropriate speeds increases the risk of severe injury and significant PDO (Pathivada et al., 2024; Sawtelle, 2023). To mitigate the frequency of occurrence for inappropriate merging during poor weather condition may involve strategic considerations; the use of technologies such as ramp metering and vehicle-to-infrastructure (V2I) communication to optimize merging processes. By controlling the rate at which vehicles enter the freeway based on real-time conditions, these technologies can reduce the likelihood of crashes at merge points. Additionally, variable message signs (VMS) can be implemented to warn drivers of poor visible conditions ahead heavy rain, fog, or snow.

Figure 14 presents SHAP summary plots for the study area, categorized into gradually increasing low-risk/class 0, volatile high-risk/class 1, and stable low-risk/class 2. Within each plot, the significant features are arranged in a descending hierarchy based on their absolute SHAP values, denoting their relative importance. The plots employ a dot representation for individual SHAP values corresponding to specific features and samples. Dots colored in red signify higher feature values, while blue dots indicate lower values. As regards volatile high-risk, several features are found to have huge effects on their occurrence. Lighting condition and the type of traffic control significantly impact whether or not a weather-related crash zone is classified as volatile high-risk zones.

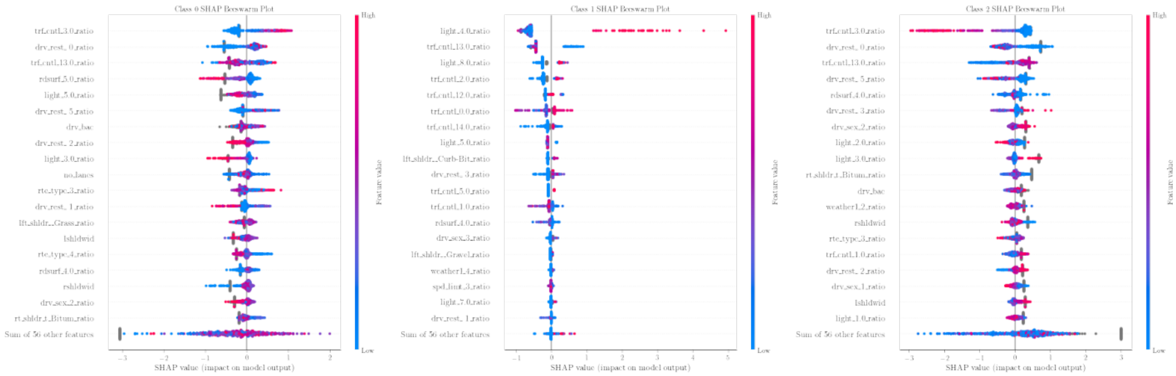


Figure 14. SHAP summary plot of crash risk patterns

3.4.6. Key Factors for Low-Risk Crash Classification

From Figure 14, one can observe how each feature has contributed to classifying crash risk zone as stable or gradually increasing low-EPDOs. Several features are common to both stable and gradually increasing low-risk zones, however, the effect varies. Features like proportion of locations with stop and go signals and proportion of drivers with no restrain used are top features in classifying a grid as both stable and gradually increasing low-risk zones. One can see that high proportion of crash zones with stop and go signals and drivers with no restraint used are associated with higher SHAP values in gradually increasing low-risk zones. However, higher proportion of crash zones with stop and go signals is associated with low SHAP values in gradually increasing low-risk zones. Generally, one can deduce that stop and go traffic signals increase the likelihood of a zone to be classified as a low-risk zone. While there is no consensus on the effect of stop and go signals on traffic crashes in literature, the efficiency of stop-and-go signals in reducing traffic crashes and improving has been substantiated by various studies (Elmitiny et al., 2010; Suh & Yeo, 2016). Cohn et al., 2020 explained how mismanaged signals or poor signal placement can inadvertently increase the risk of certain types of crashes. This nuanced understanding is further expanded by insights from Figure 15, which suggest that the stop and go traffic signal is associated with rainy weather condition. This correlation might initially appear counterintuitive, given the primary role of traffic signals in regulating and securing vehicular movement. However, this relationship can be rationalized by considering the interplay between weather-induced road conditions and traffic control mechanisms.

Low and high average driver blood alcohol (BAC) level is associated with high SHAP values. The relationship between elevated BAC levels and increased crash risk is well-established in literature. High BAC levels impair cognitive functions, reaction times, and decision-making abilities, substantially increasing the likelihood of vehicle crashes (Martin et al., 2013). Conversely, the association of zones with low BAC levels and high SHAP values may initially appear counterintuitive, given the legal BAC limits set to promote traffic safety. However, this observation can be interpreted through several lenses. Firstly, the average BAC levels might be low, however, several individuals will have BAC level above the legal permissible limit. Secondly, low BAC level may still produce subtle impairments in driving ability, particularly in individuals with low alcohol tolerance or in combination with other factors such as poor weather condition (Martin et al., 2013).

3.4.7. Interaction between Weather Condition and Crash Risk

From Figure 15, one can see that if a crash occurs in stable/gradually increasing low-risk zones, the proportion of crashes that occur as a result of rain tends to vary with increasing ratio of stop and go traffic control. The spread of SHAP values suggests a non-linear interaction between rainy conditions and the prevalence of stop-and-go signals. This suggests a varying predictive impact on low-risk zones. For instance, when the proportion of crashes that occur as a result of rain is between 0.2 to 0.3 with a high proportion of stop and go traffic control, they are less likely to be classified as low-risk crashes. A potential hypothesis could be that when the proportion of stop and go signals is low, the impact of rain on crash risk might be different compared to areas where there is a high proportion of stop and go signals, possibly due to varying traffic dynamics. For example, in areas with many stop and go signals, rainy conditions could lead to more frequent braking and reduced speeds, which may lower crash risk. Conversely, in areas with fewer such signals, drivers may be traveling at higher speeds during rain, potentially increasing crash risk. In areas identified as high-risk for traffic crashes, it is recommended to implement a stop and go traffic control system designed to regulate vehicle flow and reduce collision risk. This system should be strategically deployed at critical intersections, congested areas, and zones with a history of frequent accidents, using traffic signals, signage, and road markings to guide vehicular movement effectively. The system must be supported by comprehensive traffic studies that assess the unique needs and challenges of each area, ensuring that timing intervals are optimized to balance traffic flow with pedestrian safety. Additionally, public awareness campaigns and driver education programs should accompany the implementation to enhance compliance and effectiveness.

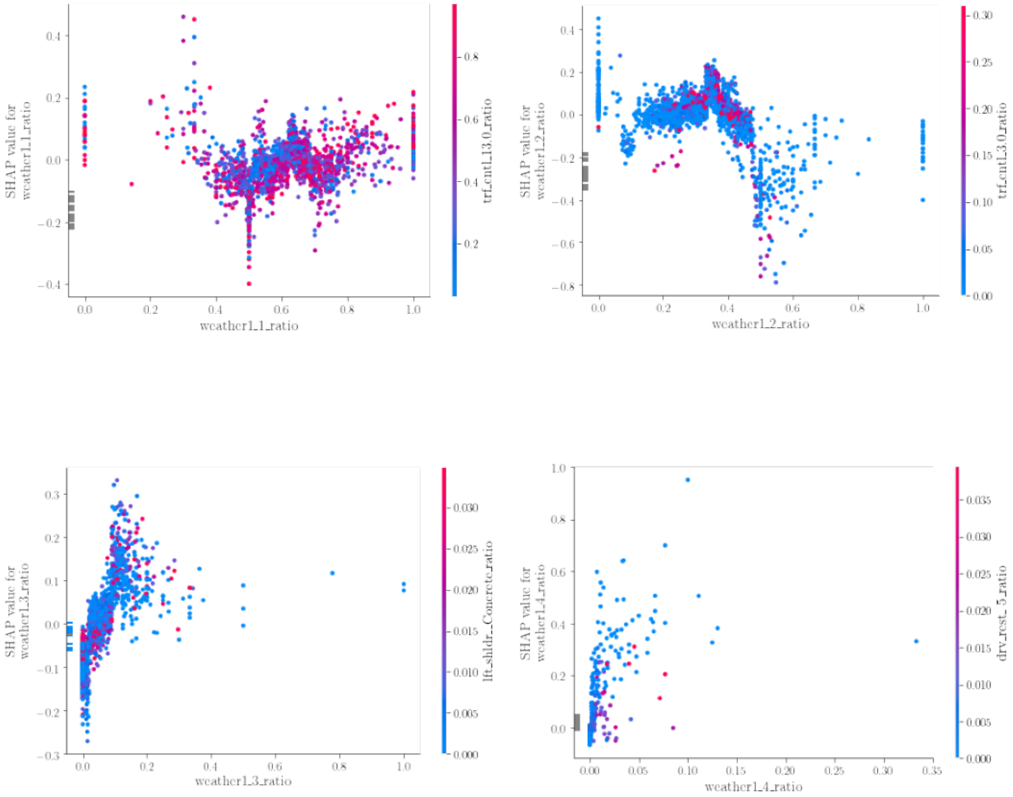


Figure 15. Dependence plot of weather conditions and crash risk

The spread of SHAP values also suggests a non-linear interaction between unclear weather conditions and the prevalence of no passing zones. When the proportion of crashes that occur as a result of unclear/cloudy weather condition is between 0.5 to 0.7 with a high proportion of no passing zones, there is a high likelihood of being classified as a high-risk zone. To bolster the result in section 3.4.6, one can conclude that majority of high-risk crash that occurred at no passing zones are as a result of unclear, cloudy, or foggy weather condition.

CHAPTER 4: EFFECTIVENESS OF CRASH DATA IMBALANCE TREATMENT IN WEATHER-RELATED CRASH SEVERITY ANALYSIS

4.1. Introduction

Weather-related vehicle crashes are particularly hazardous due to their unpredictable nature. Not only can such crashes cause serious physical harm, but they can also have a substantial economic impact, with research indicating that weather-related crashes in the United States in 2014 costing approximately $46 billion (FHWA, 2023). Furthermore, weather-related crashes are more likely to be fatal than other types of crashes (Dey et al., 2014), with statistics from the same year showing that they accounted for 21% of all fatalities (FHWA, 2023). The psychological repercussions of such crashes can be long-lasting and have a substantial economic impact in addition to reducing the quality of life of those affected (Gao et al., 2021).

To mitigate the impact of weather-related crashes, practitioners must be able to accurately predict and identify the factors associated with their severity levels. However, due to the highly imbalanced nature of crash data, traditional machine learning techniques are not robust enough to effectively classify crash severity (Gao et al., 2021; Kim et al., 2021). Thus, it is crucial to address the problem of class imbalance before predicting severity classes on weather-related crash dataset.

Naive approaches, such as using only one classifier or the majority class to predict all classes, often lead to inaccurate results (Gao et al., 2021). Two of the most popular methods are under-sampling and over-sampling techniques. In under-sampling, the number of instances from the majority weather-related crash severity class are reduced to match the frequency of the minority class (Kim et al., 2021; Chawla et al., 2002). However, there is a risk of losing vital information about the majority class. On the other hand, over-sampling involves replicating weather-related crash events from the minority class to increase their representation in the dataset. In the context of crash event, the rare occurrence of fatal/severe crash gives room for overfitting in the case of over-sampling method (He et al., 2008).

Due to the weakness of under-sampling and over-sampling, SMOTE (Chawla et al., 2002) and ADASYN (He et al., 2008) methods are used to balance the data, thus avoiding the risk of overfitting or data loss (Saarikko et al., 2020; Gaikwad & Markande, 2016). SMOTE works by generating new events that are interpolations of existing minority class events, thereby improving the diversity of instances without simply duplicating data. ADASYN, however, takes this step further by adaptively generating minority data points based on their density distribution with more weight on the most difficult-to-learn data instances. However, a significant limitation arises when dealing with a dataset that is entirely composed of nominal predictors. The method struggles because they rely on the concept of distance or interpolation between points.

Modern data sources such as weather-related crash data have resulted in complex datasets that are cumbersome to model with classic statistical methods (Strong et al., 2010; Perrels et al., 2015; Daniels et al., 2016; Musselwhite et al., 2021). On the other hand, machine learning techniques are robust enough to deal with complex datasets with high dimensional features. While these techniques are becoming popular in predicting traffic crashes, the presence of data imbalance in weather-related crash severity datasets often introduce unintended biases to resultant models, making it essential to address these imbalances before beginning the model training process (Songchitruksa & Zheng, 2010; Lakshmi et al., 2019). This problem becomes even more complex with a dataset containing predominantly nominal predictors.

Therefore, this dissertation aims to assess the efficacy of different data treatment techniques and their influence on machine learning techniques, particularly focusing on handling nominal predictors. Specifically, data treatment techniques involving two synthetic approaches, namely SMOTE-N and ADASYN-N, are investigated for nominal predictors. The dissertation employed two crash severity classification models, the bagging algorithm (Random Forest - RF) and the boosting algorithm (Extreme Gradient Boosting - XGBoost). By evaluating the performance of these models with different data resampling methods, this dissertation seeks to gain insights into how such techniques can enhance the accuracy of machine learning techniques when applied to weather-related traffic crash data.

4.2. Case Study of North Carolina (NC)

The data used in this dissertation was obtained from the HSIS database. Crashes that took place between January 1, 2015, to December 31, 2017, were extracted. Crashes are reported using case numbers and observations with the same number indicate that the vehicles involved are part of the same crash incident. To gain a thorough understanding of the crash occurrence process, Washington & Haque (2013) argued that crashes due to different causes should be modeled separately. Hence, only weather-related crashes were extracted while crashes occurring under clear weather conditions were excluded to obtain a final dataset. The final dataset only included crashes that happened under non-clear weather conditions (i.e., with cloudy, rain, fog/smog, sleet/hail/freezing rain/drizzle, severe crosswinds, or blowing sand conditions described in the crash reports). Crash severity is defined in the HSIS database as five different levels (i.e., fatal crashes, injury type class A, injury type class B, injury type class C, and no injury/PDO). For this analysis, the crash severity was re-categorized into three levels, i.e., severe injury (fatal and injury type class A), moderate injury (injury type class B and injury type class C) and PDO/no injury. A total of 238,252 weather-related crashes were recorded within the study period with 2,952 severe crashes, 71,688 moderate crashes and 163,612 PDO crashes.

The summary statistics gives information about the counts and percentages of different types of weather conditions for severe, moderate, and PDO injury cases. Cloudy weather is the most common weather condition in all three categories, with 63.4% of severe crashes, 57.6% of moderate injury crashes, and 56.6% of no injury crashes occurring under cloudy conditions. Rain is the second most common weather condition, followed by snow, fog, smog, and smoke. Sleet, hail, and freezing rain/drizzle are less common, with blowing sand and dirt being the least common weather condition for all three-severity types.

The dataset summarized in Table 6 underwent initial preprocessing to facilitate model selection and evaluation. Initially, the comprehensive dataset was partitioned into two distinct subsets: a training set and a testing set. The formation of the test dataset was accomplished through the methodology of random sampling, ensuring a distribution proportional to the size of each constituent class. Consequently, a fifth of the data, uniformly distributed across classes, was extracted to constitute the testing set. The residual 80% of the data served as the foundational set, subject to diverse treatment methodologies to generate corresponding training datasets.

Table 6. Descriptive statistics of variables for analysis

Variable	Categories	Description	Severe		Moderate		PDO	
			Count	%	Count	%	Count	%
Weather condition (Weather)	1	Cloudy	1871	63.4	41293	57.6	92663	56.6
	2	Rain	893	30.3	26828	37.4	60374	36.9
	3	Snow	41	1.4	1085	1.5	4296	2.6
	4	Fog, Smog, Smoke	104	3.5	1247	1.7	2786	1.7
	5	Sleet, Hall, Freezing Rain/Drizzle	39	1.3	1196	1.7	3324	2
	6	Severe Crosswinds	3	0.1	31	0.04	132	0.08
	7	Blowing Sand, Dirt	1	0.03	8	0.01	37	0.02
Contributing factor of the crash (contfac)	1	No contributing factors	1191	40.3	32095	44.8	75111	45.9
	2	Disregarding signs or signals	98	3.3	2124	3.0	2492	1.5
	3	Exceeded safe speed/speed limit or fail to reduce speed	548	18.6	17302	24.1	41651	25.5
	4	Improper turn or right turn on red	17	0.6	730	1.0	1979	1.2
	5	Crossed centerline, improper lane change, or use of an improper lane	222	7.5	2135	3.0	6102	3.7
	6	Overcorrected, oversteered, improper passing, or improper backing	77	2.6	1503	2.1	3019	1.8
	7	Failing to yield to the right-of-way, or driver inattention	267	9.0	9147	12.8	18870	11.5
	8	Operating too closely, aggressive driving, or alcohol use	429	14.5	3768	5.3	6578	4.0
	9	Visibility obstruction, or defective equipment	16	0.5	471	0.7	1275	0.8
	10	Other/unable to determine	87	2.9	2413	3.4	6535	4.0
Road surface condition (roadsurf)	1	Dry	1290	43.7	27329	38.1	60413	36.9
	2	Wet, presence of water (standing/moving)	1577	53.4	41617	58.1	93882	57.4
	3	Ice, snow, slush	81	2.7	2708	3.8	9248	5.7

Variable	Categories	Description	Severe		Moderate		PDO	
			Count	%	Count	%	Count	%
	4	Sand, mud, dirt, gravel, fuel, or oil	4	0.1	34	0.05	69	0.04
Functional class of road (fclass)	1	Principal arterial – interstate, freeways, and expressways	402	13.6	12617	17.6	35002	21.4
	2	Principal arterial – other	682	23.1	22955	32.0	50092	30.6
	3	Minor arterial	693	23.5	18303	25.5	39919	24.4
	4	Major collector	789	26.7	11795	16.5	24314	14.9
	5	Local	386	13.1	6018	8.4	14285	8.7
Location type (intersecti)	0	Non-intersection	2516	85.2	57486	80.2	138677	84.8
	1	Intersection	436	14.8	14202	19.8	24935	15.2
Light condition (lightnew)	1	Daylight	1727	58.5	51558	71.9	118084	72.2
	2	Dusk, and dawn	163	5.5	3582	5.0	7799	4.8
	3	Dark lighted roadway/unknown lighting	244	8.3	7146	10.0	15206	9.3
	4	Roadway not lighted	818	27.7	9402	13.1	22523	13.8
Road characteristic (roadchar)	1	Straight-leveled road	1675	56.7	50437	70.4	118417	72.4
	2	Straight-grade/hillcrest/botto m	485	16.4	12319	17.2	27794	17.0
	3	Curve-leveled/grade/hillcrest	788	26.7	8876	12.4	17221	10.5
	4	Not stated/unknown	4	0.1	56	0.1	180	0.1
Driver gender (drv_sex)	1	Male	2006	68.0	38058	53.1	92305	56.4
	2	Female	946	32.0	33630	46.9	71307	43.6
Driver age (ageis)	1	15–19 years	262	8.9	7105	9.9	16690	10.2
	2	19–69 years	2505	84.9	60725	84.7	138762	84.8
	3	≥70 years	185	6.3	3858	5.4	8160	5.0
Speed limit class (slgrp)	1	≤20 mph	5	0.2	415	0.6	1482	0.9
	2	20–30 mph* (30 mph included)	17	0.6	1064	1.5	3042	1.9
	3	30–40 mph	302	10.2	15354	21.4	35896	21.9
	4	40–50 mph	860	29.1	28690	40.0	62077	37.9
	5	50–60 mph	1495	50.6	20053	28.0	42518	26.0
	6	>60 mph	273	9.2	6112	8.5	18597	11.4
Crash type (crashtyp)	1	Ran off-road	113	3.8	2524	3.5	5590	3.4
	2	Jackknife, overturn/rollover	124	4.2	1191	1.7	1276	0.8
	3	Pedestrian/pedal cyclist	186	6.3	484	0.7	53	0.0

Variable	Categories	Description	Severe		Moderate		PDO	
			Count	%	Count	%	Count	%
	4	Animal or movable object	26	0.9	727	1.0	9287	5.7
	5	Parked vehicle or fixed object	664	22.5	9296	13.0	22011	13.5
	6	Rear-end collision	395	13.4	29527	41.2	67554	41.3
	7	Left-/right-turn crashes	340	11.5	9243	12.9	15346	9.4
	8	Head-on collision	416	14.1	1534	2.1	763	0.5
	9	Sideswipe or angle collision	599	20.3	15879	22.2	37251	2.8
	10	Other	89	3.0	1283	1.8	4481	2.7
Work zone area (workzone)	0	No	2880	97.6	69872	97.5	159385	97.4
	1	Yes	72	2.4	1816	2.5	4227	2.6
Vehicle type (vehicle)	1	Passenger car/taxi	1331	45.1	40185	56.1	90650	55.4
	2	Pickup, light truck, sports utility, or van	1228	41.6	28305	39.5	65899	40.3
	3	Commercial bus, school bus, activity bus, other bus	13	0.4	255	0.4	593	0.4
	4	Single unit truck, truck/trailer, truck/tractor, tractor doubles, semitrailer, farm equipment, or other heavy trucks	192	6.5	1847	2.6	5557	3.4
	5	Motor scooter, moped, pedal cycle, or motorcycle	179	6.1	876	1.2	185	0.1
	6	Other	9	0.3	220	0.3	728	0.4
Seasonal factors (season)	1	Spring	623	21.1	18168	25.3	44375	27.1
	2	Summer	762	25.8	17669	24.6	37989	23.2
	3	Autumn	717	24.3	15408	21.5	33642	20.6
	4	Winter	850	28.8	20443	28.5	47606	29.1
Road terrain (terrain)	1	Flat	748	25.3	13405	18.7	29761	18.2
	2	Rolling	1975	66.9	53124	74.1	120807	73.8
	3	Mountainous	229	7.8	5159	7.2	1304	8.0
Time of the day (TOD)	1	12:00 AM – 03:00 AM	193	6.5	2047	2.9	3992	2.4
	2	03:00 AM – 06:00 AM	177	6.0	1880	2.6	4381	2.7
	3	06:00 AM – 09:00 AM	412	14.0	11886	16.6	29002	17.7
	4	09:00 AM – 12:00 PM	366	12.4	9488	13.2	22060	13.5
	5	12:00 PM – 03:00 PM	445	15.1	13421	18.7	29605	18.1

Variable	Categories	Description	Severe		Moderate		PDO	
			Count	%	Count	%	Count	%
	6	03:00 PM – 06:00 PM	552	18.7	18351	25.6	41449	25.3
	7	06:00 PM – 09:00 PM	499	16.9	10164	14.2	23404	14.3
	8	09:00 PM – 12:00 PM	308	10.4	4451	6.2	9719	5.9
Day of the week (dow)	1	Sunday	412	14.0	6389	8.9	13217	8.1
	2	Monday	485	16.4	12518	17.5	29077	17.8
	3	Tuesday	410	13.9	12078	16.8	28917	17.7
	4	Wednesday	465	15.8	10533	14.7	24472	15.0
	5	Thursday	335	11.3	9869	13.8	22882	14.0
	6	Friday	449	15.2	12493	17.4	28711	17.5
	7	Saturday	396	13.4	7808	10.9	16336	10.0
Locality (locality)	1	Agricultural	1573	53.3	21185	29.6	47922	29.3
	2	Residential	597	20.2	13674	19.1	27615	16.9
	3	Commercial	760	25.7	35720	49.8	85696	52.4
	4	Institutional	8	0.3	625	0.9	1379	0.8
	5	Industrial	14	0.5	484	0.7	1000	0.6

Algorithm 2: Synthetic Minority Over-sampling Technique – Nominal (SMOTEN-N) Data Generation

```
Input: $D = \{x_1,\ y_1\},\ \ y_s\ \in Y = \{1, \ldots C\}$ // training dataset D with n samples, & p dimensional feature space
Output: Synthetic data
Calculate the distance metric for each sample using modifies value difference metric MVDM
$\delta(V_1, V_2) = \sum_{i=1}^{h} | \frac{C_{1i}}{C_1} - \frac{C_{2i}}{C_2} |$
For each $x_i$ in minority class
for $i = 1$ to $m_s$:
    for $j = i + 1$ to $m_s$:
        $\delta(V_1, V_2) = 0$
        for $f = len(p)$:
            if instance (X[0, f], nominal):
                Calculate the MVMD distance for categorical features
                $C_1 = \sum_k [X[y = m_s][k, f] = [X[i, f]$ Indicator function for $C_1$
                $C_2 = \sum_k [X[y = m_s][k, f] = [X[i, f]$ Indicator function for $C_2$
                $\delta(V_1, V_2) = \sum_{i=1}^{h} | \frac{C_{1i}}{C_1} - \frac{C_{2i}}{C_2} |$
            else:
            Calculate the Euclidean distance for numerical features
        end for
    end for
end for
Compute the distance between two feature vectors
$\Delta(X, Y) = w_x w_y \sum_{b=1}^{p} \delta(x_b, y_b)^r$
return X, Y
end function
```

4.3. Treatments for Imbalanced Data

In the domain of machine learning, there are a variety of methods that can be employed to tackle imbalanced crash data. The consequence of neglecting this discrepancy can result into a biased model, particularly for the minority class. In this dissertation, two data imbalance treatment methods were employed to assess how they influence the prediction of crash severity related to weather conditions.

4.3.1. Synthetic Minority Over-sampling Technique – Nominal (SMOTE-N)

The SMOTE, introduced by Chawla et al. (2002), addresses class imbalance by over-sampling minority instances through the creation of synthetic data points. SMOTE-N, as presented in Algorithm 2, is an extension designed for nominal datasets using a modified version of the value difference metric (MVDM) to measure the distance between categorical feature values. The MVDM considers the occurrences of values and their response classes. The distance between feature vectors is calculated using a weighted Euclidean or Manhattan distance. In the case of SMOTE-N, weights are often disregarded, as it primarily aims to balance data distribution between classes rather than direct classification.

4.3.2. Adaptive Synthetic – Nominal (ADASYN-N)

ADASYN-N is designed to address class imbalance with nominal predictors as an extension of ADASYN (He et al., 2008) initially intended for numeric predictors. In ADASYN-N, as presented in **Algorithm 3**, the aim is to generate synthetic data for the minority class in a way that prioritize events that are challenging to learn. The process with a training dataset containing n samples, where x_i represents data in a p-dimensional feature space X, and y_i is the class label. The number of synthetic data points to be generated (G) is determined based on the desired level of balance, controlled by the parameter β (ranging from 0 to 1). For each x_i in the minority class, the algorithm calculates k nearest neighbors in the p-dimensional space and computes r_i, which represents the ratio of majority events among these nearest neighbors. Higher r_i values indicate events that are more challenging to learn. These r_i values are normalized, ensuring that the sum of normalized r_i values equal 1.

Algorithm 3: Adaptive Synthetic – Nominal (ADASYN-N) Data Generation

```
Input: $D = \{x_1,\ y_1\},\ y_s\ \in Y = \{1, \dots C\}$ // training dataset D with n samples, and p dimensional feature space
Output: Synthetic data
Calculate the number of synthetic data to be generated
$G = (m_l -\ m_s) \text{ x } \beta$
Separate dataset into minority and majority classes
    $m_s$ = number of instances in minority class
    $m_l$ = number of instances in majority class
For each $x_i$ in minority class
for i = 1 to $m_s$:
    Calculate the number of majority instances in nearest neighbor
    $H_i$ = number of majority instances in nearest neighbor
    Calculate $r_i$ as ratio of majority domination in k-nearest neighbors
    $r_i = \frac{H_i}{k}$
    Normalize $r_i$
    $\widehat{r_i} = \frac{r_i}{\sum_{i=1}^{m_s} r_i}$
    Calculate number of Synthetic data to be generated
    $g_i = \ \widehat{r_i} \text{ x } G$
    Generate data by replicating $x_i$
    GenerateData $(x_i, g_i)$
end for
Generate Data function
function GenerateData$(x_i, g_i)$
    Generate data by replicating $x_i$
    for i = 1 to $g_i$:
        syntheticData = replicate $x_i$
    end for
end function
```

4.4. Feature Selection

A variable section technique called permutation feature importance (PFI) was employed as presented in Algorithm 4. The idea behind this technique, as presented in Figure 16, is to destroy a feature of interest $x_j \in X$ by perturbing it such that it becomes uninformative.

For example, observations in x$_j$ are randomly permuted where the marginal distribution P(x$_j$) stays the same.

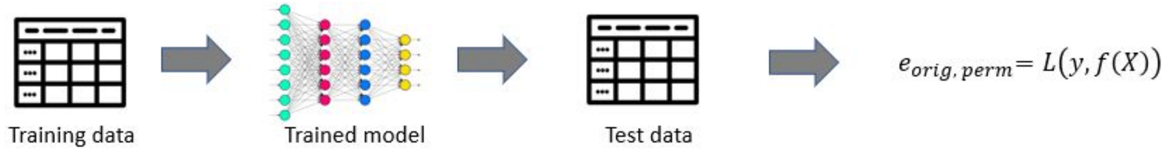


Figure 16. Permutation feature importance

Algorithm 4: Permutation Feature Importance

```
Input: trained model f; feature matrix X, target vector y, loss function L (y, f)
Output: Sorted list of features by descending permutation feature importance PFI
Calculate the original model error $e_{orig}$
$e_{orig} = L(y, f(X))$
Initialize an empty list to store feature importance values: PFI_values
For each feature j from 1 to the number of feature p:
    for j = 1 to p + 1:
  Create a permutated feature matrix $x_{perm}$ by randomly shuffling the value of feature j in the data X
  Estimate the permuted error $e_{perm}$ using the error measured:
  $e_{perm} = L(y, f(x_{perm}))$
    end for
Calculate the permutation feature importance $PFI_j$ for each feature j using:
  PFI as the difference between original error and permuted error:
  $PFI_j = e_{perm} - e_{orig}$
Append the calculated $PFI_j$ value to the PFI_value list
Sort the feature in descending order of their PFI values
end function
```

The error without permuting the features and with permuted feature values are measured. Repetition of the feature permutation was done 500 times and the average of the differences of both errors was computed. Figure 17 shows how the result of the PFI is interpreted. For example, if crash type is the most important variable, it implies that destroying information about crash type by permuting it increases the error of the model the most. This interpretation of the PFI helps to understand which features the model is most sensitive to and guides model refinement.

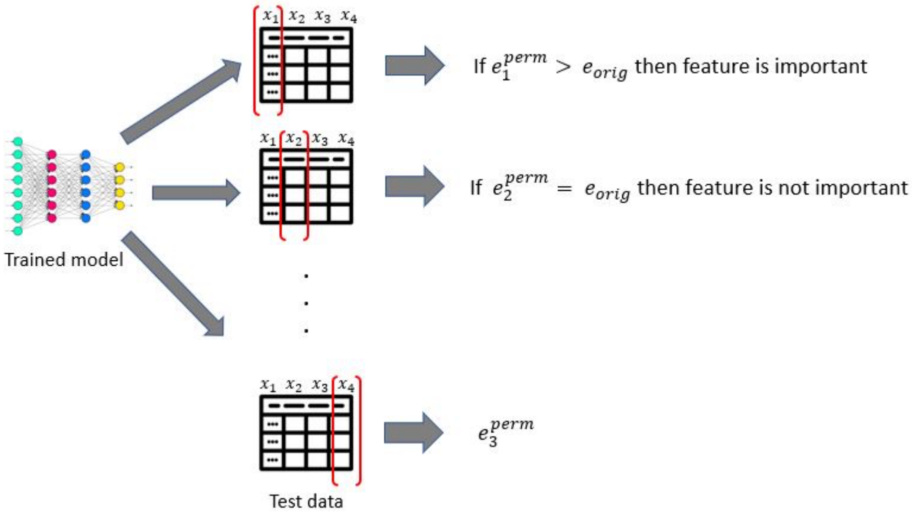

x_1 x_2 x_3 x_4
If $e_1^{perm} > e_{orig}$ then feature is important
x_1 x_2 x_3 x_4
If $e_2^{perm} = e_{orig}$ then feature is not important
Trained model
x_1 x_2 x_3 x_4
e_3^{perm}
Test data

Figure 17. Interpretation of permutation feature importance

4.5. Machine Learning Techniques

4.5.1. Extreme Gradient Boosting (XGBoost)

An XGBoost model is an ensemble of decision trees, where each subsequent tree tries to correct the errors made by the previous ones. It's an iterative process that aims to minimize a loss function. Mathematically, the prediction $\hat{y}$ of an XGBoost model with K trees for an input x can be represented by Equation 12.

$$\hat{y} = \sum_{k=1}^{K} f_k(x) \tag{12}$$

where, $f_k(x)$ is the prediction of the k^th tree.

The objective function that XGBoost tries to minimize is represented by Equation 13.

$$obj(\theta) = L(\theta) + \Omega(\theta) \tag{13}$$

where, θ represents the parameters of the model, $L(\theta)$ is the training loss function, and $\Omega(\theta)$ is a regularization term that controls the complexity of the model.

In the case of a classification problem like traffic crash severity prediction, $L(\theta)$ is typically the log loss for binary classification, or the softmax loss for multiclass classification.

The “gradient boosting” part of XGBoost comes from the fact that it trains each new tree to predict the negative gradient (or “residual”) of the loss function with respect to the current predictions. This is why it's called “gradient boosting”, as it uses gradient information to boost the performance of the ensemble. The regularization term $\Omega(\theta)$ in the objective function is what distinguishes XGBoost from regular Gradient Boosting. In XGBoost, $\Omega(\theta)$ is given by Equation 14.

$$\Omega(\theta) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j{}^2 \qquad (14)$$

where, T is the number of leaves in the tree, w_j are the scores on the leaves, γ controls the complexity of the model (the number of leaves in the trees), and λ controls the L2 regularization on the leaf scores. This regularization term helps to prevent overfitting by penalizing complex models.

4.5.2. Random Forest (RF) Model

RF is an ensemble learning method that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Suppose a RF model consists of N decision trees (Biau et al., 2012). Each tree gives a classification, and the tree “votes” for that class. The forest chooses the classification having the most votes (over all the trees in the forest). Each tree is grown as follows: 1) if the number of cases in the training set is N, then N cases are sampled at random - but with replacement from the original data. This sample will be the training set for growing the tree; 2) if there are M input variables, a number m is specified such that at each node, m variables are selected at random out of the M and the best split on these m is used to split the node. The value of m is held constant during the forest growing; 3) Each tree is grown to the largest extent possible and there is no pruning.

For a given test record, each tree in the forest gives a classification. The forest chooses the classification having the most votes (over all the trees in the forest) and in case of regression, it takes the average of outputs by different trees. Mathematically, the prediction of a RF model for an input x can be defined as shown in Equation 15.

$$\hat{\mathrm{y}} = \frac{1}{\mathrm{N}}\sum_{\mathrm{i}=1}^{\mathrm{N}} \mathrm{f}_{\mathrm{i}}(\mathrm{x}) \qquad (15)$$

where $\mathrm{f}_{\mathrm{i}}(\mathrm{x})$ is the prediction of the i^{th} decision tree.

The model would take as input features of a traffic incident (such as speed, weather condition, time of day, etc.), and output a severity class (severe injury, moderate injury, PDO). The model would be trained on a labeled dataset, and the aim would be to minimize the discrepancy between the predicted and actual labels. RF's ability to combine multiple decision trees helps it to avoid overfitting and generally results in a robust prediction performance.

4.4. Results

The results of data imbalance treatment using SMOTE-N and ADASYN-N are presented in this section. The dataset from both treatment methods is applied to RF and XGBoost. The dependent variable has three ordered levels: '1' is severe injury crash, '2' is moderate injury crash, and '3' is PDO crash.

4.4.1. Feature Extraction

Figure 18 represents the permutation feature importance as determined by the RF model. The permutation feature importance is a technique for estimating the contributions of individual features to the predictive power of a model by observing the effect on model performance by randomly permuting the values of each feature, one at a time (Fiorentini & Losa, 2020). In this study, the RF model was trained on a subset of the dataset. A representative dataset was used for this process because this technique is computationally demanding. In the plot, each bar corresponds to a specific feature in the dataset, and the length of the bar corresponds to the importance of that feature. Positive values indicate that the performance of the model decreases when the feature is shuffled, suggesting that the model relies on the feature to make accurate predictions. In other words, when the feature is perturbed or randomized (shuffled), the model’s predictability suffers because it loses access to the valuable information contained within that feature. Therefore, a decrease in performance upon shuffling implies that the feature plays a crucial role in the model’s overall accuracy. This highlights the importance of the feature in the predictive process, as it significantly contributes to the model's ability to make accurate predictions. Conversely, negative values indicate that the performance of the model improves when the feature is shuffled, suggesting that the model might be overfitting to noise in the feature.

The top three features obtained from this analysis are crash type, vehicle type, and locality suggesting that these features are the most important for predicting the severity of a crash. However, it is important to note a few caveats. First, while permutation feature importance provides a useful way to rank the importance of features, it does not provide any information about the nature of the relationship between each feature and the target variable (Ke et al., 2017). Second, it is possible that important features might appear unimportant if they are highly correlated with other features (Ke et al., 2017; Modal et al., 2020). Finally, the results might differ if the analysis were performed on the full dataset or a different sample because this analysis was performed on a sample of the original dataset.

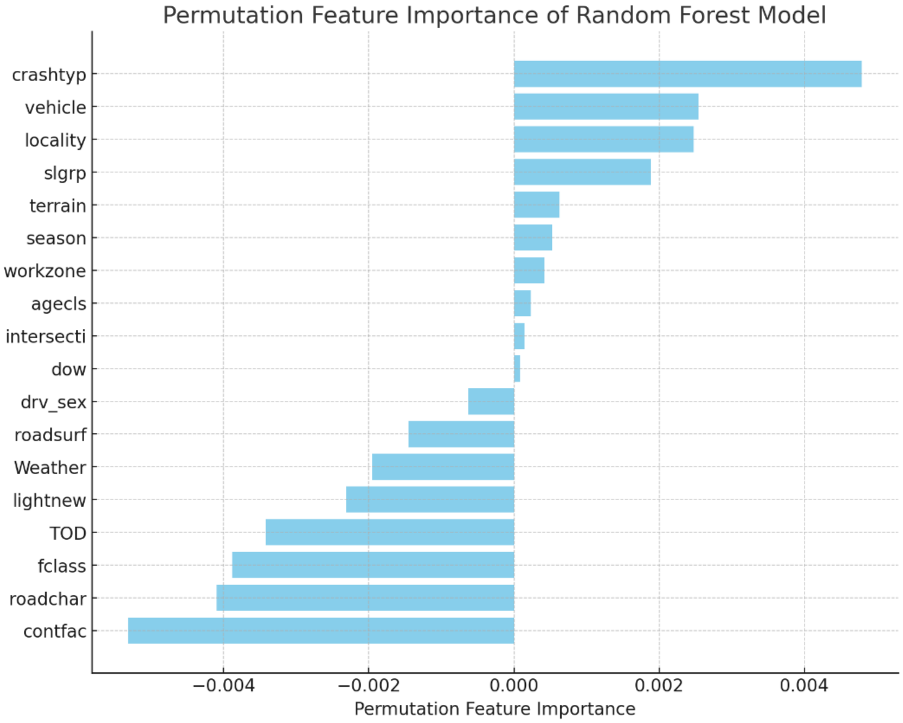


Figure 18. Permutation feature importance plot of variables described in Table 6

4.4.2. Comparing Treatments Method & Machine Learning Techniques

Table 7 presents the performance metrics for the two machine learning technique-based models, RF and XGBoost, applied to three different datasets. The metrics include accuracy, precision, recall, and F1-score. The effectiveness of the treatments in the datasets are compared across different machine learning techniques, allowing us to assess the robustness of the treatments. The confusion matrices for the RF and XGBoost models are summarized in Table 8. By doing this, the accuracy and the types of errors being made by the model are examined.

Table 7. Model performance metrics

Method	Model	Level of Crash Severity	Train			Test		
			Precision %	Recall %	F1 Score %	Precision %	Recall %	F1 Score %
CONTROL	RF	Severe Injury	99.95	89.69	94.54	69.44	6.76	12.32
		Moderate Injury	99.53	82.32	90.11	63.51	18.38	28.51
		PDO	92.65	99.84	96.11	72.32	95.93	82.47
	XGBoost	Severe Injury	99.97	93.54	96.65	9.48	10.32	9.88
		Moderate Injury	99.68	88.48	93.75	31.64	36.53	33.91
		PDO	94.62	99.92	97.20	49.98	58.19	53.77
SMOTE-N	RF	Severe Injury	99.60	99.9	99.75	22.65	10.41	14.26
		Moderate Injury	94.70	97.40	96.03	42.77	39.17	40.89
		PDO	97.51	94.40	95.93	74.08	77.53	75.77
	XGBoost	Severe Injury	99.16	99.72	99.44	17.68	18.78	18.21
		Moderate Injury	79.51	81.49	80.49	40.78	47.46	43.87
		PDO	81.21	78.74	79.96	75.31	69.64	72.36
ADASYN-N	RF	Severe Injury	98.85	99.88	99.36	13.79	19.46	16.14
		Moderate Injury	92.99	97.79	95.33	40.89	51.07	45.42
		PDO	98.19	91.89	94.94	76.16	67.32	71.47
	XGBoost	Severe Injury	96.22	99.36	97.76	10.97	25.27	15.30
		Moderate Injury	75.31	80.49	77.81	38.86	53.33	44.96
		PDO	79.71	71.41	75.33	76.13	61.96	68.32

Table 8. Confusion matrix for RF and XGBoost

		Reference											
		RF						XGBoost					
Method	Predicted	Train			Test			Train			Test		
		Severe Injury	Moderate Injury	PDO	Severe Injury	Moderate Injury	PDO	Severe Injury	Moderate Injury	PDO	Severe Injury	Moderate Injury	PDO
CONTROL	**Severe Injury**	1,984	1	0	50	15	7	1,136	12	9	76	22	12
	Moderate Injury	19	44,293	192	230	3,286	1,658	223	15,138	2,505	345	6,530	27,049
	PDO	209	9,517	122,474	460	14,576	39,281	853	38,661	120,152	319	11,325	13,885
SMOTE-N	**Severe Injury**	122,523	300	236	77	109	154	122,319	528	511	139	297	350
	Moderate Injury	58	119,521	6,662	321	7,002	9,047	184	99,960	25,572	322	8,485	11,999
	PDO	85	2,845	115,768	342	10,766	31,745	163	22,178	96,583	279	9,095	28,597
ADASYN-N	**Severe Injury**	122,054	781	634	144	389	511	121,423	2,828	1,946	187	629	889
	Moderate Injury	103	124,799	9,309	328	9,129	12,871	554	102,720	33,130	314	9,533	14,685
	PDO	49	2,034	112,723	268	8,359	27,564	229	22,066	87,590	239	7,715	25,372

4.4.3. Best Treatment Method

The focus is to assess the effect of the selected methods on the classification of crash severity into three categories: severe injury, moderate injury, and PDO. The primary evaluation metric used was the F1 score, which considers both precision and recall providing a balanced performance assessment. For the RF Model, the best-performing data imbalance treatment method was ADASYN-N, achieving a test F1 score of 45.42. ADASYN-N effectively addressed the class imbalance in case of moderate injury crash, resulting in improved predictions for this category. However, the control dataset proved most effective for handling the class imbalance specific to PDO, which has a test F1 score of 82.47. ADASYN-N was also identified as the best method for addressing the imbalance in case of severe injury crashes, achieving a test F1 score of 16.14. ADASYN-N significantly improved the model's ability to predict severe injury crashes.

For the XGBoost Model, the data imbalance treatment method that yielded the highest test F1 score of 44.96 was ADASYN-N for predicting moderate injury crashes. This data treatment method effectively handled the imbalance in the moderate injury crash category, leading to improved predictions for this severity level compared to the control dataset. SMOTE-N emerged as the best-performing method for PDO crashes, with a test F1 score of 72.36. SMOTE-N successfully mitigated the class imbalance issue, resulting in more accurate predictions for PDO crashes. SMOTE-N was identified as the most effective data imbalance treatment method for severe injury crashes, achieving a test F1 score of 18.21. SMOTE-N significantly improved the predictive performance for this severity category.

A notable consideration is that the choice of evaluation metric has a significant effect on the assessment of data imbalance treatment methods. While the F1 score was used as the primary metric, other metrics prioritizing different aspects of classification performance may lead to different rankings of the methods. For example, if recall is prioritized over precision, the best-performing data imbalance treatment method might change.

In machine learning, the difference in performance between the training and test datasets can provide valuable insights into how well the model is learning from the data. In this dissertation, the F1 score was used to evaluate this, which balances both precision and recall. A large difference between the training and test F1 scores could indicate overfitting, while a small or negative difference might suggest underfitting or a good fit, depending on the absolute scores.

4.4.4. Model Fit on Datasets

Overfitting is a common issue in machine learning where the model learns the training data so well that it includes noise or random fluctuations. This results in a model that performs extremely well on the training data but poorly on unseen test data (Figure 19). Potential overfitting was observed in some scenarios.

- The XGBoost model on the control dataset for severe injury crashes demonstrated the most pronounced case of overfitting, with an F1 score difference of 86.77.
- RF model with the SMOTE-N method for severe injury crashes and the RF model with the ADASYN-N method for severe injury crashes also showed large F1 score differences, indicating potential overfitting.

In these cases, the models achieved very high F1 scores on the training data, suggesting they were able to capture the nuances of the training data very well. However, their performance dropped significantly on the test data (Figure 19), suggesting that they may have overfit to the noise or specific patterns in the training data that do not generalize the unseen data. On the other hand, a small or negative difference in F1 scores between the training and test datasets could suggest that the model is underfitting or fitting the data well. Underfitting happens when a model is too simple to capture the complexity of the data, performing poorly on both the training and test data. If the model fits the data well, it will have good performance on both datasets. From the analysis, the following scenarios demonstrate good fitting.

- The XGBoost model with the ADASYN-N method for PDO crashes has an F1 difference of 7.01.
- The XGBoost model with the SMOTE-N method for PDO crashes has an F1 difference of 7.60.
- The RF model on the control dataset for PDO crashes has an F1 difference of 13.64.

In these cases, the models achieved reasonable F1 scores on both the training and test data, suggesting they were able to generalize well the unseen data. These observations highlight the importance of evaluating the performance of a model not only on the training data but also on unseen test data. If a model is overfitting, strategies such as simplifying the model, collecting more data, or using regularization or cross-validation might help to improve its generalization. If a model is underfitting, making the model more complex or engineering better features could help it capture the underlying patterns in the data.

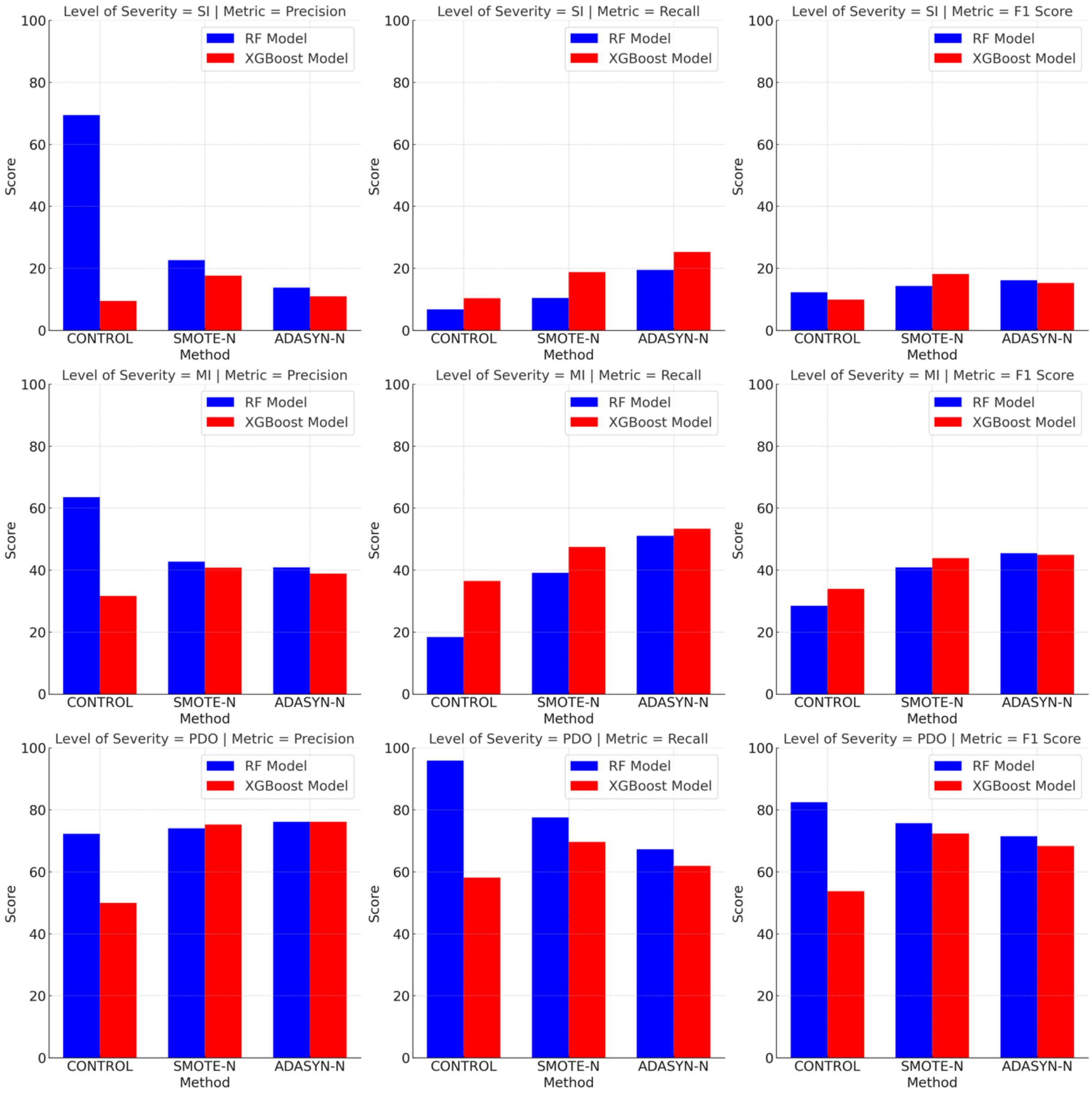


Figure 19. Performance metrics on test dataset

4.5. Discussion

The results delve into the effectiveness of data imbalance treatment methods when used on different machine learning algorithms in the context of weather-related crash severity analysis. The findings demonstrate that the choice of method significantly influences the model's ability to handle imbalanced data, and the effect varies depending on the severity category being predicted and the specific machine learning algorithm being used. One noteworthy observation is the consistent effectiveness of the ADASYN-N method for addressing moderate and severe injury crashes in both the RF and XGBoost models. ADASYN-N appears to be a robust approach for generating synthetic samples of the minority class, effectively balancing the class distribution, and enhancing the model's performance in predicting severe and moderate injury crashes. The consistent success of ADASYN-N across different algorithms indicates its potential as a reliable data imbalance treatment method for these crash severity levels. Fountas et al. (2020) argued that the generation of data for weather-related crashes involves underlying latent processes that should be considered during analysis. As these processes as essential for enhancing crash severity prediction accuracy, various statistical methods were proposed to address this issue effectively (Cai et al., 2018; Fountas et al., 2020; Fountas et al., 2021). ADASYN-N appears to address this limitation and is more resilient to the heterogeneity introduced in the dataset, consistently performing well across all the algorithms tested.

ADASYN-N's effectiveness in predicting moderate and severe injuries can be attributed to the specific strengths of the ADASYN-N approach. Unlike standard over-sampling methods, ADASYN-N adaptively adjusts the synthetic sample generation based on the learning difficulty of minority classes. This feature is beneficial for severe and moderate injury crashes, which might have more complex or less consistent patterns than PDO crashes. By focusing more on the harder-to-learn examples, ADASYN-N help create a more balanced and representative training set for these types of crashes.

The notable performance of ADASYN-N in predicting moderate injury crashes suggests that this category, while not as rare as severe injury crashes, still benefits significantly from a more balanced representation in the training data. The method's data generation process helps the model better understand the variations in moderate injury crashes, improving its predictability in this class. While both ADASYN-N and SMOTE-N improve the model's ability to predict severe injury crashes, the improvement is marginal. In addition, the model may become too attuned to the nuances of the training data i.e., there is a potential risk of overfitting, reducing its generalizability to real-world situations. In contrast, the prediction of moderate injury crashes, supported by a larger data set, tends to be more robust and generalizable.

The marginal improvement in severe crash prediction achieved in this dissertation, while seemingly small, is nonetheless significant, especially considering the high stakes involved in severe crash scenarios and the challenges presented by the highly imbalanced dataset. Even a slight enhancement in predicting severe crashes can be crucial. These crashes, though less frequent, often have the most devastating consequences, including serious fatalities or injuries (Martins et al., 2022). Therefore, any improvement in accurately identifying potential severe crashes, no matter how marginal, is valuable as it could contribute to life-saving interventions, more effective emergency response planning, and targeted safety measures (Hansson, 2022). In the scenario of weather-related crashes, a marginal improvement indicates that the model is overcoming some of the inherent biases and learning meaningful patterns related to severe crashes.

High model performance on training data is often achievable with complex methods. However, these methods may not generalize well to unseen data, which is a phenomenon known as overfitting (Salman & Liu, 2019). Conversely, a method that is too simple may not perform well even on the training data, leading to underfitting, and hence also lacking generalization (Chung et al., 2018; Salman & Liu, 2019). A pronounced case of overfitting was observed in some cases, notably the XGBoost model on the control dataset for severe injury crashes and the RF model using both SMOTE-N and ADASYN-N methods for the same crash category. This is evidenced by high F1 scores on training data but a substantial drop in performance on testing data, indicating these methods captured training data nuances, including noise, which did not generalize well to unseen/uncaptured data. This result shows that while these methods can improve model learning for underrepresented classes, they also introduce the risk of overfitting, especially when the synthetic samples do not perfectly represent the real-world data distribution (Fiorentini & Losa, 2020). In such cases, the importance of interpretability becomes paramount; understanding the model's decision-making process allows practitioners to discern whether predictions are based on genuine patterns or artifacts of the synthetic data (Antoniadi et al., 2021). Therefore, one must weigh the benefits of improved predictions against the potential loss of transparency when settling for a machine learning technique, ensuring the chosen model maintains a level of interpretability that supports reliable and actionable insights in practical applications.

Furthermore, the dissertation highlights the superior performance of the SMOTE-N method for the XGBoost model in predicting PDO crashes. This observation suggests that the SMOTE-N effectively addresses the class imbalance present in the PDO crashes category for the XGBoost algorithm. This finding emphasizes the significance of selecting data imbalance treatment methods tailored to the characteristics of the machine learning technique and crash severity type to achieve optimal performance.

CHAPTER 5: PREDICTING FUTURE WEATHER-RELATED CRASH RISK USING MACHINE LEARNING TECHNIQUE

5.1. Introduction

The fast urbanization of most United States cities has introduced both safety and sustainability challenges. These challenges are even worse in states with epileptic weather condition such as in NC. The Federal Highway Administration (FHWA) reported that 21% of all crashes are weather-related (FHWA, 2023). Between 2007 to 2016, nearly 5,400 persons were killed due to weather-related crashes making it one of the top contributing factors in traffic crashes (FHWA, 2023). Thus, intelligent transportation systems (ITS) has become an active research area given its potential to reduce crashes in poor weather conditions (Ran et al., 2012). As an essential step towards improving the ITS, weather-related crash prediction aims at projecting into the future crash status at specific location within a traffic system.

Spatial analysis of crash data is becoming more increasingly popular. In the last decades, researchers have made considerable effort in analyzing crash data at various spatial levels (Aguero-Valverde & Jovanis, 2006; Pulugurtha et al., 2007; Quddus, 2008; Plug et al., 2011; Pulugurtha et al. 2013; Ogungbire & Pulugurtha, 2024). In the most recent work (Ogungbire & Pulugurtha, 2024), a precedence was set for space-time cube theory in crash risk analysis. The preference of state DOTs to examine crash data at spatially aggregated levels, including traffic analysis zones (TAZ) (Pulugurtha et al., 2013l Bai et al., 2017) and grid-level spatial units (Ogungbire & Pulugurtha, 2024; Wu et al., 2023), is discussed to easily facilitate effective resource allocation (Roland et al., 2021). This part of the dissertation builds upon the groundwork laid in (Ogungbire & Pulugurtha, 2024) to predict potential future risk of weather-related crash events at grid level.

The availability of large datasets related to human activities in urban settings has spurred a significant increase in research on traffic incidents (Gonzalez et al., 2008; Hasan et al., 2013; Bao et al., 2017). This wealth of data offers an unprecedented opportunity to learn from historical events to better understand and predict future traffic incidents. Recent research efforts have explored integrating big data into spatially aggregated crash models (Bao et al., 2017; Xhao et al., 2024). A study by Bao et al. (2017) investigated the use of big data derived from traffic sensors and social media to improve spatially aggregated crash models. They developed a methodology that combines traditional traffic data with real-time social media analytics to predict crash hotspots in urban areas.

Early studies have been formulated to view crash prediction as either a classification or a regression problem. For example, some work aimed to predict the likelihood of a crash occurrence at specific location (Duddu & Pulugurtha, 2017; Pulugurtha et al., 2013; Gajera et al., 2023) or time periods (Pulugurtha et al., 2013). Looking at crash prediction from this lens allow researchers to identify significant predictors of crashes and quantify their impacts. Conversely some studies are focused on estimating the intensity of crashes at specific location during each time window (Pulugurtha & Mahanthi, 2016; Duddu & Pulugurtha, 2017; Najaf et al., 2018; Kalambay & Pulugurtha, 2022). Over time, the field has seen the integration of more sophisticated approaches, including time-series analysis (Feng et al., 2020; Khan et al., 2022) for understanding temporal patterns and machine learning techniques (Iranitalab & Khattak, 2017; Ogungbire et al., 2023) for capturing complex, non-linear relationships between variables. Geographic information systems (GIS) have also been applied to spatially analyze crash data and identify high-risk areas (Pulugurtha et al., 2007). These methods have evolved from simple, deterministic models to dynamic, probabilistic models that better account for the uncertainties inherent in predicting human behavior and environmental interactions.

Deep learning approach, which is gaining widespread popularity in computer vision (Loo et al., 2023), natural language processing (Valcamonico et al., 2022), artificial intelligence, and pattern recognition (Farmosa et al., 2020; Bibi et al., 2021), is now being applied in traffic safety research. This includes applications such as traffic conflict prediction (Farmosa et al., 2020; Bibi et al., 2021), near miss identification at intersections (Huang et al., 2020), estimating unsafe driving speed, and red-light violation at signalized intersection (Zhang, 2020). Deep learning distinguishes itself from traditional statistical models and other learning architectures by its ability to model complex non-linear relationships through distributed and hierarchical feature representation (Shi et al., 2015), demonstrating superior performance in predicting short-term traffic flow and speed.

The predictive potential of a spatially ensembled ConvLSTM model was utilized to predict the future state of weather-related crashes in this study. Through a rigorous comparison of the model's performance against other established models, this study aims to establish a new benchmark in weather-related crash prediction. To investigate whether the proposed framework surpasses traditional models and the standard ConvLSTM in accuracy, assessing its performance variability across different risk zones and confirming the geographical accuracy of its predictions against actual crash locations is vital.

5.2. Methodology

This section presents the data used in the study, introduces the formulation of the problem, and present the feature extraction technique.

5.2.1. Data Sources

NC was selected as the study area. The state of NC is characterized by its wide range of weather phenomena, including but not limited to snowfalls, rainfall, and wind. This diversity in weather conditions makes NC an exemplary state for examining the effects of weather, particularly precipitation, on road traffic incidents, a point made in a study by Mathew et al. (2022) in their research. The primary dataset for the investigation was sourced from the HSIS, encompassing vehicle crash records spanning 2015 through 2018. This dataset is not merely a compilation of the time and location of each crash; it also contains a few significant attributes related to the road conditions. The spatial distribution of these crash sites across NC is illustrated in Figure 20.

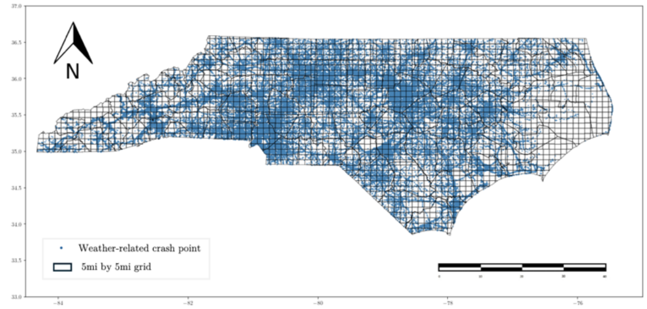


Figure 20. Spatial distribution of crashes masked by a grid layer

5.2.2. Problem Definition

The objective is to construct a predictive framework that estimates the total number of traffic crashes within specified units of a spatial grid S over distinct time intervals. This grid, denoted as $S = \{s_1, s_2, \dots, s_n\}$ comprises subdivisions, each representing an area of $d \times d$ square miles. For illustrative purposes, consider $d = 5\text{mi}$, whereby the entire geographical expanse of NC could be dissected into a grid formation of 2,045 units. Time is segmented into discrete intervals, referred to as slots, with a week being the standard length for this analysis, albeit the methodology supports adjustments in both spatial (d) and temporal (t) dimensions. Figure 21 shows the EPDO trends of the training dataset. The problem is formulated as follows.

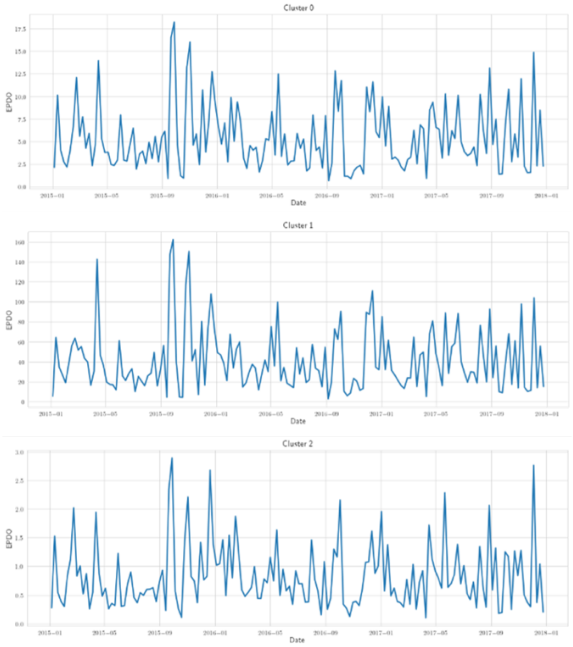


Figure 21. Training data from DTW clusters between 2015 to 2017

Provided Inputs:

- A spatiotemporal framework encapsulated by a matrix $S \times T$, where S symbolizes the spatial grid with its divisions, and T embodies the sequence of time intervals during the study period.
- A matrix A of dimensions $n \times t$, with element A_{ij} representing the crash EPDOs in spatial unit s_i at time t_j.
- A series of m matrices $\{M_1, M_2, \dots, M_m\}$, with each matrix M_k of dimensions $n \times t$, capturing distinct attributes pertinent to each grid unit s_i over the time slots t_j.
- Training data set D_{train} composed of pairs from A and the feature matrices for time slots within T_{train}, and a testing data set D_{test} containing pairs for time slots in T_{test}.

Objective:

- Formulate a model that can accurately predict the crash count matrix A for all time intervals $t_j \in T_{test}$, aiming to minimize the discrepancy between predicted and actual crash counts.

Constraints:

- The correlation between crash counts and features (M_k) varies across different spatial units.
- Crashes are presumed to occur exclusively within the confines of the road network.
- For any forthcoming timeslot t_i, the corresponding feature matrices$M_{k,ti}$ are not accessible for use in predicting $A_{i,ti}$, signifying $t_i \in T_{test}$.

5.2.3. Feature Extraction

To prepare the features for the model, the dataset was aligned with each grid s_i and week t_i combination, aggregating the data to compile a list of features. For the dependent variable, the EPDO was computed for each grid s_i for each week t_i, from January 2015 to December 2018. The process for extracting the independent features is detailed next.

The road network was mapped onto grid cells, overlaying it with a mask layer to delineate the study area. It is important to note that traffic crashes are restricted to the road network, despite the grid-based partitioning of the entire area. The risk level in each grid was assessed based on two factors: crash frequency and crash severity. For more accurate predictions, the EPDO score was normalized for each grid by the total road length within that grid, assigning null values to grids without roads. Given the stability of the road network over time, this feature is considered time-invariant.

The network mask layer was augmented by calculating and storing two additional measures for each grid cell: the average road length and the average speed limit. Further, the associated features were incorporated with road properties, which include the proportion of different traffic control types, average number of lanes, proportion of different route types, road, and annual average daily traffic (AADT) with each grid cells s_i. These features are considered time-invariant.

5.2.4. Experimental Settings

This dissertation aims to predict weather-related crashes in NC for the year 2018, based on EPDO scores and other related factors observed over the preceding three years. The dataset, comprising weekly aggregated data, translates into a sequence of 157 frames for training and 52 frames for testing. In total, the four-year dataset yielded 209 such sequences. The training set includes data from 2015 to 2017, while the final year, 2018, constitutes the testing set. Additionally, 10% of the training data was reserved for validation purposes.

The geographical scope of the study involves partitioning NC into grids measuring 5mi by 5mi. For each week of 2018, the aim is to predict a traffic crash map utilizing the proposed ConvLSTM model. It is guided by three research questions: (1) Does the framework outperform conventional predictive models and the standard ConvLSTM in accuracy? (2) What variations in performance does the proposed model exhibit across different crash risk zones, such as areas of high and low risk? (3) Do the model's predictions align spatially with actual crash locations, thereby confirming their logical validity?

To assess the models' precision, the mean squared error (MSE) and root-mean-square error (RMSE) were used as our primary metrics. Furthermore, the Cross-K function was used to evaluate the spatial correlation between the predicted outcomes and the actual data.

5.2.5. Spatiotemporal Ensembled ConvLSTM

The ConvLSTM model, an extension of the traditional LSTM, was initially developed by Shi et al (2015) for precipitation nowcasting. It is particularly suited for handling data where both spatial and temporal dimensions are crucial. Each input to the ConvLSTM network is treated as a 3D spatiotemporal tensor. The typical LSTM node is modified in the ConvLSTM to include convolution operations within its structure, as illustrated in a single ConvLSTM shown in Figure 22. Specifically, the input-to-state and state-to-state transitions in a ConvLSTM cell involve convolutional operations which output 3-D tensors. These modifications are governed by algorithm 5.

Algorithm 5: Spatial ensembled ConvLSTM

```
Initialize model parameters for each window
for training epoch in num_epochs:
    for each window in study_area:
        Extract spatiotemporal data x within window
        for each timestep t in x:
            Compute input gate:
                $i_t = \sigma(W_{xi} * X_t + W_{hi} * h_{t-1} + b_i)$
            Compute for gate:
                $f_t = \sigma(W_{xf} * X_t + W_{hf} * h_{t-1} + b_f)$
            Compute output gate:
                $o_t = \sigma(W_{xo} * X_t + W_{ho} * h_{t-1} + b_o)$
            Update cell state:
                $C_t = f_t . C_{t-1} + i_t . \tanh (W_{xc} * X_t + W_{hc} * h_{t-1} + b_c)$
            Hidden state output
                $h_t = o_t . \tanh (C_t)$
        Store last hidden state $h_t$
    Aggregate and store output from all frames
Use ensembled method to combine predictions from all windows
Evaluate model performance
Adjust parameters based on gradients and learning rate
```

To tackle the challenge of spatial heterogeneity in weather-related crash prediction, the approach involves constructing distinct LSTM models for various clusters within the study area. These clusters are determined based on the spatial heterogeneity of the data, which can reflect varying risk levels, such as high-risk or low-risk zones as shown in Figure 23. The ensemble method is then applied to integrate the results from multiple models, thereby mitigating the effects of data heterogeneity.

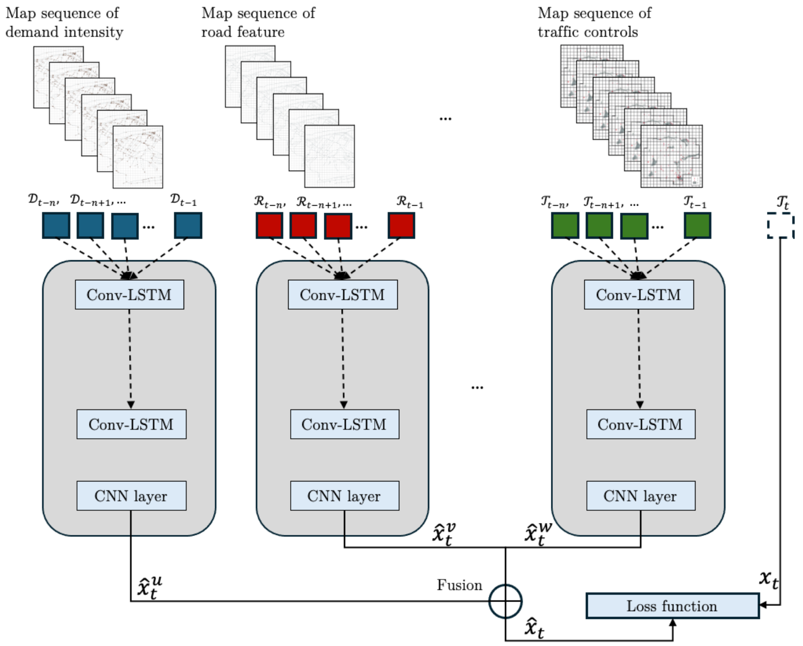

Map sequence of demand intensity
Map sequence of road feature
Map sequence of traffic controls
...
$\mathcal{D}_{t-n}, \mathcal{D}_{t-n+1}, \ldots$
$\mathcal{D}_{t-1}$
$\mathcal{R}_{t-n}, \mathcal{R}_{t-n+1}, \ldots$
$\mathcal{R}_{t-1}$
$\mathcal{T}_{t-n}, \mathcal{T}_{t-n+1}, \ldots$
$\mathcal{T}_{t-1}$
$\mathcal{T}_t$
Conv-LSTM
Conv-LSTM
CNN layer
Conv-LSTM
Conv-LSTM
CNN layer
...
Conv-LSTM
Conv-LSTM
CNN layer
$\hat{x}_t^v$
$\hat{x}_t^w$
$\hat{x}_t^u$
Fusion
x_t
Loss function
$\hat{x}_t$

Figure 22. Single ConvLSTM Architecture

A moving window technique was employed with each window measuring 32×32 grids. The study area is segmented by shifting this window from the upper-left corner at coordinates (0,0) to the bottom-right at coordinates (128, 64). The windows are moved across the grid in steps of 16 units both horizontally and vertically, ensuring comprehensive coverage and overlap across the study area. This strategy enabled the capture of localized spatial features significant for accurate crash prediction.

For each windowed region, a dedicated ConvLSTM model was trained using the local training dataset corresponding to that window. The model then performed predictions on the testing dataset for that specific region. By training individual models on localized data, the unique spatial-temporal characteristics of each region were captured, which might be crucial due to varying meteorological and traffic conditions. The final prediction for a specific grid location s_i on day t_j is computed using an ensemble method, which aggregated predictions from all models covering s_i. This aggregation is performed as a weighted average of the predictions for s_i at t_j from all significant models (Equation 16).

$$\hat{C}(s,t) = \frac{1}{\sum_{k=1}^{N} w_k} \sum_{i=1}^{N} w_i \hat{C}_i(s,t) \times I(s \in W_i) \qquad (16)$$

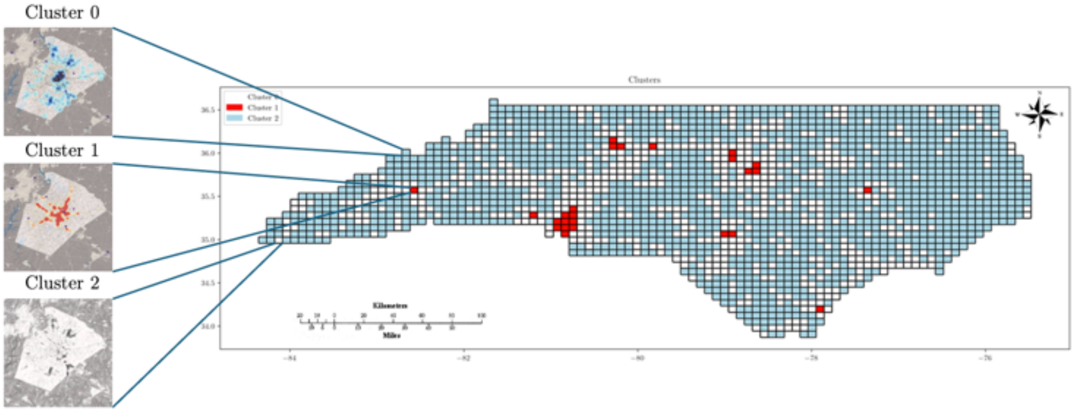


Figure 23. Distribution of frames by clusters

Here, N denotes the total number of window models that include the grid location with a weight assigned to the k^{th} window, and $I(s \in W_i)$ is an indicator function that equals 1 if s_i is within window W_i and 0 otherwise. Equal weights were considered for each model, $w_k = 1$, although optimal weights could potentially be determined through regression analysis of regional model outputs.

5.3. Results

Figure 24 presents a comparison of four predictive models: LR, ARIMA, ConvLSTM, and Spatiotemporal Ensembled ConvLSTM, using the cross-K function to measure their accuracy in forecasting weather-related crashes. The cross-K function values, plotted against ‘distance,’ serve to evaluate how closely each model’s predictions align with actual events.

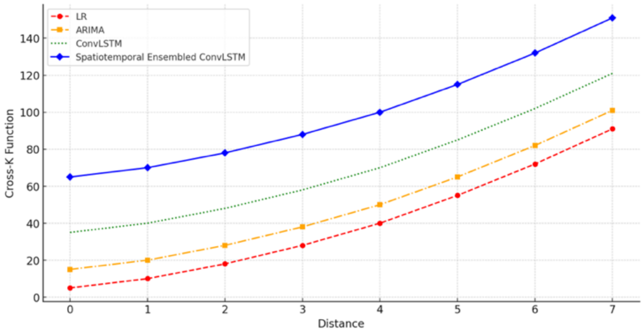


Figure 24. Cross-K function between predicted and actual weather-related crash risks

LR is observed to have the lowest performance among the four predictive models, with its cross-K function values consistently rising but lower than other models throughout the range. This shows that while LR can predict trends, its simplicity restricts its effectiveness in capturing complex patterns in weather-related crash data. ARIMA has better performance than LR, yet still falls short compared to the neural network-based models. Its ability to incorporate past values and forecast errors into future predictions does provide an edge over LR, yet it lacks the capability to effectively handle spatial or multidimensional temporal dependencies, which are crucial in the context of weather-related events.

ConvLSTM substantially outperformed LR and ARIMA, underscoring the advantages of integrating convolutional layers into LSTM networks. This architecture enables the model to capture spatial features and temporal sequences simultaneously, which is particularly beneficial for modeling scenarios like weather patterns where both spatial and temporal dynamics are significant. Spatiotemporal Ensembled ConvLSTM has the highest cross-K function values across all distances. This model combines multiple ConvLSTM models to leverage diverse spatial and temporal features more robustly, reducing the risk of overfitting to patterns and improving generalization across various scenarios.

The differences in performance can be attributed to several factors, for example, the model complexity and architecture, i.e., more complex models (ConvLSTM and spatiotemporal ensembled ConvLSTM) are designed to handle the intricacies of spatial and temporal data simultaneously, which is crucial for accurately modeling phenomena like weather-related crashes that exhibit both spatial and temporal variability. In addition, the data handling capabilities of each model plays a role. The ability of ConvLSTM to process data in both time and space allows for a more nuanced understanding of how weather conditions across different regions influence crash rates over time. Furthermore, the superior performance of the spatiotemporal ensembled ConvLSTM suggests that ensembling techniques, which combine predictions from multiple models to improve accuracy are particularly effective in dealing with complex, noisy datasets like those involving weather and traffic.

Table 9. Model performance evaluation

Model	Cluster 0		Cluster 1		Cluster 2		All regions	
	RMSE	MSE	RMSE	MSE	RMSE	MSE	RMSE	RMSE
LR	0.321	0.103	0.146	0.021	0.184	0.034	0.852	0.7259
ARIMA	0.288	0.082	0.091	0.008	0.151	0.023	0.543	0.2948
ConvLSTM	0.253	0.064	0.073	0.005	0.084	0.007	0.331	0.1096
Ensembled ConvLSTM	-	-	-	-	-	-	0.024	0.0006

Table 9 provides a comparative analysis of the mean squared error (MSE) and root mean squared error (RMSE) across different clusters for the four predictive models: LR, ARIMA, ConvLSTM, and spatiotemporal ensembled ConvLSTM. Each cluster represents different characteristics of EPDO, which measures the severity and frequency of crashes.

Cluster 0 represents gradually increasing low EPDOs. Here, LR, ARIMA, and ConvLSTM show progressively lower MSE and RMSE, indicating increasing accuracy with more sophisticated models. The improvement from LR to ARIMA and further to ConvLSTM suggests that the gradual increase in EPDO severity over time in this cluster is better modeled by algorithms that can handle time series data with trend and seasonality. However, cluster 1 represents volatile high-EPDOs. This cluster displays the lowest MSE and RMSE across all models, which may seem counterintuitive given its volatility. However, this can indicate that the models, particularly ConvLSTM, are effectively capturing the rapid fluctuations in EPDOs. The lower error metrics suggest that sophisticated models like ConvLSTM are particularly adept at managing the high variability within this cluster. Cluster 2, on the other hand, represents stable low EPDOs. Despite the stability in EPDOs, the errors (MSE and RMSE) are higher than in Cluster 1 but lower than in Cluster 0 for ConvLSTM. This might be because while the data's stability makes it easier to predict, the absolute errors remain low but perceptible, reflecting a consistent underestimation or overestimation by the models.

All region combined dataset was compared across models. When aggregating all clusters, it is notable that ConvLSTM and spatiotemporal ensembled ConvLSTM perform significantly better than simpler models. The ensemble method likely leverages individual model strengths and mitigates their weaknesses, leading to improved overall prediction accuracy. The increasing complexity and adaptability of the models (from LR to ensembled ConvLSTM) generally lead to better performance. ConvLSTM, integrating both convolutional and LSTM layers, efficiently handles spatial-temporal data, crucial for predicting EPDOs which are influenced by both spatial factors (e.g., road conditions and traffic density) and temporal factors (e.g., seasonal variations and time of day). The volatile nature of Cluster 1 might assist in model training by providing diverse scenarios for the models to learn from, which might explain the unexpectedly lower errors in this cluster compared to the more stable Cluster 2. In contrast, the stability in Cluster 2, while theoretically easier to predict, may lead to complacency in error reduction, resulting in slightly higher error metrics than Cluster 1. The superior performance of the spatiotemporal ensembled ConvLSTM in ‘All regions’ suggests that combining multiple models helps capture a broader range of patterns and anomalies in the data, thereby enhancing prediction accuracy. This is particularly beneficial when dealing with heterogeneous data across various regions.

CHAPTER 6: CONCLUSIONS & FUTURE WORK

6.1. Conclusions

The synopsis for each act of the dissertation is presented in this section.

6.1.1. A Spatiotemporal Risk Mapping of Statewide Weather-Related Traffic Crashes: A Machine Learning Technique

This dissertation focuses on how to analyze complex spatiotemporal weather-related crash pattern through a two-layered technique that captures the crash pattern similarities in time and spatial hotspot. Specifically, an unsupervised machine learning technique was used in studying these data generation pattern similarities. It was used in understanding how features impact weather-related crashes in both high and low risk crash zones using a supervised machine learning technique. The analysis presented in this study, through the application of a two-layered unsupervised machine learning technique, has helped to discern high-risk from low-risk locations and its effectiveness in identifying significant hotspots and coldspots within the clusters. The findings suggest a distinct variability in the impact of weather conditions on crash occurrences across different clusters, with unclear, foggy, and cloudy conditions playing a substantial role in all clusters, especially in Cluster C2 which exhibited the highest variability. Rain emerged as a consistent factor in crashes, particularly in Cluster C1, indicating a uniform influence across this cluster. Conversely, winter weather conditions and severe crosswinds had a lesser overall impact but showed significant spatial variability, particularly in Cluster C2.

In investigating the role of contributing factors in crash risk zones, the critical role of no passing zones, stop and go traffic controls, and roadway lighting in influencing the likelihood and severity of crashes under varying weather conditions in high-risk crash zones was identified. The consideration of these factors in statewide transportation planning and the subsequent implementation of targeted interventions could show promising result in reducing the risk of high-risk weather-related crashes.

6.1.2. Effectiveness of Crash Data Imbalance Treatment in Weather-Related Crash Severity Analysis

This dissertation successfully showcased the importance of choosing appropriate data treatment methods, specifically SMOTE-N and ADASYN-N, for handling nominal predictors in machine learning techniques applied to weather-related traffic crash severity prediction. The effectiveness of these methods varies depending on the crash severity level being predicted and the machine learning technique used.

ADASYN-N proved particularly effective in balancing class distribution and enhancing the accuracy of both RF and XGBoost models for severe and moderate injury crash predictions. In contrast, the control dataset showed notable efficiency with the RF model for predicting PDO crashes, while the SMOTE-N method significantly improved the XGBoost model's performance in the same category. This indicates that the XGBoost model benefited from a more balanced dataset provided by SMOTE-N, which likely offered more examples of the minority class to learn from. The high prediction accuracy of PDO crashes on RF model is not unusual as the model shows its bias to predicting the majority class.

The dissertation establishes a clear benchmark for practitioners in selecting suitable methods to generate synthetic samples for addressing underrepresented crash categories such as severe and moderate injury crashes. By demonstrating the varying effectiveness of SMOTE-N and ADASYN-N across different machine learning techniques and crash severity levels, this study provides valuable insights for optimizing data treatment in crash severity prediction. The insights gained from this dissertation are instrumental in guiding the development of more accurate and reliable crash severity prediction models, ultimately aiding in better informed and more effective traffic safety measures and policy decisions.

Despite these promising findings, further research is warranted to explore the applicability of the introduced synthetic data generation method to other weather-related datasets and enhance its performance in real-world scenarios. Assessing the technique's scalability and efficiency in handling large-scale datasets would be crucial, particularly for real-time applications or big data scenarios. Additionally, investigating the method's effectiveness on more complex and diverse datasets would contribute to a deeper understanding of its potential in various crash severity analysis contexts.

This dissertation demonstrates that SMOTE-N and ADASYN-N are effective data imbalance treatment methods for improving weather-related crash severity prediction in RF and XGBoost models, respectively. The effectiveness varies by crash severity category, with ADASYN-N excelling in severe and moderate injury crash predictions and SMOTE-N in PDO crash predictions. This study provides valuable guidance for researchers in choosing suitable techniques to create synthetic samples, particularly for underrepresented categories in weather-related crash severity prediction. The study paves the way for transportation agencies to develop more accurate crash severity prediction models. Such models are essential in enhancing traffic safety measures and building safer as well as more resilient roadways, especially in adverse weather conditions, ultimately protecting road users.

6.1.3. Predicting Future Weather-Related Crash Risk Using Machine Learning Technique

Several conclusions and recommendations can be made regarding the use of the spatially ensembled ConvLSTM framework for predicting weather-related crash risks. The spatially ensembled ConvLSTM framework does outperform conventional predictive models, such as LR, ARIMA, and the standard ConvLSTM model in terms of accuracy. This is evidenced by the lower MSE and RMSE values across all regions, particularly when data from different crash risk zones are aggregated. The ensemble approach effectively combines the strengths of multiple ConvLSTM models, improving prediction accuracy through a robust handling of spatial and temporal variations in the data.

The proposed model exhibits distinct performance variations across different crash risk zones. In areas of volatile high risk (Cluster 1), the model achieves the lowest MSE and RMSE, suggesting a strong capability to handle and accurately predict scenarios with high variability in crash risks. Conversely, in stable low-risk areas (Cluster 2), the model still improves upon simpler models but shows slightly higher errors than in high-risk areas, likely due to the challenges in capturing subtle variations in inherently low-risk environments.

The framework's ability to spatially align predictions with actual crash locations, especially noted in the superior performance in high-risk, volatile areas, indicates its logical validity. The integration of spatial data within the model allows it to effectively map and predict crash occurrences in relation to varying geographical and environmental factors, thereby confirming its utility and accuracy in practical applications.

6.2. Limitations

The contribution of this dissertation stems from the perspective of methodological advancement and practical application of existing methods. In terms of methodological advancement, the proposed technique to capture crash data generation pattern from a temporal standpoint using DTW-G* is novel and contribute to techniques that can be used to identify and map spatiotemporal crash risk. In terms of application, machine learning techniques were used to predict crash pattern and explain the models. One of the key limitations of the spatiotemporal risk mapping technique is its reliance on the quality and granularity of the available crash data. Data inconsistencies and missing information can affect the accuracy of the identified high-risk and low-risk zones. Furthermore, the clustering algorithm used may be sensitive to the initial parameters, which could lead to variability in the identified hotspots and coldspots. The current approach also does not fully account for dynamic changes in traffic patterns and weather conditions over time, which could further refine the spatiotemporal analysis.

In the analysis of crash data imbalance treatment, while SMOTE-N and ADASYN-N have proven effective in handling nominal predictors, their performance may vary across different datasets and machine learning techniques. The study primarily focuses on RF and XGBoost models, and the generalizability of these findings to other machine learning techniques remains to be explored. Additionally, the synthetic data generation methods, while addressing class imbalance, may introduce noise or fail to capture the true underlying data distribution, potentially affecting model performance.

The spatially ensembled ConvLSTM framework demonstrated improved predictive accuracy over conventional models. However, its complexity and computational requirements could limit its practical application, particularly for real-time prediction and large-scale datasets. The model's performance varies across different crash risk zones, suggesting a need for further optimization to handle low-risk areas more effectively. Moreover, the framework's reliance on historical data for training may not fully capture emerging trends and changes in weather patterns and traffic behaviors. Additionally, to maximize the practical benefits of the spatially ensembled ConvLSTM framework, integrating this model with real-time traffic and weather monitoring systems could provide dynamic, timely predictions that can be directly utilized for traffic management and crash prevention. Finally, development of decision-support tools that leverage the model’s outputs to provide actionable insights for urban planners and public safety officials could significantly enhance the impact of the predictive capabilities.

6.3. Future Research Opportunities

Enhancing data quality and integrating real-time data sources such as traffic flow and weather forecasts could improve the accuracy and timeliness of the spatiotemporal risk mapping. Exploring the application of more sophisticated clustering techniques and dynamic models that account for temporal variations in traffic and weather conditions would also be beneficial.

For crash data imbalance treatment, extending the evaluation to a wider range of machine learning techniques and datasets would provide a more comprehensive understanding of the effectiveness of SMOTE-N and ADASYN-N. Additionally, developing hybrid methods that combine synthetic data generation with other data augmentation techniques could further enhance model performance.

In terms of predictive modeling, simplifying the spatially ensembled ConvLSTM framework to reduce computational demands without sacrificing accuracy would be an important step. Investigating the integration of real-time data streams and adaptive learning techniques could make the model more responsive to changing conditions. Furthermore, expanding the scope of the analysis to include a wider variety of environmental and socio-economic factors would provide a more holistic view of weather-related crash risks.

REFERENCES

[1] Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., ... & Nahavandi, S. (2021). A review of uncertainty quantification in deep learning: Techniques, applications, and challenges. *Information Fusion*, 76, 243-297.

[2] Abdel-Aty, M., Ekram, A. A., Huang, H., & Choi, K. (2011). A study on crashes related to visibility obstruction due to fog and smoke. *Accident Analysis & Prevention*, *43*(5), 1730-1737.

[3] Abdel-Aty, M., Lee, J., Siddiqui, C., & Choi, K. (2013). Geographical unit-based analysis in the context of transportation safety planning. Transportation Research Part A: Policy and Practice, 49, 62-75.

[4] Aguero-Valverde, J., & Jovanis, P. P. (2006). Spatial analysis of fatal and injury crashes in Pennsylvania. *Accident Analysis & Prevention*, *38*(3), 618-625.

[5] Aguero-Valverde, J., & Jovanis, P. P. (2010). Spatial correlation in multilevel crash frequency models: Effects of different neighboring structures. *Transportation Research Record*, *2165*(1), 21-32.

[6] Ahmed, M. M., Khan, M. N., Das, A., & Dadvar, S. E. (2022). Global lessons learned from naturalistic driving studies to advance traffic safety and operation research: A systematic review. Accident Analysis & Prevention, 167, 106568.

[7] Al-Mistarehi, B. W., Alomari, A. H., Imam, R., & Mashaqba, M. (2022). Using machine learning models to forecast severity level of traffic crashes by R Studio and ArcGIS. *Frontiers in Built Environment, 8*, 860805.

[8] American Association of State Highway and Transportation Officials (AASHTO). (2010) Highway Safety Manual (HSM). Washington, DC.

[9] Amini, M., Bagheri, A., & Delen, D. (2022). Discovering injury severity risk factors in automobile crashes: A hybrid explainable AI framework for decision support. *Reliability Engineering & System Safety*, *226*, 108720.

[10] Anderson TK (2009) Kernel density estimation and K-means clustering to profile road accident hotspots. *Accident Analysis & Prevention, 41*(3):359–364

[11] Antoniadi, A. M., Du, Y., Guendouz, Y., Wei, L., Mazo, C., Becker, B. A., & Mooney, C. (2021). Current challenges and future opportunities for XAI in machine learning-based clinical decision support systems: a systematic review. *Applied Sciences*, *11*(11), 5088.

[12] Ashifur Rahman, M., Das, S., & Sun, X. (2022). Using cluster correspondence analysis to explore rainy weather crashes in Louisiana. *Transportation Research Record, 2676*(8), 159-173.

[13] Azimian, A., Pyrialakou, V. D., Lavrenz, S., & Wen, S. (2021). Exploring the effects of area-level factors on traffic crash frequency by severity using multivariate space-time models. *Analytic Methods in Accident Research*, *31*, 100163.

[14] Bao, J., Liu, P., Yu, H., & Xu, C. (2017). Incorporating twitter-based human activity information in spatial analysis of crashes in urban areas. *Accident analysis & prevention*, *106*, 358-369.

[15] Bao, J., Liu, P., & Ukkusuri, S. V. (2019). A spatiotemporal deep learning approach for citywide short-term crash risk prediction with multi-source data. *Accident Analysis & Prevention, 122*, 239-254.

[16] Bergel-Hayat, R., Debbarh, M., Antoniou, C., & Yannis, G. (2013). Explaining the road accident risk: Weather effects. *Accident Analysis & Prevention*, *60*, 456-465.

[17] Biau, G. (2012). Analysis of a random forests model. *The Journal of Machine Learning Research*, *13*(1), 1063-1095.

[18] Bibi, R., Saeed, Y., Zeb, A., Ghazal, T. M., Rahman, T., Said, R. A., ... & Khan, M. A. (2021). Edge AI-based automated detection and classification of road anomalies in VANET using deep learning. *Computational intelligence and neuroscience*, *2021*, 1-16.

[19] Böcker, L., Dijst, M., & Prillwitz, J. (2013). Impact of everyday weather on individual daily travel behaviours in perspective: a literature review. *Transport Reviews, 33*(1), 71-91.

[20] Brown, K. T. (2016). *A safety analysis of spatial phenomena about the residences of drivers involved in crashes* (Doctoral dissertation, Clemson University).

[21] Bullough, J. D., Donnell, E. T., & Rea, M. S. (2013). To illuminate or not to illuminate: Roadway lighting as it affects traffic safety at intersections. *Accident Analysis & Prevention*, *53*, 65-77.

[22] Cai, Q., Abdel-Aty, M., Sun, Y., Lee, J., & Yuan, J. (2019). Applying a deep learning approach for transportation safety planning by using high-resolution transportation and land use data. *Transportation Research Part A: Policy and Practice, 127*, 71-85.

[23] Cai, B., & Di, Q. (2023). Different forecasting model comparison for near future crash prediction. *Applied Sciences*, *13*(2), 759.

[24] Call, D. A., Medina, R. M., & Black, A. W. (2019). Causes of weather-related crashes in Salt Lake County, Utah. The Professional Geographer, 71(2), 253-264.

[25] Carter, J. V., Pan, J., Rai, S. N., & Galandiuk, S. (2016). ROC-ing along: Evaluation and interpretation of receiver operating characteristic curves. *Surgery*, *159*(6), 1638-1645.

[26] Chainey, S. P. (2013). Examining the influence of cell size and bandwidth size on kernel density estimation crime hotspot maps for predicting spatial patterns of crime. *Bulletin of the Geographical Society of Liege*, *60*, 7-19.

[27] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research, 16,* 321-357.

[28] Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785-794).

[29] Chen, T., Shi, X., Wong, Y. D., & Yu, X. (2020). Predicting lane-changing risk level based on vehicles' space-series features: A pre-emptive learning approach. *Transportation Research Part C: Emerging Technologies, 116*, 102646.

[30] Cheng, W., & Washington, S. (2008). New criteria for evaluating methods of identifying hot spots. *Transportation Research Record, 2083*(1), 76-85.

[31] Chung, Y., Haas, P. J., Upfal, E., & Kraska, T. (2018). Unknown examples & machine learning model generalization. *arXiv preprint arXiv:1808.08294*.

[32] Cohn, E. G., Kakar, S., Perkins, C., Steinbach, R., & Edwards, P. (2020). Red light camera interventions for reducing traffic violations and traffic crashes: A systematic review. *Campbell systematic reviews*, *16*(2), e1091.

[33] Daniels, N., del Pilar Guzmán Urrea, M., Rentmeester, C. A., Kotchian, S. A., Fontaine, S., Hernández-Aguado, I., ... & Viens, A. M. (2016). Resource allocation and priority setting. *Public Health Ethics: Cases Spanning the Globe*, 61-94.

[34] Das, S., Dutta, A., Jalayer, M., Bibeka, A., & Wu, L. (2018). Factors influencing the patterns of wrong-way driving crashes on freeway exit ramps and median crossovers: Exploration using 'Eclat'association rules to promote safety. *International Journal of Transportation Science and Technology*, *7*(2), 114-123.

[35] Das, S., Dutta, A., & Sun, X. (2020). Patterns of rainy weather crashes: Applying rules mining. *Journal of Transportation Safety & Security, 12*(9), 1083-1105.

[36] Das, S., Hossain, A., Rahman, M. A., Sheykhfard, A., & Kutela, B. (2023a). Case Study on the Traffic Collision Patterns of E-Scooter Riders. *Transportation Research Record*, 03611981231185770.

[37] Das, S., Vierkant, V., Gonzalez, J. C., Kutela, B., & Sheykhfard, A. (2023b). Bayesian Network for Motorcycle Crash Severity *Analysis. Transportation Research Record*, 03611981231164386.

[38] Dey, K. C., Mishra, A., & Chowdhury, M. (2014). Potential of intelligent transportation systems in mitigating adverse weather impacts on road mobility: A review. *IEEE Transactions on Intelligent Transportation Systems, 16*(3), 1107-1119.

[39] Dingus, T. A., Guo, F., Lee, S., Antin, J. F., Perez, M., Buchanan-King, M., & Hankey, J. (2016). Driver crash risk factors and prevalence evaluation using

naturalistic driving data. *Proceedings of the National Academy of Sciences, 113*(10), 2636-2641.

[40] Downs, A. (2000). *Stuck in traffic: Coping with peak-hour traffic congestion*. Brookings Institution Press.

[41] Drawve, G., Grubb, J., Steinman, H., & Belongie, M. (2019). Enhancing data-driven law enforcement efforts: exploring how risk terrain modeling and conjunctive analysis fit in a crime and traffic safety framework. *American journal of criminal justice*, *44*, 106-124.

[42] Duddu, V. R., & Pulugurtha, S. S. (2017). Modeling link-level crash frequency using integrated geospatial land use data and on-network characteristics. *Journal of Transportation Engineering, Part A: Systems*, *143*(8), 04017030.

[43] Duddu, V. R., Kukkapalli, V. M., & Pulugurtha, S. S. (2019). Crash risk factors associated with injury severity of teen drivers. *IATSS Research, 43*(1), 37-43.

[44] Elmitiny, N., Yan, X., Radwan, E., Russo, C., & Nashar, D. (2010). Classification analysis of driver's stop/go decision and red-light running violation. *Accident Analysis & Prevention*, *42*(1), 101-111.

[45] Engstrom, D. (2012). Proven Safety Countermeasures. https://highways.dot.gov/safety/proven-safety-countermeasures.

[46] Federal Highway Administration FHWA (2023). 21st century operations using 21st century technologies. How Do Weather Events Impact Roads? https://ops.fhwa.dot.gov/weather/q1_roadimpact.htm. Accessed: 28th April 2023

[47] Feng, M., Wang, X., & Quddus, M. (2020). Developing multivariate time series models to examine the interrelations between police enforcement, traffic violations, and traffic crashes. *Analytic methods in accident research*, *28*, 100139.

[48] Fiorentini, N., & Losa, M. (2020). Handling imbalanced data in road crash severity prediction by machine learning algorithms. *Infrastructures, 5*(7), 61.

[49] Formosa, N., Quddus, M., Ison, S., Abdel-Aty, M., & Yuan, J. (2020). Predicting real-time traffic conflicts using deep learning. *Accident Analysis & Prevention*, *136*, 105429.

[50] Foster, M., Brice, J. H., Shofer, F., Principe, S., DeWalt, D., Falk, R., & Ferris, M. (2011). Personal disaster preparedness of dialysis patients in North Carolina. *Clinical Journal of the American Society of Nephrology*, *6*(10), 2478-2484.

[51] Gaikwad, N., & Markande, S. D. (2016, September). Intelligent safety control for automotive systems. In IEEE 2016 International Conference on Automatic control and Dynamic Optimization Techniques (ICACDOT) (pp. 653-656).

[52] Gajera, H., Pulugurtha, S. S., Mathew, S., & Bhure, C. M. (2023). Synthesizing fatal crashes involving partially automated vehicles and comparing with fatal crashes involving non-automated vehicles. *Transportation Engineering*, *12*, 100178.

[53] Gao, L., Lu, P., & Ren, Y. (2021). A deep learning approach for imbalanced crash data in predicting highway-rail grade crossings accidents. *Reliability Engineering & System Safety*, 216, 108019.

[54] Gao, X., Jiang, X., Zhuang, D., Chen, H., Wang, S., & Haworth, J. (2023). Spatiotemporal graph neural networks with uncertainty quantification for traffic incident risk prediction. *arXiv preprint arXiv:2309.05072*.

[55] Ghahramani, Z. (2015). Probabilistic machine learning and artificial intelligence. *Nature, 521*(7553), 452-459.

[56] Ghasemzadeh, A., & Ahmed, M. M. (2019). Complementary parametric probit regression and nonparametric classification tree modeling approaches to analyze factors affecting severity of work zone weather-related crashes. *Journal of Modern Transportation, 27*, 129-140.

[57] Gonzalez, M. C., Hidalgo, C. A., & Barabasi, A. L. (2008). Understanding individual human mobility patterns. *nature*, *453*(7196), 779-782.

[58] Greibe, P. (2003). Accident prediction models for urban roads. *Accident Analysis & Prevention, 35*(2), 273-285.

[59] Grounds, M. A., & Joslyn, S. L. (2018). Communicating weather forecast uncertainty: Do individual differences matter?. *Journal of experimental psychology: applied*, *24*(1), 18.

[60] Guo, X., Wu, L., Zou, Y., & Fawcett, L. (2019). Comparative analysis of empirical bayes and bayesian hierarchical models in hotspot identification. *Transportation research record*, *2673*(7), 111-121.

[61] Hambly, D., Andrey, J., Mills, B., & Fletcher, C. (2013). Projected implications of climate change for road safety in Greater Vancouver, Canada. *Climatic Change*, 116, 613-629

[62] Hamdar, S. H., Qin, L., & Talebpour, A. (2016). Weather and road geometry impact on longitudinal driving behavior: Exploratory analysis using an empirically supported acceleration modeling framework. *Transportation Research Part C: Emerging Technologies*, 67, 193-213.

[63] Hansson, S. O. (2022). Zero visions and other safety principles. In *The Vision Zero Handbook: Theory, Technology and Management for a Zero Casualty Policy* (pp. 1-75). Cham: Springer International Publishing.

[64] Harmon, T., Bahar, G. B., & Gross, F. B. (2018). *Crash costs for highway safety analysis* (No. FHWA-SA-17-071). United States. Federal Highway Administration. Office of Safety.

[65] Hasan, S., Schneider, C. M., Ukkusuri, S. V., & González, M. C. (2013). Spatiotemporal patterns of urban human mobility. *Journal of Statistical Physics*, *151*, 304-318.

[66] Hassouna, F. M., & Al-Sahili, K. (2020). Practical minimum sample size for road crash time-series prediction models. *Advances in civil engineering*, *2020*, 1-12.

[67] Hauer, E. (1992). Empirical Bayes approach to the estimation of "unsafety": the multivariate regression method. *Accident Analysis & Prevention, 24*(5), 457-477.

[68] He, H., Bai, Y., Garcia, E. A., & Li, S. (2008, June). ADASYN: Adaptive synthetic sampling approach for imbalanced learning. *In 2008 IEEE International Joint*

Conference on Neural Networks & IEEE World Congress on Computational Intelligence) (pp. 1322-1328).

[69] Herbel, S., Laing, L., & McGovern, C. (2010). Highway safety improvement program manual: The focus is results (No. FHWA-SA-09-029). United States. Federal Highway Administration. Office of Safety.

[70] Hou, Q., Huo, X., Leng, J., & Mannering, F. (2022). A note on out-of-sample prediction, marginal effects computations, and temporal testing with random parameters crash-injury severity models. *Analytic Methods in Accident Research*, 33, 100191.

[71] HSIP, FHWA, U.S. Department of Transportation. Highway Safety Improvement Program (HSIP), 2010. http://safety.fhwa.dot.gov/hsip/

[72] Huang, H., Xu, P., & Abdel-Aty, M. (2013). Transportation Safety Planning: A Spatial Analysis Approach 4. *Transportation*, *2*(3).

[73] Huang, C., Zhang, C., Dai, P., & Bo, L. (2019). Deep dynamic fusion network for traffic accident forecasting. In *Proceedings of the 28th ACM international conference on information and knowledge management* (pp. 2673-2681).

[74] Huang, X., He, P., Rangarajan, A., & Ranka, S. (2020). Intelligent intersection: Two-stream convolutional networks for real-time near-accident detection in traffic video. *ACM Transactions on Spatial Algorithms and Systems (TSAS)*, *6*(2), 1-28.

[75] Iranitalab, A., & Khattak, A. (2017). Comparison of four statistical and machine learning methods for crash severity prediction. *Accident Analysis & Prevention*, 108, 27-36

[76] Islam, M., & Mannering, F. (2020). A temporal analysis of driver-injury severities in crashes involving aggressive and non-aggressive driving. *Analytic Methods in Accident Research*, 27, 100128.

[77] Ivan, J. N. (2004). New approach for including traffic volumes in crash rate analysis and forecasting. *Transportation Research Record*, *1897*(1), 134-141.

[78] Jackson, T. L., & Sharif, H. O. (2016). Rainfall impacts on traffic safety: Rain-related fatal crashes in Texas. *Geomatics, Natural Hazards, and Risk*, *7*(2), 843-860.

[79] Jamal, A., Zahid, M., Tauhidur Rahman, M., Al-Ahmadi, H. M., Almoshaogeh, M., Farooq, D., & Ahmad, M. (2021). Injury severity prediction of traffic crashes with ensemble machine learning techniques: A comparative study. *International journal of injury control and safety promotion*, *28*(4), 408-427.

[80] Kalambay, P., & Pulugurtha, S. S. (2022). City-oriented and inclusive bicycle-vehicle crash frequency modeling through the integration of bicycle-sharing system and other surrogates. *Transportation Research Interdisciplinary Perspectives*, *16*, 100714.

[81] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... & Liu, T. Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*, 30.

[82] Keskin, F., Yenilmez, F., Çolak, M., Yavuzer, I., & Düzgün, H. S. (2011). Analysis of traffic incidents in METU campus. *Procedia-Social and Behavioral Sciences*, 19, 61-70.

[83] Khan, G., Qin, X., & Noyce, D. A. (2008). Spatial analysis of weather crash patterns. *Journal of Transportation Engineering*, 134(5), 191-202.

[84] Khan, M. A., Etminani-Ghasrodashti, R., Kermanshachi, S., Rosenberger, J. M., Pan, Q., & Foss, A. (2022). Do ridesharing transportation services alleviate traffic crashes? A time series analysis. *Traffic injury prevention*, *23*(6), 333-338.

[85] Khanpour, A., King, M., Sheykhfard, A., & Haghighi, F. (2023). Drivers' reported crash history, sensitivity to reward and punishment, personality, and demographics: a case study in Iran. *Transportation research interdisciplinary perspectives*, *21*, 100902.

[86] Khuat, T. T., & Le, M. H. (2020). Evaluation of sampling-based ensembles of classifiers on imbalanced data for software defect prediction problems. *SN Computer Science*, *1*(2), 108.

[87] Kilpeläinen, M., & Summala, H. (2007). Effects of weather and weather forecasts on driver behaviour. *Transportation research part F: traffic psychology and behaviour*, *10*(4), 288-299.

[88] Kim, K., & Yamashita, E. Y. (2007). Using ak-means clustering algorithm to examine patterns of pedestrian involved crashes in Honolulu, Hawaii. *Journal of Advanced Transportation, 41*(1), 69-89.

[89] Kim, S., Lym, Y., & Kim, K. J. (2021). Developing crash severity model handling class imbalance and implementing ordered nature: Focusing on elderly drivers. International journal of Environmental Research and Public Health, 18(4), 1966.

[90] Kuo, P. F., Lord, D., & Walden, T. D. (2011, September). Using geographical information systems to effectively organize police patrol routes by grouping hot spots of crash and crime data. In Third International Conference on Road Safety and Simulation.

[91] Lakshmi, S., Srikanth, I., & Arockiasamy, M. (2019). Identification of traffic accident hotspots using geographical information system (GIS). *Int. J. Eng. Adv. Technol. IJEAT*, 9, 2249-8958.

[92] Ladron de Guevara, F., Washington, S. P., & Oh, J. (2004). Forecasting crashes at the planning level: simultaneous negative binomial crash model applied in Tucson, Arizona. *Transportation Research Record*, *1897*(1), 191-199.

[93] Lee, M., & Khattak, A. J. (2019). Case study of crash severity spatial pattern identification in hot spot analysis. *Transportation research record*, *2673*(9), 684-695.

[94] Levine, N. (2009). A motor vehicle safety planning support system: the Houston experience. In Planning Support Systems Best Practice and New Methods (pp. 93-111). Dordrecht: Springer Netherlands.

[95] Loo, B. P., Fan, Z., Lian, T., & Zhang, F. (2023). Using computer vision and machine learning to identify bus safety risk factors. *Accident Analysis & Prevention*, *185*, 107017.

[96] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information processing systems*, *30*.

[97] Ma, L., Yan, X., Wei, C., & Wang, J. (2016). Modeling the equivalent property damage only crash rate for road segments using the hurdle regression framework. *Analytic methods in accident research*, *11*, 48-61.

[98] Ma, X., Lu, J., Liu, X., & Qu, W. (2023). A genetic programming approach for real-time crash prediction to solve trade-off between interpretability and accuracy. *Journal of Transportation Safety & Security*, *15*(4), 421-443.

[99] Maldonado, S., & López, J. (2018). Dealing with high-dimensional class-imbalanced datasets: Embedded feature selection for SVM classification. Applied Soft Computing, 67, 94-105.

[100] Mannering, F. L., & Bhat, C. R. (2014). Analytic methods in accident research: Methodological frontier and future directions. *Analytic methods in accident research*, *1*, 1-22.

[101] Martin, T. L., Solbeck, P. A., Mayers, D. J., Langille, R. M., Buczek, Y., & Pelletier, M. R. (2013). A review of alcohol-impaired driving: The role of blood alcohol concentration and complexity of the driving task. *Journal of forensic sciences*, *58*(5), 1238-1250.

[102] Martins, R. S., Saqib, S. U., Raja, M. H. R., Gillani, M., & Zafar, H. (2022). Collision versus loss-of-control motorcycle accidents: Comparing injuries and outcomes. *Traffic injury prevention*, *23*(5), 255-259.

[103] Mathew, S., & Pulugurtha, S. S. (2022). Effect of Weather Events on Travel Time Reliability and Crash Occurrence.

[104] Mohammed, A. S., Amamou, A., Ayevide, F. K., Kelouwani, S., Agbossou, K., & Zioui, N. (2020). The perception system of intelligent ground vehicles in all weather conditions: A systematic literature review. Sensors, 20(22), 6532.

[105] Mondal, A. R., Bhuiyan, M. A. E., & Yang, F. (2020). Advancement of weather-related crash prediction model using nonparametric machine learning algorithms. SN Applied Sciences, 2, 1-11.

[106] Morris, C., & Yang, J. J. (2021). Effectiveness of resampling methods in coping with imbalanced crash data: Crash type analysis and predictive modeling. *Accident Analysis & Prevention*, *159*, 106240.

[107] Musselwhite, C., Avineri, E., & Susilo, Y. (2021). Restrictions on mobility due to the coronavirus Covid19: Threats and opportunities for transport and health. *Journal of Transport & Health*, 20, 101042.

[108] Müller, M. (2007). Dynamic time warping. Information Retrieval for Music and Motion, 69-84.

[109] Nahed, R., Nassar, E., Khoury, J., & Arnaout, J. P. (2023). Assessing the effects of geometric layout and signing on drivers' behavior through work zones. *Transportation Research Interdisciplinary Perspectives*, *21*, 100901.

[110] Najaf, P., Duddu, V. R., & Pulugurtha, S. S. (2018). Predictability and interpretability of hybrid link-level crash frequency models for urban arterials compared to cluster-based and general negative binomial regression models. *International Journal of Injury Control and Safety Promotion*, *25*(1), 3-13.

[111] Nakaya, T. (2013). Analytical data transformations in space–time region: Three stories of space–time cube: Space–time integration in geography and giscience. *Annals of the Association of American Geographers, 103*(5), 1100-1106.

[112] National Research Council NRC. (2004). Where the weather meets the road: A research agenda for improving road weather services. National Academic Press, Washington, DC.

[113] Ogungbire, A., Kalambay, P., Gajera, H., & Pulugurtha, S. S. (2023). Deep Learning, Machine Learning, or Statistical Models for Weather-related Crash Severity Prediction.

[114] Ogungbire, A., Kalambay, P., & Pulugurtha, S. S. (2024). Exploring the effect of mountainous terrain on weather-related crashes. *IATSS Research*, *48*(2), 136-146.

[115] Ogungbire, A. & Pulugurtha, S.S. (2024). A Spatiotemporal Risk Mapping of Statewide Weather-related Traffic Crashes: A Machine Learning Approach. *Arxiv*

[116] Pahl-Wostl, C. (2007). The implications of complexity for integrated resources management. *Environmental Modelling & Software, 22*(5), 561-569.

[117] Pathivada, B. K., Banerjee, A., & Haleem, K. (2024). Impact of real-time weather conditions on crash injury severity in Kentucky using the correlated random parameters logit model with heterogeneity in means. *Accident Analysis & Prevention*, *196*, 107453.

[118] PCS, FHWA U.S. Department of Transportation. Proven Safety Countermeasures (PSC), 2010. http://safety.fhwa.dot.gov/proven-safety-countermeasures/

[119] Pei, X., Wong, S. C., & Sze, N. N. (2011). A joint-probability approach to crash prediction models. *Accident Analysis & Prevention*, *43*(3), 1160-1166

[120] Perrels, A., Votsis, A., Nurmi, V., & Pilli-Sihvola, K. (2015). Weather conditions, weather information and car crashes. *ISPRS International Journal of Geo-Information, 4*(4), 2681-2703.

[121] Plug, C., Xia, J. C., & Caulfield, C. (2011). Spatial and temporal visualisation techniques for crash analysis. *Accident Analysis & Prevention*, *43*(6), 1937-1946.

[122] Prasannakumar, V., Vijith, H., Charutha, R., & Geetha, N. (2011). Spatio-temporal clustering of road accidents: GIS based analysis and assessment. *Procedia-Social and Behavioral Sciences*, 21, 317-325.

[123] Pulugurtha, S. S., Krishnakumar, V. K., & Nambisan, S. S. (2007). New methods to identify and rank high pedestrian crash zones: An illustration. *Accident Analysis & Prevention, 39*(4), 800-811.

[124] Pulugurtha, S. S., Duddu, V. R., & Kotagiri, Y. (2013). Traffic analysis zone level crash estimation models based on land use characteristics. *Accident Analysis & Prevention*, *50*, 678-687.

[125] Pulugurtha, S. S., & Mahanthi, S. S. B. (2016). Assessing spatial and temporal effects due to a crash on a freeway through traffic simulation. *Case Studies on Transport Policy*, *4*(2), 122-132.

[126] Quddus, M. A. (2008). Modelling area-wide count outcomes with spatial correlation and heterogeneity: An analysis of London crash data. *Accident Analysis & Prevention*, *40*(4), 1486-1497.

[127] Rabbani, M. B. A., Musarat, M. A., Alaloul, W. S., Rabbani, M. S., Maqsoom, A., Ayub, S., ... & Altaf, M. (2021). a comparison between seasonal autoregressive integrated moving average (SARIMA) and exponential smoothing (ES) based on time series model for forecasting road accidents. *Arabian Journal for Science and Engineering*, *46*(11), 11113-11138.

[128] Rachakonda, Y., & Pawar, D. S. (2023). Evaluation of intersection conflict warning system at unsignalized intersections: A review. *Journal of Traffic and Transportation Engineering (English edition)*.

[129] Ran, B., Jin, P. J., Boyce, D., Qiu, T. Z., & Cheng, Y. (2012). Perspectives on future transportation research: Impact of intelligent transportation system technologies on next-generation transportation modeling. *Journal of Intelligent Transportation Systems*, *16*(4), 226-242.

[130] Robbins, C. J., & Fotios, S. (2020). Motorcycle safety after-dark: the factors associated with greater risk of road-traffic collisions. *Accident Analysis & Prevention*, 146, 105731

[131] Robin, T. A., Khan, M. A., Kabir, N., Rahaman, S. T., Karim, A., Mannan, I. I., George, J. & Rashid, I. (2019). Using spatial analysis and GIS to improve planning and resource allocation in a rural district of Bangladesh. *BMJ Global Health, 4*(Suppl 5), e000832.

[132] Roland, J., Way, P. D., Firat, C., Doan, T. N., & Sartipi, M. (2021). Modeling and predicting vehicle accident occurrence in Chattanooga, Tennessee. *Accident Analysis & Prevention*, *149*, 105860.

[133] Roy, A., Cruz, R. M., Sabourin, R., & Cavalcanti, G. D. (2018). A study on combining dynamic selection and data preprocessing for imbalance learning. *Neurocomputing*, 286, 179-192.

[134] Saarikko, T., Nuldén, U., Meiling, P., & Pessi, K. (2020). Framing crisis information systems: the case of WIS. In Proceedings of the 53rd Hawaii International Conference on System Sciences.

[135] Saccomanno, F. F., Grossi, R., Greco, D., & Mehmood, A. (2001). Identifying black spots along highway SS107 in Southern Italy using two models. *Journal of Transportation Engineering, 127*(6), 515-522.

[136] Saha, S., Schramm, P., Nolan, A., & Hess, J. (2016). Adverse weather conditions and fatal motor vehicle crashes in the United States, 1994-2012. *Environmental health*, *15*, 1-9.

[137] Salman, S., & Liu, X. (2019). Overfitting mechanism and avoidance in deep neural networks. *arXiv preprint arXiv:1901.06566*.

[138] Santosh, K. C., & Gaur, L. (2022). *Artificial intelligence and machine learning in public healthcare: Opportunities and societal impact*. Springer Nature.

[139] Sattar, K., Chikh Oughali, F., Assi, K., Ratrout, N., Jamal, A., & Masiur Rahman, S. (2023). Transparent deep machine learning framework for predicting traffic crash severity. *Neural Computing and Applications, 35*(2), 1535-1547.

[140] Sawtelle, A. A. (2022). Statistical Analysis of Frequency and Severity of Lane Departure Crashes in Maine. The University of Maine.

[141] Sawtelle, A., Shirazi, M., Garder, P. E., & Rubin, J. (2023). Driver, roadway, and weather factors on severity of lane departure crashes in Maine. *Journal of Safety Research*, *84*, 306-315.

[142] Sha, P., Chen, T., Wong, Y. D., Meng, X., Wang, X., & Liu, W. (2023). Exploring key spatio-temporal features of crash risk hot spots on urban road network: A machine learning approach. Transportation Research Part A: Policy and Practice, 173, 103717.

[143] Shapley, L. S. (1953). A value for n-person games.

[144] Sheykhfard, A., Haghighi, F., Das, S., & Fountas, G. (2023). Evasive actions to prevent pedestrian collisions in varying space/time contexts in diverse urban and non-urban areas. *Accident Analysis & Prevention*, *192*, 107270.

[145] Shi, X., Chen, Z., Wang, H., Yeung, D. Y., Wong, W. K., & Woo, W. C. (2015). Convolutional LSTM network: A machine learning approach for precipitation nowcasting. *Advances in neural information processing systems*, *28*

[146] Shi, X., Wong, Y. D., Li, M. Z. F., Palanisamy, C., & Chai, C. (2019). A feature learning approach based on XGBoost for driving assessment and risk prediction. *Accident Analysis & Prevention*, 129, 170-179.

[147] Smiley, A., & Rudin-Brown, C. (2020). Drivers adapt–Be prepared for It!. *Accident Analysis & Prevention*, *135*, 105370.

[148] Soltani, A., & Askari, S. (2017). Exploring spatial autocorrelation of traffic crashes based on severity. *Injury, 48*(3), 637-647.

[149] Songchitruksa, P., & Zeng, X. (2010). Getis–Ord spatial statistics to identify hot spots by using incident management data. *Transportation Research Record, 2165*(1), 42-51.

[150] Strong, C. K., Ye, Z., & Shi, X. (2010). Safety effects of winter weather: the state of knowledge and remaining challenges. *Transport Reviews, 30*(6), 677-699.

[151] Suh, J., & Yeo, H. (2016). An empirical study on the traffic state evolution and stop-and-go traffic development on freeways. *Transportmetrica A: Transport Science*, *12*(1), 80-97.

[152] Tao, X., Li, Q., Guo, W., Ren, C., Li, C., Liu, R., & Zou, J. (2019). Self-adaptive cost weights-based support vector machine cost-sensitive ensemble for imbalanced data classification. *Information Sciences*, 487, 31-56

[153] Tarko, A. P., & Kanodia, M. (2004). Hazard Elimination Program-Manual on Improving Safety of Indiana Road Intersections and Sections; Volume 1: Research Report and Volume 2: Guidelines for Highway Safety Improvements in Indiana. https://docs.lib.purdue.edu/cgi/viewcontent.cgi?article=1583&context=jtrp.

[154] Tavenard, R., Faouzi, J., Vandewiele, G., Divo, F., Androz, G., Holtz, C., ... & Woods, E. (2020). Tslearn, a machine learning toolkit for time series data. *The Journal of Machine Learning Research, 21*(1), 4686-4691.

[155] Theofilatos, A., & Yannis, G. (2014). A review of the effect of traffic and weather characteristics on road safety. *Accident Analysis & Prevention*, 72, 244-256.

[156] Turner, J. D., & Austin, L. (2000). A review of current sensor technologies and applications within automotive and traffic control systems. Proceedings of the Institution of Mechanical Engineers, Part D: *Journal of Automobile Engineering, 214*(6), 589-614.

[157] Ungar, S. (1999). Is strange weather in the air? A study of US national network news coverage of extreme weather events. *Climatic Change, 41*(2), 133-150.

[158] Valcamonico, D., Baraldi, P., Amigoni, F., & Zio, E. (2022). A framework based on Natural Language Processing and Machine Learning for the classification of the severity of road accidents from reports. *Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability*, 1748006X221140196.

[159] Veran, T., Portier, P. E., & Fouquet, F. (2023). Interpretable hierarchical symbolic regression for safety-critical systems with an application to highway crash prediction. *Engineering Applications of Artificial Intelligence*, *117*, 105534.

[160] Vickrey, W. S. (1969). Congestion theory and transport investment. *The American economic review*, *59*(2), 251-260.

[161] Washington, S., & Haque, M. D. (2013). On the commonly accepted assumptions regarding observed motor vehicle crash counts at transport system locations. In *Transportation Research Board (TRB) 92nd Annual Meeting Compendium of Papers* (pp. 1-19). Transportation Research Board (TRB), National Academy of Sciences.

[162] Wei, Z., Zhang, Y., & Das, S. (2023). Applying explainable machine learning techniques in daily crash occurrence and severity modeling for rural interstates. *Transportation Research Record*, 03611981221134629.

[163] Wen, X., Xie, Y., Jiang, L., Pu, Z., & Ge, T. (2021). Applications of machine learning methods in traffic crash severity modelling: current status and future directions. *Transport Reviews, 41*(6), 855-

[164] World Health Organization (WHO). (2015). Global status report on road safety 2015. https://www.afro.who.int/publications/global-status-report-road-safety-2015.

[165] Wu, P., Chen, T., Wong, Y. D., Meng, X., Wang, X., & Liu, W. (2023). Exploring key spatio-temporal features of crash risk hot spots on urban road network: A machine learning approach. *Transportation research part A: policy and practice*, *173*, 103717.drvres

[166] Xu, P., & Huang, H. (2015). Modeling crash spatial heterogeneity: Random parameter versus geographically weighting. *Accident Analysis & Prevention*, *75*, 16-25.

[167] Yang, Y., Wang, K., Yuan, Z., & Liu, D. (2022). Predicting freeway traffic crash severity using XGBoost-Bayesian network model with consideration of features interaction. *Journal of Advanced Transportation*, 2022, 4257865. https://doi.org/10.1155/2022/4257865.

[168] Yijing, L., Haixiang, G., Xiao, L., Yanan, L., & Jinling, L. (2016). Adapted ensemble classification algorithm based on multiple classifier system and feature selection for classifying multi-class imbalanced data. *Knowledge-Based Systems*, 94, 88-104.

[169] Yuan, J., Abdel-Aty, M., Gong, Y., & Cai, Q. (2019). Real-time crash risk prediction using long short-term memory recurrent neural network. *Transportation Research Record, 2673*(4), 314-326.

[170] Yuan, T., da Rocha Neto, W., Rothenberg, C. E., Obraczka, K., Barakat, C., & Turletti, T. (2022). Machine learning for next-generation intelligent transportation systems: A survey. *Transactions on emerging telecommunications technologies*, *33*(4), e4427.

[171] Yoon, J., & Lee, S. (2021). Spatio-temporal patterns in pedestrian crashes and their determining factors: Application of a space-time cube analysis model. *Accident Analysis & Prevention*, *161*, 106291.

[172] Zeng, Q., Hao, W., Lee, J., & Chen, F. (2020). Investigating the impacts of real-time weather conditions on freeway crash severity: a Bayesian spatial analysis. *International Journal of Environmental Research and Public Health, 17*(8), 2768.

[173] Zhang, S. (2020). Prediction of Pedestrians' Red-Light Violations Using Deep Learning.

[174] Zhang, X., Wen, H., Yamamoto, T., & Zeng, Q. (2021). Investigating hazardous factors affecting freeway crash injury severity incorporating real-time weather data: Using a Bayesian multinomial logit model with conditional autoregressive priors. *Journal of Safety Research*, 76, 248-255.

[175] Zhang, Z., Akinci, B., & Qian, S. (2022). Inferring the causal effect of work zones on crashes: Methodology and a case study. *Analytic methods in accident research*, *33*, 100203.

[176] Zhao, S., Wang, K., Liu, C., & Jackson, E. (2019). Investigating the effects of monthly weather variations on Connecticut freeway crashes from 2011 to 2015. *Journal of Safety Research,* 71, 153-162.

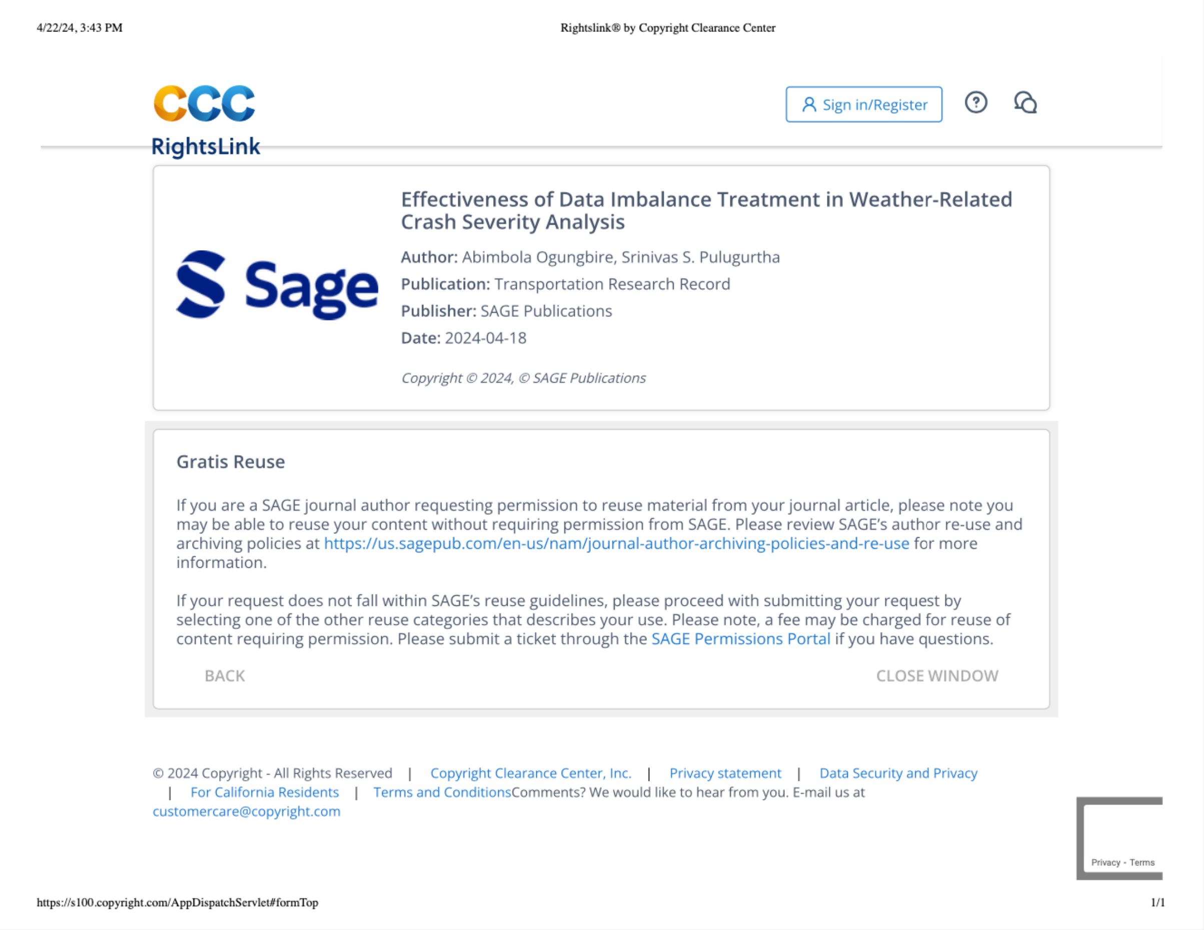
[177] Zhao, J., Liu, P., & Li, Z. (2024). Exploring the impact of trip patterns on spatially aggregated crashes using floating vehicle trajectory data and graph Convolutional Networks. *Accident Analysis & Prevention*, *194*, 107340.

APPENDIX A: BIOGRAPHY

Abimbola is a Ph.D. student of Infrastructure and Environmental Systems with domain expertise in Traffic Safety and Transportation Systems Management & Operations (TSM&O). He uses data-centric approaches to solve practical transportation issues. He has published several papers and presented his work at numerous conferences, including those organized by the Transportation Research Board (TRB) of the National Academies of Sciences, Engineering & Medicine.

His experience traverses' diverse fields with a broader engineering and data analytics background. He obtained a bachelor's degree in civil engineering, exploring potentials for improving rigid pavements using scrap steel. In his master's degree, he studied geospatial analytics, where he harnessed spatial data for transportation planning and traffic operations management. In his Ph.D., he worked with an interdisciplinary team of Data and Transportation Engineers where he applied advanced technologies and tools to Transportation, Energy, Climate, and Economic Systems.

APPENDIX B: COPYRIGHT STATEMENT



4/22/24, 3:43 PM Rightslink® by Copyright Clearance Center

Effectiveness of Data Imbalance Treatment in Weather-Related Crash Severity Analysis

Author: Abimbola Ogungbire, Srinivas S. Pulugurtha

Publication: Transportation Research Record

Publisher: SAGE Publications

Date: 2024-04-18

Copyright © 2024, © SAGE Publications

Gratis Reuse

If you are a SAGE journal author requesting permission to reuse material from your journal article, please note you may be able to reuse your content without requiring permission from SAGE. Please review SAGE's author re-use and archiving policies at https://us.sagepub.com/en-us/nam/journal-author-archiving-policies-and-re-use for more information.

If your request does not fall within SAGE's reuse guidelines, please proceed with submitting your request by selecting one of the other reuse categories that describes your use. Please note, a fee may be charged for reuse of content requiring permission. Please submit a ticket through the SAGE Permissions Portal if you have questions.

© 2024 Copyright - All Rights Reserved | Copyright Clearance Center, Inc. | Privacy statement | Data Security and Privacy | For California Residents | Terms and ConditionsComments? We would like to hear from you. E-mail us at customercare@copyright.com

https://s100.copyright.com/AppDispatchServlet#formTop 1/1