Tran-SET

Transportation Consortium of South-Central States

Solving Emerging Transportation Resiliency, Sustainability, and Economic Challenges through the Use of Innovative Materials and Construction Methods: From Research to Implementation

Modeling Crash Severity and Collision Types Using Machine Learning

Project No. 20SAUTSA36

Lead University: University of Texas at San Antonio

Final Report

January 2022

Disclaimer

The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the information presented herein. This document is disseminated in the interest of information exchange. The report is funded, partially or entirely, by a grant from the U.S. Department of Transportation's University Transportation Centers Program. However, the U.S. Government assumes no liability for the contents or use thereof.

Acknowledgements

The authors would like to thank Transportation Consortium of South-Central States (Tran-SET) for providing the essential platform and financial support to make this research possible. Also, the authors would like to thank all the Project Monitoring Committee (PMC) members for their time and constructive comments and recommendations for the progress of the project.

TECHNICAL DOCUMENTATION PAGE

1. Project No. 20SAUTSA36	2. Government Accession No.	3. Recipient's Catalog No.

4. Title and Subtitle	5. Report Date
Modeling Crash Severity and Collision Types Using Machine Learning	January 2022
	6. Performing Organization Code
7. Author(s) Dr. Amit Kumar, University of Texas at San Antonio (UTSA) Hari Krishnan Melempat Kalapurayil, UTSA	**8. Performing Organization Report No.**
9. Performing Organization Name and Address Transportation Consortium of South-Central States (Tran-SET) University Transportation Center for Region 6 3319 Patrick F. Taylor Hall, Louisiana State University, Baton Rouge, LA 70803	**10. Work Unit No. (TRAIS)**
	11. Contract or Grant No. 69A3551747106
12. Sponsoring Agency Name and Address United States of America Department of Transportation Research and Innovative Technology Administration	**13. Type of Report and Period Covered** Final Research Report Aug. 2020 – Jan. 2022
	14. Sponsoring Agency Code

15. Supplementary Notes

Report uploaded and accessible at Tran-SET's website (http://transet.lsu.edu/).

16. Abstract

Traffic safety analysis is the fundamental step for reducing economic, social, and environmental cost incurred due to traffic accidents. The essence of traffic safety is understanding the factors affecting crash occurrence, injury severity and collision type and their underlying relationships and predict-prevent future crash instances. Crash injury severity studies in past have utilized numerous statistical, econometric and Machine Learning (ML) and Artificial Intelligence (AI) tools to extract the underlying relationship between the crash causal factors and the consequent severity or collision type. The study aims to explore the Multi-Label Classification (MLC) tool from the domain of Artificial Intelligence (AI) for classification problems in the setting of traffic safety. MLC finds its application primarily in protein function, semantic scene, and music categorization problems. In the real world, multiple heterogenous subjective factors decide the extent of damage/severity of a particular crash instance. Theoretically, the traffic collision type and crash severity type can be correlated, and thus, it is intuitive to model them simultaneously. The ability of MLC to categorize an entity under analysis to more than one labels, correlated or uncorrelated, provides the approach an edge over the single-class (binary) or multi-class classification approach. The MLC based classification model was calibrated and tested using the historical crash data extracted for the state of Texas. The selection of study area was based on a link-level unsupervised principal component analysis-based clustering approach. Similar clustering approach was also tested at the county-level to understand the spatial behavior and thus transferability of the MLC approach to other key cities in the state. The performance of the proposed approach was tested, compared, and quantified with the conventional binary/multi-class classification tools used in the traffic safety domain. Inferences from the preliminary numerical analysis indicates that the proposed multi-label classification approach has promising performance compared to the traditional classification approaches, specifically found in traffic safety literatures.

17. Key Words	18. Distribution Statement
Crash injury severity, Machine Learning, Artificial Intelligence, Multi-Label Classification	No restrictions. This document is available through the National Technical Information Service, Springfield, VA 22161.

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price
Unclassified	Unclassified	55	

Form DOT F 1700.7 (8-72) **Reproduction of completed page authorized.**

SI* (MODERN METRIC) CONVERSION FACTORS

APPROXIMATE CONVERSIONS TO SI UNITS

Symbol	When You Know	Multiply By	To Find	Symbol
		LENGTH		
in	inches	25.4	millimeters	mm
ft	feet	0.305	meters	m
yd	yards	0.914	meters	m
mi	miles	1.61	kilometers	km
		AREA		
in^2	square inches	645.2	square millimeters	mm^2
ft^2	square feet	0.093	square meters	m^2
yd^2	square yard	0.836	square meters	m^2
ac	acres	0.405	hectares	ha
mi^2	square miles	2.59	square kilometers	km^2
		VOLUME		
fl oz	fluid ounces	29.57	milliliters	mL
gal	gallons	3.785	liters	L
ft^3	cubic feet	0.028	cubic meters	m^3
yd^3	cubic yards	0.765	cubic meters	m^3
		NOTE: volumes greater than 1000 L shall be shown in m^3		
		MASS		
oz	ounces	28.35	grams	g
lb	pounds	0.454	kilograms	kg
T	short tons (2000 lb)	0.907	megagrams (or "metric ton")	Mg (or "t")
		TEMPERATURE (exact degrees)		
°F	Fahrenheit	5 (F-32)/9 or (F-32)/1.8	Celsius	°C
		ILLUMINATION		
fc	foot-candles	10.76	lux	lx
fl	foot-Lamberts	3.426	$candela/m^2$	cd/m^2
		FORCE and PRESSURE or STRESS		
lbf	poundforce	4.45	newtons	N
lbf/in^2	poundforce per square inch	6.89	kilopascals	kPa

APPROXIMATE CONVERSIONS FROM SI UNITS

Symbol	When You Know	Multiply By	To Find	Symbol
		LENGTH		
mm	millimeters	0.039	inches	in
m	meters	3.28	feet	ft
m	meters	1.09	yards	yd
km	kilometers	0.621	miles	mi
		AREA		
mm^2	square millimeters	0.0016	square inches	in^2
m^2	square meters	10.764	square feet	ft^2
m^2	square meters	1.195	square yards	yd^2
ha	hectares	2.47	acres	ac
km^2	square kilometers	0.386	square miles	mi^2
		VOLUME		
mL	milliliters	0.034	fluid ounces	fl oz
L	liters	0.264	gallons	gal
m^3	cubic meters	35.314	cubic feet	ft^3
m^3	cubic meters	1.307	cubic yards	yd^3
		MASS		
g	grams	0.035	ounces	oz
kg	kilograms	2.202	pounds	lb
Mg (or "t")	megagrams (or "metric ton")	1.103	short tons (2000 lb)	T
		TEMPERATURE (exact degrees)		
°C	Celsius	1.8C+32	Fahrenheit	°F
		ILLUMINATION		
lx	lux	0.0929	foot-candles	fc
cd/m^2	$candela/m^2$	0.2919	foot-Lamberts	fl
		FORCE and PRESSURE or STRESS		
N	newtons	0.225	poundforce	lbf
kPa	kilopascals	0.145	poundforce per square inch	lbf/in^2

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS, ABBREVIATIONS, AND SYMBOLS

Notation	Mathematical Description
χ	d- dimensional instance space $\mathbb{R}^d (or\ \mathbb{Z}^d)$
$\mathcal{Y}$	Label space with q possible class labels $\{y_1, y_2, \ldots, y_q\}$
x	d-dimensional feature vector $(x_1, x_2, \ldots, x_d)^t$ $(x \in \chi)$
Y	Label set associated with x $(Y \subseteq \mathcal{Y})$
$\bar{Y}$	Complimentary set of Y in $\mathcal{Y}$
$\mathcal{D}$	Multi-label training set $\{(x_i, Y_i)\} \mid 1 \leq i \leq m$
S	Multi-label test set $\{(x_i, Y_i)\} \mid 1 \leq i \leq p$
$h(\cdot)$	Multi-label classifier h: $\chi \rightarrow 2^{\mathcal{Y}}$, where $h(x)$ returns the set of proper labels for x
$f(\cdot, \cdot)$	Real-valued function f: $\chi \times \mathcal{Y} \rightarrow \mathbb{R}$, where f (x, y) returns the confidence of y being proper label of x
$rank_f(\cdot,\cdot)$	$rank_f(x, y)$ returns the rank of y in $\mathcal{Y}$, in descending order induced from $f(x, \cdot)$
$t(\cdot)$	threshold function t: $\chi \rightarrow \mathbb{R}$, where $h(x) = \{y \mid f(x, y)\} > t(x), y \in \mathcal{Y}\}$
$\lvert \cdot \rvert$	$\lvert A \rvert$ returns the cardinality of set A
$[\![\cdot]\!]$	$[\![\pi]\!]$ returns 1 if predicate π holds, and 0 otherwise
$\phi(\cdot, \cdot)$	$\phi(Y, y)$ returns +1 if y $\in$ Y, and -1 otherwise
$\mathcal{D}_j$	Binary training set $\{(x_i, \phi(Y_i, y_j))\} \mid 1 \leq i \leq m$ derived from $\mathcal{D}$ for the j[th] class label y_i
$\Psi(\cdot, \cdot, \cdot)$	$\Psi(Y, y_i, y_k)$ returns +1 if $y_i \in Y$ and $y_i \notin Y$, and -1 if $y_j \notin Y$ and $y_k \in Y$
$\mathcal{D}_{jk}$	Binary training set $\{(x_i, \Psi(Y_i, y_j, y_k)) \mid \phi(Y_i, y_i) \neq \phi(Y_i, y_k), 1 \leq i \leq m\}$ derived from $\mathcal{D}$ for the label pair (y_i, y_k)
$\sigma_y(\cdot)$	Injective function $\sigma_{\mathcal{Y}} = 2^{\mathcal{Y}} \rightarrow \mathbb{N}$ mapping from the power set of $\mathcal{Y}$ to natural numbers ($\sigma_{\mathcal{Y}}^{-1}$ being the corresponding inverse function)
$\mathcal{D}_{\mathcal{Y}}^{\dagger}$	Multi-class (single label) training set $\{(x_i, \sigma_{\mathcal{Y}}(Y_i))\}$ $/ 1 \leq i \leq m$ derived from $\mathcal{D}$
$\mathcal{B}$	Binary learning algorithm [complexity: $F_B(m, d)$ for training: $F'_B(d)$ for (per-instance) testing]
M	Multi-class learning algorithm [complexity: $F_M(m, d, q)$ for training: $F'_M(d, q)$ for (per-instance) testing]

EXECUTIVE SUMMARY

Road safety is an important part of the social and economic wellbeing of the society. Across the globe, people use traffic infrastructures to carry out their day-to-day activities. Fatalities and injuries resulting from motor vehicle crashes presents an important public health concern globally. Traffic crashes are an important source of non-recurrent congestions, causing delays to travelers in a transportation network. The United States' annual average cost of road crashes is estimated to be around $230.6 billion or around $820 per person (*1*). In the year 2017, Texas recorded 1.38 death per 100 million miles traveled (*2*). The objective of traffic safety analysis is to ensure that people arrive at their destination without any abnormal incident or crash. In the context of traffic safety, understanding the factors affecting crash occurrence, injury severity and collision type and their underlying relationships help us in predicting future crashes, its severity and collision type under given circumstances. Naturally, crash analysis is complex due to the presence of human behavior element, which is difficult to predict and model due to the high subjective variations in people's decision making. This research primarily endeavors to analyze traffic crashes as multi-label classification problem where an instance can be mapped to multiple labels. In other words, the model proposed in this study can predict or classify the traffic crash based on the severity and collision type simultaneously through an application of supervised Machine Learning (ML) classification tool namely Multi-Label Classification (MLC) system from the domain of Artificial Intelligence. The study also incorporates an unsupervised ML tool namely, Principal Component Analysis (PCA) based clustering technique for the grouping of instances at the link-level and at the county level, to test the natural groups in the data and spatial transferability of the proposed approach. From theoretical point of view, the cluster information is extremely useful when it comes to spatial comparison of region in terms of attributes that controls the safe operation of both humans and autonomous driving features. In particular, the study tests popular multilabel learning algorithms for this simultaneous classification. The study compares the performance of proposed approach with other conventional ML classification tools used in the past for crash classification in terms of labelling accuracy and computational efficiency. The classification performance of all the conventional and proposed classification algorithms considered in this study has been benchmarked and compared in terms of prediction performance and computational efficiency. Though more comprehensive training and testing is required, the numerical result from this study indicates that the proposed approach has a promising overall classification performance compared to traditional multiclass traffic crash injury classification approaches. The numerical results and outputs from the PCA based clustering analysis of the counties in Texas with respect to crash and other related data proved to more than just auxiliary, as many essential inferences and information needed for both, pre and post modelling phase of MLC analysis was obtained from this. The insights from these safety analysis and model outputs will help in identifying critical locations/links in a transportation network. The information about critical links can be used for optimal positioning of troopers, and in prioritizing the location for frequent surveillance by traffic management centers. In particular, the model can be used for predicting the probable locations of crashes and severity types. This will allow troopers to position themselves in strategic locations. As troopers will be nearby to incident location, they are more likely to reach the incident spot in smaller time and help stabilize the victim in their golden hour. This will also help in clearing the traffic in shorter time thereby saving fuel and reducing air pollution due to congestion built by incident which otherwise may be for prolonged time. It can also help in better planning for incident management and in optimal allocation of resources/funds.

1. INTRODUCTION

Road safety is a collective responsibility that requires synergy between road users, decision-making by government, industry, non-governmental organizations, and international agencies. Fatality rate of traffic crashes on roadways is 12.4 deaths per 100,000 inhabitants, and motor vehicle crashes are a leading cause of death in the United States, with over 100 people dying every day (*1*, *3*). In 2015, more than 2.5 million drivers/passengers were admitted in emergency room for crash related injury treatment. For crashes that occurred in 2017, the cost of medical care and productivity losses associated with occupant injuries and deaths from motor vehicle traffic crashes exceeded $75 billion (*4*, *5*). In 2018, the estimated total police-reported motor vehicle crashes that involved towing of at least one passenger vehicle involved in the crash was around 2,811,185 and resulted in over 1,489,413 known injuries. Among these crashes, 2.7 percent (74,604) were crashes with injury levels rated serious or above, 33.3 percent (935,120) were crashes with moderate or minor injury levels, and 50.9 percent (1,429,853) were crashes with no injury (*6*, *7*). The estimated economic loss of all motor vehicle crashes for the state of Texas for 2019 was around $39,200,000,000 (*2*). Crash frequency prediction and risk analysis has become very popular ever since the safety was regarded as top priority aspect for Transportation planning and management. Every year, the U.S. federal government provides approximately $579 million to states for traffic safety programs (*8*).

Crash-related fatalities and injuries can only be prevented by a joint involvement from multiple sectors (transportation agencies, police, health departments, education institutions) that oversee road safety, vehicles, and the drivers themselves. Effective interventions include design of safer infrastructure and incorporation of road safety features into land-use and transport planning; improvement of vehicle safety features; improvement of post-crash care for victims of road crashes, and improvement of driver behavior, such as setting and enforcing laws relating to key risk factors and raising public awareness (*9*, *10*). One of the fundamental approach by which Vision Zero strategy aims to eliminate all traffic fatalities and severe injuries, while increasing safe, healthy, equitable mobility for all is to analyze the historical crash data which involves collecting, analyzing, and using data to understand trends and potential disproportionate impacts of traffic deaths on certain populations (*11*, *12*). In summary, crash data is important for (*10*, *13*):

- Award and target state and federal highway safety funding,
- Focus on local and state law enforcement efforts,
- Enforce existing laws to ensure driver/vehicle compliance,
- Conduct problem identification and the development of resolutions for safety programs,
- Make key legislative decisions that impact citizen safety on roadways,
- Identify high crash locations and make engineering and construction improvements to roadways,
- Educate the public on safety issues (i.e., seat belt use, aggressive driving, and speeding),
- Improve Emergency Medical Services (EMS) through processes such as training EMS personnel or the deployment of EMS units.

The analysis of traffic safety often involves categorization or classification problems. Data classification is one of the central ML tasks. Though ML classification algorithms are usually

designed and employed for single label classification tasks, most real-world problems are multilabel in nature, where an instance can have more than one class label. Literature in traffic safety domain presents two classification systems. First, categorizes crash data based on the crash severity type as: no injury, possible injury, non-incapacitating injury, incapacitating injury, and fatal injury. Second classify crash based on collision type (rear end, side swipe etc.). A common approach is to find the frequency of crashes for severity types and collision types separately (*14*). This study is motivated by the fact that traffic collision type and crash severity type may be correlated, hence it is intuitive to model them simultaneously (*14*). Modeling them separately may necessitate the need for a more complex model structure to account for cross-model-correlations (*14–16*). This research endeavors to analyze traffic crashes as multi-label classification problem where an instance can be mapped to multiple labels. In other words, the model proposed in this study can predict or classify the traffic crash based on the severity and collision type simultaneously through an application of machine learning tool namely Multi-Label Classification (MLC) system from the domain of Artificial Intelligence.

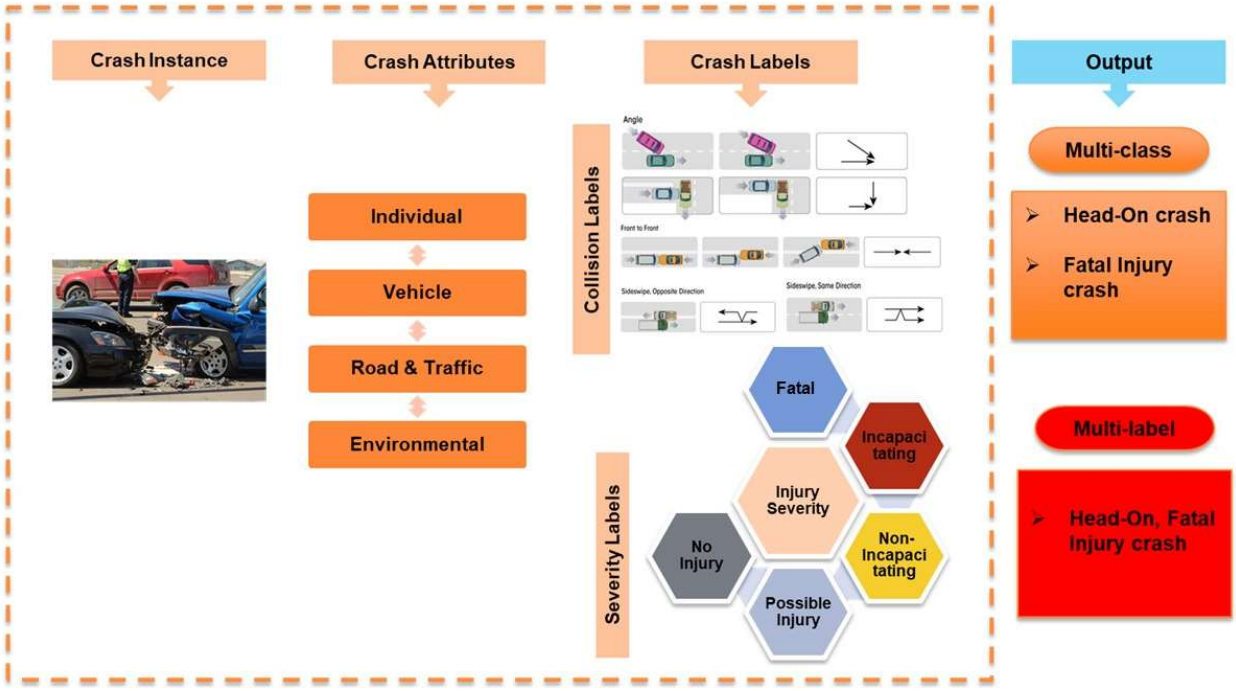


Figure 1. Multi-Label Classification-Traffic Safety Analogy

The overall structure of the report is as follows: The next section of the report covers the primary and secondary objectives/outcomes from the project, which is followed by literature review and analysis methodology respectively. Final sections of the report discuss the numerical results which is followed by conclusions that explains the preliminary inferences and significance of the performance of the proposed algorithms.

2. OBJECTIVES

This study aims at classifying the collision type and crash severity type simultaneously while capturing the correlation among them. The study reviewed past literature to understand the state of art and practice of crash classification. The study utilizes the past crash data from Texas cities for crash analysis and numerical analysis. The main objective of this project is to explore the Machine Learning Multi-Label Classification tool for classification problems in the context of traffic safety simultaneous classification of collision type and crash injury type. The ability of MLC to categorize an entity under study to more than one labels simultaneously provides it an edge over the traditional classification approaches that classify collision type and crash injury types separately. Underlying correlation between the injury severity type and collision type is leveraged using the AI tool to develop a robust classification model, and the performance of the proposed tool will be benchmarked with the conventional tools. This includes development of a person-level crash injury severity database using the past five-year historic motor vehicle crashes for the nominated target areas. Study objectives are:

- Generate the database for the predictive model framework
- Perform correlation analysis between the crash collision and injury severity type for:
    - Most severe injury person
    - Driver
    - Passenger
- Feature selection/extraction
- Numerical experiments:
    - Conventional models
    - Proposed MLC model
- Performance evaluation and comparison, and
- Result documentation.

An auxiliary objective has been added to support the primary analysis and future direction of the project. Specifically, for checking similarity within counties (natural groups) and spatial transferability of models, an unsupervised clustering technique, namely Hierarchical clustering has been added for the crash data, aggregated to the county level resolution. Attaining these objectives shall allow the research team to identify interaction of on-system and off-system segments and intersections crashes based on the collision and injury severity types and the type of person involved in the crash.

3. LITERATURE REVIEW

Crash frequency prediction and risk analysis has become very popular ever since the safety was regarded as top priority aspect for Transportation planning and management. Every year, the U.S. federal government provides approximately $579 million to states for traffic safety programs (8). Spatially aggregated crash analysis is a critical point of interest for state and federal safety and planning agency as many factors affecting crashes operate at a spatial scale (e.g. land-use policy, demographic characteristics and highway infrastructure functional class) (17). Recent developments in the subfield of spatial traffic crash analysis have enabled researchers to better understand regional crash frequency and the rate of dependency of respective causal factors. In road traffic crash, risk is a function of four elements– the amount of movement or travel within the system by different users or a given population density, the underlying probability of a crash, given a particular exposure, the probability of injury, given a crash and the outcome of injury.

The primary factors contributing to the increase in global road crash injuries is the growing number of motor vehicles. The problem is not just the growth in numbers and increase in exposure to the risk but also ensuring that appropriate road safety measures accompany this growth. The motor vehicle, along with the subsequent growth in the number of motor vehicles and in road infrastructure, has brought societal benefit but it has also led to societal cost, to which road traffic injury contributes significantly. Without proper planning, growth in the number of motor vehicles can lead to problems for pedestrians and cyclists. In fact, where there are no facilities for pedestrians and cyclists, increasing numbers of motor vehicles generally lead to reductions in walking and cycling. Speeding, drunk driving, distractions and cellphones, weather, red light accidents, time and day, driver fatigue etc. are some of the primary precursors of traffic incidents. The historic crash data analysis also exposed interesting evidence on the effect of gender on crashes. In majority of car accidents, males have been shown to have the highest risk of being subjected to high injury from crashes. From 1975-2015 the number of males died in a car accident was more twice the number of females. In 2015, over 71% of car accident deaths were males. This is a similar trend over the past decade where over 350,000 people were killed in a car accident (18, 19). In the context of traffic safety, understanding the factors affecting crash occurrence, injury severity and collision type and their underlying relationships help us in predicting future crashes, its severity and type under given circumstances. Road traffic crash results from a combination of factors related to the components of the system comprising roads, the environment, vehicles and road users, and the way they interact. Some factors contribute to the occurrence of a collision and are therefore part of crash causation. Other factors aggravate the effects of the collision and thus contribute to severity. Other factors may not appear to be directly related to road traffic injuries. Some causes are immediate, but they may be underpinned by medium-term and long-term structural causes. Identifying the risk or threat factors that contribute to road traffic crashes is important in identifying interventions that can reduce the risks associated with those factors. Risk factors and relation to traffic crashes are summarized and presented in Table 1. Due to its importance, there has been extensive research utilizing various statistical models to expose the association between risk factors and injury severity. Studies in past have utilized numerous statistical, econometric and ML tools that fit the data under investigation to extract the underlying relationship between the crash factors and crash and/or collision type.

Table 1. Risk Factors for Road Crash Injuries

Factors influencing exposure to risk	Risk factors influencing crash involvement	Risk factors influencing crash severity	Risk factors influencing post-crash outcome of injuries
• economic factors such as level of economic development and social deprivation • demographic factors such as age and sex • insufficient attention to integration of road function with decisions about speed limits, road layout and design • land-use planning practices which influence length of trip and mode of travel • mixture of high-speed motorized traffic with vulnerable road users	• inappropriate and excessive speed • fatigue • being a young male • having youths driving in the same car • being a vulnerable road user in urban and residential areas • travelling in darkness • vehicle factors – such as braking, handling and maintenance • defects in road design, layout, and maintenance, which can also lead to unsafe behavior by road users • inadequate visibility because of environmental factors (making it hard to detect vehicles and other road users) • poor eyesight of road users	• human tolerance factors • inappropriate or excessive speed • seatbelts and child restraints not used • crash-helmets not worn by users of two-wheeled vehicles • roadside objects not crash-protective • insufficient vehicle crash protection for occupants and for those hit by vehicles • presence of alcohol and other drugs	• delay in detecting crash and in transport of those injured to a health facility • presence of fire resulting from collision • leakage of hazardous materials • presence of alcohol and other drugs • difficulty in rescuing and extracting people from vehicles • difficulty in evacuating people from buses and coaches involved in crash • lack of appropriate pre-hospital care • lack of appropriate care in hospital emergency rooms.

Over the years, traffic safety professionals and researchers have identified several data characteristics and methodological issues that are critical considerations in the development and application of an appropriate statistical methodology to study such data (14, 20–27), these issues are presented in Figure 2.

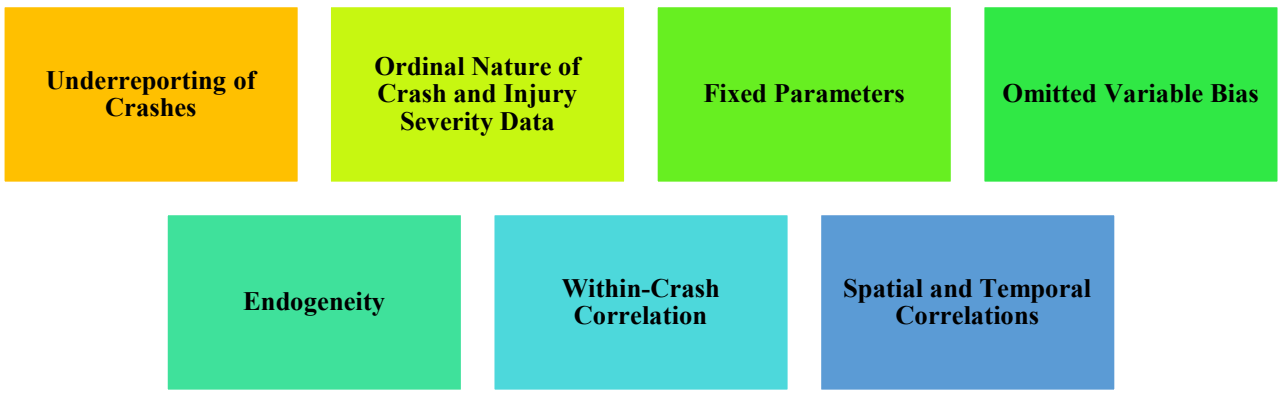


Figure 2. Challenges dealing with traffic safety data

Lord and Mannering (14) provided a comprehensive review of methodological aspects and list widely used econometric tools for investigation of crash frequency data and severity analysis. Statistical/Econometric modeling assumes a distribution of data and then extracts relationship

information between the feature and labels. The assumed distribution in statistical/econometric models may not be true about the data, thus leading to poor performance of the model estimations (28–31). Although planners and researchers typically use the traditional statistical models for classification problems, they suffer fundamental limitations in tackling multi-label classification problems. Recently, for various applications in different disciplines including transportation engineering, supervised ML classification tools have shown an edge over conventional statistical and econometric count models in terms of predictive capabilities.

3.1. Conventional ML Classification Algorithms

Logistic Regression: Target label (y_i) is modelled as a linear function of (x_i), which utilizes a standard logistic function or otherwise called sigmoid function, given in equation no. 1, that can transform the final solution to either 0 or 1. In other words, the linear combination of features, x_i is a function that spans from minus infinity to plus infinity whereas the labels vector y_i has positive discrete choices (two for binary logistic and two or more for multiclass logistic regression). The optimization criterion in logistic regression is called maximum likelihood where, rather than minimizing the average loss, the likelihood of training data is maximized (32), (33).

Decision Tree: The decision support tool which is an acyclic graph that can be used to make decisions. It consists of branch like graphs, where at each node of the graph, a specific feature *j* of the feature vector *x* is examined. There are several approaches to deploy decision tree algorithm, ID3, C4.5, CART etc. (32). ID3 algorithm is used to classify nominal valued datasets. Generally, Decision tree models are built in two steps, induction and pruning. Induction deals with the building the branches or hierarchical decision boundaries and pruning refers to the removing the unnecessary branches from the model based on the dataset (32), (33).

k-Nearest Neighbors (kNN): This is a non-parametric learning algorithm in which the KNN algorithm keeps all the training examples in the memory even after modeling and when an unlabeled instance comes, the algorithm finds *k* instances (neighbors) from the training dataset closest to the unlabeled instance and returns the majority label from these neighbors. The closeness or similarity of examples is quantified using a distance function, usually Euclidean distances (32), (33).

Support Vector Machine (SVM): The SVM algorithm finds the hyperplane in the N-dimensional space that classify the instances to its respective labels. The total number of features decides dimension of the space under consideration. Support vectors are data points that are close to the hyperplane, and they influence the position and orientation of the hyperplanes. SVM tries to find the largest separating margin between classes.

Table 2 presents some of the advantages and disadvantages of the conventional classification tools used in the study (32), (33).

Table 2. Overview of some Classification Tools

Classification tool and related works in traffic safety	Advantages	Disadvantages
Logistic Regression **Burkov(*32*), Bonaccorso(*33*), Lu et al. (*34*), Iranitalab and Khattak (*35*)**	• **Simple to implement** • **Computationally efficient**	• **Prone to overfitting** • **High reliance on proper data presentation** • **Decision surface is linear, not suited for nonlinear problems**
Decision Tree **Burkov (*32*), Bonaccorso (*33*), Iranitalab and Khattak (*35*), Yuan et al. (*36*)**	• **Computationally efficient** • **Easy interpretation of results** • **Can deal with outliers** • **Less data cleaning required**	• **Prone to overfitting** • **Not fit for continuous variables** • **Complexity increases with number of labels**
k Nearest Neighbor **Burkov(*32*), Bonaccorso(*33*), Iranitalab and Khattak (*35*), Yuan et al. (*36*)**	• **No assumption about data** • **Simple to implement** • **High accuracy** • **Versatile**	• **Computationally expensive** • **High memory requirement** • **Sensitive to irrelevant data and scale of data**
Support Vector Machine **Li et al. (*28*), Burkov(*32*), Bonaccorso(*33*), Iranitalab and Khattak (*35*), Yuan et al. (*36*)**	• **High accuracy** • **Perform well for clean small data**	• **Not suited for very large dataset, takes longer time for training** • **Less effective for data with noise**

3.2. Multi-label Classification

In multi-label learning, each object is also represented by a single instance while associated with a set of labels instead of a single label, unlike traditional supervised learning (binary or multi-class). The task is to learn a function which can predict the proper label sets for unseen instances. In multi-label classification, the examples are associated with a set of labels $Y \subseteq L$.

MLC induces a predictive model from a set of training data, which later assigns one or more labels to each new test example (*37*, *38*). Suppose $\chi=R^d$ or Z^d denotes the d-dimensional instance space, and $\mathcal{Y}=\{y_1, y_2,\ldots,y_q\}$ denotes the label space with q possible class labels. The task of multi-label learning is to learn a function $h: \chi \rightarrow 2^{\mathcal{Y}}$ from the multi-label training set $D=\{(\ldots,Y_i)\} \mid 1 \leq i \leq m$. For each multi-label example $\ldots, x_i \in \chi$ is a d-dimensional feature vector $(x_{i1}, x_{\ldots}, \ldots, x_{id})^T$ and $Y_i \subseteq \mathcal{Y}$ is the set of labels associated with x_i. For any unseen instance $(x \in \chi)$, the multi-label classifier $h(\cdot)$ predicts $h(\cdot) \subseteq \mathcal{Y}$ as the set of proper labels for x. Traditional two-class and multi-class problems can both be cast into multi-label ones by restricting each instance to have only one label. However, the generality of multi-label problem makes it more difficult to learn. An intuitive approach to solve multi-label problem is to decompose it into multiple independent binary classification problems (one per category). But this kind of method does not consider the correlations between the different labels of each instance.

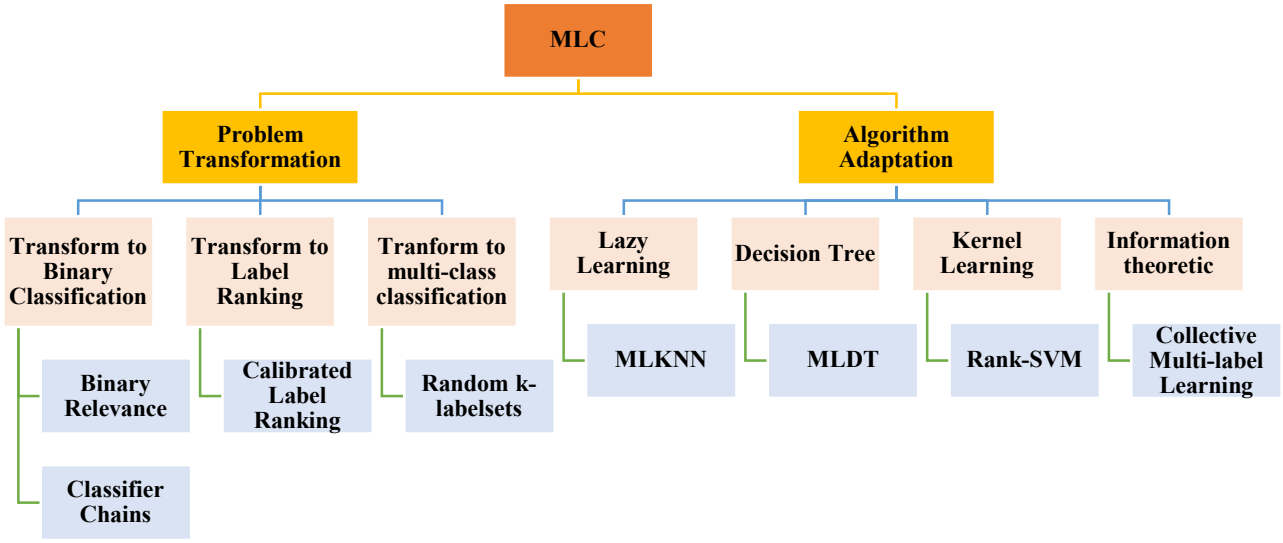


Figure 3. Few MLC algorithms

3.3. MLC Strategies

The selection of the appropriate MLC algorithm depends on the specific dataset that the algorithm is trying to classify, but the primarily depends on how well the label correlations needs to be captured. There is also a tradeoff between computational cost and classification performance while choosing the MLC strategy. Here, first, second and high order strategy are briefly discussed.

First-order strategy:

The task of multi-label learning is tackled in a label-by-label style and thus ignoring coexistence of the other labels, such as decomposing the multi-label learning problem into several independent binary classification problems (one per label). The noticeable merit of first order strategy lies in its conceptual simplicity and high efficiency. On the other hand, the effectiveness of the resulting approaches might be suboptimal due to the ignorance of label correlations.

Second-order strategy:

The task of multi-label learning is tackled by considering pairwise relations between labels, such as the ranking between relevant label and irrelevant label or interaction between any pair of labels.

As label correlations are exploited to some extent by second-order strategy, the resulting approaches can achieve good generalization performance. However, there are certain real-world applications where label correlations go beyond the second-order assumption.

High-order strategy:

The task of multi-label learning is tackled by considering high-order relations among labels such as imposing all other labels' influences on each label or addressing connections among random subsets of labels etc. High-order strategy has stronger correlation-modeling capabilities than first order and second-order strategies, while on the other hand is computationally more demanding and less scalable.

Table 3 presents summary of different MLC strategies and the order of label correlation the algorithms deal with, while training the classification models.

Table 3 Summary of Some Multi-Label Learning Algorithms (adapted from (37–45)

Algorithm	Idea	Order of Correlation	Literature Domain
Binary Relevance	Fit multilabel data to q binary classifiers	First-order	Image
Classifier Chains	Fit multilabel data to a chain of binary classifiers	High order	Image, Video, Text, Biology
Calibrated Ranking Loss	Fit multilabel data	Second-order	Image, Text, Biology
Random k-Label sets	Fit multi-label data to n multi-class classifiers	Second-order	Image, Text, Biology
ML-kNN	Fit k-nearest neighbor to multi-label data	First-order	Image, Text, Biology
ML-DT	Fit decision to multi-label data	First-order	Biology
Rank-SVM	Fit kernel learning to multi-label data	Second-order	Biology
CML	Fit conditional random field to multi-label data	Second-order	Text

In the past, multilabel classification was mainly motivated by the tasks of text categorization and medical diagnosis. Text documents usually belong to more than one conceptual class. MLC is a classification task where an instance can be simultaneously classified in more than one class. Labeled data extracted from several domains, like text, web pages, multimedia (audio, image, videos), and biology are intrinsically multi-labeled (*40*). Additionally, the number of application domains with MLC data is growing fast. Early multi-label learning studies primarily focused on the problem of multi-label text categorization. Recently, multi-label learning has attracted significant attention from ML and related communities and has been widely applied to diverse problems from automatic annotation for multimedia contents to bioinformatics, web mining, rule mining, information retrieval tag recommendation etc. (*46–49*).

In a recent application, Rivolli et al. (*50*) used MLC algorithms to recommend food truck cuisines, assuming that a person can have more than one cuisine preference, and with the same level of preference. While in multi-class classification only a single class label is predicted, in MLC, more than one class label can be simultaneously predicted. In the same way as multi-class classification tasks can be seen as a generalization of binary classification tasks, which restricts to two the number of classes, MLC can be seen as a generalization of multi-class, which restricts to one the number of predicted classes (*37*).

3.4. Clustering Analysis Segmenting crash cluster groups

The clustering analysis pipeline for this analysis was aimed to conduct a data driven Principal (PCA) based cluster analysis using Texas counties data to identify natural group based on similarity in characteristics using Hierarchical clustering tools. From theoretical point of view, this information is extremely useful when it comes to spatial comparison of region in terms of crash risk analysis and feasible transferability of efficient prediction models to similar county groups. From the perspective of planner's and policy makers, such information can assist in devising efficient crash counter measures and financial investments for necessary regions. The methodological framework presented in this study can also be used for a data exploration prior to modeling the crash risk at the aggregated spatial level.

Application of clustering techniques in the field of Transportation Engineering is not new. Papagiannakis et al. (*51*) utilized clustering technique to establish similarities in vehicle classification and axle load distributions between traffic data collection sites. Cao et al. (*52*) conducted a cluster analysis o' vessel's trajectories based on the Automatic Identification System (AIS) datasets of Wuhan Erqi Yangtze River Bridge area. Relationships and information issues regarding quality of safety data for developing nations was studied by Raihan et al.(*53*) using hierarchical clustering coupled with random forest method. Clustering is an important data preprocessing step for many of the Artificial Neural Network modeling framework. Taamneh et al.(*54*) studied the severity of road traffic crashes using a Hierarchical clustering based Artificial Neural Network model. Using clustering, Janstrup et al.(*55*) presented an integrated analysis of information about road maintenance, maintenance costs, road characteristics, crash characteristics, and geographical location that can enrich road maintenance prioritization from a traffic safety perspective. Such Macroscopic crash analysis, where crashes are aggregated to traffic analysis zones or county or ZIP code etc. are considered to quantify the impacts of socioeconomic and demographic characteristics, transportation demand and network attributes to provide countermeasures from a planning perspective (*56*). Similar analysis assists decision-makers in delivering efficient and effective resources allocation and policy analysis for priority regions. It is therefore reasonable to explore the use of spatial models of crash occurrence to better understand the implications of government policies and safety initiatives. Several planning acts have emphasized the importance of macroscopic crash analysis. Originally, the Transportation Equity Act fo^{r} the 21st Century (*57*) suggested to consider safety in the transportation planning process. Washington et al. (*58*) discussed how to incorporate safety into transportation planning at different levels. The Moving Ahead for Progress i^{n} the 21st Century Act (MAP-21 Act) (*59*) and Fixing America's Surface Transportation Act (FAST Act) (*60*) require the incorporation of transportation safety in the long-term transportation planning process.

Study Significance and Contributions

The type and extent of analysis deployed for this project, specifically the simultaneous classification of traffic crash collision and severity injury type using multi-label classification combined with the right problem transformation approach, is probably one of the first attempts in the field of traffic safety analysis. The results from the preliminary testing phase highlights the efficiency of the proposed model framework in terms of speed and accuracy. The project team has also emphasized the importance of data engineering for optimized allocation of computational memory which is key for application/ practical stages of the project, which is rarely covered in other studies. This is very crucial for many machine-learning and deep-learning algorithms that requires heavy computational horsepower. The database management framework designed for this project is reasonably efficient and highly flexible, meaning, the any road segments in Texas can be analyzed/modelled using this design, but the research also believes more improvements can be done to the efficiency side of the data, with respect to the memory usage during the data loading phase. The contributions of this study can be summarized as follows. First, it provides framework for employing MLC for the categorization problems in the traffic safety domain. Most importantly, this study is aimed at attaining better prediction benchmark for the classification models by focusing on the target rather than on modifying the algorithms. This could be very key for Spatial-Temporal transferability of ML/AI Models. The application of clustering techniques to create intuitive dendrogram visualization of the county clusters in terms of traffic safety is another new topic, the project team has explored and validated. Dendrogram produced from clustering process

is extremely useful in understanding the data. By observing the branches of the hierarchical dendrogram structure, insightful information about those county group that varies significantly with respect to explanatory features and crash types and those which does not have any effect on the same can be understood. From theoretical point of view, this information is extremely useful when it comes to spatial comparison of region in terms of crash risk analysis and feasible transferability of efficient prediction models to similar county groups. From the perspective of planner's and policy makers, such information can assist in devising efficient crash counter measures and financial investments for necessary regions. The methodological framework presented in this study can be used for a data exploration prior to modeling the crash risk at the aggregated spatial level.

4. METHODOLOGY

The primary research objective deals with simultaneous classification of crash collision type and severity level, provided accident occurred at a certain location at certain time. Motor vehicle traffic accidents can be studied by the events they contain, often by the harmful events. The first harmful event of an accident refers to the first injury or damaging producing event that occurs. The most harmful event is typically recorded for each involved unit. It refers to the event that produces the most severe injury or amount of damage for each unit. For accidents with more than four events in the sequence, it is suggested to omit the event(s) least relevant to describing the crash (*61*). The terms collision and non-collision when used to classify an accident, refer to the first harmful event of the accident being a collision event or non-collision event. Collison events in a vehicle's sequence of events is all harmful (causing injury or damage) and describe the motor vehicle striking or being struck by another vehicle, person, or object. Non-collision events in a vehicle's sequence-of-events is not necessarily harmful and describe events in the accident other than collision events.

Injury-severity data are generally represented by discrete categories such as fatal injury, incapacitating injury, non-incapacitating, possible injury, and property damage only, often referred to as the KABCO scale (*62*). The description of each level of injury severity, as per Texas regulations are:

- **fatal injury or killed (K):** succumbed due to injuries sustained from the crash, within 30 days of the crash
- **incapacitating injury (A):** severe injury which prevents continuation of normal activities; includes broken or distorted limbs, internal injuries, crushed chest, etc.
- **non-incapacitating (B):** evident injury such as bruises, abrasions, or minor lacerations which do not incapacitate.
- **possible injury:** injury which is claimed, reported, or indicated by behavior, but without visible wounds; includes limping or complaint of pain.
- **property damage only (O):** person involved in crash did not sustain an A, B, or C injury.

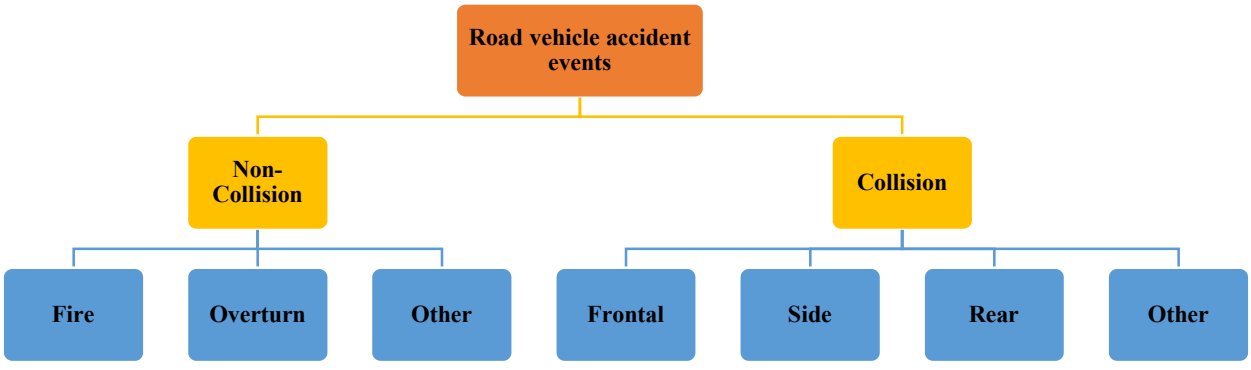


Figure 4. Crash Events

4.1. Learning Algorithms

4.1.1. Binary Relevance

The basic idea of this algorithm is to decompose the multilabel learning problem into q independent binary classification problems, where each binary classification problem corresponds to a possible label in the label space (*41*). Following the notations in (list of symbols), for the j^{th} class label y_j, Binary Relevance first constructs a corresponding binary training set by considering the relevance of each training example to y_j:

$$\mathrm{D_j}=\left\{\left(I, \phi(\mathrm{Y_i},\mathrm{y_j})\right)\right\} \mid 1 \le \mathrm{i} \le \mathrm{m} \tag{1}$$

$$\text{where, } \phi I=\begin{cases} +1, \text{ if } I \in \mathrm{Y_i} \\ -1, \text{ otherwise} \end{cases} \tag{2}$$

Next, a binary learning algorithm B is applied to induce a binary classifier $g_j: \boldsymbol{\chi} \to \mathbb{R}$, i.e., $g_j \leftarrow \mathrm{B}(\mathrm{D}_j)$. Thus, for any multi-label training example (I, Y_i), instance x_i will be involved in the learning process of q binary classifiers. For relevant label $I_j \in Y_i$, xi is regarded as one positive instance in inducing $g_j(.)$. For irrelevant label $y_j \in I_i$, x_i is regarded as one negative instance (also known as cross-training). For any unseen instance x, Binary Relevance predicts its associated label set Y by querying labeling relevance on each individual binary classifier and then combing relevant labels:

$$\mathrm{Y}=\left\{\mathrm{y_j} \mid \mathrm{g_j}(\mathrm{x}) > 0,\ 1 \le \mathrm{j} \le \mathrm{q}\right\} \tag{3}$$

Binary relevance method is a first-order approach which builds classifiers for each label separately and offers the natural opportunity for parallel implementation. The most prominent advantage of Binary Relevance lies in its extremely straightforward way of handling multi-label data, which has been employed as the building block of many state-of-the-art multi-label learning techniques.

4.1.2. Classifier Chains

The algorithm transforms the multilabel learning problem into a chain of binary classification problems, where subsequent binary classifiers in the chain is built upon the predictions of preceding ones (*42*, *63*). For q possible class labels $\{y_1, y_2, y_3, \ldots., y_q\}$, let $\tau: \{1, \ldots., q\} \to \{1, \ldots., q\}$ be a permutation function which is used to specify an ordering over them, $y_{\tau(1)} > y_{\tau(2)} > y_{\tau(3)} > \cdots .. > y_{\tau(q)}$. For the j^{th} label $y_{\tau(q)}\ (1 \le j \le q)$ in the ordered list, a corresponding binary training set is constructed by appending each instance with its relevance to those labels preceding $y_{\tau(j)}$:

$$\mathcal{D}_{\tau(j)} = \left\{\left(\left[x_i, \mathbf{pre}^i_{\tau(j)}\right], \phi\left(\Upsilon_i, y_{\tau(j)}\right)\right) \mid 1 \le i \le m\right\} \tag{4}$$

$$\text{Where } pre^i_{\tau(j)} = \left(\phi\left(\Upsilon_i, y_{\tau(1)}\right), \ldots., \phi\left(\Upsilon_i, y_{\tau(j-1)}\right)\right)^T$$

Here $\left[x_i, \mathbf{pre}^i_{\tau(j)}\right]$ concatenates vectors x_i and $\mathbf{pre}^i_{\tau(j)}$, and $\mathbf{pre}^i_{\tau(j)}$represents the binary assignment of those labels preceding $y_{\tau(j)}$on x_i $\left(\text{specifically, } \mathbf{pre}^i_{\tau(1)} = \emptyset\right)$. Then, a suitable binary classifier $\mathcal{B}$ is used to induce a binary classifier $g_{\tau(j)}: \chi \times \{-1, +1\}^{j-1} \to \mathbb{R}$, $i.e.\ g_{\tau(j)} \leftarrow$

$\mathcal{B}(\mathcal{D}_{\tau(j)})$. Or, $g_{\tau(j)}(.)$ controls whether $y_{\tau(j)}$ is a relevant label or not. For new instance x, its associated label set Υ is predicted by traversing the classifier chain iteratively.

4.1.3. Multi-label k-Nearest Neighbors (ML-kNN)

Multi-Label k-Nearest Neighbor (ML-kNN, derived from the k-nearest neighbor (kNN)) first identifies the k nearest neighbors of the test instance where the label sets of its neighboring instances are obtained. After that, maximum a posteriori (MAP) principle is employed to predict the set of labels of the test instance (*64*). This algorithm also known by the name Lazy Learning Algorithm (*45*), (*64*).

Let X be the domain of instances. Let Y be the finite set of Labels, Y= {1, 2, 3……., Q}. Let T be the training set, $T=\{(x_1, Y_1),(x_2 ,Y_2),\ldots,(x_m,I)\}$, $(I \in X, Y_i \subseteq Y)$. T is derived independently and identically from an unknown distribution.

Here, the objective of the MLC algorithm is to either find the classifier, represented by h that maps the given set of instances to label set, $h: X \rightarrow 2^Y$, or computes a real-valued function which is represented as $f: X \times Y \rightarrow R$. The algorithm output larger values for label in Y_I, or $f(x_i, y_1) > f(x_i, y_2)$ or $y_1 \in Y_i$ and $y_1 \notin Y_i$. The real-valued function is then transformed into a ranking function, such that instances with highest f value will have the lowest rank. This transformation into a ranking function can be used to derive the respective classifier h(x), for the instance x.

For the instance x, y is the vector containing the labels of x, and the l^{th} label is given by y (l), where $l \in Y$. $y(l) = 1$ when $l \in Y$ and 0 otherwise. N (x) represents the index set of the k nearest neighbor of the instance x, from the training set T. From the subset of labels from the indexed set N(x), a counting vector is defined using the following equation, which counts the number of neighbors of x belonging to l^{th} class.

$$C_x(l) = \sum_{a \in N(x)} y_{x_a}(l), l \in Y \tag{6}$$

Given the instance t from test set, the learning algorithm identify the assigned k number of neighbors, represented by N(t) from the training set T.

Let H_1^l be the event that instance t has label and H_0^l be the event that instance t does not have label. E_j^l is the event that among the k nearest neighbors of instance t, there are j instances with label l. The label set $[y_t(l)]$, of the test instances are determined using MAP- Maximum a posteriori principle using the counting vector Ct ,which is given by

(7)

The final classification label set is derived from this MAP principle by the Bayesian principle

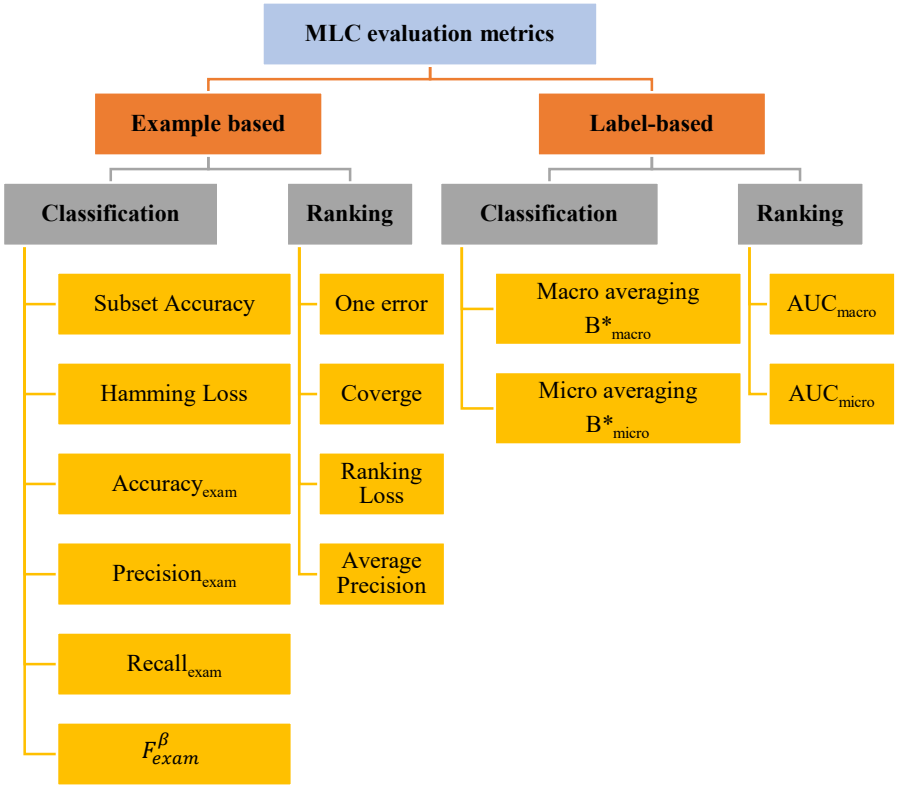
$$y_t(l) = \arg\max_{b \in \{0,1\}} \frac{P(H_b^l) P(E^l_{C_t(l)} \mid H_b^l)}{P(E^l_{C_t(l)})} \tag{8}$$

$$= \arg\max_{b \in \{0,1\}} P(H_b^l) P(E^l_{C_t(l)} \mid H_b^l) \tag{9}$$

Where $P(H_b^l)$ is the prior probability and $P(E^l_{C_t(l)} \mid H_b^l)$ is the posterior probability, and both estimated from Training set.

4.1.4. Evaluation Metrics

In traditional supervised learning, generalization performance of the learning system is evaluated with conventional metrics such as accuracy, F-measure, area under the ROC curve (AUC), etc. However, performance evaluation in multi-label learning is much complicated than traditional single-label setting, as each example can be associated with multiple labels simultaneously. Therefore, several evaluation metrics specific to multi-label learning are proposed, which can be generally categorized into two groups, i.e., example-based metrics and label-based metrics. Multi-label metrics are usually non-convex and discontinuous, in practice most learning algorithms resort to optimizing (convex) surrogate multi-label metrics. The subset accuracy evaluates the fraction of correctly classified examples, i.e., the predicted label set is identical to the ground-truth label set. Intuitively, subset accuracy can be regarded as a multi-label counterpart of the traditional accuracy metric and tends to be overly strict especially when the size of label space (i.e., q) is large.



B can be Accuracy, Precision, Recall,* $\boldsymbol{F_{\beta}}$

Figure 5. MLC Performance Evaluation

The suitability of performance metrics depends on the classification problem itself. For this study, the example-based classification performance evaluation matrices (shown in Figure **5**) were used to compare and benchmark the performance the proposed MLC classification models. A summary about these matrices is presented Table 4. Each of these metrics evaluates the learning system's performance on each test example separately, and then returning the mean value across the test set.

Table 4. Example-based Performance Metrics

Metric	Description	Equation
Subset Accuracy	evaluates the fraction of correctly classified examples, i.e., the predicted label set is identical to the ground-truth label set. Subset accuracy can be regarded as a multi-label counterpart of the traditional accuracy metric and tends to be overly strict especially when the size of label space (i.e., q) is large.	Subset Accuracy: $\frac{1}{p}\sum_{i=1}^{p}[\![h(x_i)=Y_i]\!]$
Hamming Loss	evaluates the fraction of misclassified instance-label pairs, i.e., a relevant label is missed or an irrelevant is predicted	Hamming Loss: $\frac{1}{p}\sum_{i=1}^{p}\frac{1}{q}\lvert h(x_i)\Delta Y_i\rvert$
Accuracy	evaluates the fraction of correct predictions to total predictions	Accuracy= $\frac{1}{p}\sum_{i=1}^{p}\frac{\lvert Y_i\cap h(x_i)\rvert}{\lvert Y_i\cup h(x_i)\rvert}$
Precision	evaluates the fraction of true positives to sum of true positives and false positives	Precision= $\frac{1}{p}\sum_{i=1}^{p}\frac{\lvert Y_i\cap h(x_i)\rvert}{\lvert h(x_i)\rvert}$
Recall	evaluates the fraction of true positives to sum of true positives and false negatives	Recall = $\frac{1}{p}\sum_{i=1}^{p}\frac{\lvert Y_i\cap h(x_i)\rvert}{\lvert Y_i\rvert}$
F^{β}	an integrated version of Precision and Recall with balancing Factor $\beta > 0$. B = 1 leads to the harmonic mean of precision and recall	$F_{\beta}=\frac{(1+\beta^2).\text{Precision}.\text{Recall}}{\beta^2.\text{Precision}+\text{Recall}}$

4.2. Spatial Transferability

The performance of any AI/ML model is defined by the data. The quality and quantity of the training data dictate the classification accuracy of any such algorithms. No prior assumption about the distribution of data makes it easier to make accurate generalized inferences without breaking any data-specific rules. The clustering analysis was aimed to conduct a data driven Principal (PCA) based cluster analysis using Texas counties data to identify natural group based on similarity in characteristics using Hierarchical clustering tools. The clustering framework is illustrated in Figure 6. The use of PCA based analysis is critical for macroscopic crash prediction modelling, as the analysis generally deals with small data size and large crash features, which has not been explored much in the past. Despite, being a popular algorithm for unsupervised learning for data exploration/market segment analysis etc., hardly few researchers have used this tool for crash data exploration. Apart from finding entities sharing similar attributes, clustering tools can help in separating normal data from outliers or anomalies. Thus, the main objective of this work is to treat the county level data as unsupervised learning problem to discover natural groups of similar examples or clusters within the data, or to determine how the data is distributed in the space, known as density estimation. Anticipating possible delays that may arise due to computational soundlessness, that could affect progress of the modelling analysis, we adapted the clustering approach to test/ or validate the spatial transferability of the calibrated/validated predictive models to other main regions of Texas.

Clusters or groups that share common characteristics, play an important role in how we analyze and describe the environment or system. Dendrogram produced from clustering process is extremely useful in understanding the data. By observing the branches of the hierarchical dendrogram structure, insightful information about those county group that varies significantly with respect to explanatory features and crash types and those which does not have any effect on the same can be understood. Given a set of *N* items to be clustered, and an *N*N* distance (or similarity) matrix, the basic process of hierarchical clustering consists of following steps (*65*):

1. Assign each item to its own cluster, so that if you have N items, you now have N clusters, each containing just one item. Let the distances (similarities) between the clusters equal the distances (similarities) between the items they contain.
2. Find the closest (most similar) pair of clusters and merge them into a single cluster, so that now you have one less cluster.
3. Compute distances (similarities) between the new cluster and each of the old clusters.
4. Repeat steps 2 and 3 until all items are clustered into a single cluster of size N.

In Steps 2 and 3, the algorithm deals with finding distances, which represents the similarity/dissimilarities, between cluster pairs. So, prior to clustering, it is required to determine the distance matrix that specifies the distance between each data point using some distance function. The main task of clustering/classification of observations into groups requires the computation of the distance or the (dis)similarity between each pair of observations. The result of this computation is known as a dissimilarity or distance matrix. The classical methods for distance measures are Euclidean and Manhattan distances. This study uses Euclidean distance for the generation of distance matrix. The formula for computing the Euclidean Distance (eqn. 1) is shown below:

$$\mathbf{d_{euc}}\,(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{1}$$

The AGNES clustering algorithm has multiple clustering/linkage criteria for grouping data. The linkage function takes the distance information from the calculated distance matrix and groups pairs of objects into clusters based on their similarity. The new clusters are linked to each other to create bigger clusters. This process is iterated until all the objects in the original data set are linked together in a hierarchical tree. Ward's minimum variance method is one of the most common linkage criteria used for agglomerated hierarchical clustering (see eqn. (2) and (3)). It minimizes the total within-cluster variance. Ward's method approach also performs well in separating clusters if there is noise between cluster.

$$\Delta\,(A, B) = \sum_{i \in A \cup B} \|\vec{x}_i - \vec{I}\|^2 - \sum_{i \in A} \|\vec{x}_i - \vec{m}_A\|^2 - \sum_{i \in B} \|\vec{x}_i - \vec{m}_B\|^2 \tag{2}$$

$$\text{or, } \Delta\,(A, B) = \frac{n_A\, n_B}{n_A - n_B} \|\vec{m}_A - \vec{m}_B\|^2 \tag{3}$$

where m_j is the center of cluster j, and n_j is the number of points in it. Δ is called the merging cost of combining the clusters A and B.

The main idea behind this approach is that by treating the safety data as unlabeled hierarchical clustering creates meaningful hierarchy of county groups that share common characteristics with respect to the features considered in this study. The study also utilized Principal Component Analysis to tackle the issue of overfitting due to high dimensionality by taking principal components that explains maximum variance from each feature group. One major advantage of using hierarchical clustering is its ability to provide visualization of result. The final clusters can be represented in the form of dendrograms which can help in the interpretation of the results by creating meaningful taxonomies and in post modelling results comparisons.

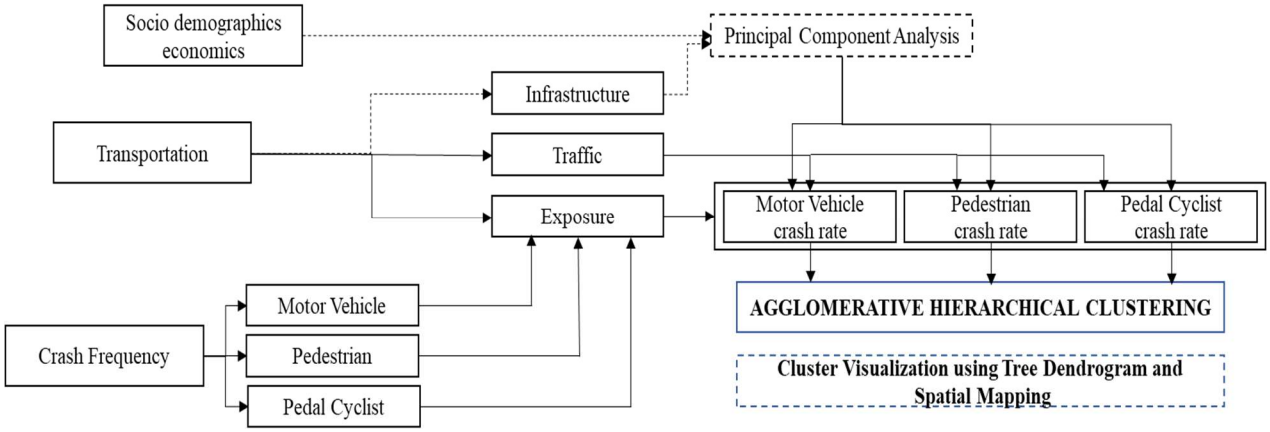


Figure 6. County-Level PCA Based Clustering Framework Approach

4.3. Study Area and Period

The study area chosen for the testing of the research framework comprises of 5 main cities. Precisely, the model calibration and validation were conducted leveraging the historic crashes data occurring in the state of Texas from 2013 to 2017. The state of Texas is in the South-Central region of the United States. At over 268,500 square miles in area, and with more than 29.1 million residents in 2020, Texas is the second largest U.S. state by area (after Alaska) and population (after California). With a population of 4,652,980, Harris County (*29.7752° N, 95.3103° W*) is the largest county in Texas, and had a recent growth rate of 13.7%. Dallas County (*32.8025° N, 96.8351° W*), Tarrant County (*32.7732° N, 97.3517° W*), Bexar County (*29.4201° N, 98.5721° W*) and Travis County (*30.2097° N, 97.6982° W*) make up the rest of the top five most populous counties in Texas, with each having populations of more than a million. Additionally, all show growth rates of between 10.6% (Dallas County) and 19.7% (Travis County), (see Figure 7) . This research targeted the top biggest cities of the state that includes, Austin, Dallas, Houston, Fort Worth, and San Antonio. As the geographic boundaries for counties are more continuous and linear compared to City (complex boundaries), the analysis boundaries for the historic crashes occurring in selected cities are carries out leveraging the county boundaries within which the respective cities belong (*66*). The spatial distribution of exposure and crash rate features are presented in Figure **8**.

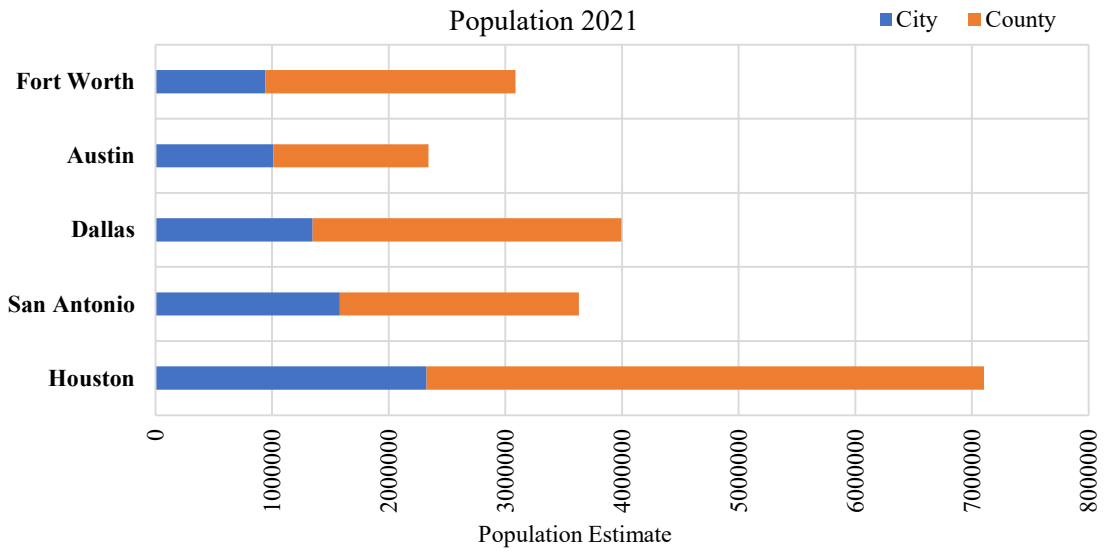
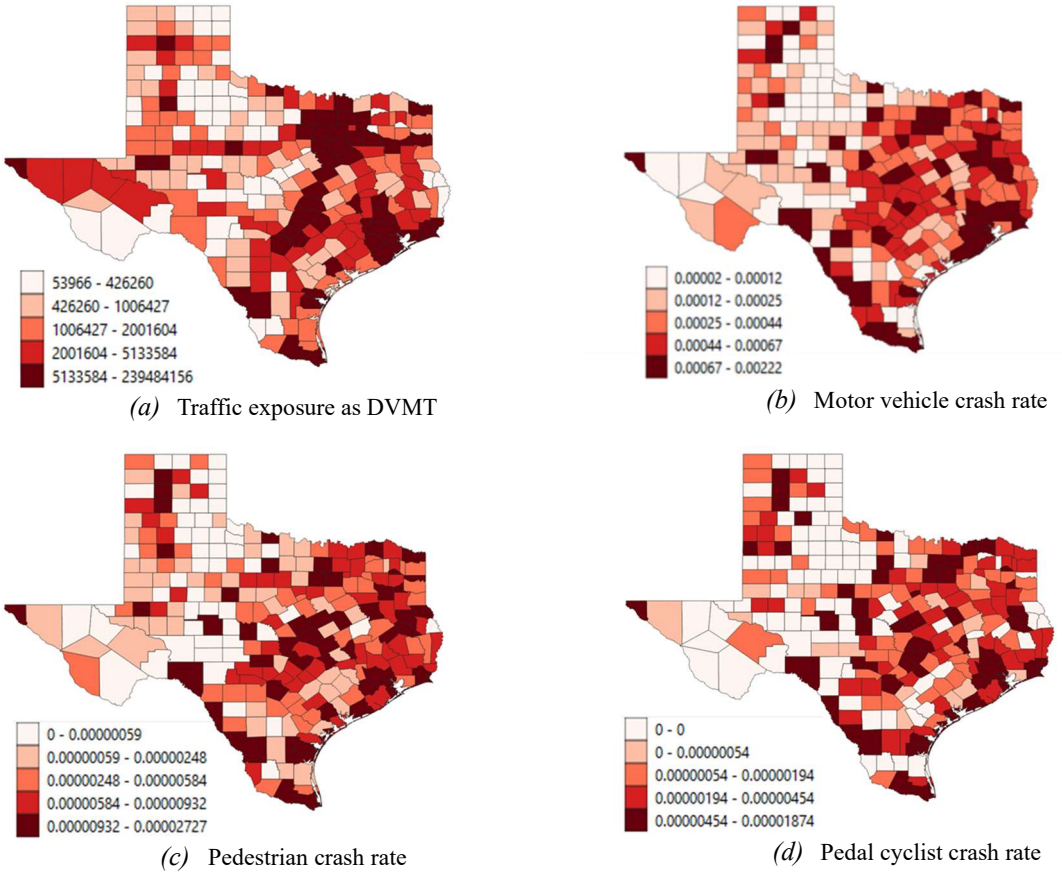


Figure 7. Population Estimate



(a) Traffic exposure as DVMT

(b) Motor vehicle crash rate

(c) Pedestrian crash rate

(d) Pedal cyclist crash rate

Figure 8. Traffic exposure and average crash rates for the Texas counties (a): Traffic Exposure, (b): Motor Vehicle Crash Rate, (c): Pedestrian Crash Rate, (d): Pedal Cyclist Crash Rate

Figure 9. Primary Study Area

4.4. Data and Data Source

Statewide motor vehicle traffic crash data provides the basic information necessary for effective highway and traffic safety efforts at any level of government: local, state, or federal. The Texas Department of Transportation (TxDOT) is the custodian of crash records for the State of Texas. Texas Transportation Code §550.062 requires any law enforcement officer who in the regular course of duty investigates a motor vehicle crash that results in injury to or the death of a person or damage to the property of any one person to the apparent extent of $1,000 or more, to submit a written report of that crash to TxDOT not later than the 10th day after the date of the crash (*67*, *68*). TxDOT collects crash reports from Texas law enforcement agencies for crashes occurring on public roadways and the state highway system. The state retention schedule for crash reports and data is 10 years plus the current year. Data for years beyond this period is unavailable. TxDOT provides guidance manual to guide and instruct peace officers in completing the Texas Peace Officer's Crash Report and the Commercial Motor Vehicle Section of the Texas Peace Officer's Crash Report as required by Section 550.063 of the Texas Transportation Code (*69*, *70*). State statutes and city ordinances govern reporting and investigation requirements. Statewide motor vehicle traffic crash data provides the basic information necessary for effective highway and traffic safety efforts at any level of government: local, state, or federal. State crash data is used to perform problem identification, establish goals and performance measures, allocate resources, determine the progress of specific programs, and support the development and evaluation of highway and vehicle safety countermeasures. Better data will lead to safer roadways. Hence, high quality data is a decisive element for effective identification of traffic safety glitches, communicate safety issues to the public and media and make better programming and resource allocation decisions. TxDOT maintains an open data portal named TxDOT Open Data Portal (*71*). This data portal provides access to numerous periodically updated GIS based transportation data that can be

explored and downloaded and has been used in this study. The TxDOT Roadway Inventory layer (*72*) is a statewide dataset that has attribute information routed to TxDOT Roadway linework. By using linear referencing tools, attribute information from the TxDOT Roadway Inventory table is located on the linework. Roadway attributes such as functional system, traffic counts, surface types among many others can be found on a roadway simply by selecting it or performing a query. The database is updated frequently by the Transportation Planning and Programming Division at TxDOT in the Data Analysis, Mapping and Reporting Branch for internal and public use. Features and information, about the transportation infrastructure and traffic relevant for the respective study regions were extracted from this data repository. The link level relevant attributes for the entire transportation link of Bexar County were intersected and extracted from the TxDOT Open Data Portal (*73*).

The database management framework was a crucial part of this research as prediction models are based on machine learning algorithms, some of them are computationally expensive. The data filters and features selection are organized in such a way that, there is ample computational memory to carry out the analysis without any interruptions. This optimized dataflow is also key for the implementation phase of this project. Figure **10** represents the database management framework used for this study. The current framework has potential for further optimization and improvements to the total modelling framework.

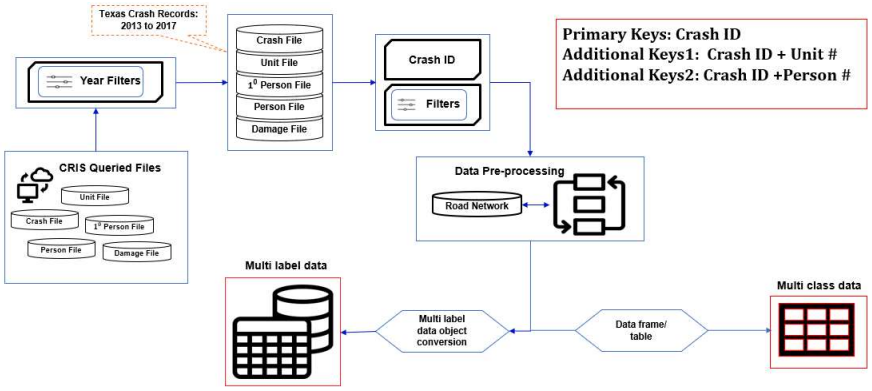


Figure 10. Database Management Framework

4.5. Model Specification

Bexar County, which encompasses the City of San Antonio, was selected as the primary testing study area. The proposed multilabel classification model for the categorization problem of crash collision and injury severity type, was calibrated using the data extracted for the primary study area for the analysis period. The spatial transferability of the proposed model framework was carried out for the rest four-study location. As crashes are relatively rare events, it is essential that a safety analysis includes an adequate time frame of study (*74*) to capture the temporal variations. The crash data for five years (2013-2017) period for all the study locations, were extracted from the crash database maintained the Texas Department of Transportation (TxDOT), called the Crash Records Information Systems (CRIS) (*67*). To summarize the modelling data information, all the crashes occurred in Urban Principal Arterial roads between the 2013 and 2017 was used for the training and testing of the classification models. The description of the crash injury severity, the collision type and the respective risk factors or features is described in the following section.

Traffic crash results from a combination of factors related to the components of the system comprising roads, the environment, vehicles and road users, and the way they interact. Some factors contribute to the occurrence of a collision and are therefore part of crash causation. Other factors aggravate the effects of the collision and thus contribute to trauma severity. Some factors may not appear to be directly related to road traffic injuries. Some causes are immediate, but they may be underpinned by medium-term and long-term structural causes. Crash risk factors or the model features that could impact the target (Collision /Injury type), is presented Table 6 and Table 5. Other features respective to the person/persons of interest or the units (vehicle involved in the car) of interest has also been considered for training the classification model. As more features creates additional data issues like overfitting, as well as increases the computational requirement standards to execute the classification algorithm.

Table 5. Categorical Risk Factors

Nominal Data	Description	Categories
Traffic Control	Type of traffic control at the scene of the crash	21
Weather Condition	The prevailing atmospheric condition reported by the officer at the time of the crash	12
Light Condition	The type and level of light that existed at the time of the crash	7
Surface Condition	The surface condition (wet, dry, etc) present at the time and place of the crash	10
Road Part	The part of the roadway on which the vehicle(s) was traveling prior to the crash	6
Entering road	Crash occurred at Entering Roads	2
Road Alignment	The geometric characteristics of the roadway at the crash site	7
Construction zone	Whether crash occurred at construction site	2
Active construction zone	Whether crash occurred at construction site with workers	2
Intersection related	Specifies whether a crash occurred at an intersection, not at an intersection, or if the presence of an intersection contributed to the crash	2
Gender	Gender of the person of interest (Most severe injury/Driver 1 injury/Driver 2 injury)	2

Table 6. Numeric Risk Factors

Feature Name	Mean	Standard Error	Median	Mode	Standard Deviation	Kurtosis	Skewness	Range	Minimum	Maximum
Crash Speed limit (mph)	52.283	0.107	55	45	15.190	3.142	-1.519	76	0	75
Number of Lanes	5.080	0.012	4	4	1.642	1.549	1.438	9	2	11
Median width plus both inside shoulders (feet)	45.492	0.172	48	48	24.347	-1.051	0.163	86	0	86
Right of way Width (feet)	304.396	0.352	300	300	49.821	2.644	-0.052	350	150	500
Roadbed Width (feet)	93.493	0.165	82	76	23.362	0.315	0.969	138	36	174
Surface Width (feet)	63.742	0.148	48	48	20.966	0.638	1.177	108	24	132
Left Shoulder Width (feet)	11.092	0.034	8	8	4.839	0.487	0.986	34	0	34
Right Shoulder Width (feet)	18.545	0.035	20	20	4.998	3.120	-0.249	36	0	36
Median Width (feet)	34.533	0.176	40	40	24.967	-1.182	0.094	77	0	77
Adjusted AADT	77633.844	277.342	70780	73838	39272.066	-0.568	0.531	159201	4279	163480
% Single Truck AADT	2.461	0.007	1.9	1.8	0.958	5.397	1.868	7.5	0.6	8.1
% Combination Truck AADT	3.457	0.026	1.5	0.6	3.637	-0.882	0.901	10.3	0.1	10.4
% Truck AADT	5.928	0.029	4	2.3	4.146	-0.669	0.874	13.8	0.7	14.5

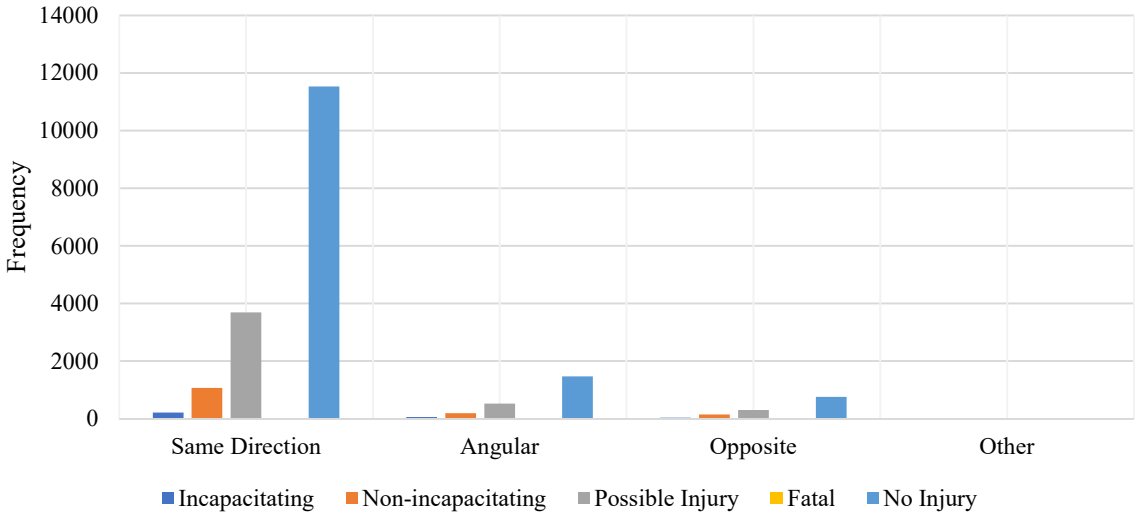


Figure 11. Target Distribution

Figure 11 shows the distribution of the model target features, crash collision type and injury severity type. The imbalance nature of the safety data can be seen from the histogram shown below. Relatively high percentage of the crashes are same direction collisions and high percentage of non-injury crashes. To account for the imbalance, the classification model or crash feature can be trained with respect to modified target types by transforming the problem from multi-class to binary, whether injured or not. This way, the models can be trained without leveraging data imputation that could create external bias to the prediction models. **Figure 12** illustrates this target transformation. The overall scope of this project is described in Table 7 and Figure 13. The detailed list of the classification models with the respective target is presented in Table 7. Figure 13 shows the various algorithms used for each of the model discussed in the Table 7.

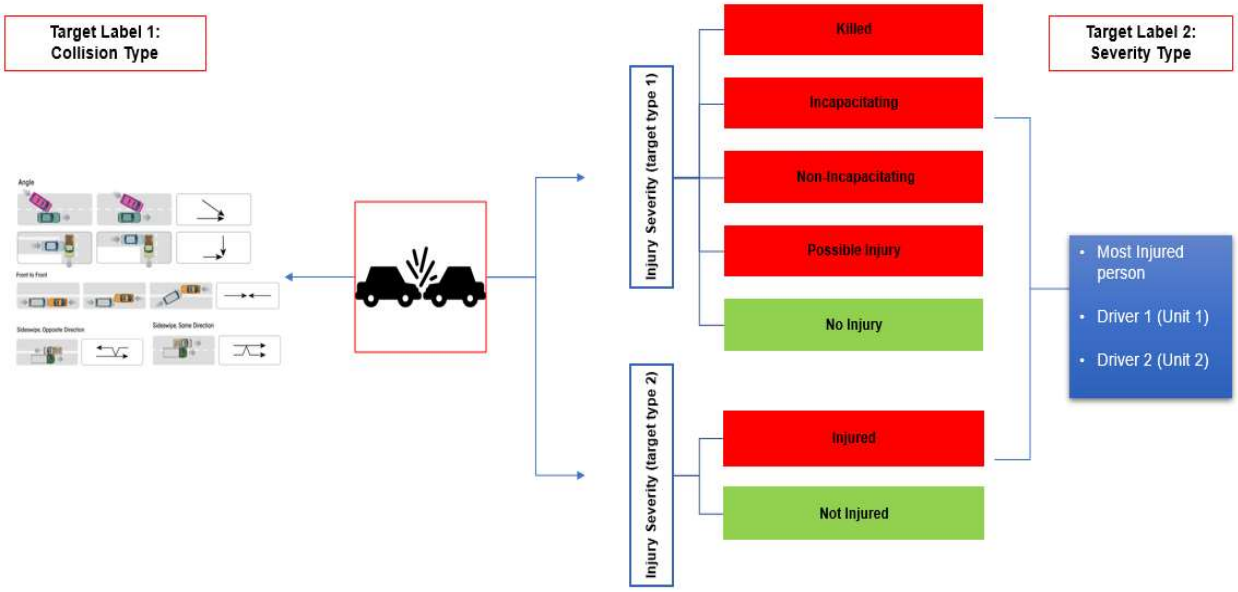


Figure 12. Target Modifications

Table 7. Model Specifications

	Sl.	Model Name	Target	Severity type
Conventional Model	1	C1	Collision type	NA
	2	C2	Most Severe Injury	Type I
	3	C3	Most Severe Injury	Type II
	4	C4	Driver Injury (unit 1)	Type I
	5	C5	Driver Injury (unit 1)	Type II
	6	C6	Driver Injury (unit 2)	Type I
	7	C7	Driver Injury (unit 2)	Type II

	Sl.	Model name	Target 1	Target 2	Target 3	Severity type
Proposed MLC Models	1	P1	Collision type	Most Severe Injury	NA	Type I
	2	P2	Collision type	Most Severe Injury	NA	Type II
	3	P3	Collision type	Driver Injury (unit 1)	Driver Injury (unit 2)	Type I
	4	P4	Collision type	Driver Injury (unit 1)	Driver Injury (unit 2)	Type II

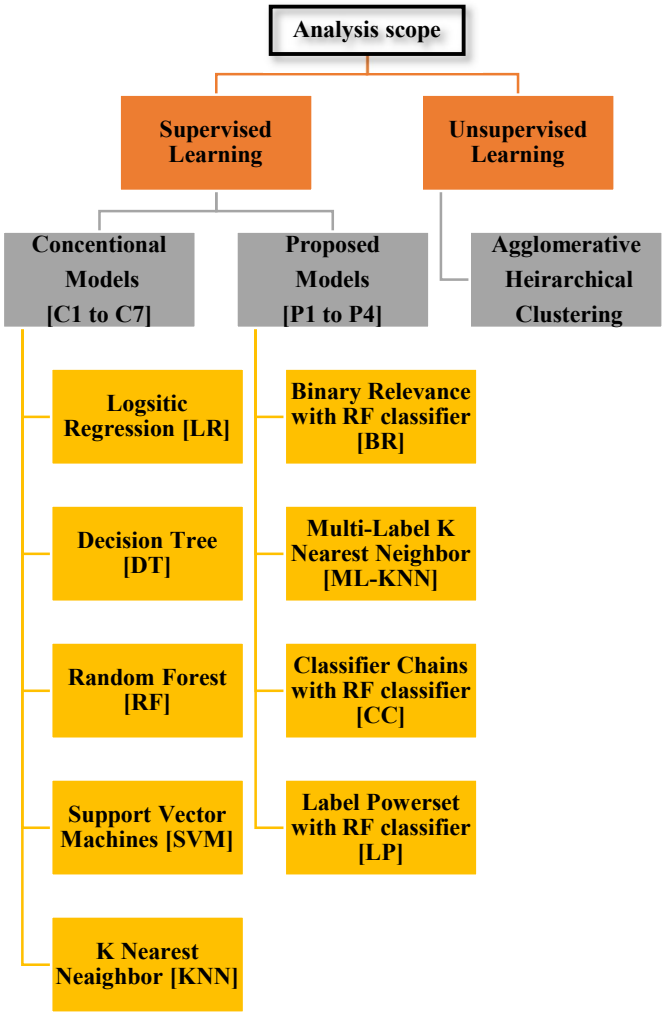


Figure 13 Overview of Tested Models

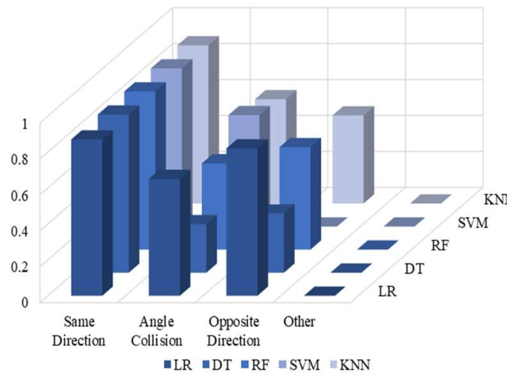
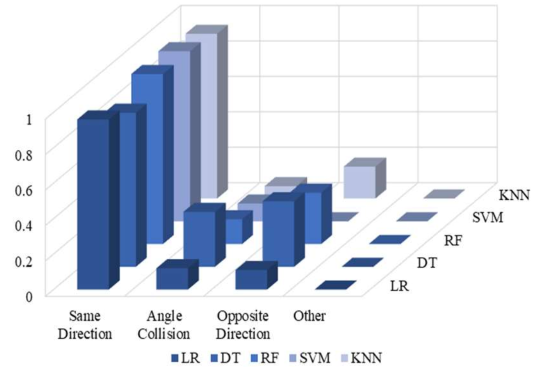
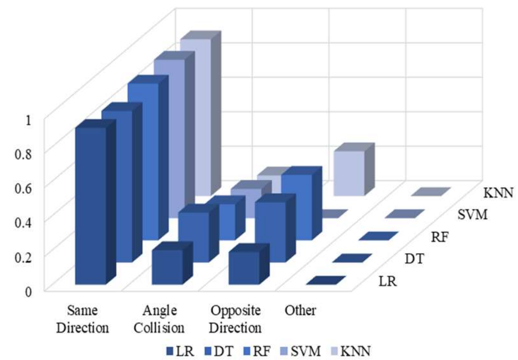
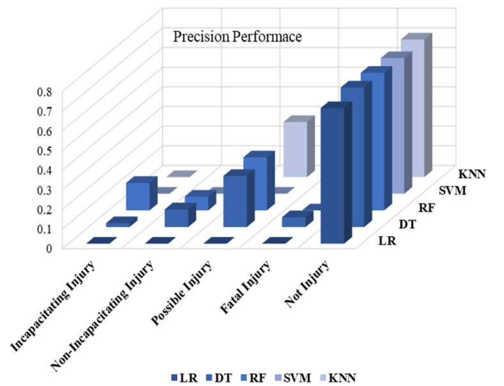
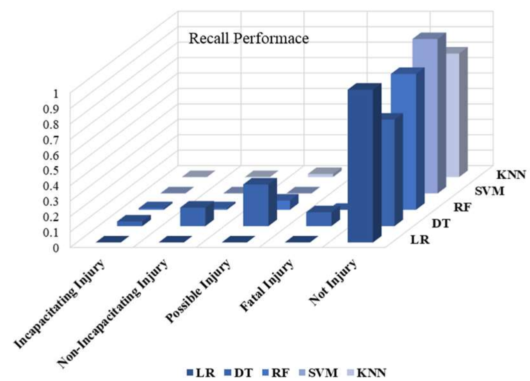
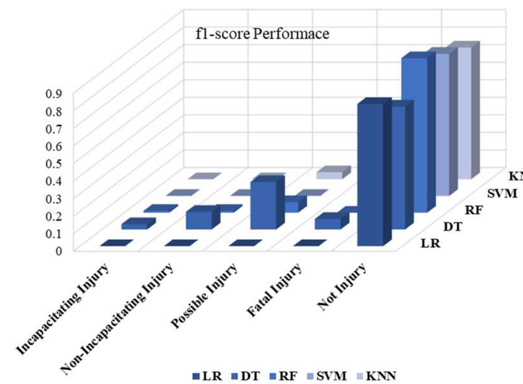
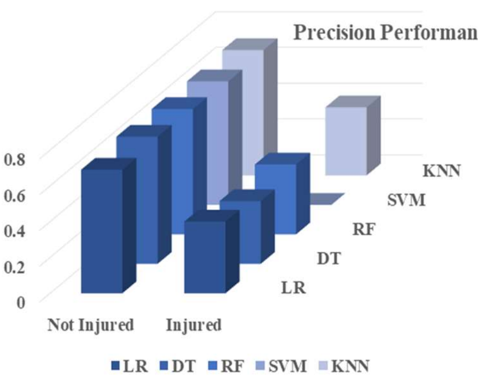
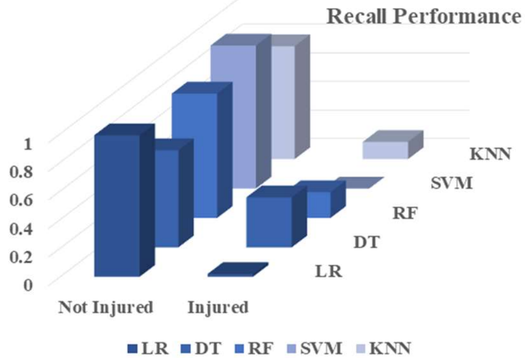
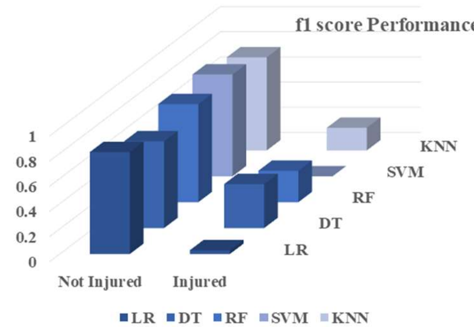
5. ANALYSIS AND FINDINGS

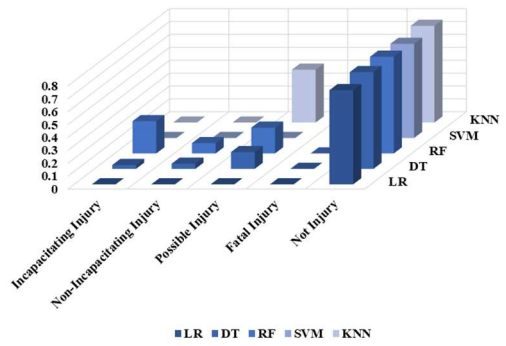
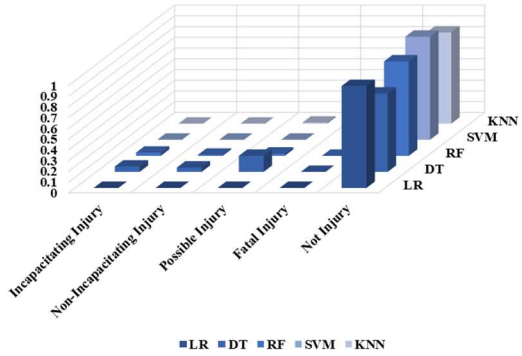
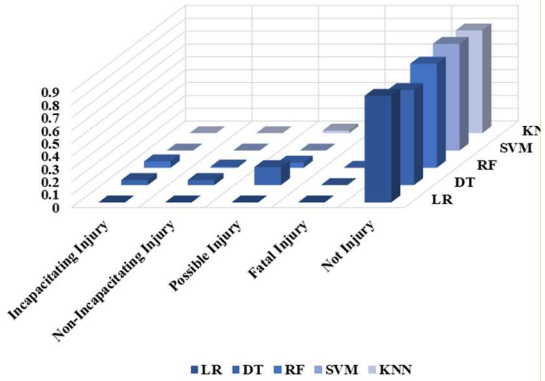
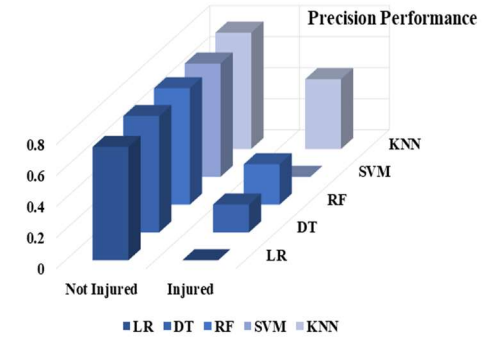
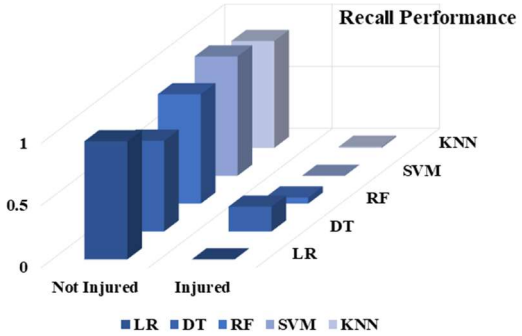
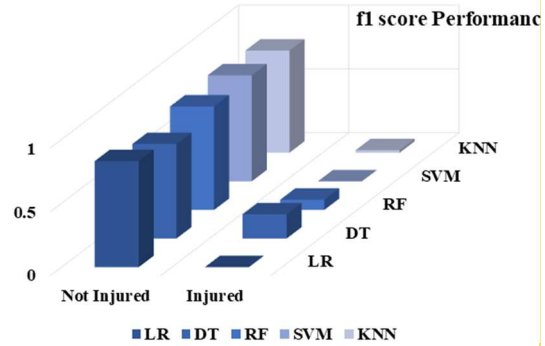
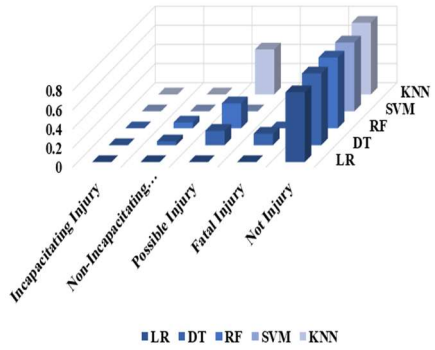
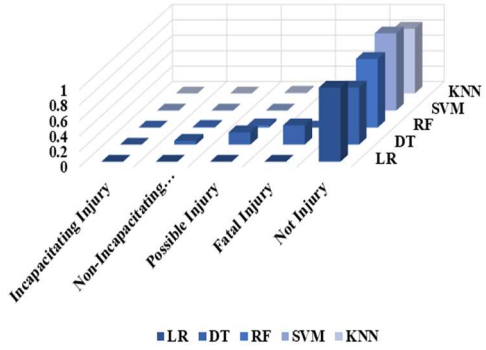
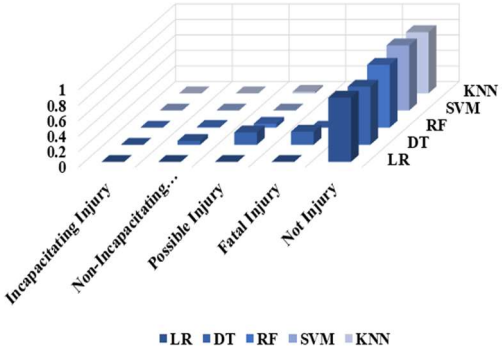
5.1. Classification Model Performance

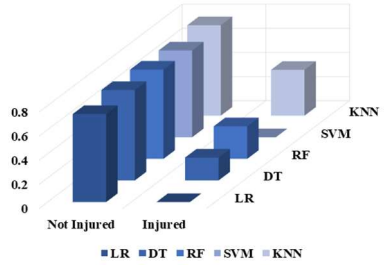
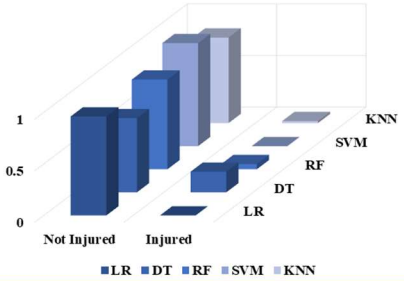
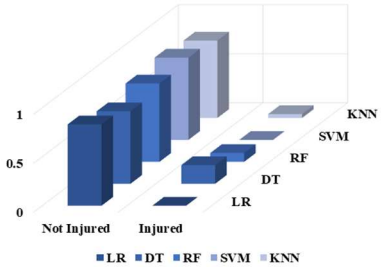
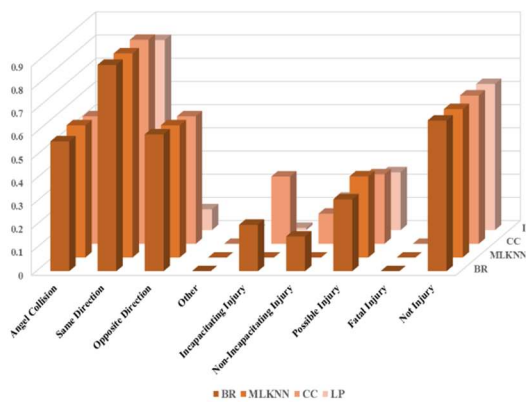
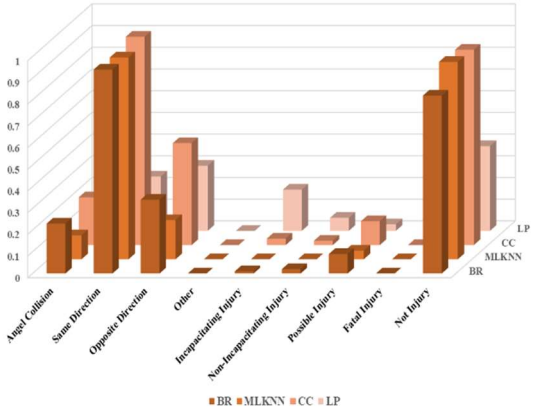
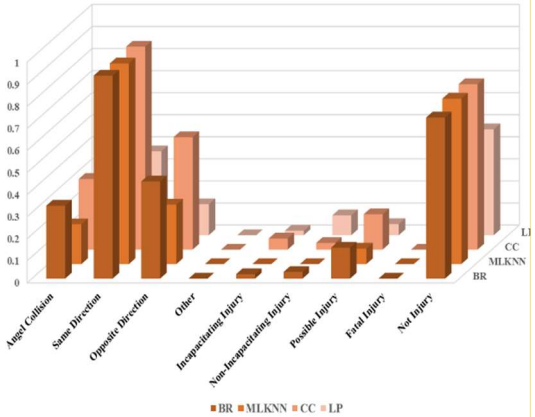
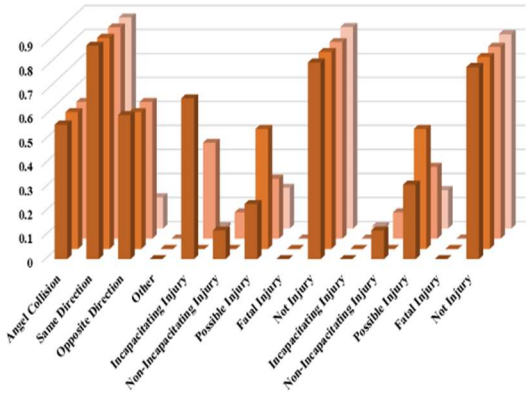
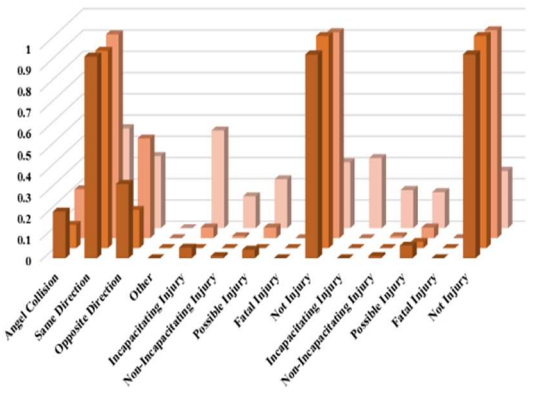
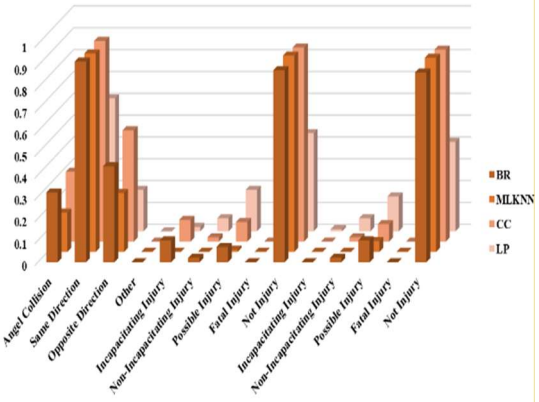
The core objective of this project is to explore the machine learning MLC tool for classification problems in the context of traffic safety simultaneous classification of collision type and crash injury type. The ability of MLC to categorize an entity under study to more than one labels simultaneously provides it an edge over the traditional classification approaches that classify collision type and crash injury types separately. Underlying correlation between the injury severity type and collision type is leveraged using the AI tool to develop a robust classification model, and the performance of the proposed tool will be benchmarked with the conventional tools [see models in Figure 13]. The following section discusses the results and inferences from the classification analysis, which is followed by the discussion of the results and inference from the clustering studies.

The proposed and conventional classification model experiment for the traffic crash collision and severity type is formulated using holdout, in which 70% of the dataset instances are used for training and 30% for test. As there are no limit on the data to train the model, removing a part of it for validation will affect the model capability up-to great extent and even poses problems such as under-fitting. K-Fold Cross Validation is a suitable method that provides ample data for training the model and leaves ample data for validation. The stability of the performance of the conventional model has been validated using repeated k-fold cross validation using with number of times cross-validator needs to be repeated is set to 10 using 5 folds. Results from the preliminary analysis has been presented as charts (shown in Figure 14**).** To be precise, the prediction performances of the conventional injury severity type classification models and the proposed multi-label simultaneous classification of collision type and injury severity type was compared and benchmarked preliminarily for label-label precision, recall and f1-score values.

The numerical results indicate that the performance of the proposed approach was comparable or even better than the conventional models (see Figure 14). The conventional models [C1 to C7] is highly impacted by the imbalance nature of the crash data. This is shown by the extreme distributions (very high-very low values) of the bars. The high performances of LR models and SVMs models are caused due to overfitting, as these model's performances are highly skewed by this imbalance, caused by the high crash rate of the "no injury" crashes. Removing the imbalance label or imputing under-represented labels can add external bias to the already complicated classification problem. This imbalance nature of the crash data can be addressed without adding bias through the multi-label classification approach as discussed in the methodology section. This is clearly visible from the distributed bars with typical variations, (shown in Figure 14 [models P1 to P4]). In other words, the decision, or the classification boundary between various injury severity type in the crash features dataspace is more separable when collision is also considered simultaneously. Though not documented, the proposed classification models outperformed the conventional models in terms of both prediction performance and computational time. This superior prediction performances of the proposed approach needs further validation though further training and testing.

Model	*Precision*	*Recall*	*F1-Score*
C1			
C2			
C3			

Model	*Precision*	*Recall*	*F1-Score*
C4			
C5			
C6			

Model	*Precision*	*Recall*	*F1-Score*
C7			
P1			
P2			

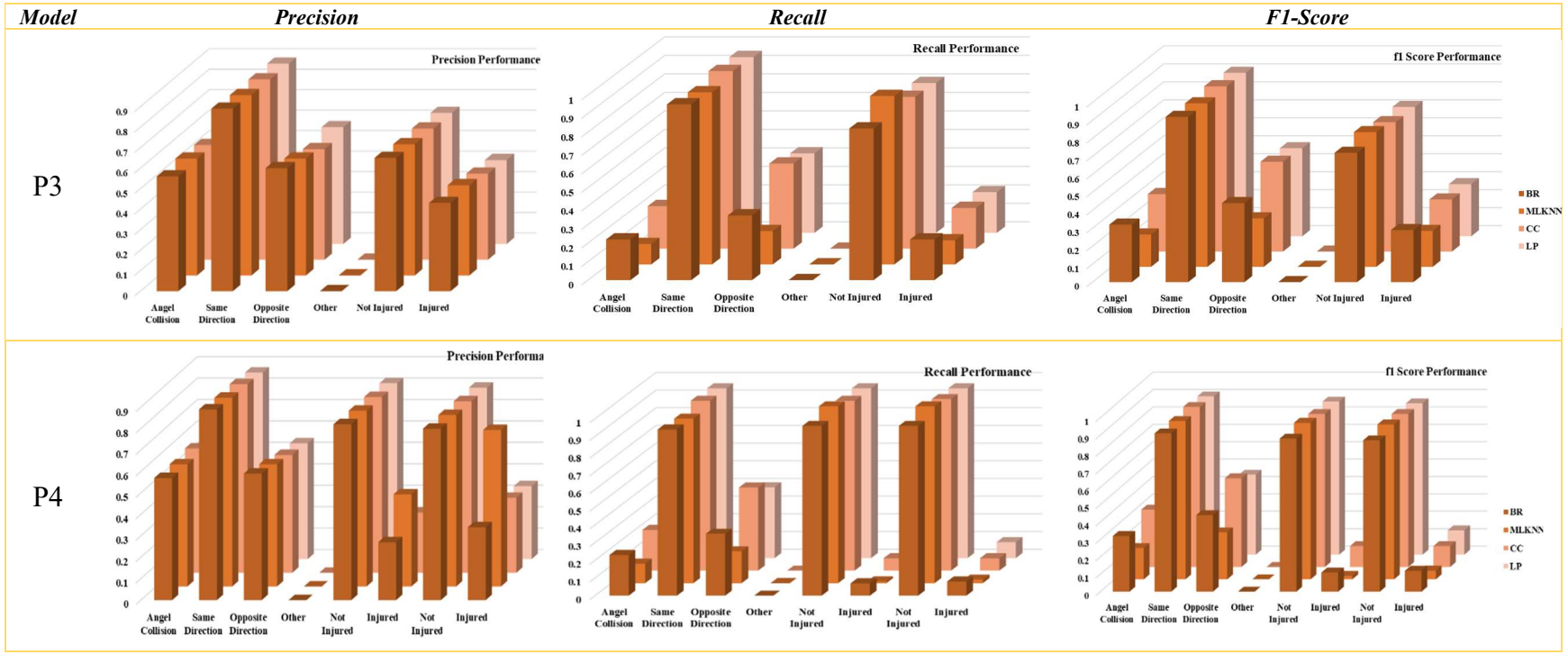


Figure 14 Prediction Performances

5.2. Agglomerative Hierarchical Clustering Results

The main advantage of Hierarchical clustering, as mentioned in the previous section, is its ability to provide representation/visualization of results using dendrogram. The 5 outputs from the principal component analysis along with traffic variables were subjected to agglomerative hierarchical clustering separately with motor vehicle crash rate, pedestrian crash rate and pedal cyclist crash rate and the computational results are summarized in Table **8**. Results from the cluster metrics shows that AGNES with 2 clusters has maximum cluster strength, i.e., the optimum number of clusters is 2.

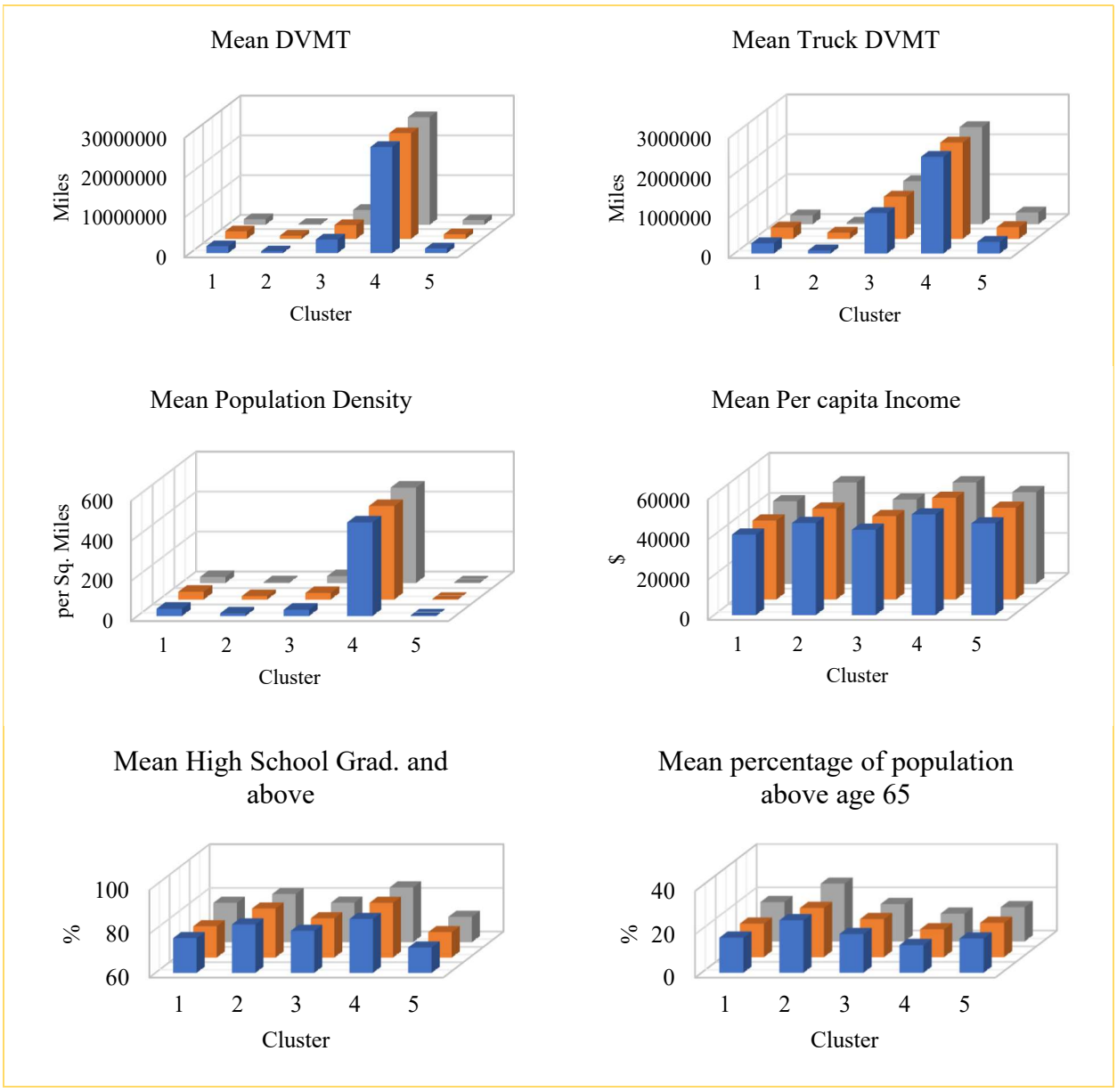
Table 8. Cluster Stability

Clusters	Motor Vehicle Crashes		Pedestrian Crashes		Pedal cyclist crashes	
	Dunn	Silhouette	Dunn	Silhouette	Dunn	Silhouette
2	0.5311	0.6377	0.5244	0.6369	0.5374	0.6369
3	0.2042	0.4146	0.2195	0.4132	0.2065	0.4104
4	0.2133	0.3809	0.2195	0.3749	0.1985	0.3947
5	0.2133	0.3652	0.2195	0.3627	0.2097	0.3727
6	0.2133	0.3554	0.1442	0.3427	0.2139	0.366

To visualize their similarities at a higher resolution, the resulting dendrogram tree was cut at the required height, which creates a hierarchy of 5 branches. Members within same branch are more similar than in different branches. Dendrogram plots from the cluster analysis results were used to identify 5 county-groups that are similar in the feature space. The resulting dendrograms was cut at a height to create 5 branches or cluster group and cluster memberships were analysed. This way, it is also possible to understand the changes in the cluster membership under each branch as we move down the dendrogram (appendix B). The number of counties in each cluster for the motor vehicle and pedestrian crash clustering is almost identical, whereas the same varies for the pedal cycle crash clusters. Specifically, cluster 1 and cluster 2 size of pedal cyclist cluster with respect to motor vehicle and pedal cyclist crash cluster shows significant variation. The cluster size and membership for Cluster 1, Cluster 2 and Cluster 3 changes for different crash types. With respects to the all counties, Cluster 5 and Cluster 4 counties are on the extreme side of the data for all crash types considerd. Counties that has significantly high values of features, as relative to other counties continue to be in the same cluster groups (Cluster 4) for all types of crashes. These counties contains few of the fast growing and popular cities in the state of texas, like Dallas, San Antonio, Austin, and Houston. The average Daily Vehicle Miles Travelled (DMVT) is relatively higher for cluster 4 compared to other groups. Similarily, counties with feature values on the lower side always cluster together (e.g. Cluster 1). Counties like Kennedy, Sterling, Brocks etc. belongs to cluster 1.

The distribution plots of average county features for the different crash types by clusters have been presented in Figure 15. Summary of Cluster Features. The bars in blue, amber and grey respectively represent motor vehicle, pedestrian and pedal cyclist crashes. The plots shown in Figure 15 presents a comprehensive picture about varying dependence of crash features on county groups crash rates, informations that are key for the modelling framework. The optimum number of clusters for the aggregated data considering the three types of crashes was found to be 2. Clusters 1, 2, 3 and 5 are identical with respect to the distribution of the features and Cluster 4 varies significantly from others. The mean value of crash features like DVMT, Truck DVMT, population

density, and road density in cluster 4 for all the crash types is significantly high relative to other clusters as explained before, whereas the education type, per capita income, and urban percent displays variation among the cluster for all the crash types.



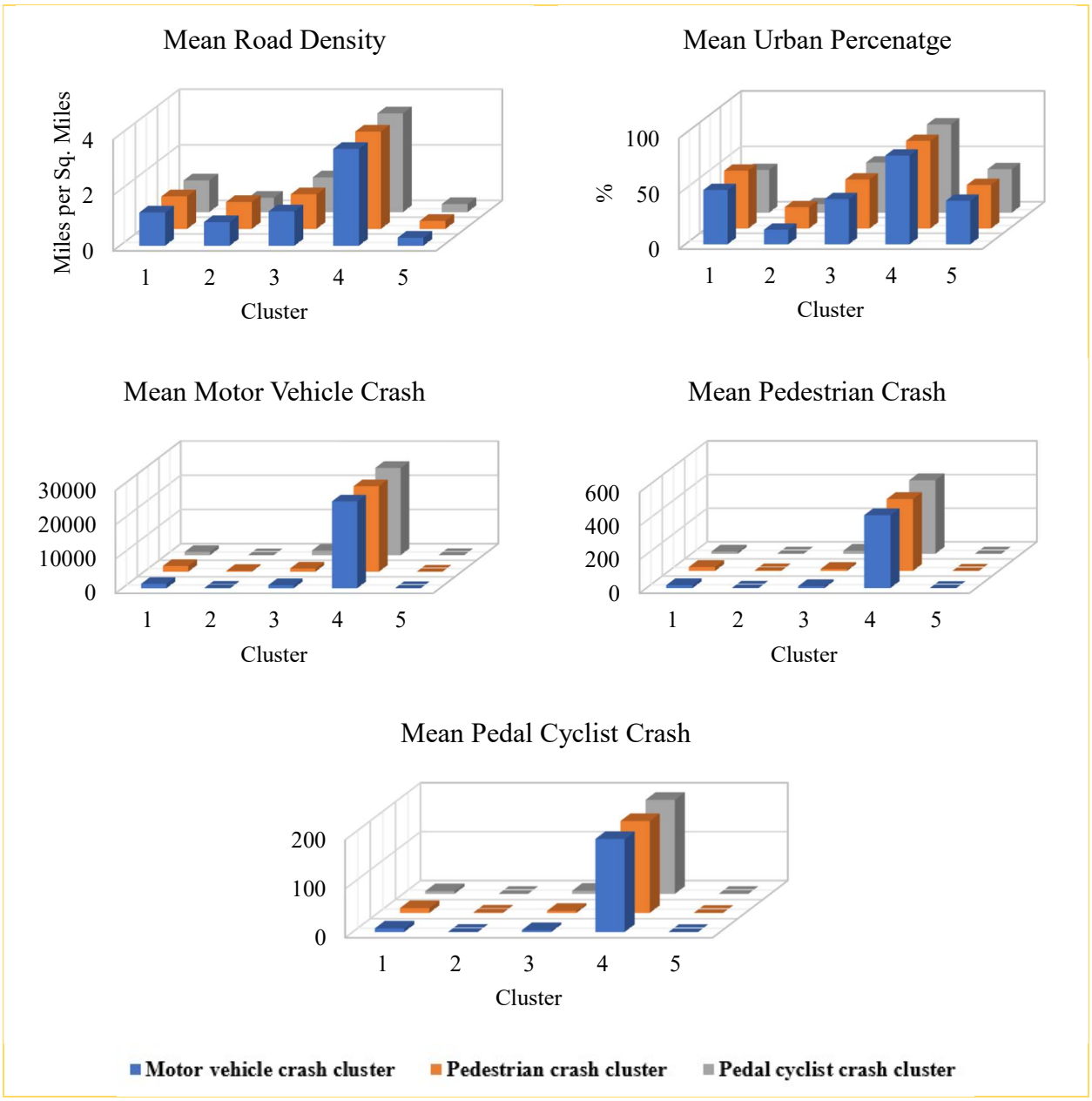


Figure 15. Summary of Cluster Features

6. CONCLUSIONS

The primary objective of this project was to explore and investigate the practicality of using Machine Learning Multi-Label Classification tool for classification problems in the context of traffic safety, in particular simultaneous classification of collision type and crash injury type. The ability of MLC to categorize an entity under study to more than one labels simultaneously provides it an edge over the traditional classification approaches that classify collision type and crash injury types separately. Underlying correlation between the injury severity type and collision type is leveraged using the AI tool to develop a robust classification model, and the performance of the proposed tool has been benchmarked with the conventional tools. The classification performance of all the conventional and proposed classification algorithms considered in this study has been benchmarked and compared in terms of prediction performance and computational efficiency.

Primary challenge of the project was the lack of literature on both mathematical and computational materials that discusses MLC algorithms specific to traffic safety domain. Although MLC is an evolving concept, it has gained popularity only in recent years for its intuitiveness and simplicity. Specifically, there is lack of materials on domain/algorithm specific hyperparameters tuning and their limits for multi-label classification side of ML-AI domain. This means more customized testing is needed to optimize such hyperparameters that affects model performance. Though showing promising preliminary results, the MLC models also had to deal with challenges such as the inherent labels dependencies, the computational complexity related of the model's inference, the large dimensions of the (input/output) spaces and the imbalance label representation where negative labels massively outnumber positive ones. Available computational packages for MLC classification are relatively new compared to the multi-class model packages, which are typically extensively updated and upgraded frequently. This could impact the direct comparison of both modelling approaches.

Though more comprehensive training and testing is required, the numerical result from this study indicates that the proposed approach has a promising overall classification performance compared to traditional multiclass traffic crash injury classification approaches. The primary output of this project is a new research framework for the fundamental injury severity classification and traffic safety analysis problems.

6.1. Future Direction

An interesting future extension of this research will be to conduct additional training and testing of the models with the modified hyperparameters and parallelly narrow down the optimized metric for performance evaluation and benchmarking. Research may be conducted on tuning the model codes to minimize the package dependencies, thus increasing flexibility over the parameters. Besides the model approaches of multi-label algorithms discussed in this report, a third category of meta-models distinguish itself as *ensemble multi-label models* where multi-label models are based on the top of a committee of single multi-label models with the goal of combining their outputs as a single prediction. This group of models aims to enhance the generalization ability of single models by combining multiple ones to accomplish jointly one common task. The improvement of performances within this family of methods relies on the concept of diversity, stating that a good ensemble is a committee of models in which misclassified instances are different from one individual model to another. Such models need to be explored for crash analysis.

REFERENCES

1. CDC - Motor Vehicle Injury - Topics - Did You Know - STLT Gateway. https://www.cdc.gov/publichealthgateway/didyouknow/topic/vehicle.html. Accessed Jul. 15, 2020.
2. Texas Motor Vehicle Crash Statistics - 2019. https://www.txdot.gov/government/enforcement/annual-summary.html. Accessed Jul. 15, 2020.
3. Road Safety Facts. Association for Safe International Road Travel.
4. Traffic Safety Facts Annual Report Tables. https://cdan.nhtsa.gov/tsftables/tsfar.htm. Accessed Jul. 14, 2020.
5. Cost Data and Prevention Policies | Motor Vehicle Safety | CDC Injury Center. https://www.cdc.gov/motorvehiclesafety/costs/index.html. Accessed Jul. 15, 2020.
6. CrashStats - NHTSA - DOT. https://crashstats.nhtsa.dot.gov/#!/. Accessed Jul. 14, 2020.
7. lynn.greenbauer.ctr@dot.gov. Crash Investigation Sampling System. *NHTSA*. https://www.nhtsa.gov/crash-data-systems/crash-investigation-sampling-system. Accessed Jul. 15, 2020.
8. Ecola, L., B. S. Batorsky, J. S. Ringel, J. Zmud, K. Connor, D. Powell, B. G. Chow, C. Panis, and G. S. Jones. How to Get the Biggest Impact from an Increase in Spending on Traffic Safety. *Montana*, Vol. 1, 2015, p. 7.
9. Mohan, D. Road Safety in Less-Motorized Environments: Future Concerns. *International Journal of Epidemiology*, Vol. 31, No. 3, 2002, pp. 527–532. https://doi.org/10.1093/ije/31.3.527.
10. Abdulhafedh, A. Road Traffic Crash Data: An Overview on Sources, Problems, and Collection Methods. *Journal of Transportation Technologies*, Vol. 07, 2017, pp. 206–219. https://doi.org/10.4236/jtts.2017.72015.
11. Vision Zero SA > Home. https://www.visionzerosa.com/. Accessed May 31, 2021.
12. What Is Vision Zero? https://visionzeronetwork.org/about/what-is-vision-zero/. Accessed May 31, 2021.
13. Why Crash Data Is Important. Mass Crash Report Manual.
14. Lord, D., and F. Mannering. The Statistical Analysis of Crash-Frequency Data: A Review and Assessment of Methodological Alternatives. *Transportation Research Part A: Policy and Practice*, Vol. 44, No. 5, 2010, pp. 291–305. https://doi.org/10.1016/j.tra.2010.02.001.
15. Miaou, S.-P., and J. J. Song. Bayesian Ranking of Sites for Engineering Safety Improvements: Decision Parameter, Treatability Concept, Statistical Criterion, and Spatial Dependence. *Accident Analysis & Prevention*, Vol. 37, No. 4, 2005, pp. 699–720.
16. Geedipally, S. R., and D. Lord. Investigating the Effect of Modeling Single-Vehicle and Multi-Vehicle Crashes Separately on Confidence Intervals of Poisson–Gamma Models. *Accident Analysis & Prevention*, Vol. 42, No. 4, 2010, pp. 1273–1282.
17. Aguero-Valverde, J., and P. Jovanis. Spatial Analysis of Fatal Injury Crashes in Pennsylvania. *Accident; analysis and prevention*, Vol. 38, 2006, pp. 618–25. https://doi.org/10.1016/j.aap.2005.12.006.
18. Fatality Facts 2019: Males and Females. *IIHS-HLDI crash testing and highway safety*. https://www.iihs.org/topics/fatality-statistics/detail/males-and-females. Accessed Aug. 17, 2021.
19. Top 7 Causes of Car Accidents - 2020 Statistics. After Car Accidents, Jan 31, 2019.

20. Savolainen, P. T., F. L. Mannering, D. Lord, and M. A. Quddus. The Statistical Analysis of Highway Crash-Injury Severities: A Review and Assessment of Methodological Alternatives. *Accident Analysis & Prevention*, Vol. 43, No. 5, 2011, pp. 1666–1676. https://doi.org/10.1016/j.aap.2011.03.025.
21. Hauer, E. The Frequency–Severity Indeterminacy. *Accident Analysis & Prevention*, Vol. 38, No. 1, 2006, pp. 78–83. https://doi.org/10.1016/j.aap.2005.07.001.
22. Hauer, E., and A. S. Hakkert. Extent and Some Implications of Incomplete Accident Reporting. *Transportation research record*, Vol. 1185, 1988, pp. 1–10.
23. Elvik, R., and A. Mysen. Incomplete Accident Reporting: Meta-Analysis of Studies Made in 13 Countries. *Transportation Research Record*, Vol. 1665, No. 1, 1999, pp. 133–140. https://doi.org/10.3141/1665-18.
24. Savolainen, P., and F. Mannering. Probabilistic Models of Motorcyclists' Injury Severities in Single-and Multi-Vehicle Crashes. *Accident Analysis & Prevention*, Vol. 39, No. 5, 2007, pp. 955–963.
25. Paleti, R., N. Eluru, and C. R. Bhat. Examining the Influence of Aggressive Driving Behavior on Driver Injury Severity in Traffic Crashes. *Accident Analysis & Prevention*, Vol. 42, No. 6, 2010, pp. 1839–1854. https://doi.org/10.1016/j.aap.2010.05.005.
26. McFadden, D., and K. Train. Mixed MNL Models for Discrete Response. *Journal of Applied Econometrics*, Vol. 15, No. 5, 2000, pp. 447–470. https://doi.org/10.1002/1099-1255(200009/10)15:5<447::AID-JAE570>3.0.CO;2-1.
27. Train, K. E. *Discrete Choice Methods with Simulation*. Cambridge University Press, 2009.
28. Li, Z., P. Liu, W. Wang, and C. Xu. Using Support Vector Machine Models for Crash Injury Severity Analysis. *Accident Analysis & Prevention*, Vol. 45, 2012, pp. 478–486. https://doi.org/10.1016/j.aap.2011.08.016.
29. Mussone, L., A. Ferrari, and M. Oneta. An Analysis of Urban Collisions Using an Artificial Intelligence Model. *Accident Analysis & Prevention*, Vol. 31, No. 6, 1999, pp. 705–718. https://doi.org/10.1016/S0001-4575(99)00031-7.
30. Delen, D., R. Sharda, and M. Bessonov. Identifying Significant Predictors of Injury Severity in Traffic Accidents Using a Series of Artificial Neural Networks. *Accident Analysis & Prevention*, Vol. 38, No. 3, 2006, pp. 434–444. https://doi.org/10.1016/j.aap.2005.06.024.
31. Ma, J., and K. M. Kockelman. Bayesian Multivariate Poisson Regression for Models of Injury Count, by Severity. *Transportation Research Record*, Vol. 1950, No. 1, 2006, pp. 24–34. https://doi.org/10.1177/0361198106195000104.
32. Burkov, A. *The Hundred-Page Machine Learning Book*. Leanpub, 2018.
33. Bonaccorso, G. *Machine Learning Algorithms: Popular Algorithms for Data Science and Machine Learning, 2nd Edition*. Packt Publishing Ltd, 2018.
34. Lu, T., Z. Dunyao, Y. Lixin, and Z. Pan. The Traffic Accident Hotspot Prediction: Based on the Logistic Regression Method. Presented at the 2015 International Conference on Transportation Information and Safety (ICTIS), 2015.
35. Iranitalab, A., and A. Khattak. Comparison of Four Statistical and Machine Learning Methods for Crash Severity Prediction. *Accident Analysis & Prevention*, Vol. 108, 2017, pp. 27–36. https://doi.org/10.1016/j.aap.2017.08.008.
36. Yuan, Z., X. Zhou, T. Yang, and J. Tamerius. Predicting Traffic Accidents Through Heterogeneous Urban Data : A Case Study. 2017.
37. Rivolli, A., and A. C. de Carvalho. The Utiml Package: Multi-Label Classification in R. *R J.*, Vol. 10, No. 2, 2018, p. 24.

38. Tsoumakas, G., I. Katakis, and I. Vlahavas. Mining Multi-Label Data. In *Data Mining and Knowledge Discovery Handbook* (O. Maimon and L. Rokach, eds.), Springer US, Boston, MA, pp. 667–685.
39. Zhang, M.-L., and Z.-H. Zhou. A Review on Multi-Label Learning Algorithms. *IEEE transactions on knowledge and data engineering*, Vol. 26, No. 8, 2013, pp. 1819–1837.
40. Tsoumakas, G., and I. Katakis. Multi-Label Classification: An Overview. *International Journal of Data Warehousing and Mining (IJDWM)*, Vol. 3, No. 3, 2007, pp. 1–13.
41. Boutell, M. R., J. Luo, X. Shen, and C. M. Brown. Learning Multi-Label Scene Classification. *Pattern recognition*, Vol. 37, No. 9, 2004, pp. 1757–1771.
42. Read, J., B. Pfahringer, G. Holmes, and E. Frank. Classifier Chains for Multi-Label Classification. Berlin, Heidelberg, 2009.
43. Fürnkranz, J., E. Hüllermeier, E. L. Mencía, and K. Brinker. Multilabel Classification via Calibrated Label Ranking. *Machine learning*, Vol. 73, No. 2, 2008, pp. 133–153.
44. Tsoumakas, G., and I. Vlahavas. Random K-Labelsets: An Ensemble Method for Multilabel Classification. 2007.
45. Zhang, M.-L., and Z.-H. Zhou. ML-KNN: A Lazy Learning Approach to Multi-Label Learning. *Pattern Recognition*, Vol. 40, No. 7, 2007, pp. 2038–2048. https://doi.org/10.1016/j.patcog.2006.12.019.
46. McCallum, A. K. Multi-Label Text Classification with a Mixture Model Trained by EM. 1999.
47. Schapire, R. E., and Y. Singer. BoosTexter: A Boosting-Based System for Text Categorization. *Machine learning*, Vol. 39, No. 2, 2000, pp. 135–168.
48. Ueda, N., and K. Saito. Parametric Mixture Model for Multitopic Text. *Systems and Computers in Japan*, Vol. 37, No. 2, 2006, pp. 56–66.
49. Kaneda, Y., N. Ueda, and K. Saito. Extended Parametric Mixture Model for Robust Multi-Labeled Text Categorization. 2004.
50. Rivolli, A., L. C. Parker, and A. C. P. L. F. de Carvalho. Food Truck Recommendation Using Multi-Label Classification. Cham, 2017.
51. Papagiannakis, A. T., M. Bracher, and N. C. Jackson. Utilizing Clustering Techniques in Estimating Traffic Data Input for Pavement Design. *Journal of Transportation Engineering*, Vol. 132, No. 11, 2006, pp. 872–879. https://doi.org/10.1061/(ASCE)0733-947X(2006)132:11(872).
52. Cao, J., M. Liang, Y. Li, J. Chen, H. Li, R. W. Liu, and J. Liu. PCA-Based Hierarchical Clustering of AIS Trajectories with Automatic Extraction of Clusters. Presented at the 2018 IEEE 3rd International Conference on Big Data Analysis (ICBDA), 2018.
53. Raihan, M. A., M. Hossain, and T. Hasan. Data Mining in Road Crash Analysis: The Context of Developing Countries. *International Journal of Injury Control and Safety Promotion*, Vol. 25, No. 1, 2018, pp. 41–52. https://doi.org/10.1080/17457300.2017.1323929.
54. Taamneh, M., S. Taamneh, and S. Alkheder. Clustering-Based Classification of Road Traffic Accidents Using Hierarchical Clustering and Artificial Neural Networks. *International Journal of Injury Control and Safety Promotion*, Vol. 24, No. 3, 2017, pp. 388–395. https://doi.org/10.1080/17457300.2016.1224902.
55. Janstrup, K. H., M. Møller, and N. Pilegaard. A Clustering Approach to Integrate Traffic Safety in Road Maintenance Prioritization. *Traffic Injury Prevention*, Vol. 20, No. 4, 2019, pp. 442–448. https://doi.org/10.1080/15389588.2019.1580700.

56. Cai, Q. Integrating the Macroscopic and Microscopic Traffic Safety Analysis Using Hierarchical Models. *Electronic Theses and Dissertations, 2004-2019*, 2017.
57. Houston, R. W. The Transportation Equity Act for the 21st Century. *Institute of Transportation Engineers. ITE Journal*, Vol. 68, No. 7, 1998, p. 45.
58. Washington, S. *Incorporating Safety into Long-Range Transportation Planning*. Transportation Research Board, 2006.
59. MAP-21 - Moving Ahead for Progress in the 21st Century Act | FMCSA. https://www.fmcsa.dot.gov/mission/policy/map-21-moving-ahead-progress-21st-century-act. Accessed Jul. 14, 2020.
60. The Fixing America's Surface Transportation Act or "FAST Act" | US Department of Transportation. https://www.transportation.gov/fastact. Accessed Jul. 14, 2020.
61. FMCSA Crash Data Collection Resource Home Page. https://ai.fmcsa.dot.gov/DataQuality/CrashCollectionTraining/index.html. Accessed Aug. 11, 2021.
62. Administrations, F. H. KABCO Injury Classification Scale and Definitions.
63. Read, J., B. Pfahringer, G. Holmes, and E. Frank. Classifier Chains for Multi-Label Classification. *Machine learning*, Vol. 85, No. 3, 2011, pp. 333–359.
64. Zhang, M.-L., and Z.-H. Zhou. A K-Nearest Neighbor Based Algorithm for Multi-Label Classification. *GrC*, Vol. 5, 2005, pp. 718–721.
65. Johnson, S. C. Hierarchical Clustering Schemes. *Psychometrika*, Vol. 32, No. 3, 1967, pp. 241–254.
66. 2021 World Population by Country. https://worldpopulationreview.com/. Accessed Oct. 10, 2021.
67. Crash Reports and Records. https://www.txdot.gov/driver/laws/crash-reports.html. Accessed Jul. 23, 2020.
68. Texas Department of Transportation. https://www.txdot.gov/content/txdot/en.html. Accessed May 31, 2021.
69. Texas Transportation Code Section 550.063 - Report on Appropriate Form (2019). https://texas.public.law/statutes/tex._transp._code_section_550.063. Accessed May 31, 2021.
70. *Instructions to Police for Reporting Crashes*. Texas Department of Transportation, Austin, TX, 2020.
71. TxDOT Open Data Portal. http://gis-txdot.opendata.arcgis.com/. Accessed Jul. 23, 2020.
72. TxDOT Roadway Inventory. https://txdot.maps.arcgis.com/sharing/rest/content/items/843ebe994c114961a855ec76ddcde086. Accessed Jul. 28, 2021.
73. CRIS Query. https://cris.dot.state.tx.us/public/Query/app/home. Accessed Jun. 27, 2021.
74. Roadway Safety Information Analysis - Safety | Federal Highway Administration. https://safety.fhwa.dot.gov/local_rural/training/fhwasa1210/s3.cfm. Accessed Jul. 9, 2021.

APPENDIX A: Classification Performance Evaluation [Conventional Models]

Model Name	Target	support	LR			DT			RF			SVM			KNN		
			Precision	Recall	f1-score	Precision	Recall	f1-score	Precision	Recall	f1-score	Precision	Recall	f1-score	Precision	Recall	f1-score
C1	Same Direction	4956	0.87	0.96	0.91	0.88	0.87	0.88	0.88	0.96	0.91	0.88	0.96	0.92	0.88	0.93	0.91
	Angle Collision	675	0.65	0.12	0.2	0.27	0.31	0.29	0.48	0.14	0.21	0.62	0.1	0.17	0.58	0.07	0.12
	Opposite Direction	379	0.82	0.11	0.19	0.33	0.37	0.35	0.57	0.29	0.38	0	0	0	0.49	0.18	0.26
	Other	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	micro avg.	6016	0.86	0.81	0.83	0.77	0.77	0.77	0.85	0.82	0.84	0.87	0.8	0.84	0.87	0.79	0.83
	macro avg.	6016	0.58	0.3	0.32	0.37	0.39	0.38	0.48	0.34	0.38	0.37	0.27	0.27	0.49	0.29	0.32
	weighted avg.	6016	0.84	0.81	0.78	0.78	0.77	0.78	0.81	0.82	0.8	0.79	0.8	0.77	0.82	0.79	0.78
	samples avg.	6016	0.81	0.81	0.81	0.74	0.77	0.75	0.82	0.82	0.82	0.8	0.8	0.8	0.79	0.79	0.79
C2	Incapacitating Injury	94	0	0	0	0.02	0.03	0.03	0.14	0.01	0.02	0	0	0	0	0	0
	Non-Incapacitating Injury	422	0	0	0	0.09	0.12	0.1	0.07	0.01	0.01	0	0	0	0	0	0
	Possible Injury	1358	0	0	0	0.26	0.27	0.27	0.27	0.06	0.1	0	0	0	0.28	0.02	0.04
	Fatal Injury	11	0	0	0	0.05	0.09	0.06	0	0	0	0	0	0	0	0	0
	Not Injury	4131	0.69	0.99	0.81	0.71	0.69	0.7	0.7	0.88	0.78	0.69	1	0.81	0.7	0.8	0.75
	micro avg.	6016	0.69	0.68	0.69	0.53	0.54	0.54	0.67	0.62	0.65	0.69	0.69	0.69	0.69	0.55	0.61
	macro avg.	6016	0.14	0.2	0.16	0.23	0.24	0.23	0.24	0.19	0.18	0.14	0.2	0.16	0.2	0.16	0.16
	weighted avg.	6016	0.47	0.68	0.56	0.55	0.54	0.55	0.55	0.62	0.56	0.47	0.69	0.56	0.54	0.55	0.52
	samples avg.	6016	0.68	0.68	0.68	0.47	0.54	0.49	0.62	0.62	0.62	0.69	0.69	0.69	0.55	0.55	0.55
C3	Not Injured	4131	0.69	0.99	0.81	0.71	0.68	0.69	0.7	0.87	0.78	0.69	1	0.81	0.7	0.79	0.74
	Injured	1885	0.4	0.02	0.03	0.35	0.35	0.35	0.39	0.18	0.25	0	0	0	0.38	0.12	0.18
	micro avg.	6016	0.68	0.68	0.68	0.59	0.58	0.58	0.66	0.65	0.66	0.69	0.69	0.69	0.66	0.58	0.62
	macro avg.	6016	0.54	0.5	0.42	0.53	0.52	0.52	0.55	0.53	0.51	0.34	0.5	0.41	0.54	0.46	0.46
	weighted avg.	6016	0.6	0.68	0.57	0.59	0.58	0.59	0.6	0.65	0.61	0.47	0.69	0.56	0.6	0.58	0.57
	samples avg.	6016	0.68	0.68	0.68	0.56	0.58	0.57	0.65	0.65	0.65	0.69	0.69	0.69	0.58	0.58	0.58
C4	Incapacitating Injury	37	0	0	0	0.03	0.05	0.04	0.25	0.03	0.05	0	0	0	0	0	0
	Non-Incapacitating Injury	179	0	0	0	0.04	0.04	0.04	0.08	0.01	0.01	0	0	0	0	0	0
	Possible Injury	716	0	0	0	0.13	0.15	0.14	0.2	0.02	0.04	0	0	0	0.41	0.01	0.02

	Fatal Injury	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Not Injury	4264	0.73	0.95	0.83	0.75	0.73	0.74	0.75	0.88	0.81	0.73	0.96	0.83	0.75	0.85	0.8
	micro avg.	5200	0.73	0.78	0.75	0.61	0.62	0.62	0.74	0.72	0.73	0.73	0.79	0.76	0.75	0.7	0.72
	macro avg.	5200	0.15	0.19	0.17	0.19	0.19	0.19	0.25	0.19	0.18	0.15	0.19	0.17	0.23	0.17	0.16
	weighted avg.	5200	0.6	0.78	0.68	0.63	0.62	0.63	0.64	0.72	0.67	0.6	0.79	0.68	0.67	0.7	0.66
	samples avg.	5200	0.68	0.68	0.68	0.5	0.54	0.51	0.63	0.63	0.63	0.68	0.68	0.68	0.6	0.6	0.6
C5	Not Injured	4263	0.73	0.95	0.83	0.75	0.73	0.74	0.75	0.88	0.81	0.73	0.96	0.83	0.75	0.86	0.8
C5	Injured	937	0	0	0	0.18	0.2	0.19	0.26	0.05	0.08	0	0	0	0.45	0.01	0.02
C5	micro avg.	5200	0.73	0.78	0.76	0.63	0.63	0.63	0.73	0.73	0.73	0.73	0.79	0.76	0.75	0.71	0.73
C5	macro avg.	5200	0.37	0.48	0.41	0.46	0.47	0.46	0.51	0.46	0.44	0.37	0.48	0.42	0.6	0.43	0.41
C5	weighted avg.	5200	0.6	0.78	0.68	0.65	0.63	0.64	0.66	0.73	0.68	0.6	0.79	0.68	0.7	0.71	0.66
C5	samples avg.	5200	0.67	0.67	0.67	0.52	0.55	0.53	0.63	0.63	0.63	0.68	0.68	0.68	0.61	0.61	0.61
C6	Incapacitating Injury	35	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
C6	Non-Incapacitating Injury	187	0	0	0	0.04	0.05	0.05	0.06	0.01	0.01	0	0	0	0	0	0
C6	Possible Injury	726	0	0	0	0.15	0.16	0.16	0.26	0.03	0.05	0	0	0	0.47	0.01	0.02
C6	Fatal Injury	4	0	0	0	0.12	0.25	0.17	0	0	0	0	0	0	0	0	0
C6	Not Injury	4312	0.73	0.96	0.83	0.75	0.74	0.75	0.74	0.89	0.81	0.72	1	0.84	0.75	0.84	0.79
C6	micro avg.	5264	0.73	0.79	0.76	0.63	0.63	0.63	0.73	0.73	0.73	0.72	0.82	0.76	0.75	0.69	0.72
C6	macro avg.	5264	0.15	0.19	0.17	0.22	0.24	0.22	0.21	0.18	0.17	0.14	0.2	0.17	0.25	0.17	0.16
C6	weighted avg.	5264	0.6	0.79	0.68	0.64	0.63	0.64	0.65	0.73	0.67	0.59	0.82	0.68	0.68	0.69	0.65
C6	samples avg.	5264	0.69	0.69	0.69	0.51	0.55	0.53	0.64	0.64	0.64	0.72	0.72	0.72	0.61	0.61	0.61
C7	Not Injured	4312	0.73	0.96	0.83	0.75	0.72	0.74	0.74	0.87	0.8	0.72	1	0.84	0.75	0.83	0.79
C7	Injured	952	0	0	0	0.19	0.2	0.19	0.27	0.05	0.09	0	0	0	0.38	0.02	0.04
C7	micro avg.	5264	0.73	0.78	0.76	0.64	0.63	0.63	0.72	0.73	0.73	0.72	0.82	0.76	0.75	0.68	0.71
C7	macro avg.	5264	0.37	0.48	0.42	0.47	0.46	0.47	0.5	0.46	0.45	0.36	0.5	0.42	0.57	0.42	0.41
C7	weighted avg.	5264	0.6	0.78	0.68	0.65	0.63	0.64	0.66	0.73	0.67	0.59	0.82	0.68	0.68	0.68	0.65
C7	samples avg.	5264	0.69	0.69	0.69	0.53	0.55	0.53	0.63	0.64	0.64	0.72	0.72	0.72	0.6	0.6	0.6

APPENDIX B: Classification Performance Evaluation [Proposed Models]

Target		Support	Binary Relevance with Random Forest Classifier			ML_KNN			Classifier Chains with Random Forest Classifier			Label Powerset with Random Forest Classifier		
			Precision	Recall	f1-score	Precision	Recall	f1-score	Precision	Recall	f1-score	Precision	Recall	f1-score
P1	Angel Collision	511	0.56	0.23	0.33	0.57	0.11	0.18	0.55	0.22	0.32	0.25	0.25	0.25
	Same Direction	3654	0.89	0.94	0.92	0.88	0.93	0.91	0.88	0.96	0.92	0.82	0.25	0.38
	Opposite Direction	319	0.59	0.34	0.44	0.57	0.18	0.27	0.55	0.47	0.51	0.09	0.3	0.14
	Other	1	0	0	0	0	0	0	0	0	0	0	0	0
Collision Type and Most Severe Injury	Incapacitating Injury	80	0.2	0.01	0.02	0	0	0	0.29	0.03	0.05	0.01	0.19	0.02
	Non-Incapacitating Injury	375	0.15	0.02	0.03	0	0	0	0.13	0.02	0.03	0.14	0.06	0.09
	Possible Injury	1186	0.31	0.09	0.14	0.35	0.04	0.07	0.3	0.11	0.16	0.25	0.03	0.05
	Fatal Injury	6	0	0	0	0	0	0	0	0	0	0	0.33	0
	Not Injury	2838	0.65	0.82	0.73	0.64	0.91	0.75	0.64	0.9	0.75	0.63	0.39	0.48
	micro avg.	8970	0.74	0.68	0.71	0.75	0.68	0.71	0.72	0.72	0.72	0.26	0.26	0.26
	macro avg.	8970	0.37	0.27	0.29	0.34	0.24	0.24	0.37	0.3	0.3	0.24	0.2	0.16
	weighted avg.	8970	0.67	0.68	0.66	0.66	0.68	0.64	0.66	0.72	0.67	0.59	0.26	0.34
	samples avg.	8970	0.74	0.68	0.7	0.73	0.68	0.7	0.72	0.72	0.72	0.26	0.26	0.26
P2	Angel Collision	511	0.56	0.22	0.32	0.57	0.11	0.18	0.57	0.23	0.32	0.15	0.35	0.21
	Same Direction	3654	0.89	0.95	0.92	0.88	0.93	0.91	0.88	0.96	0.92	0.88	0.47	0.61
	Opposite Direction	319	0.6	0.35	0.44	0.57	0.18	0.27	0.57	0.47	0.51	0.13	0.34	0.19
	Other	1	0	0	0	0	0	0	0	0	0	0	0	0
Collision Type and Driver 1 Injury Type and Driver 2 Injury	Incapacitating Injury	37	0.67	0.05	0.1	0	0	0	0.4	0.05	0.1	0.01	0.46	0.02
	Non-Incapacitating Injury	160	0.12	0.01	0.02	0	0	0	0.11	0.01	0.02	0.04	0.15	0.06
	Possible Injury	629	0.23	0.04	0.07	0.5	0	0.01	0.25	0.05	0.09	0.17	0.23	0.19
	Fatal Injury	4	0	0	0	0	0	0	0	0	0	0	0	0
	Not Injury	3655	0.82	0.96	0.88	0.82	1	0.9	0.82	0.97	0.89	0.84	0.31	0.45
	Incapacitating Injury	30	0	0	0	0	0	0	0	0	0	0.01	0.33	0.01

Type	Non-Incapacitating Injury	182	0.12	0.01	0.02	0	0	0	0.11	0.01	0.02	0.03	0.18	0.06
	Possible Injury	699	0.31	0.06	0.1	0.5	0.03	0.05	0.3	0.05	0.08	0.16	0.17	0.16
	Fatal Injury	3	0	0	0	0	0	0	0	0	0	0	0	0
	Not Injury	3571	0.8	0.96	0.87	0.8	1	0.89	0.8	0.98	0.88	0.81	0.27	0.41
	micro avg.	13455	0.82	0.79	0.8	0.83	0.8	0.81	0.81	0.81	0.81	0.33	0.33	0.33
	macro avg.	13455	0.37	0.26	0.27	0.33	0.23	0.23	0.34	0.27	0.27	0.23	0.23	0.17
	weighted avg.	13455	0.75	0.79	0.75	0.76	0.8	0.74	0.74	0.81	0.76	0.71	0.33	0.43
	samples avg.	13455	0.81	0.79	0.8	0.82	0.8	0.81	0.81	0.81	0.81	0.33	0.33	0.33
P3	Angel Collision	511	0.56	0.22	0.32	0.57	0.11	0.18	0.56	0.23	0.32	0.48	0.25	0.33
	Same Direction	3654	0.89	0.95	0.92	0.88	0.93	0.91	0.88	0.96	0.92	0.88	0.95	0.91
	Opposite Direction	319	0.6	0.35	0.44	0.57	0.18	0.27	0.54	0.46	0.5	0.57	0.43	0.49
	Other	1	0	0	0	0	0	0	0	0	0	0	0	0
Collision Type and Most Severe Injury	Not Injured	2838	0.65	0.82	0.72	0.64	0.91	0.75	0.64	0.82	0.72	0.64	0.81	0.72
	Injured	1647	0.43	0.22	0.29	0.44	0.13	0.2	0.42	0.22	0.29	0.41	0.22	0.29
	micro avg.	8970	0.73	0.71	0.72	0.74	0.7	0.72	0.72	0.72	0.72	0.72	0.72	0.72
	macro avg.	8970	0.52	0.43	0.45	0.52	0.37	0.38	0.51	0.45	0.46	0.5	0.44	0.46
	weighted avg.	8970	0.7	0.71	0.69	0.7	0.7	0.66	0.69	0.72	0.69	0.68	0.72	0.69
	samples avg.	8970	0.73	0.71	0.71	0.73	0.7	0.71	0.72	0.72	0.72	0.72	0.72	0.72
P4	Angel Collision	511	0.57	0.23	0.32	0.57	0.11	0.18	0.58	0.23	0.33	0.5	0.23	0.31
	Same Direction	3654	0.89	0.94	0.91	0.88	0.93	0.91	0.88	0.96	0.92	0.87	0.96	0.91
	Opposite Direction	319	0.59	0.35	0.44	0.57	0.18	0.27	0.55	0.47	0.51	0.54	0.4	0.46
	Other	1	0	0	0	0	0	0	0	0	0	0	0	0
Collision Type and Driver 1 Injury Type and Driver 2 Injury Type	Not Injured	3655	0.82	0.96	0.88	0.82	1	0.9	0.82	0.96	0.88	0.82	0.96	0.88
	Injured	830	0.27	0.07	0.11	0.43	0.01	0.02	0.28	0.07	0.12	0.3	0.08	0.13
	Not Injured	3571	0.8	0.96	0.87	0.8	1	0.89	0.8	0.97	0.88	0.8	0.96	0.87
	Injured	914	0.34	0.08	0.12	0.73	0.02	0.05	0.35	0.07	0.12	0.34	0.09	0.14
	micro avg.	13455	0.81	0.8	0.8	0.83	0.8	0.81	0.81	0.81	0.81	0.8	0.8	0.8
	macro avg.	13455	0.54	0.45	0.46	0.6	0.41	0.4	0.53	0.47	0.47	0.52	0.46	0.46
	weighted avg.	13455	0.75	0.8	0.76	0.79	0.8	0.74	0.75	0.81	0.76	0.75	0.8	0.76
	samples avg.	13455	0.81	0.8	0.8	0.82	0.8	0.81	0.81	0.81	0.81	0.8	0.8	0.8

APPENDIX C: Agglomerative Hierarchical Clustering

Table 9 Clustering Data

Name	Description	Min	Max	Mean	Std. Dev.
	Crash Frequency (2015,2016,2017 average)				
Pedestrian	Crash involving Pedestrians	0	5031	84.48	420.56
Motoveh	Motor vehicle crashes	2	303034	5110.42	23330.20
Cyclist	Crash involving pedal cyclists	0	1943	36.76	167.77
Exposure features					
DVMT	Daily vehicle miles traveled (in thousands)	53.97	239484.16	6093.72	20973.16
Traffic Features					
trf_intnsty	Traffic Intensity (DVMT/Road Length)	107.15	15768.85	2454.029	2508.98
trck_intnsty	Truck Traffic Intensity (Truck DVMT/Road length)	23.38	2653.49	451.93	427.7
	Road Infrastructure				
IS_prcnt	Interstate Road %	0.00	30.20	2.75	4.92
FrEx_prcnt	Freeway and Expressway %	0.00	8.57	0.54	1.39
PA_prcnt	Principal Arterial %	0.00	68.39	7.40	7.04
MinA_prcnt	Minor Arterial %	0.00	26.62	6.12	4.34
MaCol_prcnt	Major Collector Road %	1.41	55.65	16.53	6.04
MinCol_prcnt	Minor Collector Road %	0.00	22.89	6.21	4.15
Loc_prcnt	Local Road %	18.76	76.71	60.42	9.35
NoMed_prcnt	% of Road section with No Median	31.61	100.00	92.06	7.59
Unprtcd_prcnt	% of Road section with Unprotected Median	0.00	66.60	5.59	7.05
curbd_prcnt	% of Road section with Curbed Median	0.00	7.85	0.41	1.05
PosBar_prcnt	% of Road section with Positive Barrier Median	0.00	14.43	1.95	2.95
FrExway_nlane	Average Number of Lanes on Freeway Expressway	0.00	3.573	0.60	1.14
PA_nlane	Average Number of Lanes on Principal Arterial	0.00	4.00	2.40	1.01
MinorA_nlane	Average Number of Lanes on Minor Arterial	0.00	4.22	2.29	0.61
Maj_col_nlane	Average Number of Lanes on Major Collector	1.86	2.49	2.07	0.10
Min_col_nlane	Average Number of Lanes on Minor Collector	0.00	2.73	1.98	0.29
Local_nlane	Average Number of Lanes on Local Roads	1.93	2.02	2.00	0.01
Road_den	Road density Miles/ Sq. Mile	.11	11.60	1.51	1.45
	Socio Demographics and Economics features				
age17_under	Percentage of population of age 17 and under	8.51	35.99	24.22	3.83
age65_older	Percentage of population of age 65 and older	8.61	35.61	17.82	5.24
age85_older	Percentage of population of age 85 and older	0.00	4.99	2.15	0.80
HSG_over	% High School Graduate or higher	48.50	93.90	79.47	8.12
Bach_Over	% Bachelor s Degree or higher	3.00	50.20	18.26	7.46
unemp_rate	Unemployment Rate (%)	1.90	11.70	4.47	1.53
PC_income	Per Capita Income ($)	19801	130461.0	43856.59	12504.32
pop_den	Population Density Per Sq Mile	0.12	2718.00	104.80	305.71
urban_prcnt	% of Urban Area	0.00	99.31	44.48	31.90

Figure 16 Dendrogram Spatial mapping of cluster members