Prioritize Winter Crash Severity Influencing Factors in US Midwestern for Autonomous Vehicle

Shenghong Dai

sdai37@wisc.edu

University of Wisconsin Madison

Abstract: Adverse weather conditions in winter have significant impacts on crash occurrences and risks. Human drivers can adjust driving styles based on the context information of the surrounding traffic and environments. Similar schemes should be considered and designed into autonomous vehicle (AV) control systems. However, most of the existing autonomous vehicle control systems do not have effective mechanisms to deal with extreme weather conditions. There are very limited numbers of research works have focused on the risk factors influencing crash severity affected by winter precipitation. In this study, we aim to find out how different weather conditions relate to crash severities and what are the most influencing risk factors for autonomous vehicles. We utilized three-year crash data of the state from Wisconsin for this study. We evaluated the performance of three statistical prediction models and compared the importance of relevant factors with all crashes and crashes affected by winter precipitation. Evaluation results showed that different weather conditions have a significant influence on crash risk factors. Finally, we prepared a prioritized list of variables that has potential significant impacts on autonomous vehicles safety. Our findings might be useful for designing the control system to improve AV safety under adverse weather conditions.

***Index Terms*—Crash Severity, Winter Precipitation, Machine Learning, Transportation Safety, Autonomous Vehicles**

I. INTRODUCTION

Existing research studies and historical crash data show that crash rates increase dramatically in the winter time in the Midwest of the US [1]. In the year of 2015, over 6.29 million motor vehicle crashes occurred in the U.S. Among these crashes, there were 35,485 fatal crashes, over 2.44 million injury crashes. Even more fatal crashes happened in 2016. The number of fatal crashes increased to 34,439 [2]. An estimation report published by Eisenberg, D.,[3] showed that about 45,000 injury crashes and 150,000 property damage only crashes are caused by rain and snow each year. Another study showed that factors like snow, sleet and freezing rain could a risk factor need to be considered when evaluating fatal crashes. More than 27,000 fatalities based on the crash reports between 1975 and 2011 were caused by or related to winter precipitation [4].

A recent study conducted by the RAND research showed that putting autonomous vehicles (AV) on the road could save hundreds of thousands of lives over time even if they are not flawless [5]- [6]. It could cost hundreds of thousands of lives before achieving perfection or nearly flawless for AVs, according to RAND researchers. One of the main challenges currently we are facing is how to deal with different conditions, including both surrounding traffic conditions and weather conditions. In this work, we focus on how different weather conditions affect crash severities and how to improve the robustness and safeness of autonomous vehicle systems. Existing works mainly focus on how to design efficient algorithms to allow self-driving vehicles handle different weather conditions. Aldibaja et. al proposed an improved localization model which can achieve a better mapping and localization accuracy in snow-wet environments [7]-[8]. Existing solutions change driving schemes when bad or extreme weather conditions are detected [9]-[11]. Hence, one prerequisite of existing solutions is to detect weather condition first then apply corresponding changes to operate the vehicle. In this work, we studied how different influencing factors affect self-driving vehicles and prioritize crash severity influencing factors for autonomous vehicles. To achieve this goal, we first studied historical human driving data and then filtered out a list of factors that has impact on self-driving vehicles. Besides, we also discussed how autonomous vehicles could avoid potential crashes and improve safeness under bad/extreme weather conditions.

Several research reports about the influence of adverse weather condition such as winter precipitation on transportation safety and traffic operation has been published previously [12]-[21]. Fatal crashes under the adverse weather conditions have been examined in order to find the relative crash risks all over the world. For instance, Andrey et al. [13] demonstrated that drivers failed to handle the driving task under unfamiliar environmental conditions. Also, drivers were more likely to maintain high speed on rural road and road with high posted speed limit under the condition of winter precipitation through a study which covered 23 Canadian Cities. A large-scale research covered the crash data from all 50 states of the United States from 2000 to 2007 showed the crash risk can be significantly increased under snow and rain weather condition [1]. Other study showed the injury and fatal crashes were highly elevated because of winter precipitation [3]. In winter, if the crashes have not been cleared promptly, there is a higher chance of continuous crashes happen [22]-[23]. Moreover, research conducted by Qiu and Nixon [24] found that crash rate could increase up to 84% more during snowfall.

Winter precipitation could lead to the decrease of visibility as well as the friction between tires and road surface. In addition, crash rates are more likely to increase since adverse weather causes driving hazards such as poor friction, reduced visibility, and tough operation. Besides the environmental factors, research about driving behavior in winter weather condition was also conducted. For instance, it was suggested that drivers were not qualified enough to take actions to avoid the risk during adverse weather conditions [25]. In addition, foggy weather, icy and wet road surface can also affect drivers' average speed, and time to collision [26].

Although previous studies have confirmed the crash risk is highly related to winter precipitation, there are still uncertain issues that were not explained in detail. For example, are there any differences in relevant risk factors between normal weather and winter weather condition? If there are, what are these differences? Moreover, why the relevant risk factors are different for crashes happened in normal and winter precipitation weather conditions? If the difference is reduced by the effective measurement, will the crash rate decrease too? This study investigated these uncertain issues through comparing the significant levels of relevant factors with all crashes and crashes caused by winter precipitation, using three-year crash data from Wisconsin Traffic Operations and Safety Laboratory (TOPS). Then the different influential variables, the counterpart between the whole crash dataset and the filtered dataset of winter precipitation, were compared and discussed. The different risk factors between the all crashes and winter precipitation related crashes were identified and investigated in this study.

The rest of the content is organized as follows. Section 2 introduces the crash data collection and preparation for the winter weather related crashes and total crashes respectively. The crash prediction models and a two-step analysis procedure are introduced in section 3. Data analysis results comprises section 4, in order to improve the prediction performance, three statistical models were utilized and compared. Moreover, the importance of the influential factors of these two crash datasets (winter weather related crashes and total crashes) were ranked and compared. The significant factors associated with crashes were highlighted and discussed. The relations between crashes and winter precipitation and the limitation of this research are concluded in section 5.

II. Data Description and Preparation

Wisconsin is one of the most representative state which traffic condition and safety could be heavily affected by winter adverse weather. The average seasonal snowfall varies from about 30 inches at Beloit to over 100 inches in northern Iron County along the steep western slope of the Gogebic Range. Average annual duration of snow cover ranges from 85 days in the southernmost part of Wisconsin to more than 140 days in the northernmost part along Lake Superior. This study is conducted based on three-year crash data collected in Wisconsin from 2012 to 2014, provided by TOPS in the University of Wisconsin-Madison and Wisconsin Department of Transportation (WisDOT).

There are two datasets in this study, the first is the whole reported crash dataset (WCD) which has all crashes in the state, the second dataset contains crashes only related to winter precipitation (WPCD). After eliminating incomplete and erroneous records, in total, the WCD included 291,821 crashes. And there are 31,018 records in WPCD.

Each dataset consists of four subsets: 1) general crash information regarding occurrence time, region, speed, detailed location, accident types, crash severities, driving directions; 2) environmental information when crash happened including road types, light conditions, speed limits, surface condition of the road, traffic control measures, traffic condition and road geometric characteristics; 3) vehicle data explaining vehicle type and actions, manner in which participants collided in the crash, and number of vehicles involved; 4) driver data containing driver characteristics, alcohol involved and driver behavior [23],[27].

The response variable, crash severity, is defined by WisDOT as three levels: fatal accident (FAT), injury occurred (INJ), and property damage only (PD). Additional effort was made to improve the prediction. For example, ACCDTYPE, the categorized variable descripting the type of crash based on the first harmful event, was re-categorized as three groups: SV (single vehicle involved crash), MV (multiple vehicles involved crash), and PV (bicycle & pedestrian interacting with vehicle crash). The descriptive statistics of all the variables are listed in Table I.

TABLE I
VARIABLE DEFINITION AND DATA DESCRIPTION

Variable	Description	Crash severity and frequencies							
		FAT		INJ		PD		Total	
		WCD	WPCD	WCD	WPCD	WCD	WPCD	WCD	WPCD

General crash information								
Crash time								
Morning (6:00am-12:00pm)	24.36%	40.54%	27.50%	34.47%	29.28%	32.58%	28.74%	32.96%
Afternoon (12:00pm-6:00pm)	33.77%	25.68%	44.51%	37.91%	42.04%	36.79%	42.72%	36.98%
Evening (6:00pm-0:00am)	22.98%	21.62%	19.58%	19.39%	19.11%	20.73%	19.27%	20.48%
Night (0:00am-6:00am)	18.89%	12.16%	8.41%	8.23%	9.56%	9.89%	9.27%	9.58%
Crash region								
Northwestern	16.47%	24.32%	10.01%	13.05%	11.12%	13.95%	10.83%	13.80%
Northeastern	16.47%	16.22%	17.02%	18.54%	16.32%	17.67%	16.53%	17.83%
Southwestern	24.64%	18.92%	22.52%	23.21%	24.40%	23.95%	23.86%	23.80%
Southeastern	27.06%	18.92%	41.23%	31.63%	38.54%	30.89%	39.27%	31.00%
Center area	15.36%	21.62%	9.21%	13.57%	9.61%	13.54%	9.52%	13.56%
Municipality type (crash occurred)								
City	25.47%	20.27%	57.16%	42.09%	59.76%	45.72%	58.83%	44.96%
Town	67.96%	75.68%	33.01%	47.92%	29.02%	42.88%	30.37%	43.92%
Village	6.57%	4.05%	9.83%	9.99%	11.22%	11.40%	10.79%	11.11%
Primary travel direction								
West	24.12%	12.16%	24.40%	23.85%	23.89%	22.69%	24.04%	22.89%
South	25.64%	20.27%	25.42%	26.85%	25.96%	27.41%	25.80%	27.33%
North	27.37%	31.08%	25.24%	25.70%	25.58%	26.22%	25.49%	26.13%
East	22.87%	36.49%	24.94%	23.60%	24.57%	23.67%	24.67%	23.65%
Intersection Distance (numerical value)								
Within 50 feet	63.11%	48.65%	82.25%	70.82%	83.68%	75.10%	83.16%	74.22%
50-500 feet	35.36%	45.95%	17.12%	28.13%	15.77%	24.00%	16.26%	24.84%
More than 500 feet	1.52%	5.41%	0.63%	1.05%	0.55%	0.91%	0.58%	0.94%
Accident type								
Single vehicle involved	46.23%	35.14%	30.46%	43.41%	39.75%	57.35%	37.08%	54.64%
Multiple vehicles involved	45.74%	56.76%	64.17%	54.90%	60.11%	42.62%	61.22%	45.00%
Bicycle/pedestrian involved	8.03%	8.11%	5.37%	1.69%	0.14%	0.03%	1.70%	0.36%
Location of first harmful event in relation to a roadway								
On road way	71.83%	83.78%	85.28%	77.45%	85.03%	74.46%	85.03%	75.05%
Off road way	28.17%	16.22%	14.72%	22.55%	14.97%	25.54%	14.97%	24.95%
Crash location								
Intersection related	29.00%	8.11%	49.27%	34.48%	41.40%	34.80%	43.63%	34.67%
Non-intersection related	71.00%	91.89%	50.73%	65.52%	58.60%	65.20%	56.37%	65.33%
Environmental information at crash occurrence								
Road type								
Highway	69.76%	78.38%	52.72%	66.11%	47.42%	59.52%	49.08%	60.82%
Others	30.24%	21.62%	47.28%	33.89%	52.58%	40.48%	50.92%	39.18%
Road curvature								
Curve	26.57%	42.97%	13.61%	17.09%	13.23%	19.34%	13.41%	18.92%
Straight	73.43%	57.03%	86.39%	82.91%	86.77%	80.66%	86.59%	81.08%
Road grade								
Rural highway	39.24%	54.05%	19.51%	34.23%	18.10%	30.06%	18.61%	30.91%
Urban highway	11.49%	12.16%	23.54%	21.20%	22.12%	20.29%	22.48%	20.44%
Rural street	33.77%	21.62%	19.93%	19.64%	19.61%	20.92%	19.78%	20.68%
Urban street	15.50%	12.16%	37.02%	24.93%	40.17%	28.73%	39.13%	27.97%
Road design								
Not physically divided	77.72%	72.97%	64.01%	57.35%	63.89%	58.46%	64.00%	58.29%
One-way traffic	1.04%	0.00%	3.02%	1.84%	4.47%	3.42%	4.03%	3.11%
Physically divided	21.25%	27.03%	32.97%	40.81%	31.64%	38.12%	31.98%	38.61%

Road condition								
Dry road	78.69%	1.35%	72.16%	0.91%	62.36%	0.62%	65.29%	0.68%
Wet	11.00%	2.70%	13.81%	7.74%	13.32%	4.87%	13.45%	5.42%
Snow	7.20%	82.43%	9.99%	80.24%	18.16%	85.17%	15.73%	84.23%
Icy	3.11%	13.51%	4.05%	11.11%	6.17%	9.33%	5.53%	9.68%
Light condition								
Daylight	50.38%	44.68%	73.38%	64.86%	76.53%	69.81%	75.57%	68.94%
Dusk	7.00%	2.13%	6.25%	6.15%	5.44%	5.35%	5.66%	5.48%
Dark	42.63%	53.19%	20.37%	28.99%	18.03%	24.84%	18.77%	25.59%
Road vertical								
Flat	77.58%	67.03%	27.50%	79.28%	29.28%	77.57%	29.00%	77.90%
Hill	22.42%	32.97%	27.50%	20.72%	29.28%	22.43%	28.73%	22.10%
Traffic control								
None	81.18%	93.24%	62.40%	75.12%	68.67%	75.89%	66.91%	75.79%
Sign	11.76%	1.35%	13.77%	8.67%	12.55%	9.35%	12.90%	9.20%
Signal	7.06%	5.41%	23.83%	16.21%	18.78%	14.75%	20.19%	15.01%
Posted speed								
25 or less MPH	9.83%	8.11%	24.78%	16.26%	32.36%	23.24%	30.05%	21.87%
30-55 MPH	80.62%	68.92%	67.16%	62.12%	57.48%	56.49%	60.41%	57.59%
60 or more MPH	9.55%	22.97%	8.06%	21.62%	10.15%	20.27%	9.54%	20.54%
Vehicle data								
Number of vehicles in crash								
Single	53.01%	36.49%	34.07%	42.70%	31.34%	52.63%	32.24%	50.69%
Two or more	46.99%	63.51%	65.93%	57.30%	68.66%	47.37%	67.76%	49.31%
Vehicle type								
Light	74.74%	71.62%	83.67%	78.87%	78.95%	79.04%	80.31%	78.99%
Heavy	25.26%	28.38%	16.33%	21.13%	21.05%	20.96%	19.69%	21.01%
First harmful event collision manner								
No collision	51.83%	39.19%	33.45%	43.61%	36.87%	55.63%	35.95%	53.30%
Sideswipe	4.71%	5.41%	7.80%	9.11%	16.42%	11.99%	13.85%	11.42%
Rear	6.57%	4.05%	27.27%	19.07%	24.86%	15.10%	25.47%	15.83%
Head	36.89%	51.35%	31.49%	28.21%	21.85%	17.28%	24.73%	19.45%
Drivers data								
Driver age								
24 or less	23.11%	20.27%	31.31%	30.32%	36.29%	35.81%	34.77%	34.72%
25-64	61.73%	62.16%	58.69%	61.39%	54.86%	58.23%	56.01%	58.84%
65 or more	15.16%	17.57%	10.00%	8.28%	8.86%	5.97%	9.22%	6.43%
Driver gender								
Male	75.92%	60.81%	58.32%	51.61%	62.75%	55.96%	61.52%	55.14%
Female	24.08%	39.19%	41.68%	48.39%	37.25%	44.04%	38.48%	44.86%
Driver seatbelt use								
Seatbelt is used	54.53%	58.86%	84.43%	89.59%	82.79%	89.80%	83.12%	89.70%
Seatbelt not used	45.47%	41.14%	15.57%	10.41%	17.21%	10.20%	16.88%	10.30%
Driving action								
Going straight	44.13%	64.38%	38.28%	50.77%	38.79%	43.53%	38.67%	44.98%
Changing lanes	2.36%	2.74%	2.78%	3.42%	4.53%	3.61%	4.00%	3.57%
Turning	5.49%	0.00%	17.27%	10.30%	18.32%	13.69%	17.94%	13.00%
Merging	18.42%	15.07%	7.95%	10.18%	7.32%	11.09%	7.56%	10.93%
Decelerating or stopped	4.93%	2.74%	11.30%	13.24%	14.75%	14.98%	13.67%	14.61%
Accelerating	24.67%	15.07%	22.42%	12.09%	16.28%	13.10%	18.16%	12.91%

Note. There are only classified variables in Table I. However, the original data, which contents both numerical and classified variables, are served in the statistical models. “Intersection Distance” means the distance between the crash location and the nearest intersection.

III. METHODS

A. Research Design

In this study, the data analysis is divided into two steps. For the first step, in order to guarantee a higher predictive accuracy and trustful inputs for the second step, three statistical models including Random Forest (RF), Support Vector Machine (SVM), and Lasso Regression were developed and compared for each crash severity prediction model with the two datasets. These three models have been widely used in traffic safety studies [27] as they could process high dimensional data with high predictive accuracy. There are both categorical and numerical variables in the pre-processed data, this data can be directly used by RF and Lasso models. While, for SVM models, the numerical variables in the original data need to be converted to categorical data as shown in Table I. In addition, each model has its advantages as well as disadvantages. For instance, RF models randomly select a subset of features to split at each node when growing a tree so that overfitting is supposed to avoid. SVM models lack the capability of automatically choosing the relevant factors Lasso models are good at selecting factors and measuring their coefficients. As a result, the efficient prediction model and variable importance are achieved.

In the second step, the importance (significant level) of relevant factors of the WCD and WPCD are separately ranked. By comparing the same factor of different datasets, significant levels are discovered. Next, the reasons leading to the differences are discussed.

B. Model Description and Specification

(a) Random Forest (RF)

RF algorithm is a useful approach for regression or classification [28]. RF algorithm is consisted by a series of tree predictors, which is the combination of two machine learning methods: bagging and random feature selection. Within bagging, each tree is trained on a bootstrap sample of the training data, and predictions are made by majority vote of trees or average of trees, depending on the setting. RF algorithm provides an advanced over bagged trees by way of a random forest small tweak that decreases the correlates with the trees. As in bagging, RF build a number of decision trees on bootstrapped training samples. But when building these decision trees, each time a split in a tree is considered, a random sample of m predictors is chosen as split candidates from the full set of p predictors.

The RF algorithm could be implemented by the “Random Forest” package in R software, according to the Breiman and Cutler's original Fortran code [29].

(b) Support Vector Machine (SVM)

SVM algorithm is a supervised learning method for data classification in machine learning and statistical learning. It is a kernel-based classifier. In SVM, an optimal hyperplane is chosen to separate most of the training samples into two classes [30]-[31]. It employs kernel functions to map original data to a feature space of higher dimensions and locate an optimal separating hyperplane there. SVM can be written equivalently as the following optimization criterion:

$$\min_{\alpha,\alpha_0} \sum_{i=1}^{N}(1 - y_i f(x_i)) + \frac{\lambda}{2}\alpha^T K\alpha \qquad (1)$$

where the loss function is the hinge loss and K is the matrix of kernel evaluations for all pairs of training features. In this study, linear kernel is used with fine Gaussian Kernel (γ=0.1) setting. The SVM algorithm was implemented by R software.

(c) Lasso Regression

The Lasso algorithm [32] is a form of regularized or “penalized” regression, where L1 regularization is introduced into the standard multiple linear regression procedure, using a compound cost function to optimize the regression coefficients:

$$\beta^* = \underset{\beta}{\operatorname{argmin}} \|y - X\beta\|^2 + \lambda\|\beta\|_1 \qquad (2)$$

where $\|\cdot\|_2$ and $\|\cdot\|_1$ denote the l_1 norm and l_2 norm of vectors respectively. And λ is a tuning parameter. The objective function of the Lasso consists of two parts. One is the empirical loss function and the other is the penalty which incorporates the sparsity prior of data. The singularity of the L1 at the origin induces sparse solutions. The efficient algorithm solving process makes Lasso widely used. Particularly, Friedman et al. developed the coordinate descent algorithm to obtain the solution of the Lasso. Coordinate descent algorithm updates one coordinate at a time and makes it computational efficient.

In this study, the glmnet (2.0-5) package was implemented in R, which optimizes model fitting parameters using a coordinate descent algorithm [33]. Because single leave-one-out cross-validation cycles with the coordinate descent algorithm is applied on each fold in order to find regularization parameters. The smallest average mean squared errors would be across all folds.

C. Variable Importance Ranking and Comparison

The two datasets (WCD and WPCD) with total 24 predictor variables were imported for variable importance analysis. The variable with the largest overall importance is scored 100, and all other variables have their scores relatively scaled to the best performing variable and ranged downwards toward negative 100 as shown in Table III. Also, in Table III, crash risk factors were divided into four sections. In each section, the same factors of the WCD and WPCD are listed next to each other for comparison.

IV. RESULTS AND DISCUSSION

A. Predictive Performance

Cross-validation (CV) is the one of the most common methods to estimate the model prediction error [34]. Model prediction errors of RF, SVM, and Lasso were 1.64, 1.71, and 1.63 respectively. The RF and Lasso model had the closed CV prediction errors which were substantially lower than the SVM. This indicated that the RF and Lasso models could provide better predictive accuracy than the SVM model.

The superiority of the RF and Lasso are also shown in Table II. The predictive performance of three models described in section 3 (RF, SVM, Lasso) was evaluated. The winter precipitation related crash data was best predicted with Lasso (85.05% and 88.17%). And the whole crash data was best predicted with RF (87.85% and 98.45%).

Lasso and RF models both showed a superior performance. However, as described in section 3.2, the variables were able to be befittingly selected and their specific coefficients were derived through Lasso. Finally, the variable importance was measured according to the coefficients of Lasso model.

TABLE II

RESULTS COMPARISON OF THREE PREDICTION MODELS

Percentage of correctly predicted instances	RF		SVM		Lasso	
	Test	Training	Test	Training	Test	Training
The whole crash data	87.85	98.45	83.71	89.43	82.48	84.87
Winter precipitation related crash data	82.42	87.84	80.33	82.81	85.05	88.17

Note. Testing dataset is 20% of the whole dataset according to the default statistic method.

B. Variable Impact Analysis and Comparison

These two different datasets, each with 24 predictor variables, were imported for variable importance analysis. The relative variable importance is listed in Table III. All 24 predictor variables were selected in the WCD, while, 17 predictor variables were maintained in the WPCD. To make the comparison at a normalized scale, the most relevant factor (Accident type) was valued equal to 100 referring to section 3.3. The rest variables were calculated according to their coefficient values. In this study, these variable importance scores were used for quantitative interpretations of variable influence or significant levels. The value in "Gap" is the absolute value of variables' differences between WCD and WPCD models.

In order to make a logical interpretation, the influential variables are discussed with four separate groups according to the four subsets of original dataset. Table III shows that there are more changes of variables impacts in "Environmental information at crash occurrence" between WCD and WPCD. Two of three variables in "Vehicle data" also make a great difference. Only one single variable in the rest of the two groups shows a difference over 10.

TABLE III

VARIABLE IMPORTANCE SCORE RANKING

Subset	Variable	Score		
		WCD	WPCD	Gap (>10)
General crash information	Accident type	100.000	100.000	
	Location of first harmful event in relation to a roadway	2.969	-	
	Primary travel direction	1.946	10.154	
	Crash location	1.518	21.745	20.227
	Crash region	0.588	-	
	Municipality type (crash occurred)	-0.564	-	
	Crash time	0.191	-	
	Intersection Distance (numerical value)	0.036	0.019	
Environmental information at crash occurrence	Road type	33.044	27.461	
	Road curvature	20.520	40.028	19.508
	Traffic control	17.737	-	17.737
	Road grade	-10.640	-6.938	
	Road vertical	6.453	37.493	31.04
	Light condition	4.495	-	
	Road condition	3.361	21.159	17.798
	Road design	2.827	-	
	Posted speed	0.820	1.086	
Vehicle data	First harmful event collision manner	21.199	17.827	
	Vehicle type	-20.675	4.030	24.705
	Number of vehicles in crash	2.395	26.289	23.894

Drivers data	Driver seatbelt use	-14.105	-15.535	10.914
	Driver gender	9.158	20.072	
	Driver age	0.594	0.828	
	Driving action	-0.361	-4.677	

Note. "-" means no obvious relevant.

(a) General crash information

There are six variables under General crash information in Table III. However, five of them have slight relevant (scores from -0.564 to 2.969) with crash severity in WCD, three of them, Location of first harmful event, Crash region, and Municipality type even have no relevant at all in WPCD dataset.

Obviously, Accident type is the most important factor contributing to crash severity outcomes for both WCD and WPCD. But there were more Multiple vehicles involved instances (WCD is 45.74% and WPCD is 56.76%) involving in FAT crashes in WPCD shown in Table I. It is found that the fatal crashes are more likely to involve with multiple vehicles under winter precipitation weather conditions. Also, this means there is a bigger chance to injure more drivers and passengers as well as to cause traffic chaos.

A significant difference of relevant of Crash location in winter precipitation crash data could be found in Table I and III Non-intersection related FAT crash rate increases from 71.0% (WCD) to 91.89% (WPCD) and INJ crash rate raise from 50.73% (WCD) to 65.52% (WPCD). In addition, the variable importance score of Crash location is 21.745 in WPCD which is 13.32 times more comparing to WCD's 1.518. This is reasonable since winter precipitation, such as rain, snow and ice may force drivers to be more cautious near intersections. On the contrary, most drivers keep the same driving habits and speed in winter precipitation when not near an intersection, ignoring that more effort is needed to maintain normal vehicle operations. As a result, it is relatively safer at intersection than the non-intersection parts of road under winter precipitation.

Similar conclusions could also be drawn on travel direction, as it is shown in Table I that directions West (36.49%) and North (31.08%) tend to induce more fatal crash, with a probability 5%-10% higher than that for its counterpart. Also, variable importance score of Primary travel direction in WPCD (10.154) is 5 times more than the counterpart in WCD (1.946). Since more than 71% winter winds are either north winds or west winds (National Weather Service Climate). The winter precipitation interacting with the winds may have an effect on transportation safety. It is suspicious that there are more FAT crashes in these two directions. Therefore, further study is needed for more evidences about the relationship of crashes with travel directions under adverse weather conditions.

(b) Environmental information at crash occurrence

There are nine variables associated with Environment information at crash occurrence in Table III. Four variables' absolute value of WCD are above 10 (from 10.640 to 33.044), and the same results apply to WPCD (from 21.159 to 40.028). Table I shows that winter precipitation is likely to result in FAT crash or INJ crash on the highway, with the increasing rate from 69.76% (WCD) to 78.38% (WPCD) for FAT and from 52.72% (WCD) to 66.11% (WPCD) for INJ. In which, rural highway is affected most, with FAT crash rising from 39.24% (WCD) to 54.05% (WPCD) and INJ crash rising from 19.51% (WCD) to 34.23% (WPCD). In addition, Table III shows that both Road type and Road grade maintain a similar high relevance with crash severity.

It is found in Table III that the importance score of Road curvature is changing from 20.520 to 40.028. The worse situation can be seen with Road vertical, increasing from 6.453 to 37.493, and on Road condition, increasing from 3.361 to 21.159. All these three significant changes are associated with poor road friction. Driver responses are insufficient to offset the risk of driving on adverse road surface, which make vehicle handling more difficult with reduced traction, ultimately leads to the increase of crash rates.

Traffic control is also a main factor related to crash severity in WCD, with a score of 17.737 in Table III. On the contrary, Traffic control was not selected by Lasso model in WPCD. Table I shows 93.24% of fatal crash happened in none traffic control sections. In normal weather condition, some drivers take risk to violate the traffic controls like signal and sign. However, this group of drivers are supposed to comply with the traffic controls in winter precipitation condition.

(c) Vehicle data

"Vehicle type" have a negative score equal to -20.675, which means light vehicles are more likely to experience fatal crash than the heavy vehicles in WCD. But as the result of winter precipitation, the "Vehicle type" get a positive score of 4.030, which means the heavy vehicles are more relative with serious crash. It is known that heavy vehicles have larger inertia than light vehicles, plus the poor road friction, they are hardly flexibility controlled. in Further studies should discuss if individual low speed limit is required for heavy vehicles.

The different roles that the "Number of vehicles in crash" play in influencing crash severity between WCD and WPCD cannot be neglected. It is shown through comparisons in Table III that more than two vehicles are likely to be involved in crash, with the score from 2.395 to 26.289. As a result, crashes of multiple vehicles are more likely to appear under winter precipitation condition with an increasing severity. This conclusion is consistent with the growth rate of FAT crash (from 46.99% of WCD to 63.51% of WPCD).

(d) Driver data

As is shown in Table III, there is little gap between scores of WCD and WPCD in terms of Driver seatbelt use and Driver age, with the coincident results show in Table I. Age and seatbelt usage barely show diverse impacts between normal weather and winter precipitation.

In contrast, different genders react differently when driving in winter precipitation (score from 9.158 to 20.072). Female drivers are more likely to be implicated by winter precipitation than male drivers. According to Table I, female drivers have higher fatal crashes. To be highlighted, the score of Driving action has been enlarged 12.96 times. Table I shows, when driving straight, 20.25% more FAT crashes happened in WPCD compared with WCD. After rechecking the Going straight FAT records, it is found that 71.84% crashes are associated with higher posted speed limits. Therefore, drivers are less likely to reduce speed when going straight in highway with high posted speed limits even during precipitation. Also, another possible explanation is drivers overestimate their driving skills, at the same time underestimate the risk of crash during frequently experienced weather conditions.

C. Influencing Variable Impacts on Autonomous Vehicles

The intrinsic features of autonomous vehicles help them avoid making unnecessary mistakes and hence improving the safeness of the vehicle. For example, drivers are not needed anymore, so all the driver related factors will not contribute to the seriousness of the crash. We summarized the influence level of different factors when applied to self-driving vehicles in Table IV. As shown in the table, the negative impacts of all the variables will be reduced since self-driving vehicles strictly follow traffic rules and take as much information as possible into consideration. With the help of LiDAR, cameras and other sensors, autonomous vehicles are able to keep track of multiple vehicles at the same time. Besides, self-driving vehicles have a full view of their operating status, hence they are more aware of the driving status than human drivers [35]-[37]. Due to abovementioned reasons, the negative effects caused by different crash variables that are under general crash information, environmental information and vehicle data categories can be reduced. Obviously, self-driving vehicles are not an answer for all the issues. These is also a list of variables that requires further study. When the vehicle is at road intersections and there are multiple vehicles and people involved, how to choose appropriate actions that can minimize the severity of the accident is also an open question for automobile companies. Furthermore, current technologies are not mature enough to allow autonomous vehicles driving in real world environments, especially when the light condition is not good and the road condition is uncertain. Researchers and companies need to put lots efforts in these fields to make vehicles that can further reduce or even eliminate fatal accidents.

TABLE IV
VARIABLE IMPORTANCE SCORE RANKING

Subset	Variable	Self-Driving vehicle	
		Impact Reduced	Need Further Study
General crash information	Accident type	Yes	Yes
	Location of first harmful event in relation to a roadway	Yes	-
	Primary travel direction	Yes	Yes
	Crash location	Yes	Yes
	Crash region	Yes	-
	Municipality type (crash occurred)	Yes	-
	Crash time	Yes	Yes
	Intersection Distance (numerical value)	Yes	-
Environmental information at crash occurrence	Road type	Yes	-
	Road curvature	Yes	-
	Traffic control	Yes	-
	Road grade	Yes	-
	Road vertical	Yes	-
	Light condition	Yes	Yes
	Road condition	Yes	Yes
	Road design	Yes	-
	Posted speed	Yes	-
Vehicle data	First harmful event collision manner	Yes	-
	Vehicle type	Yes	-
	Number of vehicles in crash	Yes	Yes
Drivers data	Driver seatbelt use	Yes	-
	Driver gender	Yes	-
	Driver age	Yes	-

Driving action	Yes	-

By comparing the score between WCD and WPCD, we suggest the following prioritized list as a guidance for researchers in this community and automobile makers to help them improve when designing algorithms for self-driving vehicles. The score difference is very large between WCD and WPCD for road condition hence this is a variable that requires further study. The score differences of crash location and number of vehicles are also significant. These two variables directly affect the severity level of car accidents, so we suggest focusing on these two variables as well. Last but the least, although there is no comparison between WCD and WPCD of the light condition variable, this is a key factor to autonomous vehicles since it directly relates to how well the vehicle can understand of the surrounding environments. We put the light condition variable in our prioritized list as well.

V. Conclusion

Crashes under the winter precipitation condition are a major source of fatal crashes. In this work, we conducted in depth research on crash severity distribution and heterogeneous factor impacts on crash severities in winter precipitation are of practical importance. We analysis three-year crash data to evaluate how different factors affect the severity levels of crashes, especially under extreme winter weather conditions. Based on our analysis results, we prepared a prioritized list with several variables that needs to be considered for autonomous vehicles in order to reduce the severity level of crash or avoid accidents.

Random Forest, SVM, and Lasso models are popular and effective prediction methods which have been widely used in traffic safety investigations. Based on the two 3-year crash datasets (the whole crash dataset and winter precipitation crash dataset) from Wisconsin Traffic Operations and Safety Laboratory (TOPS) at the University of Wisconsin- Madison and Wisconsin DOT, Random Forest, SVM, and Lasso models are developed and applied to predict crash severity. The crash injury severities are classified as three levels: fatal accident (FAT), injury occurred (INJ), and property damage only (PD). Then the prediction performance of these three models are compared and examined. Lasso model not only shows a standout outcome but also offers a detailed coefficients and parameters. As a result, referring to the results of Lasso model, variable importance is scored and ranked. According to the four original data subsets (General crash information, Environmental information at crash occurrence, Vehicle data, and Driver data), the different variable impacts, the counterpart between the whole crash dataset and winter precipitation crash dataset, are compared and discussed. Finally, the reason why there are different risk factors between the whole database and winter precipitation crash dataset was analyzed.

Differences of variable impacts have taken place between the two datasets (WCD and WPCD). It is found that variables in Environmental information at crash occurrence have the most significant effects on driver injuries and fatalities in winter precipitation crashes. Road surface condition is the most relevant factor that causes fatal crash, especially for these road segments with more horizontal curves and terrain changes. Dynamic electric waring screens are suggested to be applied far away before the curves and hills. In further study, the optimal position of waring sign will be researched. Moreover, factors in “Vehicle data” also have a great impact. Heavy vehicles are much harder to operate on wet or icy road, as a result, they experience a high fatal crash rate. Variety posted speed limits should be set in slow lane for heavy vehicles like bus, coach, and truck during and after snowfall. At the same time, overtaking should be forbidden through curve, merging and mountain area, especially for heavy vehicles. Multiple vehicles involved crashes aggravate the crash severity under winter precipitation weather condition. Drivers are always too self-overestimated to keep enough safety gap with front vehicle even under precipitation. Female drivers have a higher likelihood to suffer injuries and deaths when driving in winter precipitation. Female drivers may underestimate the risk of driving under adverse weather condition. Improper driving actions, including sudden acceleration, deceleration, brake and turn, also increase injury or fatality potential. Based on evaluation results, we believe future autonomous vehicles can reduce the severity level of crashes and even avoid potential crashes from happening since efficient algorithms work better than human drivers and lots of unnecessary mistakes can be avoided. Future studies are also needed to continue discovering the influence factors which are unique to self-driving vehicles.

Acknowledgement

This work was supported in part by the Jilin Province Transportation Science and Technology Plan Project (2014-1-8). Thanks the generous support provided by the TOPs lab at UW-Madison.

References

[1]. Abdel-Aty, M., K. Haleem. Analyzing angle crashes at unsignalized intersections using machine learning techniques. *Accident Analysis and Prevention*. 2011: 43(1): 461-470.

[2]. National Highway Traffic Safety Administration, U.S. Department of Transportation. Quick Facts 2016, DOT HS 812 451. October 2017.

[3]. Eisenberg, D., K. E. Warner. Effects of snowfalls on motor vehicle collisions, injuries, and fatalities. *American journal of public health*.2005: 95(1): 120-4.

[4]. Black, A., Mote T. Effects of Winter Precipitation on Automobile Collisions, Injuries, and Fatalities in the United States. *Journal of Transport Geography*. 2015:48:165-175.

[5]. Melissa Bauman. (2017, Nov 7) "Why Waitng for Perfect Autonomous Vehicles May Cost Lives". [Online]

Available: https://www.rand.org/blog/articles/2017/11/why-waiting-for-perfect-autonomous-vehicles-may-cost-lives.html

[6]. Qi, B., Zhao, W., Zhang, H., Jin, Z., Wang, X., & Runge, T. (2019, July). Automated Traffic Volume Analytics at Road Intersections Using Computer Vision Techniques. In 2019 5th International Conference on Transportation Information and Safety (ICTIS) (pp. 161-169). IEEE.

[7]. Aldibaja, M., Naoki S. and Keisuke Y. "Robust intensity-based localization method for autonomous driving on snow–wet road surface." *IEEE Transactions on Industrial Informatics* 13, no. 5 (2017): 2369-2378

[8]. Aldibaja, M., Noaki S., and Keisuke Y. "Improving localization accuracy for autonomous driving in snow-rain environments." *In 2016 IEEE/SICE International Symposium on System Integration (SII)*, pp. 212-217. IEEE, 2016

[9]. Qiu, H., Ahmad, F., Bai, F., Gruteser, M., and Govindan, R.. "Augmented vehicular reality: Enabling extended vision for future vehicles." *In Proceedings of the 18th International Workshop on Mobile Computing Systems and Applications*, pp. 67-72. ACM, 2017.

[10]. Zang, S., Ding, M., Smith, D., Tyler, P., Rakotoarivelo, T. and Kaafar. M. A. "The Impact of Adverse Weather Conditions on Autonomous Vehicles: How Rain, Snow, Fog, and Hail Affect the Performance of a Self-Driving Car." *IEEE Vehicular Technology Magazine* 14, no. 2 (2019): 103-111.

[11]. Lee, U. U., Jung, J. J., Shin, S. S., Jeong, Y. Y., Park, K. K., and Shim, D. D. H. "EureCar turbo: A self-driving car that can handle adverse weather conditions." *In 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 2301-2306. IEEE, 2016

[12]. Chang, L.-Y., H.-W. Wang. Analysis of traffic injury severity: an application of non-parametric classification tree techniques. *Accident analysis and prevention*. 2006:38(5): 1019-27.

[13]. Andrey, J., M. Christie, S. Michaels. Toward a National Assessment of the Travel Risks Associated with Inclement Weather. 2005.

[14]. Agarwal, M., T. H. Maze, R. Souleyrette. Impacts of weather on urban freeway traffic flow characteristics and facility capacity. *In Proceedings of the 2005 mid-continent transportation research symposium*, pp. 18-19. 2005.

[15]. Roh, H.-J., S. Datla, S. Sharma. Effect of Snow, Temperature and Their Interaction on Highway Truck Traffic. *Journal of Transportation Technologies*.2013: 3(1): 24-38.

[16]. Li, Z., Y. Li, P. Liu. Development of a variable speed limit strategy to reduce secondary collision risks during inclement weathers. *Accident Analysis & Prevention*. 2014:72:134-145.

[17]. Zhao, W., Xu, L., Xi, S., Wang, J. and Runge, T. A Sensor-Based Visual Effect Evaluation of Chevron Alignment Signs' Colors on Drivers through the Curves in Snow and Ice Environment. *Journal of Sensors*, 2017a.

[18]. Zhao, W., Xu, L., Bai, J., Ji, M. and Runge, T. Sensor-based risk perception ability network design for drivers in snow and ice environmental freeway: a deep learning and rough sets approach. *Soft computing*, 2017b: pp.1-10.

[19]. Zhao, W., Xu, L., Dong S. Z., Qi, B., and Qin, L. "Improving transfer feasibility for older travelers inside high-speed train station." Transportation Research Part A: Policy and Practice 113 (2018): 302-317.

[20]. Hao, Y., Xu, L., Qi, B. Wang, T., and Zhao, W. "A Machine Learning Approach for Highway Intersection Risk Caused by Harmful Lane-Changing Behaviors." In CICTP 2019, pp. 5623-5635. 2019.

[21]. Liu, P., Qi, B., & Banerjee, S. Edgeeye: An edge service framework for real-time intelligent video analytics. In Proceedings of the 1st International Workshop on Edge Systems, Analytics and Networking, ACM, 2018 NY, USA, 1-6.

[22]. Ma, X., Ding, C., Luan, S., Wang, Y. and Wang, Y. Prioritizing influential factors for freeway incident clearance time prediction using the gradient boosting decision trees method. *IEEE Transactions on Intelligent Transportation Systems*, 2017,18(9), pp.2303-2310.

[23]. Ding, C., Ma, X., Wang, Y. and Wang, Y., 2015. Exploring the influential factors in incident clearance time: disentangling causation from self-selection bias. *Accident Analysis & Prevention*. 2015:85:58-65.

[24]. Qiu, L., and W. Nixon. Effects of adverse weather on traffic crashes: systematic review and meta-analysis. *Transportation Research Record Journal of Transportation Research Record*. 2008:2055(2055):139-146.

[25]. Andrey, J. Weather Information and Road Safety. *Institute for Catastrophic Loss Reduction*. 2001.

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

[27]. Chen, C., G. Zhang, Z. Qian. Investigating driver injury severity patterns in rollover crashes using support vector machine models. *Accident Analysis & Prevention*. 2016:90:128-139.

[28]. Breiman, L. Random Forests - Springer. *Machine Learning*. 2001:45(1):5-32.

[29]. Liaw, A., M. Wiener, M. Andy Liaw. Breiman and Cutler's Random Forests for Classification and Regression Description Classification and regression based on a forest of trees using random inputs. 2015.

[30]. Li, Z., P. Liu, W. Wang. Using support vector machine models for crash injury severity analysis. *Accident Analysis & Prevention*. 2012:45:478–486.

[31]. Yu, R., M. Abdel-Aty. Utilizing support vector machine in real-time crash risk evaluation. *Accident Analysis & Prevention*.2013:51:252-259.
[32]. Tibshirani, R. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*. 196:58(1): 267–288.
[33]. Friedman, J., T. Hastie, N. Simon, et al. Package “glmnet” Title Lasso and Elastic-Net Regularized Generalized Linear Models. 2016.
[34]. John Lu Z Q. The elements of statistical learning : data mining, inference, and prediction. *Journal of the Royal Society*. 2010:173(3):693-694.
[35]. Zhao, W., Yin, J., Wang, X., Hu, J., Qi, B., & Runge, T. (2019). Real-time vehicle motion detection and motion altering for connected vehicle: algorithm design and practical applications. Sensors, 19(19), 4108.
[36]. Qi, B., Liu, P., Ji, T., Zhao, W., & Banerjee, S. (2018, December). DrivAid: Augmenting driving analytics with multi-modal information. In 2018 IEEE Vehicular Networking Conference (VNC) (pp. 1-8). IEEE.
[37]. Zhao, W., Yin, J., Wang, X., Hu, J., Qi, B., & Runge, T. (2019). Real-time vehicle motion detection and motion altering for connected vehicle: algorithm design and practical applications. Sensors, 19(19), 4108.