2023 International Wireless Communications and Mobile Computing (IWCMC) | 979-8-3503-3339-8/23/$31.00 ©2023 IEEE | DOI: 10.1109/IWCMC58020.2023.10182749

Exploring the Risk Factors Influencing the Road Accident Severity: Prediction with Explanation

Mohammed Akallouch*, Khalid Fardousse*, Afaf Bouhoute*, Ismail Berrada†
*Faculty of Sciences Dhar EL Mahraz, Sidi Mohamed Ben Abdellah University, Morocco
{firstname.lastname}@usmba.ac.ma
†School of Computer Sciences, Mohammed VI Polytechnic University, Morocco
{firstname.lastname}@um6p.ma

***Abstract*—Prediction and analysis of the severity of road traffic accidents have become a top priority and concern for road traffic safety. Reducing the accident risk and traffic congestion caused by accidents is still one of the most critical issues in the intelligent transportation system (ITS). This paper presents an effort to analyze the factors that influence the severity of accidents with different environmental experiences. This study is based on a public dataset of 2.25 million cases of car traffic accidents in the United States. Methodologically, we have adopted gradient boosting algorithms (i.g. XGBoost, LightGBM, and CatBoost), which have been recognized as robust classifiers, especially in handling feature values of categorical variables. In addition, we employed the SHapley Additive explanation (SHAP) technique to understand the black box outputs. This study identifies risk factors, including weather, time, spatial, and road analysis. Also, a detailed interpretation of models output was considered to understand geospatial safety effects at the micro, meso, and macro levels. The experiment results suggest that the importance of risk factors varies across levels in different severities. These results filled research gaps by conducting a longitudinal analysis of environmental factors in accident severity. Meanwhile, it can provide helpful insights for personal travelers and government agencies.**

***Index Terms*—XAI, Traffic Accident, Accident severity, Road Traffic, ITS**

I. Introduction

Rapid urbanization and the increase in the number of vehicles and drivers have resulted in many serious traffic-related issues (e.g. congestions, crashes), making traffic safety and control a challenging task. According to the World Health Organization (WHO), around 1.24 million people were killed and about 50 million people were injured in traffic accidents annually (i.e., over 3,300 people are killed and 137,000 are injured every day) [1]. Even with the governments efforts to address this worldwide problem and improve road safety, the results are still far from satisfactory. The ability to analyze and predict potential accidents (e.g., how, when, or where) has now become a necessity for transportation administrators as well as individual travelers, especially with the high availability of traffic data.

Road traffic accident analysis is recently an active research topic in the road safety research, and is therefore among the key elements of intelligent transportation systems. Identifying the crash-contributing factors and investigating their impact on accident frequency or severity are initial steps in road safety improvement. The factors influencing the frequency and severity of road accidents are mainly divided into three categories [2]: driver behavior (i.e., drink-driving, speed, failure to understand signs, etc.), vehicle characteristics (i.e., vehicle type, vehicle age, tire condition, etc.), and environmental conditions (i.e., geometric characteristics of the road, traffic flow, weather condition, etc.). Throughout the literature, researchers have associated road accidents with road geometric conditions and road user behavior. Rolison et al. have conducted a comprehensive study, using questionnaire-based data, to investigate the human factors in traffic accidents [3]. As a result of their investigation, the authors presented a detailed discussion of the main driver-related factors contributing to accidents, and their relation with driver characteristics such as age and gender. In addition to accident frequency, researchers have also investigated factors contributing to accident severity.

Over the last years, various approaches have been adopted for accident severity analysis and risk factors identification. Traditional approaches were typically based on statistical models [4]–[6]. The statistical approaches are not efficient enough to handle complex data, especially when the data are high dimensional with multiple independent variables. To overcome these limitations, recent works started applying Machine Learning (ML) for accident severity prediction [7], [8]. However, one major constraint of most ML techniques is that they typically work like a black-box. This makes the results of analysis hard to interpret, since the relation between accident severity and descriptive variables is not directly defined. In recent years, explainability of AI/ML models has become a serious issue. Explainable Artificial Intelligence (XAI) methods have been proposed to enable interpretability of ML models and provide a better understanding of the relational link between model outputs (i.e. predictions) and inputs (i.e. explanatory variables) [9]–[11].

In this paper, we propose a new analytical study to predict traffic accident severity and investigate in-depth the factors affecting it at three spatial levels (i.e. micro, meso and macro). To this end, a large dataset of accident data [12] with a diversity of factors (e.g. spatial, time and environment data) is used. The accident severity prediction is formulated as a classification task, where the goal is to predict the level of severity. Since we are dealing with a classification problem,

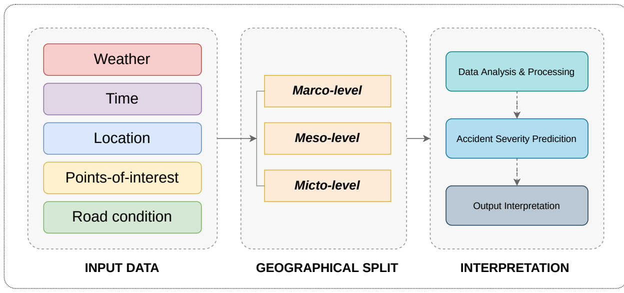


Fig. 1: Dataset description

we implemented boosting algorithms, which are commonly known as strong classifiers. Then, to explain the model predictions and analyze the contributing factors, the well known XAI technique, Shapley Additive exPlanations (SHAP) is applied. The SHAP approach was developed for the global and local explanation of ensemble trees for binary and multiclass classifiers to improve prediction outputs' accuracy [10].

To summarize, the main contributions of the paper are as follows:

- A large scale dataset of US traffic accident, containing different types of factors, namely spatial, time, weather, and road information is used. An additional dataset (PeMS) is used to investigate the impact of traffic flow on the accident severity.
- A comparison of three boosting algorithms, namely LightGBM, XGBoost and CatBoost, for predicting accident severity levels is performed.
- An in-depth analysis to investigate, at the micro, meso and macro levels, the effect of various factors on the accident severity level, using SHAP, is presented.
- An analysis by month to extract relevant factors in each month

II. Methodology

For an efficient analysis that significantly impacts the final decision made by authorities, there are two main approaches: reactive and proactive. Our paper is in the second class and implies anticipating the major factors of most accidents while reactive strategies deal with an unexpected situation after it happens. However, to design an effective accident risk analysis system, two main components should be considered: the prediction and the interpretation. Prediction deals with constructing efficent models that can predict the accident severity and then interpret these models' outputs to extract the major factors.

Technically, the approach adopted for analyzing the risk factors affecting the accident severity level comprises two major steps. First, three gradient boosting based algorithms are used for predicting the severity level. These algorithms are widely known to be among the most powerful and robust classifiers, especially when handling categorical features. Then, SHAP is used to visualize and assess the impact of each risk factor on the accident severity level. In our experiments, the SHAP values are derived using TreeSHAP, a method provided by the SHAP for tree-based models. Considering the important role of geographical zone in traffic analysis, our approach analyzes the risk factors at three different geographical levels, namely the micro-level, the meso-level, and the macro-level. These levels have been distinguished as follows:

- Macro-level: At this large level of analysis, traffic accident data of the whole country is analyzed without considering any other spatial data (e.g. state, street). We focus on time analysis, weather analysis, and points of interest.
- Meso-level: At this intermediate level of analysis, in addition to the data considered at the macro-level, additional features about geographical regions, such as state, county, and city, are added. This provides risk factors analysis by geographical region.
- Micro-level: This level of analysis covers a more narrow area since it provides analysis considering also street data. The street feature containing unique IDs was considered to analyze the accidents places.

At each of the aforementioned level, the classification algorithms are applied for predicting the severity level and the SHAP values for analyzing the contribution of the used features. In this analysis, we evaluated three classification algorithms that belong to the family of gradient boosting decision trees (GBDTs), namely eXtreme Gradient Boosting (XGBoost), Light Gradient Boosting Machine (LightGBM) and Categorical Boosting (Catboost).

III. Datasets

As most other works consider only existing standard data without taking into account the impact of traffic flow information, which is decisive in predicting and exploring accident risk, we consider traffic flow as an essential component in our work. We used the Caltrans Performance Measurement (PeMS) dataset to combine traffic flow with US-Accidents in the state of California (Due to the limit of PeMS data).

A. US-Accident Dataset

US-Accidents Dataset [12] is a large-scale public dataset of car accidents that occurred in 49 states of the United States (US). The dataset consists of about 2.25 million cases of traffic accidents, that took place between 2016 and 2020. The data was collected from different data providers, using multiple APIs that provide streaming traffic incident data. This data provided by these APIs is captured in a continuous way by a variety of entities (e.g. department of transportation, traffic sensors and cameras). Each accident in the dataset is described by different attributes, which are divided into 4 categories of information: location, time, weather and road information (Fig.1).

B. PeMS Dataset

The traffic flow study was analyzed by using a dataset from the Caltrans Performance Measurement System (PeMS) that was collected from highways [13]. The PeMS traffic flow dataset presents real-time traffic data on highways collected by various sensors across the major areas of all California. The dataset includes over ten years of historical analytical data, allowing planners, engineers, and the government to precisely analyze system performance throughout most California metropolitan freeways and provide informed suggestions for possible changes. In addition, the dataset integrates a wide variety of information, including vehicle classification, average traffic flow, average speed, VMT, VHT, and VHD.

IV. Experiments and Results

In this section, the experiments performed for severity prediction and risk factors analysis are presented.

TABLE I: Models performance

Model	Level	Precision	Recall	F-score
XGboost	`Micro-Level`	0.81	0.83	**0.80**
	`Meso-Level`	0.78	0.81	**0.78**
	`Macro-Level`	0.77	0.80	**0.77**
	`Traffic Flow Analysis`	*0.89*	*0.92*	*0.90*
CatBoost	`Micro-Level`	0.76	0.80	0.75
	`Meso-Level`	0.74	0.78	0.72
	`Macro-Level`	0.74	0.78	0.69
LightGBM	`Micro-Level`	0.79	0.77	0.69
	`Meso-Level`	0.77	0.76	0.67
	`Macro-Level`	0.59	0.74	0.65

A. Model performance assessment

As mentioned before the accident severity is predicted using three ML classifiers. For model development, we used the stratified sampling method to split the data into the training (80%), and the testing (20%) sets. The resulting training (resp. testing) set for each level (i.e. micro, meso and macro) consists of a total of 1m (resp. 26k) samples. In order to find the optimal hyper-parameters for the used models, 5-fold cross-validation was performed. To assess the performance of the models, well-known evaluation metrics are used, namely weighted-average precision, recall, and F1-score. These metrics are calculated using Equations 1, 2 and 3.

$$Precision = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} \tag{1}$$

$$Recall = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} \tag{2}$$

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{3}$$

We ran each model three times (i.g., three levels) and reported the average results. To better compare different classifiers, Table I shows the average results of each model at each level, by separately weighted-average precision, recall, and F1-score.

The performance results are presented in Table I. Comparing the scores values at the different levels, we can notice small differences across the levels with the highest score always achieved at the micro level. The F1-scores obtained for each model show that the best values are achieved by XGBoost (0.80,0.78 and 0.77 for levels micro, meso and macro respectively), followed by CatBoost (0.75, 0.72 and 0.69 for levels micro, meso and macro respectively) and LightGBM (0.69, 0.67 and 0.65 for levels micro, meso and macro respectively). Given the optimal performance of XGboost compared to CatBoost and LightGBM, we choose to use it for the risk factor analysis perforemed in the next subsection.

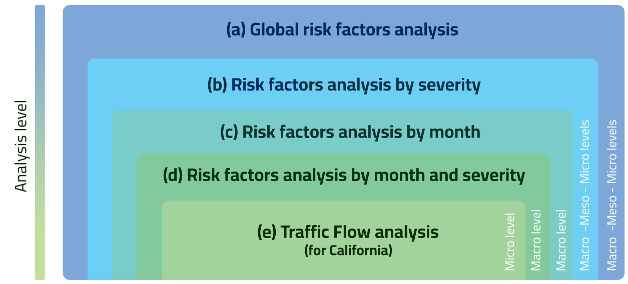


Fig. 2: Factors analysis methodology

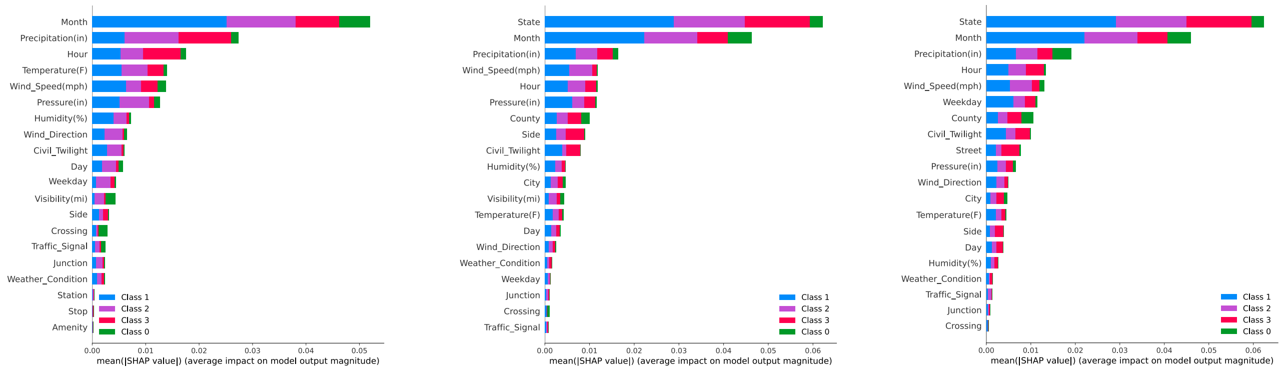
B. Model Interpretation for risk factors analysis

To perform an efficient analysis and understand the risk factor in different scenarios, we use six strategies to analyze the risk factors. Figure. 2 summarizes our analysis strategies. First, we start with a global factors analysis on three levels, macro, meso, and micro. In this phase, we extract the significant global risk factors influencing road accidents. Then we focus on factors that affect each accident's severity at all levels. In addition, we have provided an analysis by month to extract features during the year for both global and by severity (most serious severity). Finally, an additive study was provided for the impact of traffic flow on road accidents.

1) Global risk factors analysis: As mentioned earlier, the analysis conducted in this study uses SHAP to interpret the prediction results of the XGBoost model. At this point of our analysis, we are interested in a global interpretation of the model results, through which we try to answer the following two questions: At each level, (1) What are the important factors leading to accidents? What are major factors influencing accident severity?. These questions can be answered by analyzing the plots of SHAP feature importance illustrated in Fig.3. These SHAP plots show the magnitude of the attribution of each feature (risk factor) to the accident severity prediction model's outputs. The different colors represent the various accident severity levels. The risk factors are ranked in descending order (from top to bottom) based on the mean absolute SHAP values.

Comparing the feature importance at the macro-level illustrated in Fig.3a, we can see that the month and precipitation are the most significant factors contributing to accident occurrence, followed by hour, temperature, pressure, wind speed, humidity, weekday, wind direction, and day. On the other hand, we can notice that factors related to points of

Fig. 3: Global features importance of different levels.



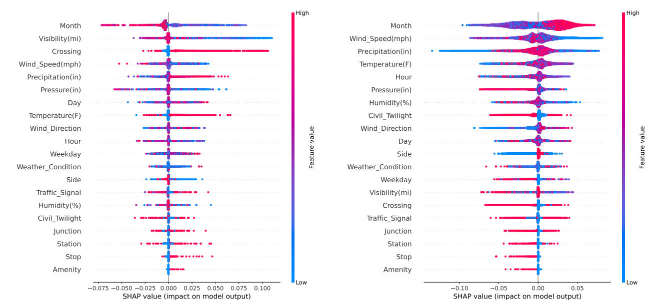
(a) Macro level. (b) Meso level. (c) Micro level.

interest (e.g. crossing, station, stop, and amenity) have the most negligible impact on accident severity. The plots for the meso and micro levels (Fig.3b and Fig.3c) shows similar results. The most significant factor is the state feature followed by month, precipitation, hour, weekday, pressure, side, and county. We also notice an influence of the street information on the accident severity at the micro-level. Summarizing the presented SHAP results, we can conclude that factors related to time, weather and geographical location are most likely to contribute to accident severity while factors representing information about points of interest make less contribution.

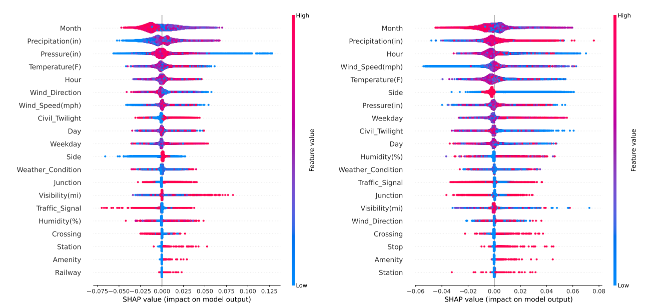
2) Analysis of risk factors by severity class: In this section, we will highlight in detail the factors that contributed to the accident for each severity class for all levels were mentioned in previous study. Fig.4, Fig.5, and Fig.6 showing in severity factors in detail for macro-level, meso-level, and micro-level, respectively. These figures are represented by the mean absolute SHAP values (i.e., importance). The color bar in each sample provides more details on how each risk factor affects the accident severity. In addition, the red dots mean higher values for a risk factor, and the blue dots indicate lower risk.

a) Macro level: By plotting the SHAP values of the macro level in Fig.4a, Fig,4b, Fig.4c, and Fig.4c of severity, 0, 1, 2, and 3, respectively, we provide a detailed look at which factors are most important and understand them. We focus only on the top features to analyze the relevant factors for each severity. Fig.5 suggests that the impact (e.g., blue and red dots) of the month feature on accident severity is different for severity levels. However, the high month value (e.g., November and December) is relevant at severity class 1, while the low month value is more important at severity levels (0, 2, and 3). The visibility factor is suggested in Fig.4a, the lower value of visibility (e.g., weak visibility) has an important impact at level 0. In addition, the higher value of precipitation (i.g, heavy rainfall) is more important at level 3 of the severity. In comparison, the lower precipitation value (i.g, weak rainfall) decreases the accident severity. Taking the atmospheric pressure factor into consideration, the smallest pressure value negatively affects the severity level (the higher

Fig. 4: Macro level.



(a) Macro level: Severity class 0 (b) Macro level: Severity class 1

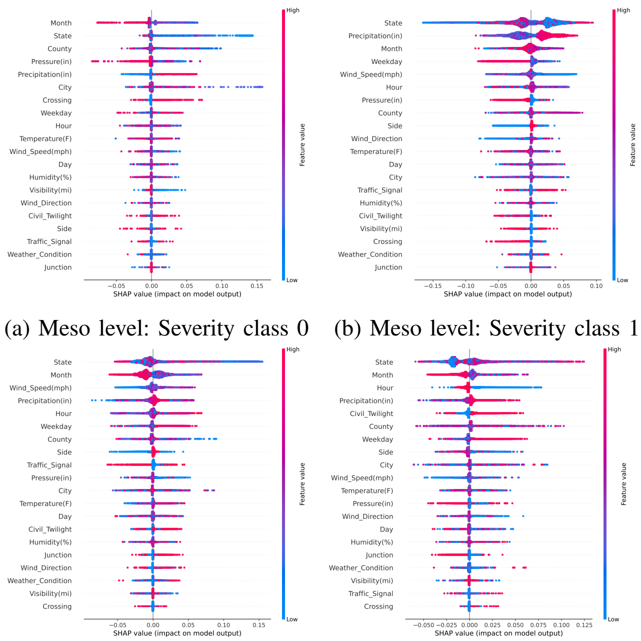


(c) Macro level: Severity class 2 (d) Macro level: Severity class 3

values affect the severity). The crossing feature is another factor, especially at the severity of 0. Regarding the Fig.4a, we can notice the high value of crossing (i.e., the accident occurred next to a crossing) has a high impact on accident severity. Furthermore, the hour feature can be considered an essential aspect of time factors. However, when the hour values decrease (e.g., from 23:00 to 6:00), the SHAP values increase and stabilize at a positive value (i. e., reducing the severity risk).

b) Meso level: Fig.5 shows the influence of each feature at the meso-level, which has extra information about geographical features, including state, county, and city. Regarding geographic factors, Fig.5a, Fig.5b, Fig.5c, and Fig.5d suggest

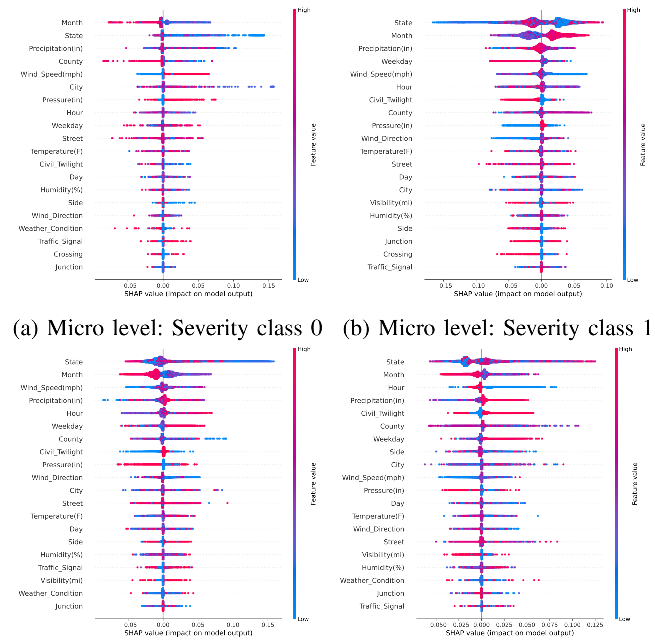
Fig. 5: Meso level.



(a) Meso level: Severity class 0 (b) Meso level: Severity class 1

(c) Meso level: Severity class 2 (d) Meso level: Severity class 3

Fig. 6: Micro level.



(a) Micro level: Severity class 0 (b) Micro level: Severity class 1

(c) Micro level: Severity class 2 (d) Micro level: Severity class 3

Factor/Month	Jan.	Feb.	Mar.	Apr.	May	Jun.	Jul.	Aug.	Sep.	Oct.	Nov.	Dec
Precipitation	H	H	H	H	M	M	M	M	M	H	H	H
Temperature	M	H	H	H	H	M	M	H	H	M	L	M
Wind speed	M	M	H	H	M	M	M	M	H	H	H	M
Hour	H	M	M	M	M	H	H	H	H	H	M	M
Humidity	M	M	M	M	H	H	M	M	M	M	M	M
Weekday	M	M	M	M	M	H	H	M	M	M	M	M
Day	M	M	M	M	H	M	M	H	M	M	M	M
Wind direction	M	M	M	M	M	M	M	M	M	M	M	L
Side	M	M	M	M	M	M	M	M	M	M	M	M
Civil twilight	M	M	M	M	M	M	H	M	M	M	M	H
Visibility	M	M	M	M	M	M	M	M	M	L	H	H
Weather	M	M	M	M	M	M	M	M	L	M	M	M
Junction	L	M	M	M	M	M	M	M	L	L	L	L
Pressure	H	H	M	M	M	M	M	M	M	M	M	M
Traffic signal	M	M	M	M	M	M	M	M	L	L	L	L
Crossing	L	L	M	M	M	M	M	M	L	L	L	L
Station	L	L	L	L	L	L	L	L	L	L	L	L
Stop	L	L	L	L	L	L	L	L	L	L	L	L
Railway	L	L	L	L	L	L	L	L	L	L	L	L
Amenity	L	L	L	L	L	L	L	L	L	L	L	L

TABLE II: Analysis global factor by month. H: High; M: Medium; L: Low.

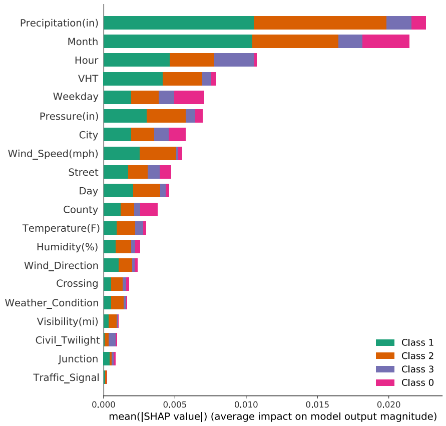
that state, county, and city also have a high impact on accident severities. Comparing these Shaply values with the previous model's results, the model's performance increases from macro-level to meso-level, confirming that geographic factors have an essential influence. As previously analyzed on the macro-level, time and weather factors most affect the accident severity.

c) Micro level: The SHAP summary plot in Fig .6 describes crash severity and risk factors at the micro-level. Like the macro and meso levels, the micro-level has additional information (i.g., street information). Besides the factors at the meso-level, Fig.6a, Fig.6b, Fig.6c, and Fig.6d show that the street is an additional factor that has weak influence in the accident severity.

3) Factors risk analysis by month: Based on the previous analysis of different scenarios, the month factor was an important factor that should be considered and discussed in detail. In this section, we provide a detailed discussion and analysis of the possible factors that are commonly relevant in each month. Firstly, we have calculated the SHAP mean values for each month by splitting the data by month. To better classify the factors that influence the traffic accident, we may cast them into the following categories: High (the top 3 factors), Medium (Medium factors), and Low (the lowest factors). Moreover, a summary table is presented to visualize and describe all risk factors during the year (Table. II). By analyzing the table, we can notice that the precipitation significantly impacts traffic accidents, which is basically the highest factor during the year (from January to April and from October to December). Extreme temperature value (low in winter and high in summer) is also significant, especially n (from February to May, August, and September). In addition, hour (i.e., time factors) has a high impact on traffic accident severity which is very high from June to October and medium impact in other months. Weather factors like wind speed, humidity, and atmospheric pressure are high for many months.

4) Traffic Flow analysis: In this section, we focus on the impact of traffic on accident severity. Traffic flow is another important factor to be considered in road safety analyses, especially in expressways. This study uses Caltrans Performance Measurement (PeMS) [13] and US-Accident datasets to analyze California's traffic flow factors. PeMS provides real-time traffic data on freeways, collected by multiple sensors across the major areas of all California. We match the accident

Fig. 7: Global feature importance of traffic flow analysis.



instance with the traffic information by using accident information (e.g., location, time). The traffic flow is presented by VHT, which is the number of vehicle-hours spent by travelers. VHT is calculated by:

$$VHT = \frac{\text{Flow} \times \text{Length}}{\text{Speed}} \tag{4}$$

The total number of accidents in this study is 130k cases over California state. Since PeMS is covered only for California, this analysis is presented as a Micro-level study. The classification performance of the XGBoost model of traffic flow analysis is presented in Table I. In Figure 7, the SHAP summary plot shows the features which are important of the traffic flow study. As we have analyzed previously, precipitation, month, hour, and weekday are the most important factors influencing accident severity. In addition to these, VHT is also a major factor that causes an accident.

V. Discussion and Conclusion

Identifying and understanding the influential factors for traffic accident severity is important for reducing future accidents and saving human lives. This research studied the influencing factors of traffic accidents in the United States by using a large sample of 2.5m observations. The major contributions of this paper are as follows. First, we adopted XGBoost, LightGBM, and CatBoost models to predict the accident severity. Second, we trained the models on three different scenarios by considering geographical levels at the micro, meso, and macro levels. Third, the explanations were interpreted using the SHAP technique to identify critical factors that affect traffic accident severity. Exploring the influential factors was analyzed based on two phases. First, we perform a global analysis at three levels. Then, we provide a detailed analysis of accident severity (e.g., four classes). Based on the evaluation results, we can draw the following conclusions:

- In the prediction of accident severity, XGboost achieved the best result at all levels (F1 score: 0.80, 0.78, and 0.77). The score has improved from level to level, which means the geographical features are essential to predict the accident severity.
- Time factors (i.e., month, hour, weekday), weather factors (i.e., precipitation, pressure, and civil twilight), and geographical factors (i.e., state, county, city, and street) are the most factors that causes an accident. The importance of these features varies for levels with different geographical features. However, the points-of-interests features have a low impact on accident.
- Analyzing factors by each month can help road users and authorities take some action to prevent further accidents in each individual month. Our results show that precipitation was the most critical factor, especially from January to April and December. In addition, time factors (hour, day, and weekday) are the top factors in the summer (from May to August). The Wind speed was the principal factor in September, October, and November.
- Traffic flow information (i.e., VHT) has improved the prediction score (F1: 0.90). However, traffic flow condition affects traffic accidents and causes serious injuries and fatalities.

For coming work, it would be interesting to include extra factors when available (e.g., traffic flow, road condition, and driver information) to improve model performance and analyze their effects on accident severity.

References

[1] "Global status report on road safety," *World Health Organization*, 2015.
[2] G. Zhang, K. K. Yau, and G. Chen, "Risk factors associated with traffic violations and accident severity in china," *Accident Analysis & Prevention*, vol. 59, pp. 18–25, 2013.
[3] J. J. Rolison, S. Regev, S. Moutari, and A. Feeney, "What are the factors that contribute to road accidents? an assessment of law enforcement views, ordinary drivers' opinions, and road accident records," *Accident Analysis & Prevention*, vol. 115, pp. 11–24, 2018.
[4] V. Shankar and F. Mannering, "An exploratory multinomial logit analysis of single-vehicle motorcycle accident severity," *Journal of safety research*, vol. 27, no. 3, pp. 183–194, 1996.
[5] A. Khorashadi, D. Niemeier, V. Shankar, and F. Mannering, "Differences in rural and urban driver-injury severities in accidents involving large-trucks: an exploratory analysis," *Accident Analysis & Prevention*, vol. 37, no. 5, pp. 910–921, 2005.
[6] P. Savolainen and F. Mannering, "Probabilistic models of motorcyclists' injury severities in single-and multi-vehicle crashes," *Accident Analysis & Prevention*, vol. 39, no. 5, pp. 955–963, 2007.
[7] X. Wang and S. H. Kim, "Prediction and factor identification for crash severity: comparison of discrete choice and tree-based models," *Transportation research record*, vol. 2673, no. 9, pp. 640–653, 2019.
[8] L.-Y. Chang and H.-W. Wang, "Analysis of traffic injury severity: An application of non-parametric classification tree techniques," *Accident Analysis & Prevention*, vol. 38, no. 5, pp. 1019–1027, 2006.
[9] L. S. Shapley, "17. a value for n-person games," 1953.
[10] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.
[11] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee, "From local explanations to global understanding with explainable ai for trees," *Nature machine intelligence*, vol. 2, no. 1, pp. 56–67, 2020.
[12] S. Moosavi, M. H. Samavatian, S. Parthasarathy, R. Teodorescu, and R. Ramnath, "Accident risk prediction based on heterogeneous sparse data: New dataset and insights," in *Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 2019, pp. 33–42.
[13] C. Chen, "Freeway performance measurement system (pems)," 2003.