Interactive Visualization and Prediction of U.S. Traffic Accidents

Chao Chen, Zhengxi Qu, Guang Wen Ren, Linna Su, Chang Xu, Qiaochu Zhu

1 Introduction

Traffic accidents have been a leading cause of deaths and severe injuries in the U.S., lead to heavy financial losses, and have been a key public safety challenge. Today many map applications highlight routes with heavy traffic or give an alert after an incident happens, but don't provide traffic accident predictions in advance. Academic studies that focus on predictions don't usually visualize their findings in an interactive way and often only focus on certain parts of the U.S. Many predictive models also fail to capture subtle nuances in the data such as complex interactions among predictors, nonlinearity, non-normal distribution, etc. Products that can derive insights from past traffic accidents, predict traffic accidents or accident-prone areas with wide geographic coverage, user-friendly visualizations and reasonably high accuracy can potentially save lives.

2 Problem Definition

In an attempt to help reduce road traffic accidents, we aim to create an interactive product that helps users gain insights from historical traffic accident patterns and assess the risk of experiencing traffic accident in terms of probability and severity based on user inputs and relevant web-scrapped data. Probability ranges between 0 to 1 and severity has 4 possible distinct values: 1 (LowImpact), 2(Minor), 3(Moderate), and 4(Serious), as defined by Bing Maps and closely related to the impact on traffic flow. We will:

1. Visualize historical road traffic accidents across 49 states in the U.S. from January 2016 to December 2021 based on a countrywide traffic accident dataset put together by Moosavi et al[15], on a scalable interactive map using Tableau and other visualization tools, showing traffic accident hotspots, patterns and trends.
2. Analyze the relation between time, weather, road conditions and road traffic accidents in the U.S., and predict the probability of the occurrence of high severity traffic accidents (severity level 3-4) as well as predict the severity of an accident. The predictions will be based on 3 machine learning models which are initially trained using data from January 2016 to December 2020 and tested using 2021 data, but will be updated on a monthly basis as new data comes in.
3. Feed real time data to our predictive model, but due to resource constraints, only real time weather data is used for this project.
4. Offer user-friendly interactive visualizations.

3 Literature Survey

Research has shown that there are 1.35 million people killed on roadways globally each year [1]. Better visualization of traffic accident data can help government devise more effective prevention and safety plans [2].

Many researchers have used statistical models that assume linearity or certain pre-defined statistical distributions to predict traffic accidents. Kidando et al. studied the impact of travel time reliability on crash severity and used traditional Gaussian distribution based random-effect logistic regression model and non-parametric Dirichlet random-effect logistic regression [3]. Savolainen et al. reviewed more than 20 methodologies to predict the crash severity, most of which are probit or logit-based models [4]. Rakha et al. pointed out that traffic crash distribution is highly skewed and not normal[5], and many researchers prefer using General Linear Models with Poisson or Negative Binomial error structure. Abdulhafedh studied interstate highway (I-90) in Minnesota and used Poisson regression and Negative Binomial regression to predict traffic crash frequency and achieved model fit greater than 50% in both training and testing [6]. Eisenberg did an in-depth analysis of the impact of precipitation on traffic accidents in US also using negative binomial regression [7]. All these models mentioned above have several limitations. One limitation of Poisson regression is it assumes that mean must be equal to the variance and does a poor job at

handling over-dispersed data. While Negative Binomial regression relaxes this constraint, it is not good at dealing with under-dispersed data [8]. We try to use more advanced models to avoid these limitations.
In recent years, more studies have adopted advanced machine learning techniques. A team from China designed a novel Deep Forest algorithm to predict traffic accident severity using UK road traffic data which resulted in around 93% accuracy [9]. Abdulhafedh used an Artificial Neural Network model for crash frequency prediction and model fit was close to 70% 6. Lu et al. used convolutional neural network to predict traffic accidents on US I-15 highway and achieved accuracy of 78% [10]. Xu et al used random forest model to select predictors, and then developed two genetic programming models to predict the crash occurrence under congested and uncongested traffic conditions for I-880N freeway in California[11]. Silva et al. overviewed and compared major groups of machine learning models used for crash prediction and found neural network and decision tree models have both shown encouraging results, but there have been criticisms around the "black box" nature of the former [12]. He et al. studied four metropolitan areas in U.S., used kernel density estimation method to predict probability of traffic accidents, and created a high-resolution accident risk map with satellite imagery, GPS trajectories and road maps [13]. Shaik et al. compared accident prediction studies around the world and found recurrent neural network and convolutional neural network outperform single- and multi-layer perceptron neural network [14]. Moosavi et. al. integrated different data sources to create a countrywide traffic accident dataset. They also proposed a deep neural network-based accident prediction model using long-short-term-memory and produced a F1 score >0.8 for 5 out of 6 U.S. cities [15,16]. Chen et. al. collected large and heterogeneous traffic accident data in Japan and used a deep architecture with denoise autoencoder layers and logistic regression layer to predict the impact of human mobility on traffic accident risk [17]. Lin et. al. proposed a novel variable selection method based on frequent pattern tree algorithm to identify important variables in real time traffic accident risk prediction models [18]. Kumar et. al. designed a framework with K-modes clustering algorithm, association rule mining using APriori algorithm to analyze accident patterns for various road accidents in India [19]. Moosavi et. al. proposed new processes to discover short-term and long-term patterns in geo-spatiotemporal data and adopted a tree pattern mining approach to reveal propagation patterns [20]. Ren et al. proposed a deep learning model, a long and short-term memory (LSTM) model to predict the frequency of 1km × 1km grid cells based on the history of the past 100 hours [21]. Many of these studies, however, focused on small-scale datasets, like certain parts of the U.S. like one state, certain part of the highway and don't provide interactive visualization which our project will provide.

4 Proposed Method

4.1 Intuition and Innovations

Our product is better than what's currently available for the following reasons:
1. We focus on providing users the analytics and tools to avoid future traffic accidents, and giving them assessment of the risk and early warnings in advance.
2. We combine 3 machine learning models: Logistic Regression, Random Forest and Multi-Layer Perceptron Neural Network, to predict the occurrence of severe traffic accidents in locations across the U.S., users can zoom in and out or type in a search keyword via our user interface to customize the result.
3. To improve model robustness, we resampled our data to make it more balanced and applied various feature engineering and overfitting reduction techniques such as regularization, bootstrapping, randomization, among others.
4. We visualize historical road traffic accidents data in an interactive map where users can easily navigate
5. Our prediction model will also be fed with real time data to make real time predictions.
6. Our analytics and prediction models have a wide geographic coverage: the whole U.S., and have the flexibility to be further expanded.

4.2 Methodology

4.2.1 Data Collection

1) Historical Traffic Accidents Data:

The U.S. road traffic accident dataset is a nationwide car accident dataset constructed by Moosavi et al[15], covering 49 states in the U.S., and available from Kaggle. The data has over 2.8 million records from January 2016 to December 2021. Each record contains a wide range of information including severity of an accident (a number between 1 and 4 to indicate the severity of the accident as explained in section 2), location, time, temperature, air pressure, wind speed, humidity, visibility, road conditions etc.

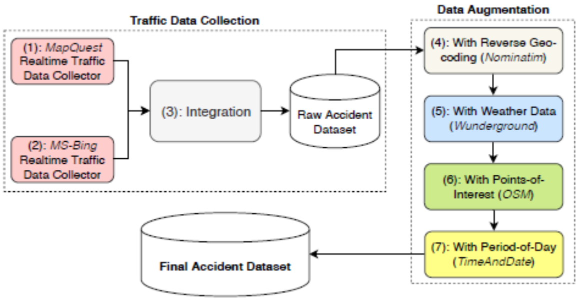

Traffic Data Collection
(1): MapQuest Realtime Traffic Data Collector
(2): MS-Bing Realtime Traffic Data Collector
(3): Integration
Raw Accident Dataset
Data Augmentation
(4): With Reverse Geo-coding (Nominatim)
(5): With Weather Data (Wunderground)
(6): With Points-of-Interest (OSM)
(7): With Period-of-Day (TimeAndDate)
Final Accident Dataset

Figure 1: Data Integration and Augmentation Process by Moosavi et al[15]

2) Realtime Weather Data:

The real-time weather data is acquired from OpenWeather API. As users input a location/zipcode, recent weather data of that location will be captured and used to make predictions.

3) Base Canvas:

We use Mapbox maps as the base canvas of our interactive visualizations. Traffic accidents data and predictions are projected onto the scalable map to show traffic accident hotspots.

4.2.2 Model Training, Validation and Testing

We leveraged functions from the Scikit-Learn library to construct, train, validate and test the models.

1) Data resampling and processing: since our dataset is highly imbalanced: over 90% data points are non-severe (severity = 1 or 2) accidents; we down sampled these data and over sampled the severe accidents data by bootstrapping. We relabeled severity 1 and 2 as 0 (non-severe accidents) and severity 3 and 4 as 1 (severe accidents) to train our models to predict the probability of a severe traffic accident happening. However for predicting the severity of an accident, the relabeling step is skipped.

2) Model Selection

- Logistic Regression: L2 regularization, regularization parameter = 1/10
- Random Forest: n_estimators = 100, criterion for split 'gini', min_samples_split = 100, max_features = 0.8
- Multi-Layer Perceptron Neural Network: solver = stochastic gradient descent, hidden layer size = (100,), activation function = 'relu', max iteration = 200
- Each model will generate a prediction and the associated probability, and predictions from 3 models will be averaged to produce the final forecast

2) Training Parameters Selection: 10-fold cross validation

3) In sample vs Out of sample: 2016-2020 vs 2021

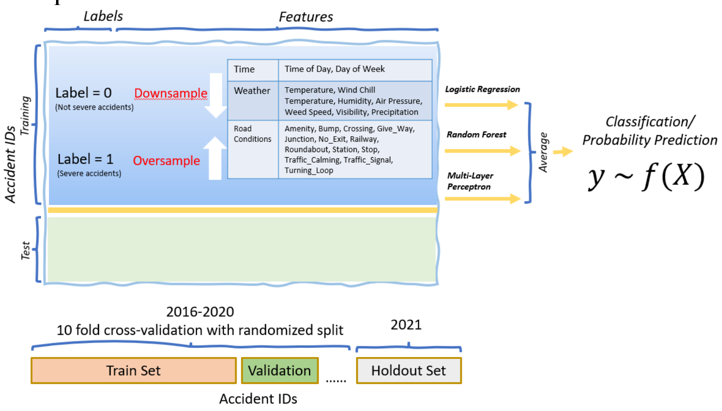

Labels
Features
Accident IDs
Training
Test
Label = 0
(Not severe accidents)
Downsample
Label = 1
(Severe accidents)
Oversample
Time
Time of Day, Day of Week
Weather
Temperature, Wind Chill Temperature, Humidity, Air Pressure, Weed Speed, Visibility, Precipitation
Road Conditions
Amenity, Bump, Crossing, Give_Way, Junction, No_Exit, Railway, Roundabout, Station, Stop, Traffic_Calming, Traffic_Signal, Turning_Loop
Logistic Regression
Random Forest
Multi-Layer Perceptron
Average
Classification/ Probability Prediction
y ~ f(X)
2016-2020
10 fold cross-validation with randomized split
2021
Train Set
Validation
......
Holdout Set
Accident IDs

Figure 2: Machine Learning Models, Training, Validation and Testing Process

4.3 User Interface Demonstration

Our product is a web application consisted of 3 pages: Historical Accident Analytics Dashboard, Severe Accident Probability Heatmap and Accident Severity Prediction, and built using Tableau, MapBox, Flask, HTML, CSS, Javascript and related libraries. Users can easily toggle back and forth between these pages.

4.3.1 Interactive visualization of historical traffic accidents, as shown in Figure 3:

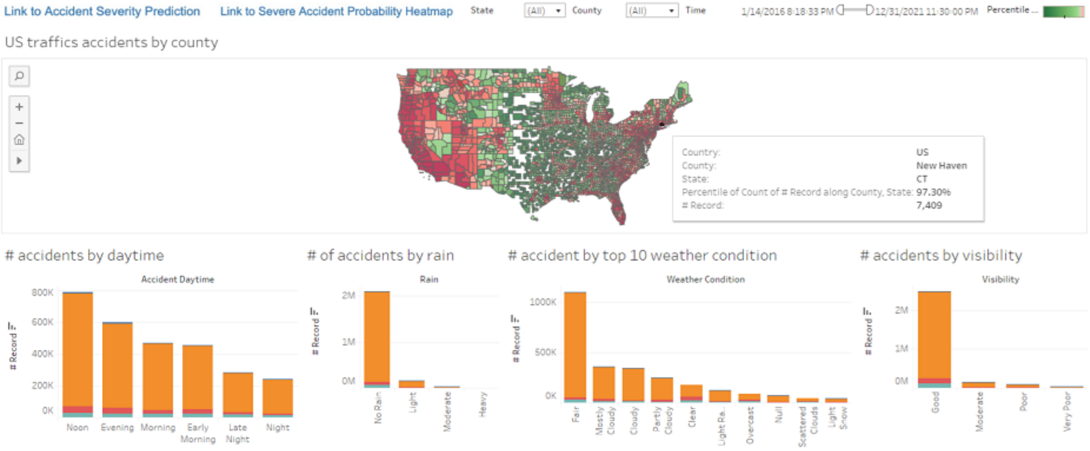
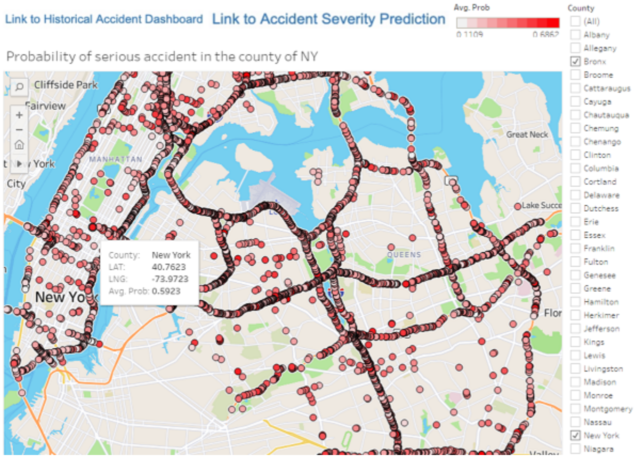
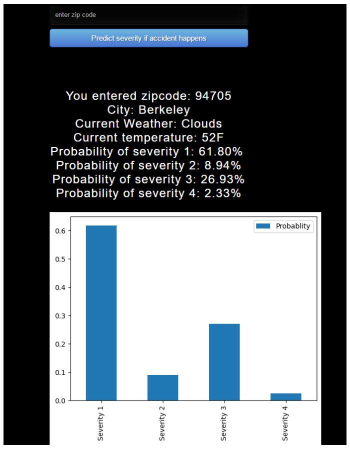


Figure 3: Historical Accident Analytics Dashboard

4.3.2 Accident heatmap with predicted probability of severe accidents happening across different locations, New York City is shown in Figure 4 as an example.





```
You entered zipcode: 94705
City: Berkeley
Current Weather: Clouds
Current temperature: 52F
Probability of severity 1: 61.80%
Probability of severity 2: 8.94%
Probability of severity 3: 26.93%
Probability of severity 4: 2.33%
```

Figure 4: Severe Accident Probability Heatmap

Figure 5: Accident Severity Prediction

4.3.3 Webpage with traffic accident severity prediction after user typing in search keyword (zipcode in our example), as shown in Figure 5.

4.4 User Manual

4.4.1. Historical Accident Analytics Dashboard

- Upper half of the page shows a heatmap of road traffic accidents throughout the U.S.
- Lower half of the page displays charts and statistics describing historical patterns and trends.
- There are dropdown menus and time scale at the top right where users can select the state, county and time range they are interested in to customize the view.
- Users can also click on a specific location in the map or content in the charts to customize the view, and clicking it again will return to the previous view.
- Mouseover on any specific location in the map or bars in the chart, detailed information related to the selection will be shown in a tooltip.
- Users can also zoom in and out the map.

4.4.2. Severe Accident Probability Heatmap

- Shows traffic accident hotspots in the map with predicted probability of having a severe traffic accident, each hotspot is shown as a circle and the darker the color, the higher the probability is.
- Mouseover any hotspot on the map, detailed information relating to that location will be shown in a tooltip.
- Users can zoom in and zoom out on the heatmap to get the information they want.
- Users can also customize the view by (un)checking the location boxes to the right of the map.

4.4.3. Accident Severity Prediction

- Users can type in a zipcode in the search box, location, real-time weather information and predicted probability for each severity category will be shown below.

Moreover, our product can also be integrated with navigation applications to give travelers alert if they are going to experience high probability of having severe traffic accident.

5 Experiments/Evaluation

5.1 Overview

Our goal is to find main drivers of road traffic accidents, especially severe accidents, and predict them in advance. To evaluate our model performance, we can look at the following:

- Accuracy: number of correct predictions / number of total predictions
- ROC Curve (Receiver Operating Characteristic Curve) and AUC (Area Under Curve)
- F1 score: 2･(precision ･ recall)/(precision + recall), where precision = True Positives/(True Positives + False Positives) and recall = True Positives/ (True Positives + False Negatives)
- Consistency between in sample and out of sample performance.

Usually higher accuracy, AUC, F1 score and more consistent performance in out of sample period indicate better model performance.

5.2 Experiments and Observations

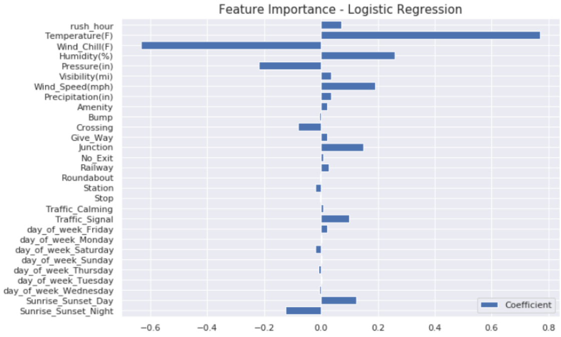
The way how features and models are constructed can have a large impact on model performance. There are more than 40 columns in the original dataset that can be used as features/predictors, but we can’t use all of them directly. First of all, the features need to be standardized, otherwise we may get very misleading results, in our case, our models failed to converge and crashed. There’re also some variables which shouldn’t be included in the model, for example, the binary variable *Turning_Loop* is False across all data points, so it won’t have any predictive power. Also, some variables are highly correlated, such as *Temperature(F)* and *Wind_Chill(F)* (correlation = 0.9) as well as *Traffic_Calming* and *Bump* (correlation = 0.8), which is problematic when using linear prediction models. Therefore, we used L2 regularization to shrink the weights of these variables. Resampling the data is also key as most of the data points are minor or low impact accidents, if we don’t make the data balanced, the model will predict 0 most of the time with great accuracy, but precision and recall will be extremely bad.

Many of the above-mentioned issues such as collinearity, not being normal are less problematic for random forest and neural network models. Also random forest and neural network are supposed to do a better job at capturing more complex non-linear patterns in the data. As seen in Table 1, which summarizes performance metrics of all three models we used, both random forest and multilayer perceptron model achieved much higher accuracy in test period than our logistic regression model. Comparing performance between train and test periods, none of the three models had significant decay in accuracy, which is not surprising as we used multiple techniques such as resampling, regularization, randomization, bootstrapping etc. to reduce overfitting. By using the average prediction of the 3 models and applying to the whole dataset, we achieved an overall prediction accuracy of 89% and AUC of 0.61. However, in terms of AUC and F1 score, all three models didn’t do a great job, especially in the test period. We think one potential reason is in more recent period with COVID and a lot of changes people’s travel/driving behaviors, our models haven’t taken those into account. Also the data we use doesn’t include some key factors that will affect the probability of experiencing a traffic accident such as vehicle and driver characteristics, driving speed, traffic conditions, distance with nearby vehicles etc.

Model	**Accuracy**		**ROC_AUC**		**F1**	
	Train	**Test**	**Train**	**Test**	**Train**	**Test**
Logistic Regression	0.60	0.55	0.60	0.50	0.29	0.03
Random Forest	0.82	0.68	0.80	0.53	0.54	0.04
Multi-Layer Perceptron	0.87	0.98	0.51	0.50	0.03	0.01

Table 1: Model Performance Metrics

Among the factors we considered, weather conditions, whether it's rush hours or at a junction are some of the top features that affect the probability of having a severe traffic accident.



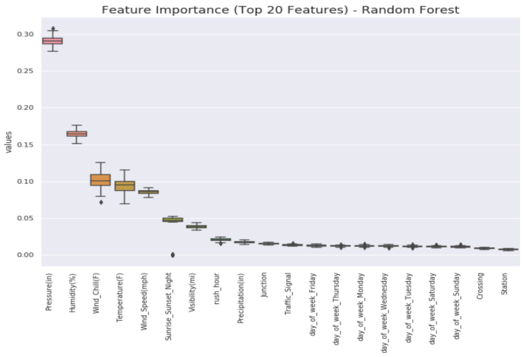


Figure 6: Feature Importance – Logistic Regression

Figure 7: Feature Importance – Random Forest

5.4 User Survey

To follow up, we will also randomly select 50-100 users to do a paid survey, and make sure participants are diversified and representative of the true user population. Sample survey questions are shown below:

1. How old are you?
2. Gender
3. How many years of driving experience do you have?
4. How useful do you find our application in helping you avoid traffic accidents? (Scale [1-5]. 1 being not useful at all, 5 being extremely useful)
5. How easy do you find it to use our application (Scale [1-5]. 1 being least easy, 5 being easiest)?
6. How much do you like using our product compared to other map/traffic applications? (Scale [1-5]. 1 being least, 5 being most)
7. What information do you want to know but is now shown in our application?
8. What suggestions do you have for us?

6 Conclusions and Discussion

In this project, our team combined 1 linear and 2 non-linear machine learning models to predict the probability of experiencing traffic accident of different severity levels and used Tableau, Flask, Javascript to provide visualizations and statistical information based on U.S. traffic accident data from 2016 to 2021, and delivered good prediction accuracy and user-friendly interactive visualizations. We focused on providing users the analytics and tools to avoid future traffic accidents, and giving them assessment of the risk and early warnings in advance, which is a function not available in common map applications. There are a few directions for us to further improve this product, such as further improving our machine learning model design, incorporating more real-time data in addition to weather information, adding vehicle and driver characteristics and traffic conditions, more closely studying new relevant factors that may affect traffic accidents in the post-COVID era. We would also like to do a user survey to collect feedback. Moreover, we'll try to integrate our product with navigation map applications to give travelers alert in real time if there's high probability that they'll encounter severe traffic accident.

** All team members have contributed a similar amount of effort.*

Bibliographies

1. Organization, W. H. Global status report on road safety 2018: summary. (2018).
2. Babar, M. *et al.* Road traffic accident data analysis and its visualization. *researchgate.net* **9**, 1603–1614 (2021).
3. Kidando, E., Moses, R., Ozguven, E., and, T. S.-J. of traffic & 2019, undefined. Incorporating travel time reliability in predicting the likelihood of severe crashes on arterial highways using non-parametric random-effect regression. *Elsevier*.
4. Savolainen, P. T., Mannering, F. L., Pankow, C., Lord, D. & Quddus, M. A. The statistical analysis of highway crash-injury severities: a review and assessment of methodological alternatives. *Elsevier* (2011).
5. Rakha, H., Arafeh, M., Abdel-Salam, A. G., Guo, F. & Flintsch, A. M. Linear regression crash prediction models: Issues and proposed solutions. *Efficient Transportation and Pavement Systems: Characterization, Mechanisms, Simulation, and Modeling - Proceedings of the 4th International Gulf Conference on Roads* 241–255 (2008) doi:10.1201/9780203881200-29/LINEAR-REGRESSION-CRASH-PREDICTION-MODELS-ISSUES-PROPOSED-SOLUTIONS-RAKHA-ARAFEH-ABDEL-SALAM-GUO-FLINTSCH.
6. Abdulhafedh, A.**,** Crash frequency analysis. *Journal of Transportation Technologies, 2016, 6, 169-180.*
7. D. E.-A. analysis & prevention and undefined 2004, “The mixed effects of precipitation on traffic crashes,” *Elsevier*, 2003, doi: 10.1016/S0001-4575(03)00085-X
8. Abdulhafedh, A.**,** Road crash prediction models: different statistical modeling approaches. *Journal of Transportation Technologies, 7, 190-205. doi: 10.4236/jtts.2017.72014.*
9. Gan, J. *et al.* An alternative method for traffic accident severity prediction: using deep forests algorithm. *hindawi.com*.
10. Wenqi, L., Dongyu, L., International, Y. M.-2017 2nd I. & 2017, undefined. A model of traffic accident prediction based on convolutional neural network. *ieeexplore.ieee.org*.
11. Xu, C., Wang, W., Intelligent, P. L.-I. T. on & 2012, undefined. A genetic programming model for real-time crash prediction on freeways. *ieeexplore.ieee.org* **14**, (2013).
12. Silva, P., Andrade, M., transportation, S. F.-J. of traffic and & 2020, undefined. Machine learning applied to road safety modeling: A systematic literature review. *Elsevier*.
13. He, S. *et al.* Inferring high-resolution traffic accident risk maps based on satellite imagery and GPS trajectories. *openaccess.thecvf.com*.
14. Shaik, M., Islam, M., Studies, Q. H.-A. T. & 2021, undefined. A review on neural network techniques for the prediction of road traffic accident severity. *Elsevier*.
15. Moosavi, S., Samavatian, M. H., Parthasarathy, S. & Ramnath, R. A countrywide traffic accident dataset. *arxiv.org (2019)*.
16. Moosavi, S., Samavatian, M. H., Parthasarathy, S., Teodorescu, R. & Ramnath, R. Accident risk prediction based on heterogeneous sparse data: New dataset and insights. *dl.acm.org* 10 (2019) doi:10.1145/3347146.3359078.
17. Chen, Q., Song, X., Yamada, H., on, R. S.-T. A. conference & 2016. Learning deep representation from big and heterogeneous data for traffic accident inference. *aaai.org*.
18. Lin, L., Wang, Q., Emerging, A. S.-T. R. P. C. & 2015, undefined. A novel variable selection method based on frequent pattern tree for real-time traffic accident risk prediction. *Elsevier* **55**, 444–459 (2015).
19. Kumar, S. & Toshniwal, D. A data mining framework to analyze road accident data. *Journal of Big Data* **2**, (2015).
20. Moosavi, S., Samavatian, M. H., Nandi, A., Parthasarathy, S. & Ramnath, R. Short and long-term pattern discovery over large-scale geo-spatiotemporal data. *dl.acm.org* **19**, 2905–2913 (2019).

21 Ren, H., Song, Y., Wang, J., … Y. H.-2018 21st I. & 2018, undefined. A deep learning approach to the citywide traffic accident risk prediction. *ieeexplore.ieee.org*.