Real-time Accident Severity Prediction with Spatiotemporal Data

Yasas Senarath
Information Sciences and Technology
George Mason University
Virginia, US
ywijesu@gmu.edu

Abinash Vasudevan
Information Sciences and Technology
George Mason University
Virginia, US
avasudev@gmu.edu

Shiva Ram Kaushil Pabba
Data Analytics Engineering
George Mason University
Virginia, US
spabba2@gmu.edu

***Abstract*—It is important to reduce the number of road accidents to mitigate human and financial damages. One crucial step towards that is finding out the severity of a road traffic accidents. An abundance of information on road traffic data has enabled the creation of models for the prediction of various factors related to traffic accidents. The contributions of this paper include the creation of processed datasets for accident severity prediction, exploring novel features for creating accident severity prediction models, and advanced analysis of the impact of features in each model created. The knowledge obtained from accident severity prediction models could be transferred to downstream tasks such as accident risk prediction.**

This study presents a history-aware model for the prediction of accident severity. The accident severity is the effect of road accidents on road traffic. This research uses a grid-based approach to take the history of the most recent accident severity for predicting current severity along with other environmental and spatial-temporal features. We introduce the history-aware-grid-based approach as Spatial-Time-Series (STS) feature as it considers both grid (space) and history (time) to generate a series of event history. We show that the inclusion of this additional feature improves the baseline model performance for all datasets identified.

***Index Terms*—Big Data, Traffic Analysis, Accident Severity, Time Series Analysis**

I. Introduction

It is well known that road traffic accidents are a serious health problem. It is a strong indication of traffic-related accidents and injuries that improving road safety is a key priority nationwide. Unfortunately, traffic collisions cause many injuries or deaths. Global status report on road safety [1] states that in 2016 there were 1.35 million traffic deaths, on average 3,700 people lose their lives every day on the roads.. Additionally, [1] reports that road traffic accidents have become the leading cause of death of people aged between 5 and 29 in 2016. The World Health Organization (WHO) has estimated that there were around 39,888 fatalities in the United States in 2016 [1]. Road accidents cost countries 3 percent of their domestic gross output on average. Moreover, the annual United States road crash statistics show that road crashes cost $230.6 billion per year [2] and $380 million in direct medical costs. Road accidents are inevitable and can be prevented. Significant intervention is necessary to enforce these steps and reach all potential global targets. Therefore, it is important to reduce the number of road accidents to mitigate human and financial damages.

It is a strong indication from traffic-related accidents and injuries that improving road safety is a key priority nationwide. We can save lives through gathering and storing traffic data on a wide scale, big data may be used to enhance road safety, reduce traffic accidents. If it is possible to predict accidents and their severity before an accident occurs, then we can significantly minimize the number of fatalities and injuries. With big data Traffic safety analysis, Driving behavior, Real-time accident prediction, Road health behavior can be predicted. In addition to saving lives, accident prediction can help in optimizing traffic routes to minimize time and cost. Big data allows accident events to be more easily detected and track driver activity while in traffic also recognize taking into consideration improvements. The importance of determining traffic accidents has led to many research studies on using different environmental and temporal data. According to [3], past studies on traffic accidents are divided into three groups: analysis of effects environmental stimuli, prediction of accident frequency, and prediction of accidents.

Computers may generate a statistical map of collisions by gathering details on automobile accidents, such as where, where, and why they occur with attributes like traffic, location, weather, POI and time. Accident prediction models have utilized data from sources such as satellites, cameras and road-network properties to do the prediction [4], [5]. However, the datasets used by those were limited to either part of the road network or a city. The models in the literature were not appropriate to predict real-time data [4]. Moreover, some approaches were overly simplistic to accurately predict the accidents [6], [7]. These shortcomings have led to the creation of the dataset named US-Accidents [3]. US-Accidents contain traffic data of Contiguous United States. Additionally, [3] describes a neural network-based model for accident risk prediction. However, their research does not focus on determining accident severity.

Hypothesis: prediction of real-time traffic accident severity can be improved with various environmental conditions, location information and time.

Our goal is to build and evaluate a multi-task model (M) that takes environmental conditions, location information and

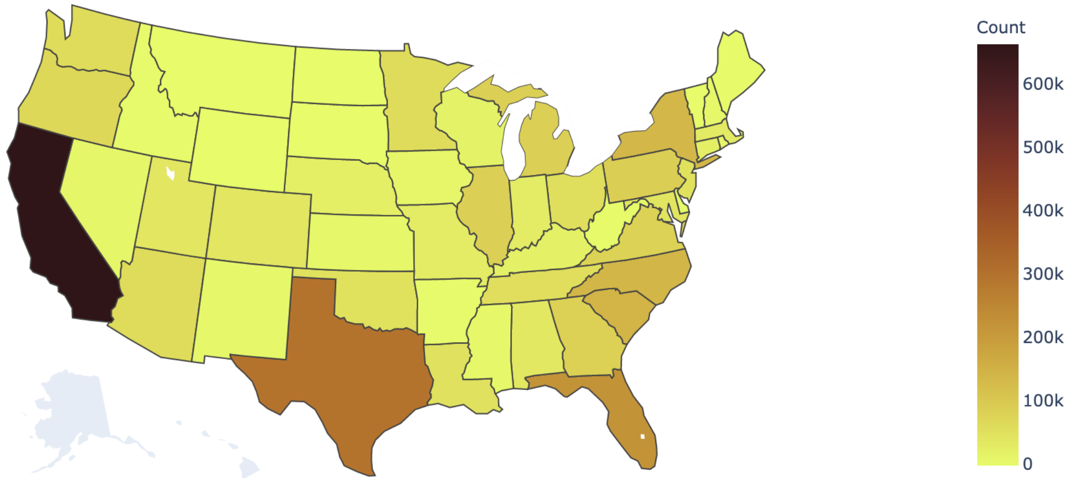


Fig. 1: Number of accident reports by US state

time as features and predict the accident risk and the severity in a geographic region for a determined time interval.

This research considers weather conditions such as visibility, temperature, and humidity as the environmental conditions. We will consider traffic-events and points-of-interests (POI) as location information. Examples of traffic event types are lane-blocked, broken-vehicle, congestion, etc. Examples for POI are railway, roundabout, etc. Information may determine, in a government context, that a path needs to be revamped based on driver patterns and traffic. Accident severity determines the impact of the event on the traffic (delays).

The contributions of this paper are as follows:

- Introduce a novel machine learning model for Severity Prediction using environmental and spatiotemporal information.
- Provide a processed and clean dataset for severity prediction using data collected in past research.

The next sections of this paper are organized as follows, the overview of the related work is included in section 2. Section 3 describes about the dataset that is used in this research. Section 4 lays out the foundation for proposed methodology by analyzing attributes provided in the dataset. Next, Section 5 presents the methodology we used for accident severity prediction. Section 6 presents the evaluation results of the methodologies proposed. Finally, Section 7 will conclude the paper by providing the summary of the paper and future work.

II. Literature

The literature splits in to three sections based on the analysis task performed on the traffic accident data.

A. *Analysis of Environmental Factors on Accidents*

Accidents do not happen in a completely stochastic fashion; a number of variables affect their occurrence. These studies focus on analyzing how each accident characteristic change according to some environmental factor. The research offers insights into and underlines the possible impacts of environmental factors that influence road traffic accidents. Sunrise, sunset, climate, muddy road surfaces and wind speeds, uphill, downhill path could be some of the environment factors. Death odds in the scene were higher in muddy, overcast, foggy, snowy, humid and stormy conditions than in clear weather. [8] has studied the factors affecting the detection of severity of the accidents with environmental aspects. [9] has studied the effects of weather on the traffic accidents. Although, these researches provide insights on the factors they cannot be applied in real time systems.

B. *Prediction of Accident Frequency*

Predicts the number of accidents occurring in a region in a given period of time [4]. We can predict a mean frequency most likely for a given location such as the type of intersection and In order to estimate the likelihood that an event will occur, it will be of a certain kind. Many have employed deep learning techniques (LSTMs/CNN) to train these models [4]. Although predicting accident frequency is helpful, this is mostly performed with non-real-time systems since it uses non real time data.

C. *Prediction of Accident Risk*

Research in this category try to predict the possibility of an accident in real time [7] to identify the risk of accidents. They employ a Bayesian network to model the relationship between the accident risk and features such as weather, visibility, traffic volume, speed and occupancy information [7]. However, their approach is limited by the number of samples. Chen et al. [10] uses mobility and traffic accident data for real-time traffic accident inference. They have used grid based approach with time window approach to predict the accident risk.

#	Attribute Name	Description	Example
1	ID	Unique identifier for an accident record	A-1
2	Source	Indicates from which source the accident report is got. In our dataset is either got from MapQuest traffic or Microsoft Bing Map Traffic.	MapQuest
3	TMC	Traffic Message Channel is a code that provides description about the event	201
4	Severity	The severity of the accident.	3
5	Start_Time	Shows the start time of the event in its local time zone	2/8/16 7:44
6	End_Time	Shows the end time of the accident in its local time zone	2/8/16 8:14
7	Start_Lat	Gives the latitude in GPS coordinate of the start point	39.747753
8	Start_Lng	Gives the longitude in GPS coordinate of the start point	-84.205582
9	End_Lat	Gives the latitude in GPS coordinate of the end point	40.84992
10	End_Lng	Gives the longitude in GPS coordinate of the end point	-73.94408
11	Distance	Gives the length of the road affected by the accident. It is given in miles	0.01
12	Description	Natural language description about the accident	Accident on I-75 Southbound at Exits 52 52B US-35. Expect delays.
13	Number	Street number in the address field	376
14	Street	Street name in the address field	N Woodward Ave
15	Side	Side of the street with respect to the address	R
16	City	Shows the city from the address field	Dayton
17	County	Shows the county in the address field	Montgomery
18	State	Shows the state in the address field	OH
19	Zipcode	Shows the zipcode in the address field	45417-2476
20	Country	Shows the country in the address field	US
21	Time_Zone	Tells the time zone of the location the accident happened	US/Eastern
22	Airport_Code	Denotes the airport weather station which is the closest to the accident location	KDAY
23	Weather_Timestamp	Shows the timestamp of weather reading record	2/8/16 7:38
24	Temperature	Shows the temperature at the location of the accident in Fahrenheit	35.1
25	Wind_Chill	Shows the wind chill in Fahrenheit	31
26	Humidity	Shows the humidity at the location of the accident in percentage	100
27	Pressure	Shows the pressure at the location of the accident in inches	29.66
28	Visibility	Shows the visibility at the location of the accident in miles	7
29	Wind_Direction	Shows the direction of wind at the location of the accident	WSW
30	Wind_Speed	Shows the speed of wind at the location of the accident in miles per hour	3.5
31	Precipitation	Shows the precipitation at the location of the accident in inches, if any	0.03
32	Weather_Condition	Tells about the weather condition in the location of the accident like rain, snow, fog, etc.	Light Rain
33	Amenity	POI annotation in Boolean value which indicates the presence of amenity in a nearby location	FALSE
34	Bump	POI annotation in Boolean value which indicates the presence of a bump in a nearby location	FALSE
35	Crossing	POI annotation in Boolean value which indicates the presence of a crossing in a nearby location	FALSE
36	Give_Way	POI annotation in Boolean value which indicates the presence of a yield sign in a nearby location	FALSE
37	Junction	POI annotation in Boolean value which indicates the presence of a junction in a nearby location	FALSE
38	No_Exit	POI annotation in Boolean value which indicates the presence of a no exit in a nearby location	FALSE
39	Railway	POI annotation in Boolean value which indicates the presence of a railway in a nearby location	FALSE
40	Roundabout	POI annotation in Boolean value which indicates the presence of a roundabout in a nearby location	FALSE
41	Station	POI annotation in Boolean value which indicates the presence of a station in a nearby location	FALSE
42	Stop	POI annotation in Boolean value which indicates the presence of a stop in a nearby location	FALSE
43	Traffic_Calming	POI annotation in Boolean value which indicates the presence of traffic calming in a nearby location	FALSE
44	Traffic_Signal	POI annotation in Boolean value which indicates the presence of a traffic signal in a nearby location	FALSE
45	Turning_Loop	POI annotation in Boolean value which indicates the presence of a turning loop in a nearby location	FALSE
46	Sunrise_Sunset	Shows the period of the day either day or night based on Sunrise/sunset	Night
47	Civil_Twilight	Shows the period of the day either day or night based civil twilight	Day
48	Nautical_Twilight	Shows the period of the day either day or night based on nautical twilight	Day
49	Astronomical_Twilight	Shows the period of the day either day or night based on astronomical twilight	Day

TABLE I: Descriptions of Attributes

Data Source	Number of Records (% of Records)
# MapQuest	1,702,565 (75.9%)
# Bing	516,762 (23%)
# Reported by Both	24,612 (1.1%)
Total Accidents	**2,243,939**

TABLE II: Summary of Data Source of Dataset

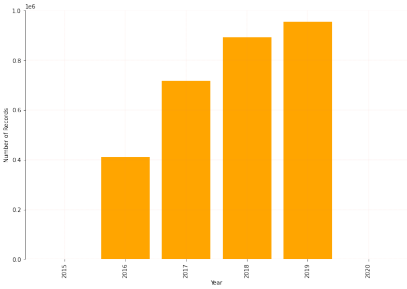


Fig. 2: Number of accident reports by Year

There are many other studies on identifying accident risks. However, they will be categorized in to one of these categories. It is important to note that we are trying to predict the accident severity in real time along with the lines of research performed in Section II-A and Section II-C.

III. Dataset

This study will utilize the dataset introduced in [3]. It contains accident data collected from 49 states of the United States during February 2016 to December 2019 period. In addition to traffic events (accidents) it contains the weather and POI information corresponding to each traffic record. Table IV summarizes the attribute types and the number

Statistic	Value
count	953630
mean	83
std	1009
min	-35
25%	29
50%	59
75%	89
max	422640

TABLE III: Summary of Accident Duration

Attribute Name	Number of Attributes
Traffic	10
Location	8
Weather	10
POI	13
Time	4
Total Attributes	**45**

TABLE IV: Number of Attributes for Each Feature Type

of attributes belonging to that type available in the dataset. Table II demonstrate the number of records obtained from each source. In addition, Figure 4 shows a geographical heat map with the color key indicating the number of accident records.

This research only uses a sample of data from [3] to reduce the computation load when analysing the data. Figure 2 shows the distribution of accident reports by year. To minimize the effect of sampling on the usability of accident severity prediction model, this study focuses on utilizing the data from year 2019 to model traffic accident severity since most number of records appeared in that year and it is the latest available accident reports.

A. *Attributes*

Table I shows the available attributes and their descriptions. In addition, it includes a column to indicate an example value from our dataset for each attribute.

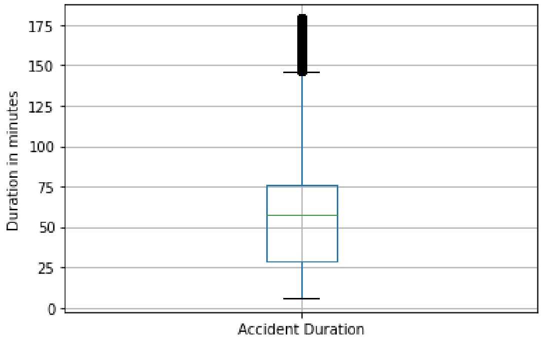


Fig. 3: Distribution of Accident Duration

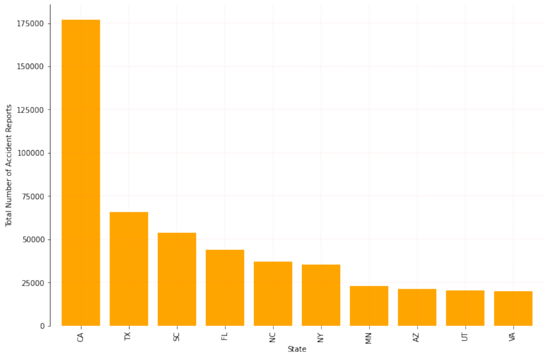


Fig. 4: Number of accident reports by US state

B. *Removing Outliers*

Prior to the data processing it is essential to remove the outliers from the data to prevent anomalous data interfering with the models. This study proposes to find anomalies in data by using '*start time*' and '*end time*' attributes of the dataset. The reason for selecting those attributes for outlier detection was due to its relationship with our target variable. The target variable '*severity*' is defined as the delay caused by the accident to the traffic. First, the accident duration is calculated by taking the difference between the '*start time*' and '*end time*' attributes. The summary statistics of the accident duration is indicated in Table III. Interquartile range (IQR) is then calculated using the Equation 1. The Q1, Q3 represents first (25%) and third (75%) quartiles accordingly. The IQR of the accident duration is $IQR_{ad} = 60$ for this data. The lower and upper outliers 0, 179 were identified by using Equation 2. It should also be noted that this step removes invalid data records that contain negative accident duration since time duration can only be positive. Figure 3 shows the distribution of accident duration after filtering out outliers.

$$IQR = Q3 - Q1; \tag{1}$$

$$\begin{aligned} LowerLimit &= min(0, Q1 - 1.5 \times IQR); \\ UpperLimit &= Q3 + 1.5 \times IQR; \end{aligned} \tag{2}$$

C. *Data Selection by State*

In order to build models that are specific to each state we would require our data to be presented by each state. Our goal here is to select only the states having high number of accident records for our analysis. This reduces the number of records that we have to deal at a time, reducing the total number of data points to process. We assume that each state will have its own unique distribution for accident prediction.

Figure 4 shows a barchart illustrating the total number of accidents reported from different states. It is observed that following states have the maximum number of records: California (CA), Texas (TX), South Carolina (SC), Florida (FL) and North Carolina (NC). Further simplifying the constraints this study will only be using these five states for building the models.

D. *Analysis of Target Variable*

The target variable for the problem we are trying to solve is the '*severity*' of accidents. The Definition 3.1 defines the term severity used in the context of the problem.

Attribute Type	Attributes		
Meta Data	Source		
Location	Start_Lng	Start_Lat	Distance(mi)
	County	Side	
Time	Hour	Weekday	Timezone
	Sunrise_Sunset		
Environmental	Temperature(F)	Pressure(in)	Wind_Direction
	Humidity(%)	Visibility(mi)	Weather_Condition
Point-Of-Interest (POI)	Amenity	Give_Way	Railway
	Bump	Junction	Roundabout
	Crossing	No_Exit	Station
	Traffic_Calming	Traffic_Signal	Stop
	Turning_Loop		

TABLE V: Selected Attributes for Training the Models

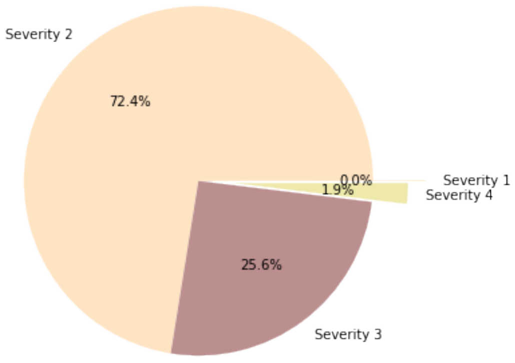


Fig. 5: Accident Severity Distribution

Definition 3.1: Accident Severity is a discrete number between 1 and 4, where 1 indicates the least impact on traffic and 4 indicates a significant impact on traffic.

Figure 5 shows the distribution of the categories in the original dataset from [3]. It is observed that majority of severity values are either 2 or 3.

IV. Traffic Severity Prediction Model

This section outlines the approach taken to model the severity of accidents in the 5 states of the US that have the maximum number of accident records.

The task of severity detection can be modeled as a multi-class classification model as four labels indicate the severity. Therefore, it is required to train a classifier to identify each category.

After processing our data using the methods identified in Section III, the first step is to further process the features for building the models. The following subsections provide information on the processes used to further process the data.

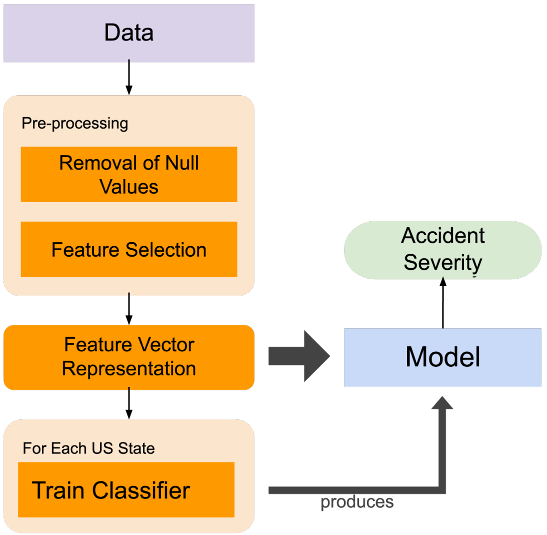


Fig. 6: High Level Traffic Severity Prediction Process

A. Additional Pre-processing

1) Null Values: In order to train a model on the data, it is required to remove null values from the attributes used for analyzing the data. We drop all *null* values of the features used in training the model.

2) Feature Selection: The next step in processing the data to be used in the severity prediction is to select the best features that determine the severity. This is required to improve the performance of the system. We perform a manual feature selection from the available features. Table V shows the attributes that were used in our modeling.

B. Feature Vector Representation

This step converts features selected from Step IV-A2 to a numerical vector format that all classifiers can understand. All categorical features available in our dataset are one-hot encoded. The process of one-hot encoding converts each value

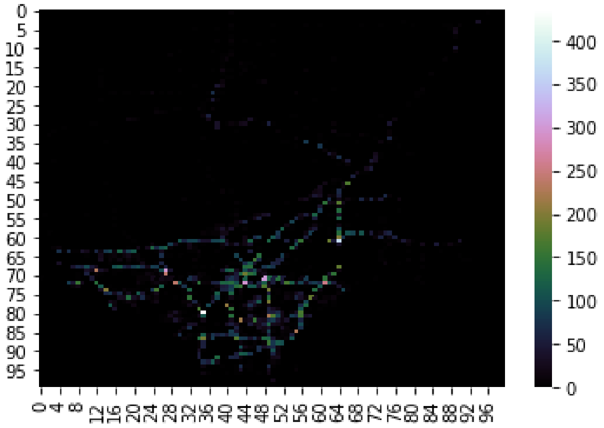


Fig. 7: The frequency of accidents for each grid in Los Angeles county.

in a categorical attribute to a separate feature that indicates the availability of that category as a binary feature.

C. Data Selection by State

As described in Section III-C we extract the features of each state before training the classifier to predict the severity of each case report.

D. Classifier

The last step of the approach is to train a machine learning model to predict the severity. A set of machine learning classifiers have been used in our approach for severity prediction. Figure 6 visualizes the process for training classifiers and predicting the final output.

The baseline models do not use temporal information for predicting the severity. We use following common machine learning algorithms to model the severity:

- Logistic Regression
- Decision Trees
- Random Forest

Algorithm 1: Psudacode for Time series Feature Extraction

```
def extract_time_series_data(data):
    for grid in set(data.grids):
        ids = data.filter(grid==grid).ids
        severity = data[ids].severity
        for i in 1 to 3:
            prv = lagged(severity, i)
            data[ids].previous_{i} = prv
    return data
```

E. Spacial Time Series Features

We experimented with spacial-time-series data apart from the features aforementioned in § IV-A2 which were given in the dataset. The process of feature extraction is provided in Algorithm 1.

The first step in spacial time series data extraction is to divide each state into blocks to identify the accidents that are near-by that could help in identifying the severity. We used United States National Grid (USNG) System [1] to identify the grids up to 1000 meter precision. Then we grouped by the time series data for each grid to identify the severity of the most recent accidents (up to 3 events). The *lagged* function in Algorithm 1 is used to induce a step delay in the sequence data for each grid in order to obtain previous events as features.

F. Training Models

This section will describe the parameters and other specifications used in training the models. A model was created by training classifiers for each state as indicated in the Algorithm 2.

All classifiers we used the default hyper-parameters in scikit-Learn [11]. For the Decision-Tree classifier, we use '*gini*' index as the splitting criterion and use '*best*' strategy to split each node. The Decision-Tree classifier is left to split until it finds a pure leaf (a group containing only one category) or until it reaches the *minimum samples* of 1 for the leaf. Our random forest classifier uses *100* trees and similar to Decision Tree it uses '*gini*' index to measure the quality of a split. Furthermore, Random Forest uses the same number of minimum sample in the leaf as our Decision-Tree.

Algorithm 2: Code used for Training the Models

```
def train(clf, data):
    features = [...features in §IV-A2]
    ohe = OneHotEncoder()
    models = Dictionary()
    for state in data.states:
        features ← data.state.features
        x ← ohe.encode(features)
        y ← data.state.labels
        model[state] ← clf.train(x, y)
    return models

def train_all():
    clfs ← [...classifiers in §IV-D]
    data ← load_processed_dataset()
    for clf in clfs:
        train, test ← split(data)
        models ← train(clf, train)
        results = evaluate(models, test)
        show results
```

V. Evaluation

Our primary evaluation is based on train test data. We separate 20% of data as test data and 80% of data as train data. Train data will be used for training the models. For testing the models, we will be using the test data. We will use Precision 4, Recall 5, and F1-Score 6 to evaluate our models in addition to Accuracy 3.

$$Accuracy = \frac{True\ Positives + True\ Negatives}{Total\ Number\ of\ Records} \tag{3}$$

[1] https://www.fgdc.gov/usng

State	Classifier	Features	Accuracy	Precision	Recall	F1
CA	Logistic Regression	BS	82.0%	79.0%	82.0%	80.0%
		BS+STS	87.6%	87.0%	88.0%	87.0%
	Decision Trees	BS	82.0%	88.0%	88.0%	88.0%
		BS+STS	87.8%	88.0%	88.0%	88.0%
	Random Forest	BS	89.0%	89.0%	89.0%	89.0%
		BS+STS	91.0%	91.0%	91.0%	90.0%
TX	Logistic Regression	BS	76.0%	76.0%	78.0%	76.0%
		BS+STS	83.2%	82.0%	83.0%	82.0%
	Decision Trees	BS	84.0%	84.0%	84.0%	84.0%
		BS+STS	82.5%	82.0%	83.0%	83.0%
	Random Forest	BS	84.0%	84.0%	84.0%	84.0%
		BS+STS	87.1%	87.0%	87.0%	87.0%
SC	Logistic Regression	BS	84.0%	82.0%	84.0%	80.0%
		BS+STS	89.5%	89.0%	90.0%	89.0%
	Decision Trees	BS	89.0%	89.0%	89.0%	89.0%
		BS+STS	88.3%	88.0%	88.0%	88.0%
	Random Forest	BS	89.0%	88.0%	89.0%	88.0%
		BS+STS	91.5%	92.0%	92.0%	91.0%
FL	Logistic Regression	BS	71.0%	69.0%	71.0%	69.0%
		BS+STS	78.1%	76.0%	78.0%	77.0%
	Decision Trees	BS	79.0%	79.0%	79.0%	79.0%
		BS+STS	78.0%	78.0%	78.0%	78.0%
	Random Forest	BS	82.0%	82.0%	82.0%	82.0%
		BS+STS	84.0%	84.0%	84.0%	83.0%
NC	Logistic Regression	BS	85.0%	82.0%	85.0%	82.0%
		BS+STS	87.0%	84.0%	87.0%	84.0%
	Decision Trees	BS	88.0%	88.0%	88.0%	88.0%
		BS+STS	86.0%	87.0%	86.0%	86.0%
	Random Forest	BS	89.0%	88.0%	89.0%	87.0%
		BS+STS	90.0%	89.0%	90.0%	89.0%

TABLE VI: Evaluation results provided using Accuracy, Precision, Recall and F1 Scores for each state.

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives} \tag{4}$$

$$Recall = \frac{True\ Positives}{True\ Positive + False\ Negatives} \tag{5}$$

$$F1 = \frac{2 \cdot precision \cdot recall}{precision + recall} \tag{6}$$

Table VI shows the evaluation results of the trained models on the test-set. It presents the Accuracy, Precision, Recall and F1 Scores for predicting the accident severity for each state considered. The presented values for Precision, Recall and F1 scores are weighted based on the support for each category as our dataset is highly unbalanced.

VI. Results and Discussion

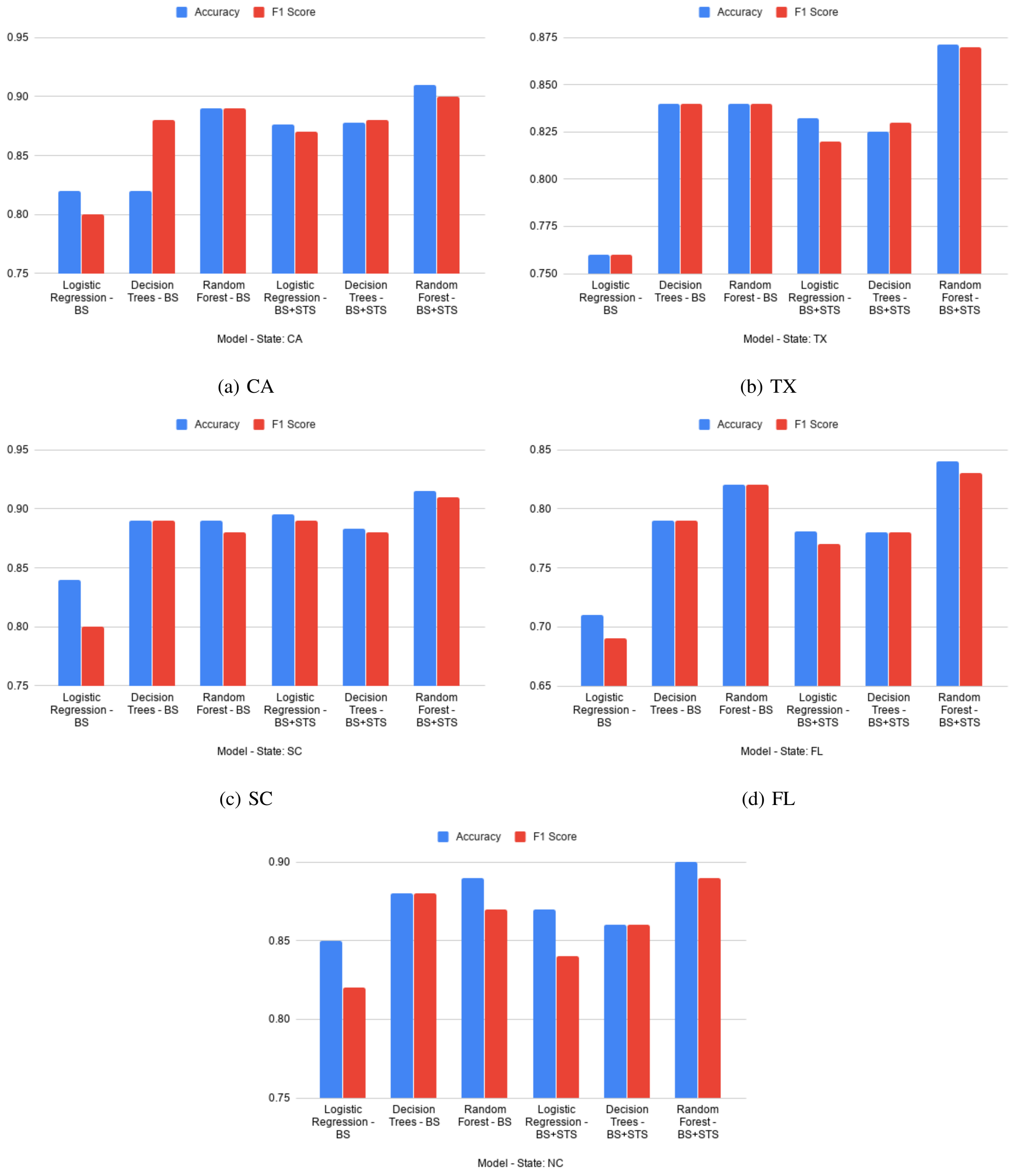
Table VI shows the variation of performance of different classifiers for each state. In addition to the classifier algorithms, the Table VI illustrates the improvement of those algorithms after using spacial-time-series features we proposed in Section IV-E. The column values BS and STS in Table VI represents basic attributes in Table V and Spacial Time Series features in Section IV-E.

Overall we see there is a consistent variation of performance of different classifier algorithms and features in each state. We will discuss these improvements for each state in subsections.

A. Model Performance of Each State

1) California: California is the state with the largest number of accident records for the year 2019 in our dataset. Bar-chart in Figure 8a shows performance of models trained for California. We observed that the increasing performance variation of classification algorithms are in the order: Logistic Regression, Decision Tree, and Random Forest. Furthermore, the addition of spacial-time-series data provides an improvement of 9%, 0%, 1% for algorithms Logistic Regression, Decision Tree, and Random Forest in order.

2) Texas: The state of Texas had more than sixty-five thousand accident records for the year 2019 in our dataset. Figure 8b shows the variation of F1 Scores and Accuracy.



(a) CA

(b) TX

(c) SC

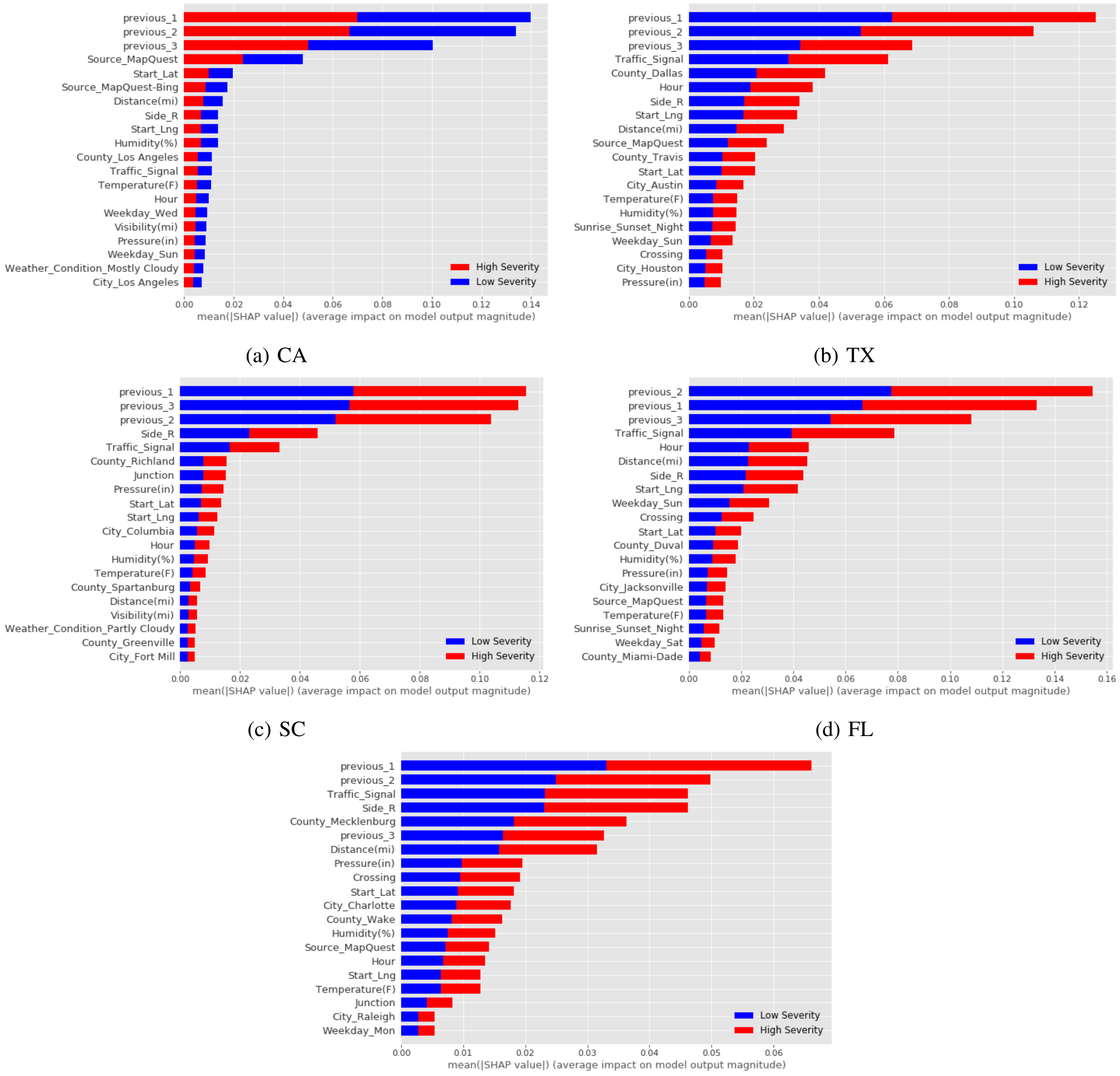
(d) FL

(e) NC

Fig. 8: Evaluation Results for Models by State

We observed that the increasing performance variation of classification algorithms are in the order: Logistic Regression, Decision Tree, and Random Forest. The addition of spacial-time-series data has provided an improvement of 8% and 4% for Logistic Regression and Random Forest Regression. In the case of the Decision Tree Regression model, the performance for STS model reduced by 1%.

3) South Carolina: South Carolina had more than fifty-three thousand accident records for the year 2019 in our dataset. Figure 8c shows the variation of model performance for South Carolina. We observed that the increasing performance variation of classification algorithms are in the order:



(a) CA

(b) TX

(c) SC

(d) FL

(e) NC

Fig. 9: Feature Contribution for Models by State

Logistic Regression, Decision Tree, and Random Forest. The addition of spacial-time-series data has provided an improvement of 11% and 3% for Logistic Regression and Random Forest Regression. In the case of the Decision Tree Regression model, the performance for STS model reduced by 1%.

4) Florida: Florida had a little more than forty-four thousand accident records for the year 2019 in our dataset. The model performance of the Florida Model is included in Figure 8d. We observed that the increasing performance variation of classification algorithms are in the order: Logistic Regression, Decision Tree, and Random Forest. The addition of spacial-time-series data has provided an improvement of 12% and 1% for Logistic Regression and Random Forest Regression. In the case of the Decision Tree Regression model, the performance for STS model reduced by 1%.

5) North Carolina: North Carolina had more than thirty-seven thousand accident records for the year 2019 in our dataset. The performance of the North Carolina model is indicated in Figure 8e. We observed that the increasing performance variation of classification algorithms are in the order: Logistic Regression, Decision Tree, and Random Forest. The addition of spacial-time-series data has provided an improvement of 2% for both Logistic Regression and Random Forest Regression each. In the case of the Decision Tree Regression model, the performance for STS model reduced by 2%.

B. Overall Model Performance

The three models that were used for predicting the severity were Logistic Regression, Decision Tree Regression, and Random Forest Regression. The states that were used for the analysis are California, Texas, South Carolina, Florida, and North Carolina. For each of the states, the three regression models were used and the accuracy and F1 score we calculated in two different methods. In the first method, only the base features are used. In the second method, spacial-time-series data is added.

The value of accuracy and F1 score has improved in the following order: Logistic Regression, Decision Tree Regression, and Random Forest Regression. With the addition of spacial-time-series data, the performance increased for Logistic Regression and Random Forest Regression. In the case of the decision tree algorithm, the accuracy reduced. This is because, in the decision tree model, the number of trees used to predict is one and therefore the accuracy is less. Of all the three regression models, the Random Forest Regression model performed the best on the whole with better accuracy in most of the states.

C. Feature Contribution

This section provides details on the contribution of each feature for the final model performance.

In order to perform feature contribution analysis, we used algorithms proposed by Lundberg et al. [12]. The SHAP tool provides the Shapely values, which could be used to explain the contribution of the attributes for the predictions [13]. The tool takes a game theory approach to find the Shapely values.

Figure 9 provides the contribution of features for each model by state. The x-axis represents the mean SHAP values for a sample of data records with the sample size 100 in the testing set. The features identified by $previous_1$, $previous_2$, and $previous_3$ are STS features lagged by 1, 2 and 3 events respectively.

We observe that the most contributed feature in Figure 9 is the $previous_n$ features. In four out of five $(4/5)$ models we observe that all three $previous_n$ severity features dominate other features. Out of the three $previous_n$ severity features the latest severity feature ($previous_1$) dominates other $previous_n$ severity features. We see that the introduction of Spacial-Time-Series (STS) has contributed towards the prediction.

Interestingly, we see that the source contributes to the model output and is included in top features in many models provided in Figure 9. We believe that the reason for this is the veracity of the data source. By providing the source we present the algorithm to model the accuracy of the data better.

VII. Conclusion

Road traffic accident modeling is a crucial requirement to reduce the effects of traffic accidents such as traffic delays and more serious concerns such as injuries to human life. In this study, we present an approach to model the severity of traffic accidents using various environmental factors and road conditions. The knowledge learned in this model could be used for practical applications to predict the severity of prevailing road accidents or to transfer knowledge for downstream tasks of accident risk prediction. The contributions of this paper include processed and cleaned dataset and machine learning models for accident severity prediction. Moreover, we identify the contribution of features on the models to describe how our features affect the final model and the importance of each of those features on final predictions.

In future research, we hope to integrate more sophisticated methods of presenting the features to the algorithm. The most important change should be how we present the location information to the training algorithm. While providing the coordinates directly to the training algorithm proved useful in predicting the accident severity, we are yet to explore the possibility of using a graph-based approach for presenting the location information.

A. Replicability

This paper provides detailed descriptions of the models, processing steps, and the parameters used in relevant sections. Additionally, the code and the processed dataset for severity prediction is provided in https://github.com/ysenarath/traffic-accident-prediction.

References

[1] WHO, “Global status report on road safety 2018,” 2018. [Online]. Available: https://www.who.int/publications-detail/global-status-report-on-road-safety-2018

[2] ASIRT, “Road Safety Facts,” Apr. 2020. [Online]. Available: https://www.asirt.org/safe-travel/road-safety-facts/

[3] 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 - SIGSPATIAL ’19*. Chicago, IL, USA: ACM Press, 2019, pp. 33–42. [Online]. Available: http://dl.acm.org/citation.cfm?doid=3347146.3359078

[4] A. Najjar, S. Kaneko, and Y. Miyanaga, “Combining satellite imagery and open data to map road safety,” in *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[5] Z. Yuan, X. Zhou, T. Yang, J. Tamerius, and R. Mantilla, “Predicting traffic accidents through heterogeneous urban data: A case study,” in *Proceedings of the 6th International Workshop on Urban Computing (UrbComp 2017), Halifax, NS, Canada*, vol. 14, 2017.

[6] C. C. Ihueze and U. O. Onwurah, “Road traffic accidents prediction modelling: An analysis of Anambra State, Nigeria,” *Accident Analysis & Prevention*, vol. 112, pp. 21–29, Mar. 2018.

[7] L. Lin, Q. Wang, and A. W. Sadek, “A novel variable selection method based on frequent pattern tree for real-time traffic accident risk prediction,” *Transportation Research Part C: Emerging Technologies*, vol. 55, pp. 444–459, Jun. 2015.

[8] F. L. Mannering, V. Shankar, and C. R. Bhat, “Unobserved heterogeneity and the statistical analysis of highway accident data,” *Analytic Methods in Accident Research*, vol. 11, pp. 1–16, Sep. 2016. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S2213665716300100

[9] D. Eisenberg, “The mixed effects of precipitation on traffic crashes,” *Accident Analysis & Prevention*, vol. 36, no. 4, pp. 637–647, Jul. 2004.

[10] Q. Chen, X. Song, H. Yamada, and R. Shibasaki, “Learning deep representation from big and heterogeneous data for traffic accident inference,” in *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[11] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, “Scikit-learn: Machine learning in python,” *Journal of machine learning research*, vol. 12, no. Oct, pp. 2825–2830, 2011.

[12] 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. 2522–5839, 2020.

[13] L. S. Shapley, “A value for n-person games,” *Contributions to the Theory of Games*, vol. 2, no. 28, pp. 307–317, 1953.