World Academy of Science, Engineering and Technology
International Journal of Transport and Vehicle Engineering
Vol:18, No:8, 2024

Probabilistic Crash Prediction and Prevention of Vehicle Crash

Lavanya Annadi, Fahimeh Jafari

***Abstract*—**Transportation brings immense benefits to society, but it also has its costs. Costs include the cost of infrastructure, personnel, and equipment, but also the loss of life and property in traffic accidents on the road, delays in travel due to traffic congestion, and various indirect costs in terms of air transport. This research aims to predict the probabilistic crash prediction of vehicles using Machine Learning due to natural and structural reasons by excluding spontaneous reasons, like overspeeding, etc., in the United States. These factors range from meteorological elements such as weather conditions, precipitation, visibility, wind speed, wind direction, temperature, pressure, and humidity, to human-made structures, like road structure components such as Bumps, Roundabouts, No Exit, Turning Loops, Give Away, etc. The probabilities are categorized into ten distinct classes. All the predictions are based on multiclass classification techniques, which are supervised learning. This study considers all crashes in all states collected by the US government. The probability of the crash was determined by employing Multinomial Expected Value, and a classification label was assigned accordingly. We applied three classification models, including multiclass Logistic Regression, Random Forest and XGBoost. The numerical results show that XGBoost achieved a 75.2% accuracy rate which indicates the part that is being played by natural and structural reasons for the crash. The paper has provided in-depth insights through exploratory data analysis.

***Keywords*—**Road safety, crash prediction, exploratory analysis, machine learning.

I. Introduction

WITH the widespread use of sensors, coupled with the latest developments in wireless technologies, Intelligent Transport Systems (ITS) have become commonplace. The focus is collecting data from the distributed sensors, archiving it, analyzing it, transforming it into actionable knowledge, and disseminating it through multiple transportation applications that provide planning, mobility, and safety assistance. Data are critical to the transportation sector and all modes of transport, and now transport operators have access to a vast amount of information that they can use to optimize performance, service, efficiency, and safety. Data are continuously collected in the ever-increasing number of sensors, remote sensors, cameras, microphones, wireless sensor networks, and mobile devices. Due to the wide availability of data through multiple technologies, datasets can rapidly become too large and complex to be processed using traditional data analysis. Still, with increasing transportation demands, data play an increasingly important role in transportation management and administration. Due to the expansion of traffic and detectors, an increase in data is manifest in the volume of available traffic information, which is constantly growing.

This paper uses an expanded accident dataset to study the specific variables, which are weather-related and road structure phenomena that are being caused for crashes. It will take a probabilistic approach and classify the crash percentage of vehicles using Multinomial Logistic Regression, Random Forest and XGBoost techniques. These techniques are part of supervised machine learning algorithms of classification. They consider both natural reasons for crashes and man-made structures. Natural reasons for the crash are weather conditions, GPS location, temperature, pressure, wind speed, wind direction, precipitation, and other man-made systems like, Bump Roundabout, Turning loop etc. Overall, this paper considers random factors that belong to natural and manmade causes. Many machine algorithms can tackle this problem, from supervised learning to reinforcement learning. We preferred supervising learning techniques over unsupervised because of the availability of sufficient amounts of data, and the probabilistic models are classification models.

The rest of this paper is organized as follows: Section II discusses the related works. Section III introduces the problem statement. Section IV is devoted to the research methodology underlying the system model and the notations in our analysis. Section V describes the details regarding data collection, data pre-processing, and feature engineering in this research. Section VI discusses the multiclass classifiers used for this problem. Section VII represents numerical results from exploratory data analysis and multiclass classification models. Section VIII provides deep insights based on numerical results. Section IX discusses the contributions of this paper. Finally, Section X concludes this paper and gives ideas for future works.

II. Related Works

Authors in [1] have worked on a Smart watch-based driver vigilance indicator with the Kernel-Fuzzy-C_means-Wavelet method. They found that sensors at the fingertips give better data collection and then proposed and tested different AI methodologies. In [2], the authors proposed a system that can predict and avoid crashes in real-time with past data. In this paper, different data analytics are performed on a crash database of 14 years from UK DoT, which contains a million rows and 32 categories and presents Peak Accident Time, Day, Week, Location, etc. The authors in [3] surveyed various technologies utilized in the automotive industry, ranging from vehicular networks to artificial intelligence (AI). They

Fahimeh Jafari is with University of East London, United Kingdom (e-mail: f.jafari@uel.ac.uk).

described how decision trees and artificial neural networks (ANN) could be applied to data gathered from diverse sensors. The paper delved into the interaction between vehicles and the environment and message-sharing systems. Furthermore, it explored different AI methods and how they can be integrated with vehicle networks.

Reference [4] represents a generative model that solves GPS errors without knowing environmental conditions and uses the information inside a vehicle for Automated Vehicles. It proposes a model to reduce GPS residual errors, which mimics the environmental conditions and information inside the vehicle only. Authors [5] researched predicting automated driving vehicles' trajectories during cut-in lanes. They provided a model predictive controller and built a probabilistic trajectory of lane cut-in using previously trained data with interaction awareness capability of other vehicles approaching.

Reference [6] discusses the application of AI in mitigating the geo-hazards risk for vehicles in the Mountains of Beijing. The paper mainly contributes to applying AI in different categories, e.g., how well the information extracted from Images can be used to mitigate more crash drivers. Authors [7] propose a deep learning-based mapping approach for predicting city-scale road safety maps from raw satellite imagery. They investigate the usage of satellite imagery for Road safety and the detection of problematic structures that can cause road crashes.

Reference [8] worked on predicting crashes using neural networks by taking vehicle, driver, and road characteristics data. It explains the complex and nonlinear relationships between road crashes and their reasons by taking huge parameters. It helps vehicle drivers be more aware of hotspots and their conditioning in moving through them. If the model is updated every time, it will yield better results. Authors [9] provide a public representative support system to identify the unwanted things on the road to remove or change the structure that causes many traffic and accident problems. They explain how public authorities can spend their funds on effective road safety. They use a new technique called 'Concordance analysis' to rank the structure or part of the road. Some other works on this topic investigate the critical factors of road accidents [10]-[13]. Some safety models minimize or predict accident factors [14]-[17].

This paper uses an expanded accident dataset to study the specific variables, which are weather-related and road structure phenomena that are being caused by crashes. It will take a probabilistic approach and classify the crash percentage of vehicles using three different multiclass classification techniques.

III. Problem Statement

As per our problem statement, we aim to predict the probability of a crash of a vehicle given by natural and man-made interventions. There are many other reasons to establish a reason for a crash. It may be because of the driver's misguided attention, drunk driving, traffic rule-breaking, etc. However, the scope of this study is limited to the natural and structural causes of the crash.

IV. Research Methodology

We obtained a dataset from the US government to build a predictive model for crash probability. This dataset comprises 2.7 million instances of crash data spanning a particular timeframe, along with an observational analysis of various random variables linked to each crash. There are two distinct methodologies for conducting research: qualitative and quantitative. Researchers can gather more information by combining qualitative and quantitative methods, mainly when data cannot be observed and measured directly. In our case, while the collected dataset includes many observed values of natural and structural phenomena, the probability of which category they belong to is not directly mentioned. Therefore, we will utilize the quantitative method to obtain further information.

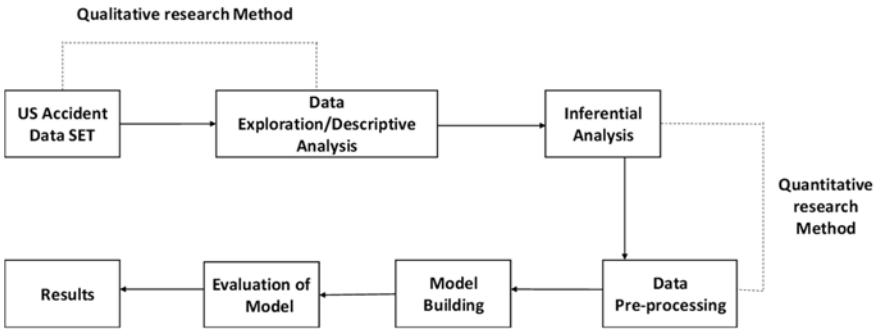


Fig. 1 Research Pipeline

Fig. 1 represents the entire pipeline of research. The dataset collected for the United States Government website has been used for our study. It has been explored in depth to analyze the descriptive statistics as Exploratory Data Analysis by applying qualitative research methodology. However, we need pre-processing and some inferences to assume any relationship between the data. Our task is to predict the crash probability in ten bins of 100-width. After performing the necessary Quantitative Analysis for inference-making, data were pre-processed using those inferences.

V. Data Collection

A dataset is taken from the United States government website, which contains 47 different variables that were collected during crashes of 2.7 million occurrences. The parameters include *temperature*, *pressure*, *wind temperature*, *humidity*, *precipitation*, *wind speed*, *wind direction*, *weather type*, *GPS coordinates*, *time of crash*, *traffic congestion* due to *crash*, *bump*, *crossing*, *give way, junction, no exit, traffic signal, etc.*

This dataset is perfect for studying our area, but no dataset can give complete information as usual. We need more data for a comprehensive analysis, so prescriptive analytics were performed on this dataset using exploratory data analysis.

A. Data Pre-processing

Data pre-processing is the most critical task before building any model because the right kind of data should be appropriately chosen before estimating any parameters. It involves data exploration, drawing inferences, data cleaning, and feature engineering to be done to data given as input to the model. Our problem statement specifies the need to predict the likelihood of a crash based on natural and structural causes. The dataset requires cleaning and inferring certain conclusions, which are outlined in Subsections V *B* and *C*. The qualitative method gave us collected data. Quantitative methods will allow us to do the rest. From the given dataset, the most influential factors for our research are *Start_Latitude, Start_Longitude, Side, Temperature, Pressure, Humidity, Visibility, Wind Speed, Wind direction, Wind chill, Weather Type, Precipitation, Turning loop, Give Way, No exit, Traffic Signal, Roundabout, Junction, Crossing, Bump*. These features include the data suited for calculating our probability for data to be estimated for natural and structural reasons. The rest of the data can be used for data exploration, but these columns are concluded for our Problem statement.

B. Data Cleaning

Data cleaning is the process where unwanted and unnecessary data types that can interfere with the model are removed. A dataset was extracted containing all the variables mentioned above. This dataset holds multiple Not a Number values (Nan), and some have no Types. These rows are removed from the dataset to execute a clear model.

C. Feature Engineering

Now that our dataset is ready by cleaning it, we need to engineer the necessary features to be given as input to the Machine Learning model. Feature engineering is a technique that is both analytical and quantitative. It ensures that the data fed to the model are correctly encoded, and enough features are selected.

For our problem statement, we need to predict the probability of a vehicle crash given the inputs from the above-selected columns. However, the dataset does not contain any prior probability, and the whole variables in a dataset point to the one class that is Crash probability 100%. For this reason, a whole new column of different classes must be engineered. We chose to give the name of that column as Probability class which contains the following classes: 0-10, 11-20, 21-30, 31-40, 41-50, 51-60, 61-70, 71-80, 81-90, 91-100.

Here are the steps performed for feature engineering:

1. As all data points are evidence of a crash, we assigned the highest probability class to them. That is 91-100.
2. We calculated the Mean, Range, and Median of all columns separately for every column.
3. We calculated the frequency of each point in the column relative to other values in the column. That gives us the probability of that value in relation to other values. We suppose there are N rows in the column and value B occurred I times in the column. The probability of finding B in that column is given as:

   $P(B) = (B*I)/N$ ----- For every value in every column

4. By using Range and Mean ($\boldsymbol{\mu}$) we created multiple rows that vary randomly for every column.
5. We calculated the expected value for each row by multinomial probability mass distribution function as shown in the diagram presented in Fig. 2.

$$f(x_1, \dots, x_k; p_1 \dots, p_k) = \propto \frac{\Gamma(\sum_i x_i + 1)}{\prod_i \Gamma(x_i + 1)} \prod_{i=1}^{k} p_i^{x_i} \qquad (1)$$

6. Then, we calculated the expected value of each row by multiplying the data value with its relative frequency and assigned that score to every row.
7. All the values that are near the mean of the original points are assigned the same probability of 91-100. Recommended by the score, the rest of the rows are allotted with different class labels from 0-10 to 81-90.
8. By doing so, a new column of classes is generated.

VI. Multiclass Classification

As explained before, we need to use multiclass classification due to the nature of our problem. Given a dataset, it can be dissected or trisected into training and testing samples with the already observed class it belongs to. While the model trains on training samples, the testing set is used to predict unseen data. This lets us know how to evaluate the model that is trained. Model implementation has the following stages:

1. Dataset Division stage
2. Data Encoding Stage
3. Learning Stage
4. Calculating cost function, Minimizing cost function.
5. Calculating probabilities
6. Predicting Stage

Dataset division stage: During this stage, data provided with 'n' number of samples are divided into testing and training sets with 0.33 and 0.66 ratios, respectively.

Data Encoding Stage: During this stage, data should be properly encoded to give it as input to the model. Datasets may contain two types of variables: numerical, which can be quantified, and categorical, which primarily consists of text. As ML models cannot comprehend text data, label encoding is

employed to transform the categorical data into numerical values. One-hot encoding technique is then used to carry out this process. This technique converts all categorical values into numbers and then inputs them into the model.

Learning stage: During this stage, the Machine learns from the training samples we provide, which carries its observed output from the collected data. Every example row will be fed to the model. The model learns from each example by calculating and minimizing the cost function.

Calculating and minimizing cost function: Every training sample carries its observed value when some line is fit to predict the output. The machine will compare it with the observed output and calculate its differences. It is called the Cost function. Minimizing this cost function gives us the best line for the samples fed to the SoftMax function.

Calculating probabilities and *Prediction Stages* depends on the Classification models used for the problem. In this paper, we have applied three common multiclass classifiers as in Subsection *A*.

A. Multinomial Logistic Regression

Multinomial Logistic regression is the most suitable algorithm for many classification problems because it can work on any large dataset. It is prescribed mostly when there are multiple classes in the data that are needed to be predicted. Logistic regression is like linear regression, but instead of predicting a certain value, it predicts a certain probability to which class it belongs. The formula to calculate the logistic regression is:

$$P(z/x) = 1/(1 + exp\ (-z)) \quad (2)$$

The learning and evaluation process of logistic regression is illustrated in Fig. 2. To address the nature of the problem, we have employed the Multinomial or One vs Rest classifier, wherein probabilities are computed through One vs Rest. The aim is to determine the probability that the given sample data belong to all the categories listed in the dataset and then classify the sample into the category with the highest probability.

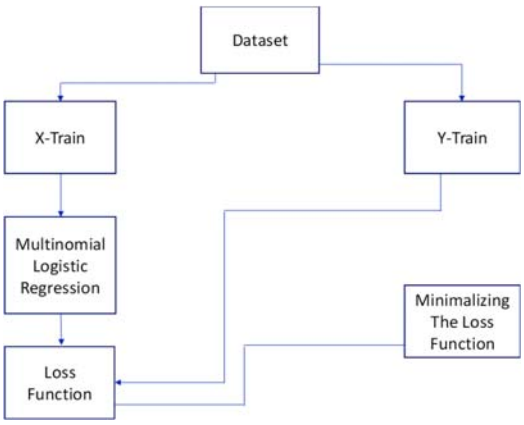


Fig. 2 Learning and Evaluation of Logistic Regression

During the Prediction Stage, the model will receive the test dataset one by one to make predictions and determine its performance. The probability prediction process is as follows: Let *K* be the number of classes to be calculated for the prediction. *P1, P2, P3… Pk* are the probabilities that the model calculates. Instead of calculating the probability for one class, it follows One vs Rest. The model then calculates the probability of belonging to every class in the dataset. If *I* is the sample, the test data model will calculate the corresponding probabilities.

$$P(I/x) = Max(P1/(P1 + P2 + P3 + \ldots + PK)) \quad (3)$$

Then, the model calculates the probabilities for *P2… Pk*, and identifies the class with the highest value.

B. Random Forest

The Random Forest algorithm is one of the most popular machine learning algorithms comprising Decision Trees. The more trees it has, the more sophisticated the algorithm is. It selects the best result from the votes polled by the trees, making it robust. This creates numerous branches of decision trees randomly to determine the class probability. It uses a multi-node approach from where the decision branches are created based on the requirement.

To achieve precise predictions, random forests employ a multitude of decision trees. While there is a common notion that having numerous trees might lead to overfitting, it appears not to be a drawback. This is because only the optimal prediction (the one with the most votes) is selected from the potential output classes, ensuring seamless, dependable, and adaptable executions. An important strength of Random Forests is that they can perform well in the case of missing data. According to their construction principle, not every tree uses the same features. If there is any missing value for a feature during the application, there usually are enough trees remaining that do not use this feature to produce accurate predictions. On the other hand, when applied to regression problems, Random Forests have the limitation that they cannot exceed the range of values of the target variable used in training. Thus, Random Forests may perform poorly with data that are out of the range of the original training data [18].

C. XGBoost

XGBoost is a decision-tree-based ensemble Machine Learning algorithm that uses a gradient-boosting framework. This is an implementation of gradient boosting that pushes the edges of computing power for boosted tree algorithms. It was developed primarily to enhance machine learning performance and computational speed. The intuition behind it is that the best possible next model will minimize the prediction error when combined with previous models. For this next model, the key concept is to set target outcomes to minimize errors. Depending on how much each case's prediction changes the overall prediction error, the target outcome for that case will vary [19]:

- The next target outcome of a case is a high value when a small change in the prediction leads to a large drop in error. The error will be reduced by using a model that predicts close to its targets.
- The next target outcome of a case is zero when a small

change in the prediction leads to no change in error. A change in this prediction does not reduce the error.

VII. Numerical Result

Results of the research can be divided into two sections:

- Results from Exploratory Data Analysis (EDA)
- Results from Multiclass Classification Models

A. Results from EDA

We map all the GPS coordinates from the dataset the United States map as Fig. 3, where a red spot indicates the site of accidents.

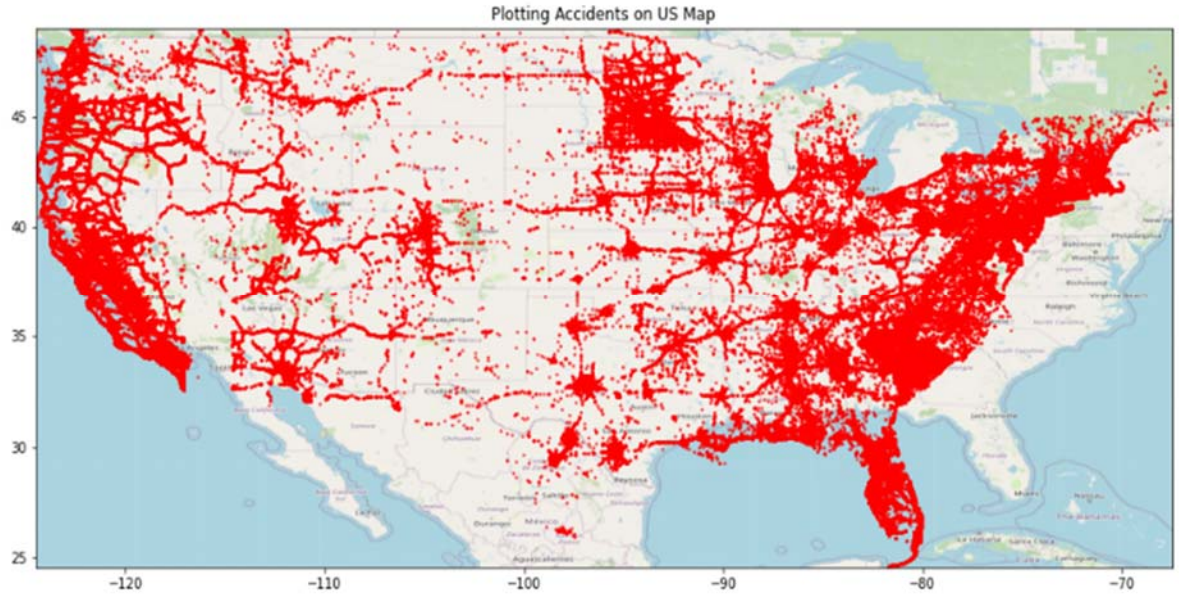


Fig. 3 Plotting accidents from the dataset on a US map [20]

TABLE I
Descriptive Facts of Location Latitude and Longitude in Degrees

	Range	Mean	Median	Max
Latitude	24.446	36.53	36.099	49.002
Longitude	57.510	-96.426	-91.166	-67.113
	Min	Variance	Standard deviation	
Latitude	49.002	25.139	5.0139	
Longitude	-124.62	315.208	17.754	

Fig. 5 indicates that Severity Type 2 crashes are causing the highest congestion, and severity Type 3 crashes are followed by that. Severity is an ordinal number. A higher number indicates heavy traffic congestion and a lower number indicates low traffic congestion due to crashes. Severity Type 2 is followed by Severity Type 3, which shows these large types of traffic congestion. Traffic congestion affects every other person who has been traveling on the road.

Fig. 3 is plotted by the coordinates given in the dataset and mapped from open-street.org using geopandas. This figure clearly shows that accident density is higher in coastal regions of the United States, especially California and Florida.

Start_lat and *Start_lng* are the area features that correspond to the observed values of location latitude and longitude in GPS, respectively. They represent where latitude and longitude accidents have occurred. Kernel Density Estimation is used to measure the probability of finding cumulative data. It is also called cumulative distribution. It is measured to create new data points that follow the same type of distribution. Table I shows some of the descriptive facts for these features.

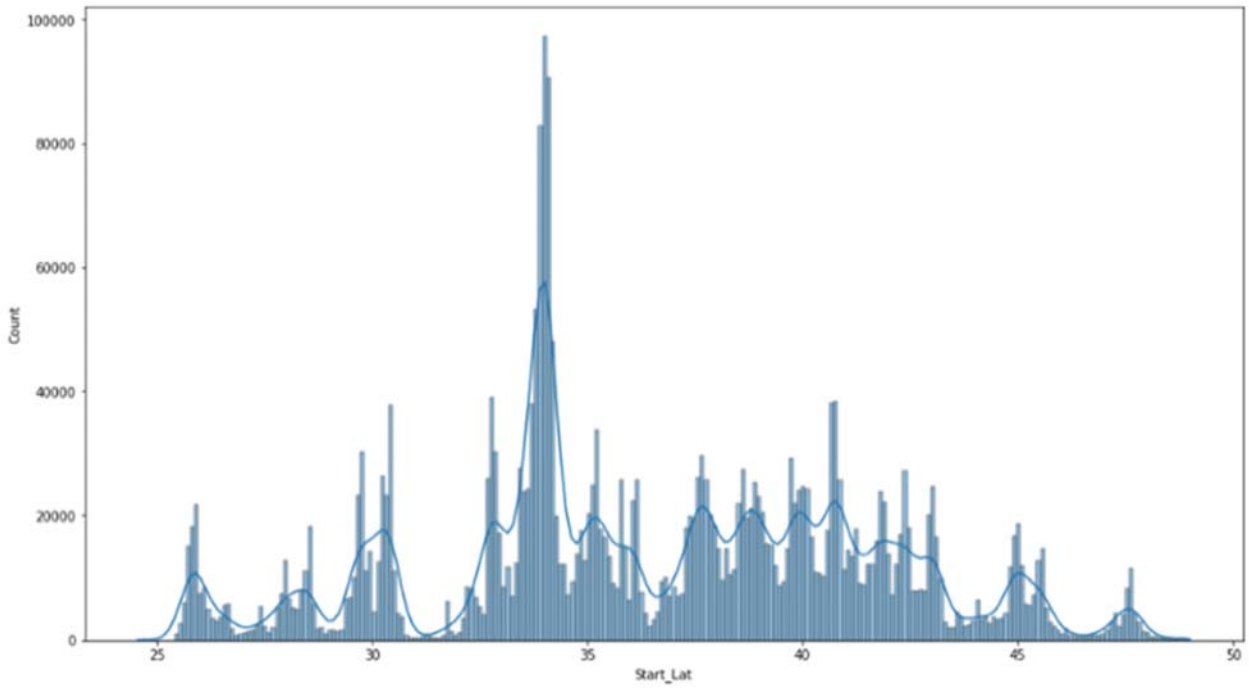
We can see from Fig. 4 that more accidents occur at 34 degrees start latitude and -117 degrees longitude. Mid-level latitudes have seen a greater number of accidents than either side.

Four types of severity indicate a scale on how severe the traffic has been congested. It represents the effect that other people on the road felt.

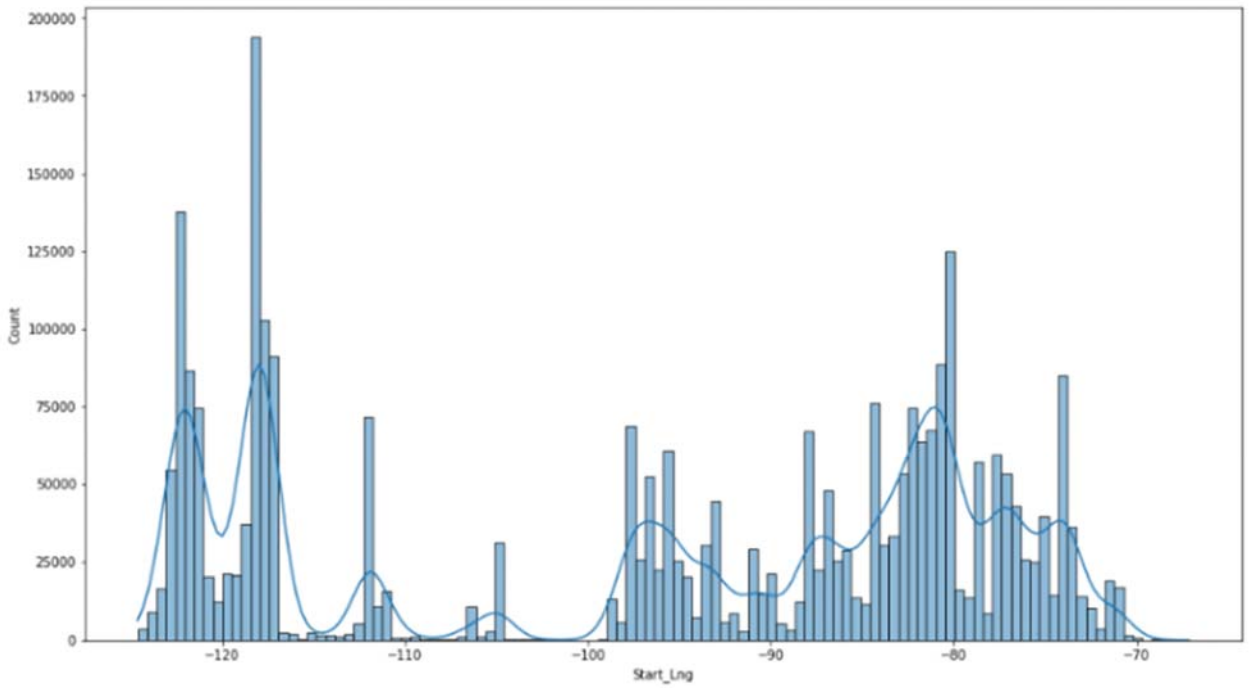
End_lat and *End_lng* are the area features that correspond to the observed values of location latitude and longitude in GPS, respectively. These features denote the point where traffic congestion concludes based on latitude and longitude. These are numerical variables. So, we can calculate the information represented in Table II through quantitative methods.

TABLE II
Descriptive Facts of Congestion End Location Latitude and Longitude in Degrees

	Range	Mean	Median	Max
Latitude	-24.519	-36.517	-36.058	-49.075
Longitude	- 57.514	-96.203	- 91.051	-67.109
	Min	Variance	Standard deviation	
Latitude	-24.555	-25.166	-5.0166	
Longitude	-124.62	-311.865	-17.659	



(a)



(b)

Fig. 4 Histogram of accident location with KDE: (a) Location latitude; (b) Location longitude

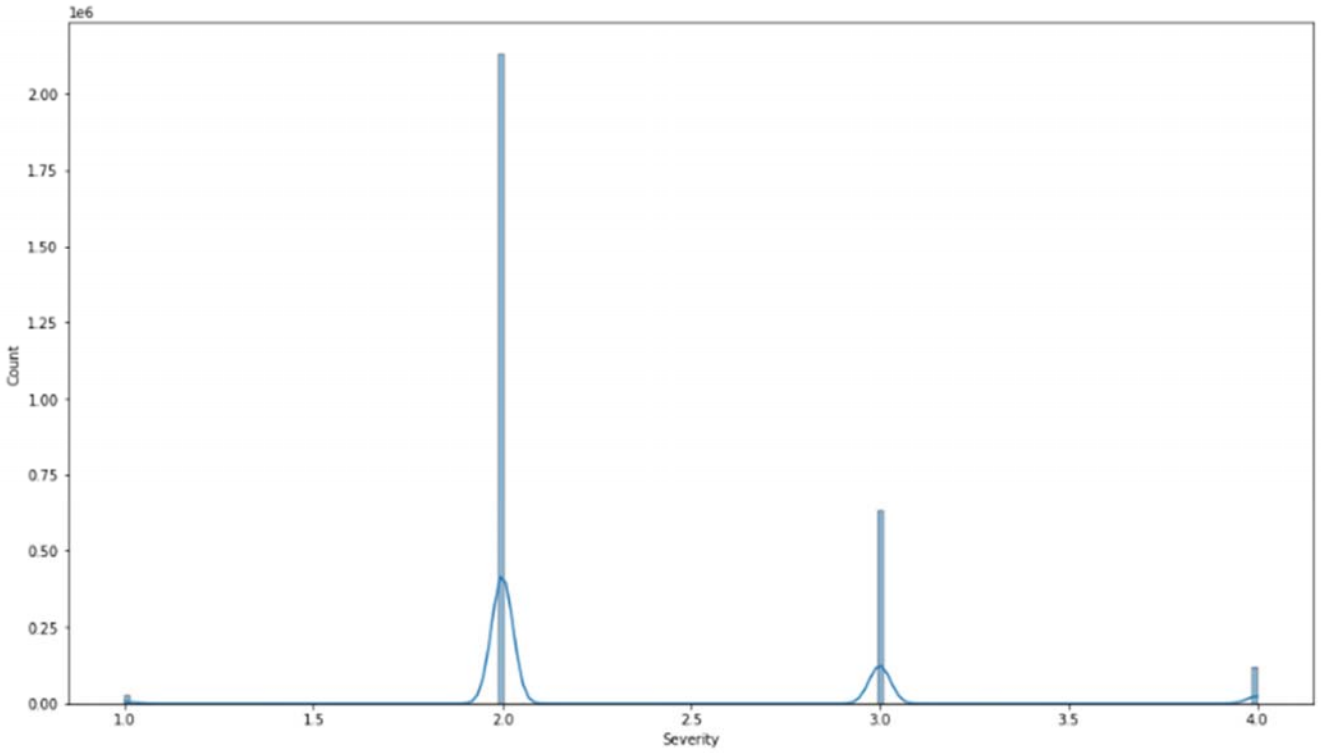


Fig. 5 Histogram of Severity with its KDE

We can see from Fig. 6 that crash data spread from -24 degrees longitude to -49 degrees latitude on Globe. It converges with United States Geographical coordinates.

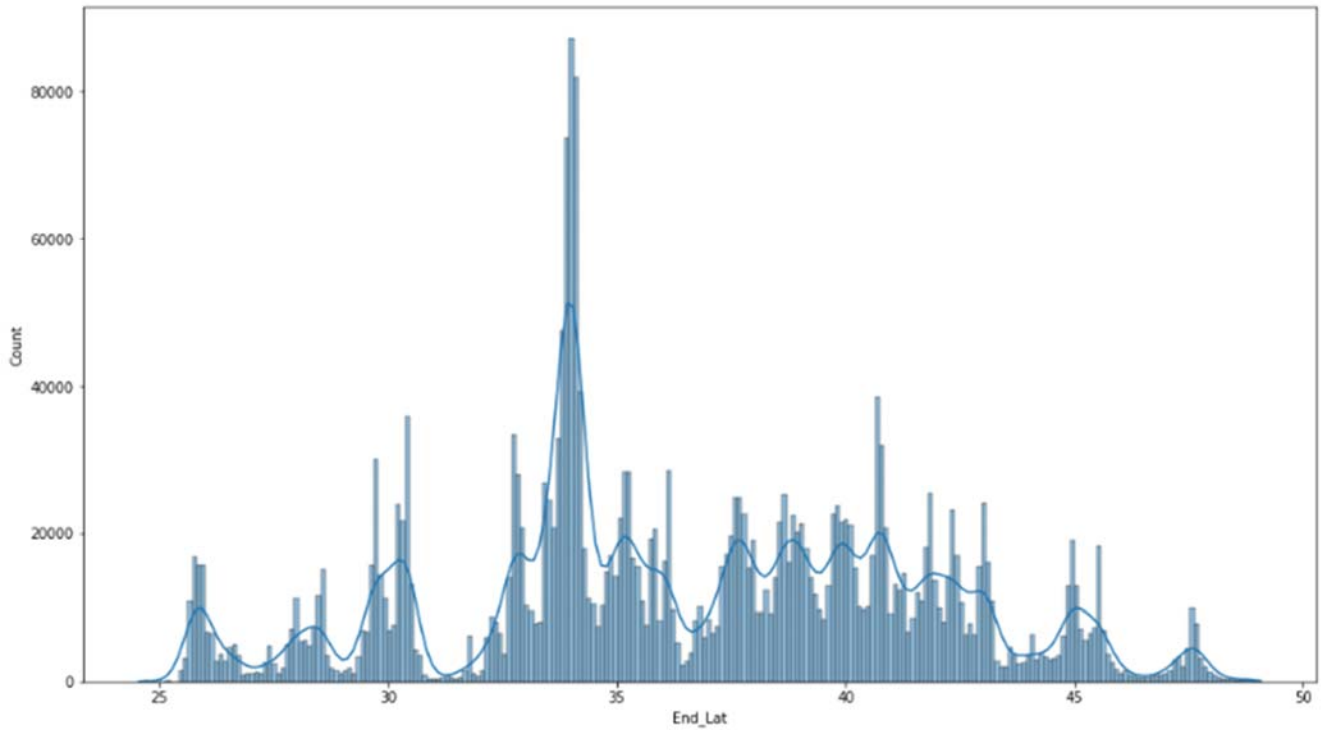
Analyzing the distance of traffic congestion shows that the maximum number of crashes causes no traffic congestion, which indicates roads are in pretty good condition or Authorities may have worked efficiently to clear the traffic. The highest traffic congestion happened with 333 miles of traffic jammed. It may indicate a very big accident that entirely cut the road commute.

As shown in Fig. 7, more accidents occurred on the Right Side of the road than on the left. The right side in the United States indicates the regular traveling traffic rule. It means accidents by the wrong route are far less than accidents on the regular traveling side, 'Right Side'. There might be so many underlying causes for this. We require additional data to conduct a thorough analysis.

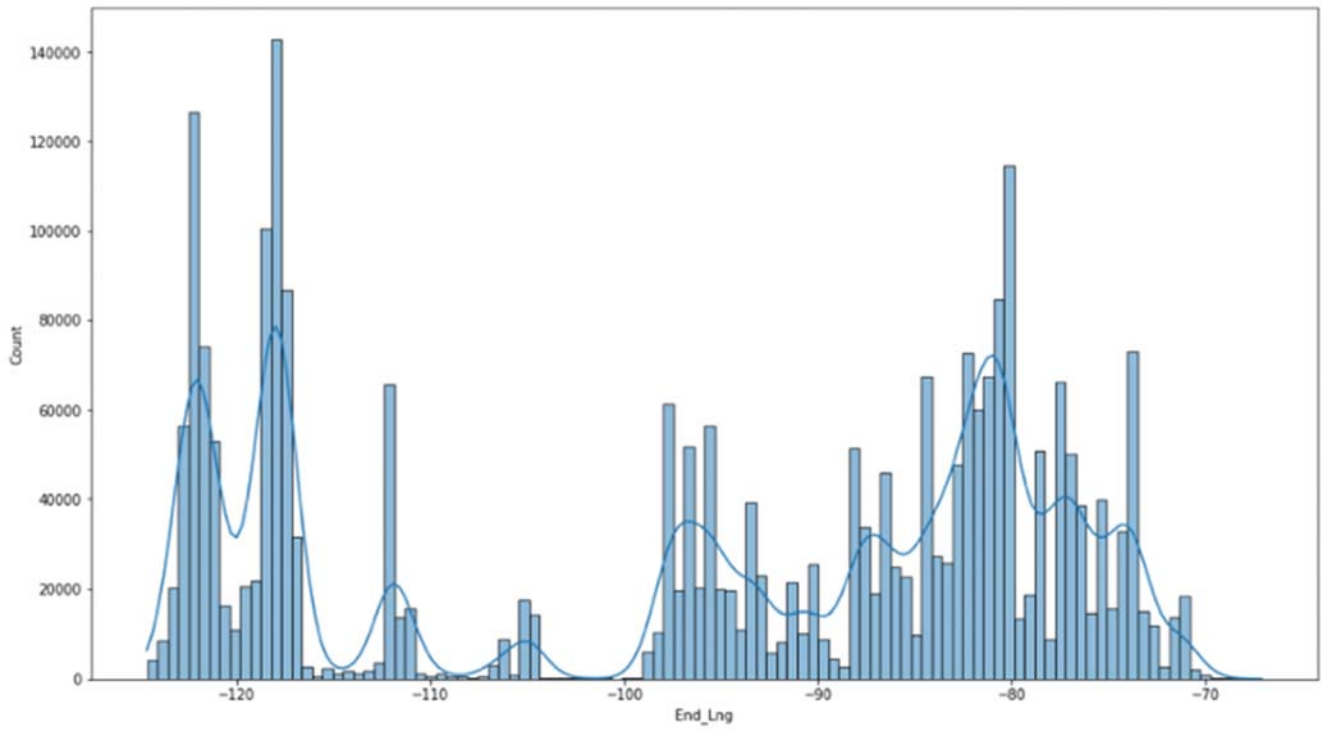
Fig. 8 (a) shows that California (CA) and Florida (FL) have more crashes than the rest of the states, while Vermont (VT), North Dakota (ND), South Dakota (SD) and Wyoming (WY) have very few crashes. There are many reasons for this type of outcome ranging from road structures and conditions to population density and traffic volume, but it indicates that CA and FL had more crashes. Fig. 8 (b) shows that more crashes happened in the US/Eastern Time Zone, which belongs to DC and Florida. Coastal areas come Second, and the Central Region has a medium number of crashes. Mountain Region has a significantly smaller number of crashes compared to other Time zones.

Fig. 9 (a) indicates the histogram of windchill for every accident with its kernel density estimation. As expected, the figure follows a bimodal distribution with two peaks. Most accidents happen between two ranges, one between 10 degrees Fahrenheit to 50 degrees Fahrenheit and another range from 50 degrees Fahrenheit to 100 degrees Fahrenheit. At extreme temperatures, very few accidents were recorded. It can be due to other causes like low traffic volume at extreme values.

Fig. 9 (b) shows humidity values recorded during and at crash locations. Humidity is the percentage of moisture in cubic cm of air. It is measured in percentages. It correlates with temperature. If the temperature decreases, humidity can convert into rain. This is one of the Natural factors that can be considered for our problem statement. The figure shows that the humidity distribution is left-tailed and skewed to the right. As Humidity increases, the number of crashes also increases. The observed values tell us there are more accidents when there is 100% humidity and 90% humidity. When Humidity is zero, the number of accidents is less. It is just a correlation between Humidity and crashes. It is not causation; some underlying factors might also be involved.



(a)



(b)

Fig. 6 Histogram of end location of accident congestion with KDE: (a) Location latitude; (b) Location longitude

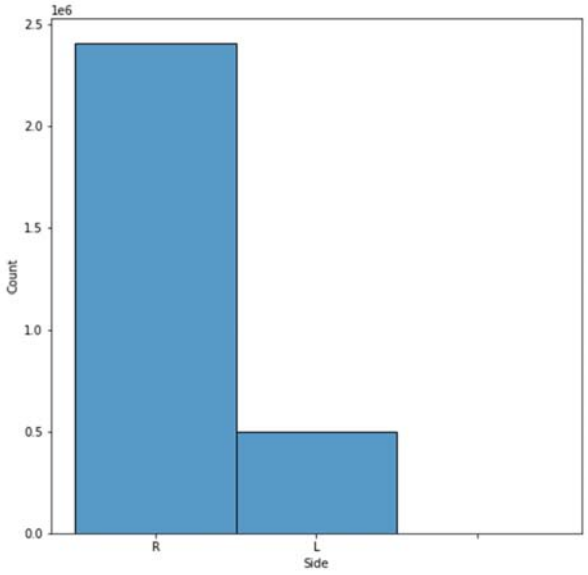


Fig. 7 Histogram of Side column

Fig. 10 (a) shows the histogram of recorded pressure observations during crash locations. Pressure is a force caused by the atmosphere in a particular area. It is inversely correlated with temperature and directly correlated with Humidity. Pressure can affect the person driving the vehicle. The human body is made to withstand only a certain amount of pressure. If it falls below or becomes high, it will have severe consequences on a person's health, so it must be somehow related to the accident as one of the natural factors. As shown in Fig. 10 (a) distribution of this feature follows the single peak distribution close to normal distribution but not precisely. It has one peak at 30 inches of pressure. It tells us that more accidents happened at 30 inches of pressure, and entire vehicle crashes occurred in the range of 27.5 to 31.5 approximately. The human body can withstand 50 inches of sudden impact pressure and gradual pressure up to 400 inches. So, Pressure has very little relevance here until and unless it affects other variables in some correlated way.

Fig. 10 (b) indicates the histogram of observed visibility values during crash events. Visibility is measured in the distance measured in miles in this dataset. Visibility can affect the probability of a crash severely. It can lead to many crashes in the same place and can also be a cause of traffic congestion. From the figure, it is obvious that most of the accidents are caused when visibility is between 0-10 miles. Surprisingly, crash values are higher when visibility is at 7-10 than when there is zero visibility. It confirms that other causes for accidents might influence the crash probability more than Visibility. When visibility is far higher, there are very few crashes. This indicates that column values can be a minor factor in deciding a crash.

Fig. 11 (a) shows the values of wind direction during crashes. Wind direction may not cause the crash because the direction is less critical when the wind flows slower. As shown in the figure, there is no clear pattern in the distribution; it lies somewhere between the Uniform distribution except in two cases where a greater number of accidents happened in the calm wind direction. A small number of accidents happened during variable wind and eastern wind direction. This can be a minor factor from natural causes to decide the crash probability. Wind speed can affect crash probability when it blows against or alongside the vehicle during travel. This entirely depends upon how fast the wind is blowing.

Fig. 11 (b) shows the histogram of precipitation, which contains the values of precipitation measured in inches for each observed crash incident. Precipitation can be of many types, such as Rain, Drizzle, Snow, etc. It is usually measured in centimeters or inches. Precipitation can affect the crash probability in many ways like reducing visibility and making roads sloppy. By analysing this, we can understand how precipitation affects crash probabilities. Some of the quantitative are presented in Table III.

TABLE III
DESCRIPTIVE FACTS OF PRECIPITATION

Range	Mean	Median	Max
24.0	0.011	0.0	24.0
Variance	Standard deviation		Mode
0.023	0.154		0.0

The precipitation data span a range of 24 inches, with a minimum of 0 inches and a maximum of 24 inches. The mode and median are both 0.0, and the mean is approximately 0. Despite the wide range, the maximum value of 24 falls outside the last quarter of the interquartile range, indicating it could be an outlier. Considering this, we may choose to exclude it from our analysis. Most accidents occurred within the range of 0 to 1 inch of precipitation. When factoring in visibility data, it appears that precipitation may not be a significant contributing factor to accidents.

By analyzing data on weather conditions, it becomes evident that in fair weather, accidents are primarily attributed to spontaneous errors by commuters and other variables. The next highest incidence of accidents occurs in cloudy weather, which, when combined with precipitation, can significantly reduce visibility on the roads, thereby increasing the likelihood of accidents. The fewest accidents occurred during thunder/hail, heavy smoke, sleet/windy, and rain and sleet weather conditions. It is important to note that this feature consists of categorical values. Therefore, a qualitative approach may offer deeper insights compared to a quantitative one. Table IV shows the number of accidents versus some effective structural parameters.

TABLE IV
THE NUMBER OF ACCIDENTS VS SOME EFFECTIVE STRUCTURAL PARAMETERS

	Amenity	Bump	Crossing	Give Way	Junction
No	2875240	2906031	2687681	2898390	2630533
Yes	31370	579	218929	8220	276077
	No Exit	Railway	Roundabout	Station	Stop
No	2902752	2880683	2906468	2848700	2861156
Yes	3858	25927	142	57910	45454
	Traffic Calming		Traffic Signal		Turning Loop
No	2905303		2452945		2906610
Yes	1307		453665		≅0

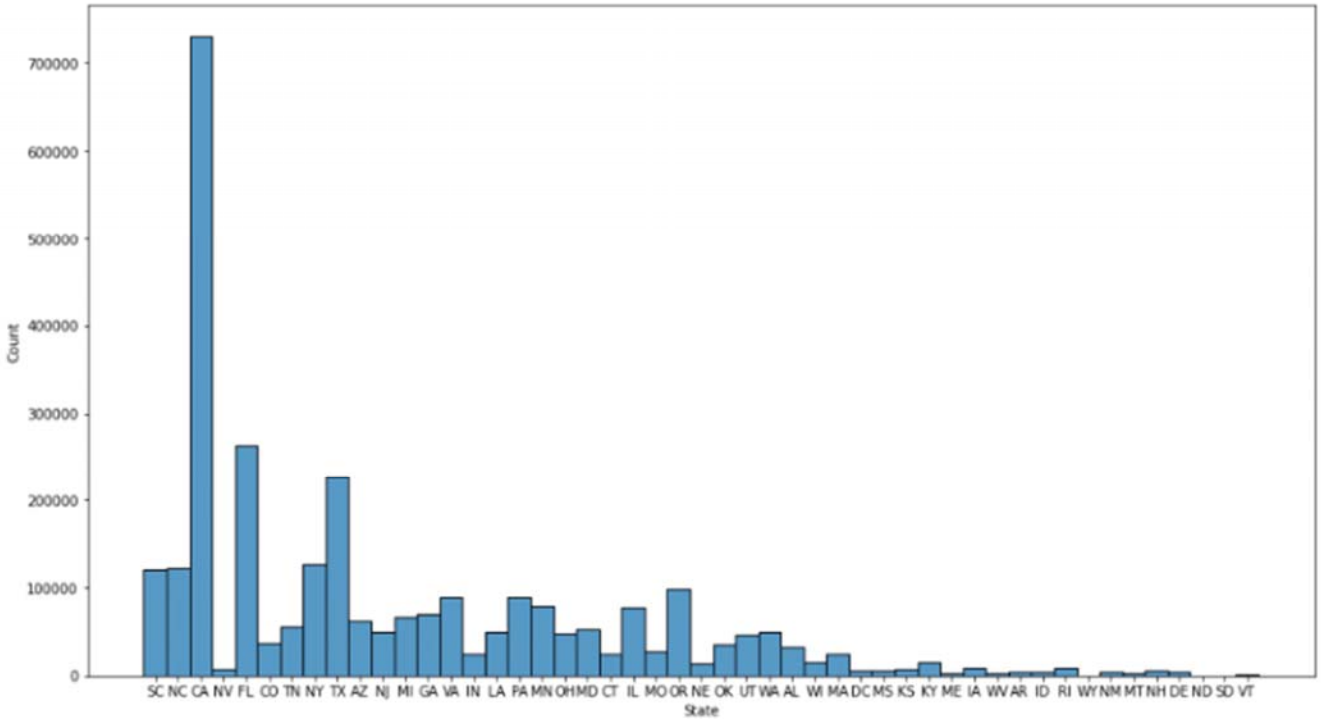


Fig. 8 (a) Histogram of States

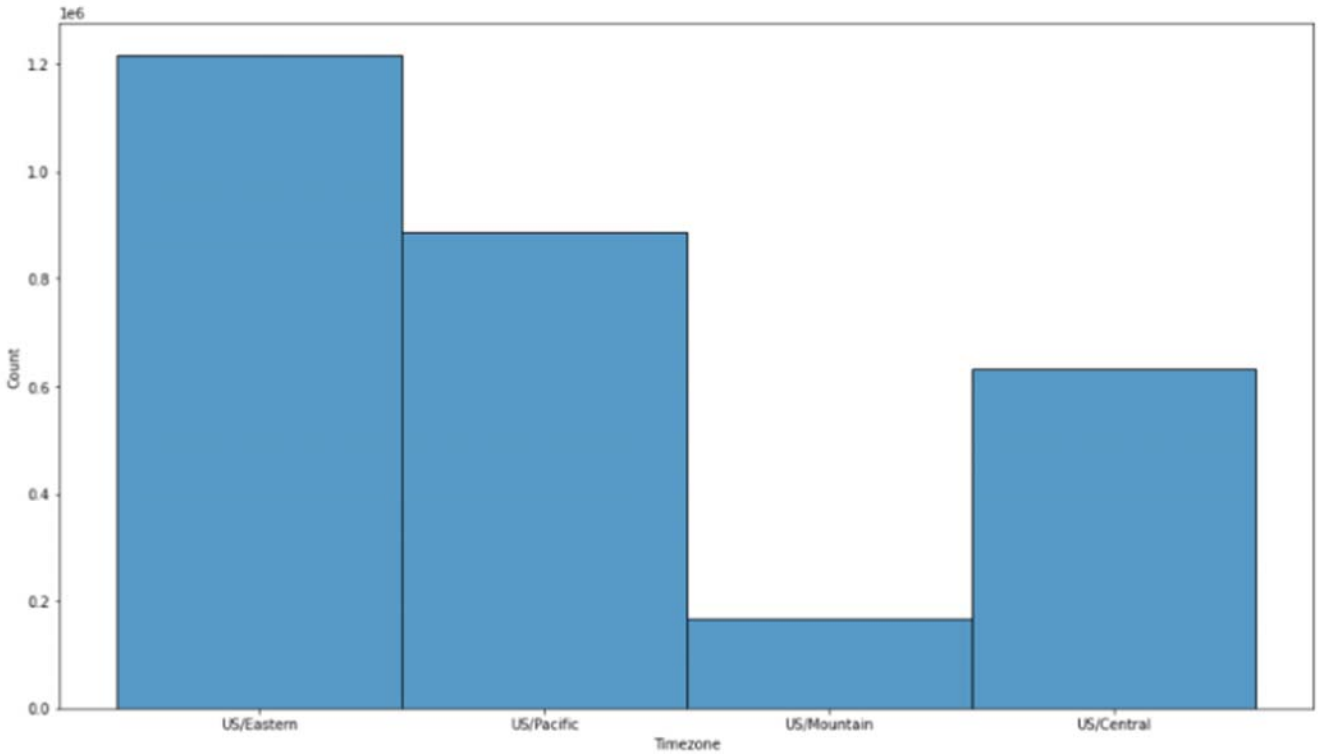


Fig. 8 (b) Histogram of the Time zone

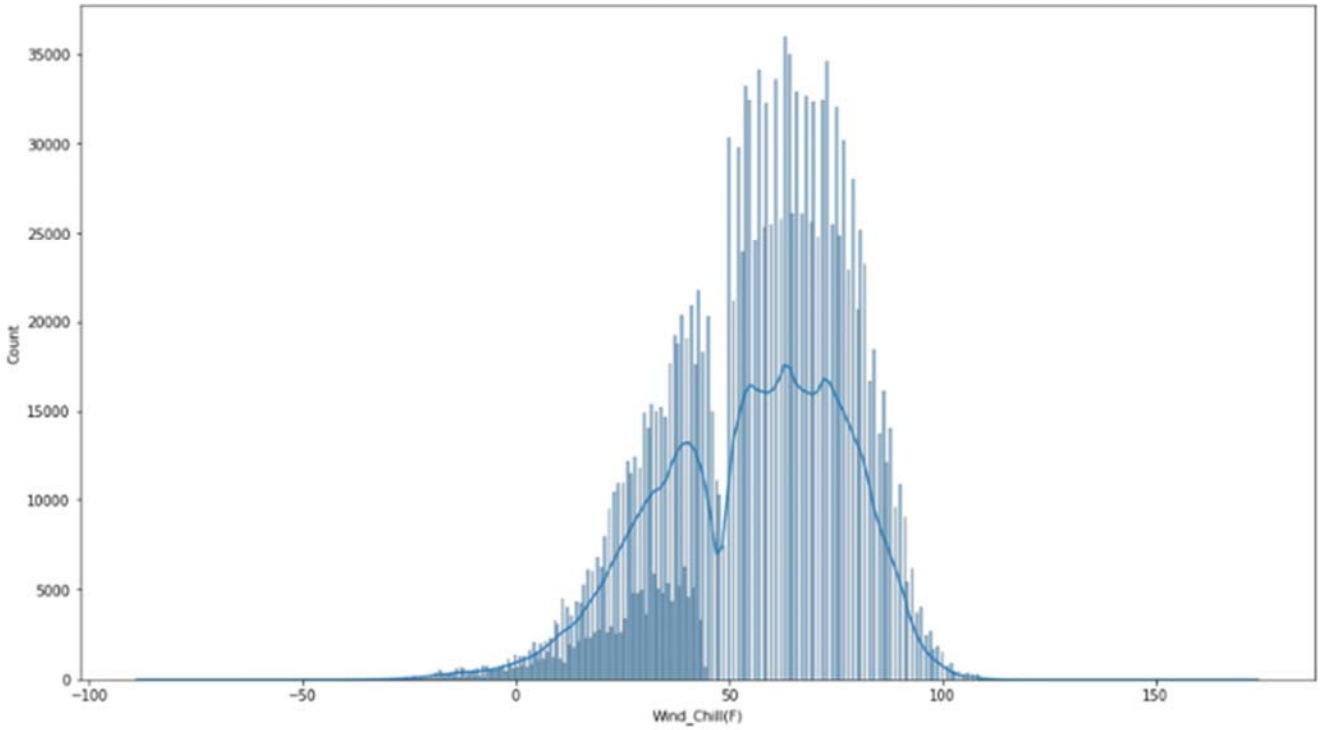


Fig. 9 (a) Histogram of windchill (F)

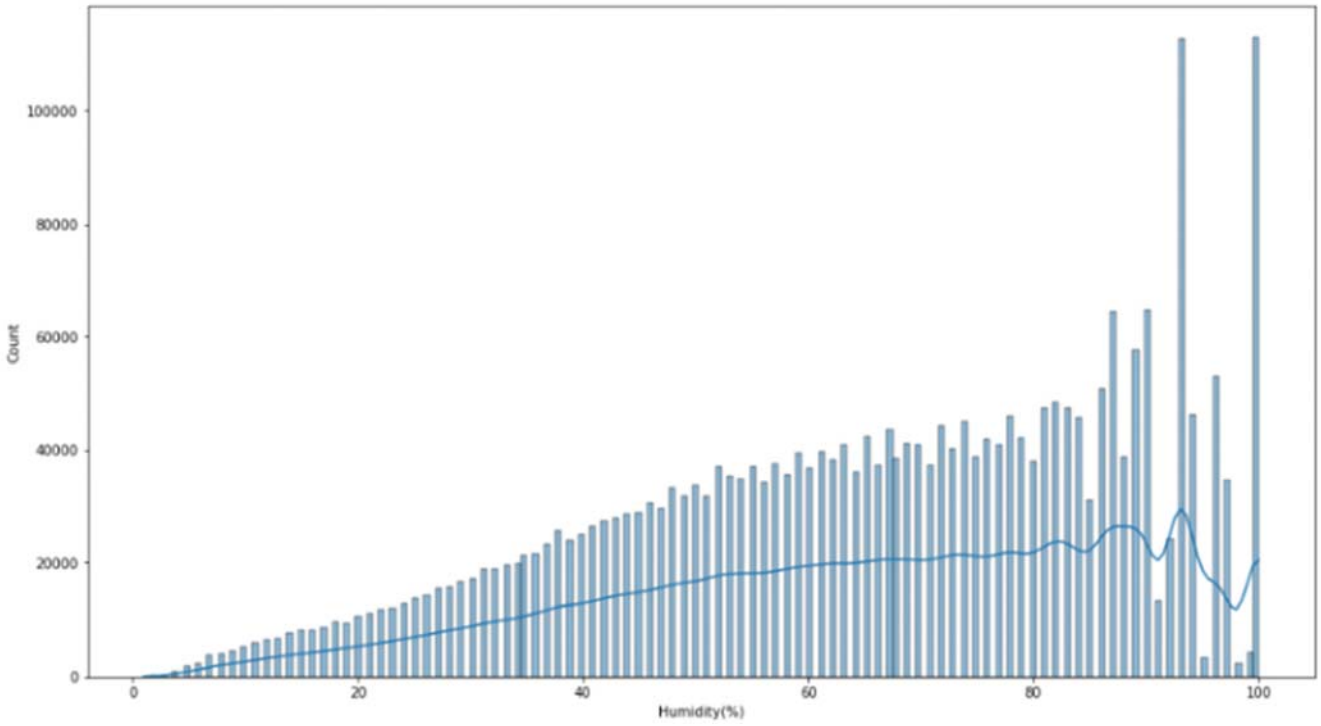


Fig. 9 (b) Histogram of Humidity (%)

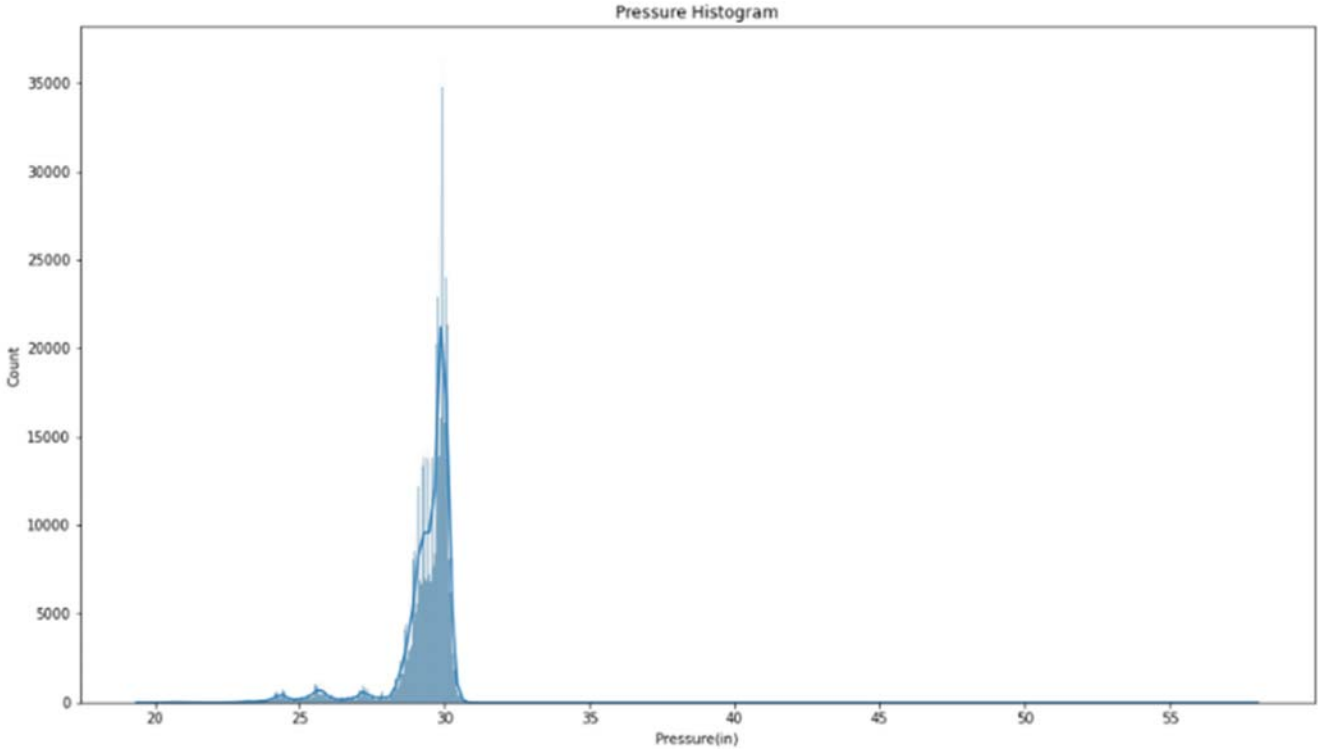


Fig. 10 (a) Histogram of pressure (in)

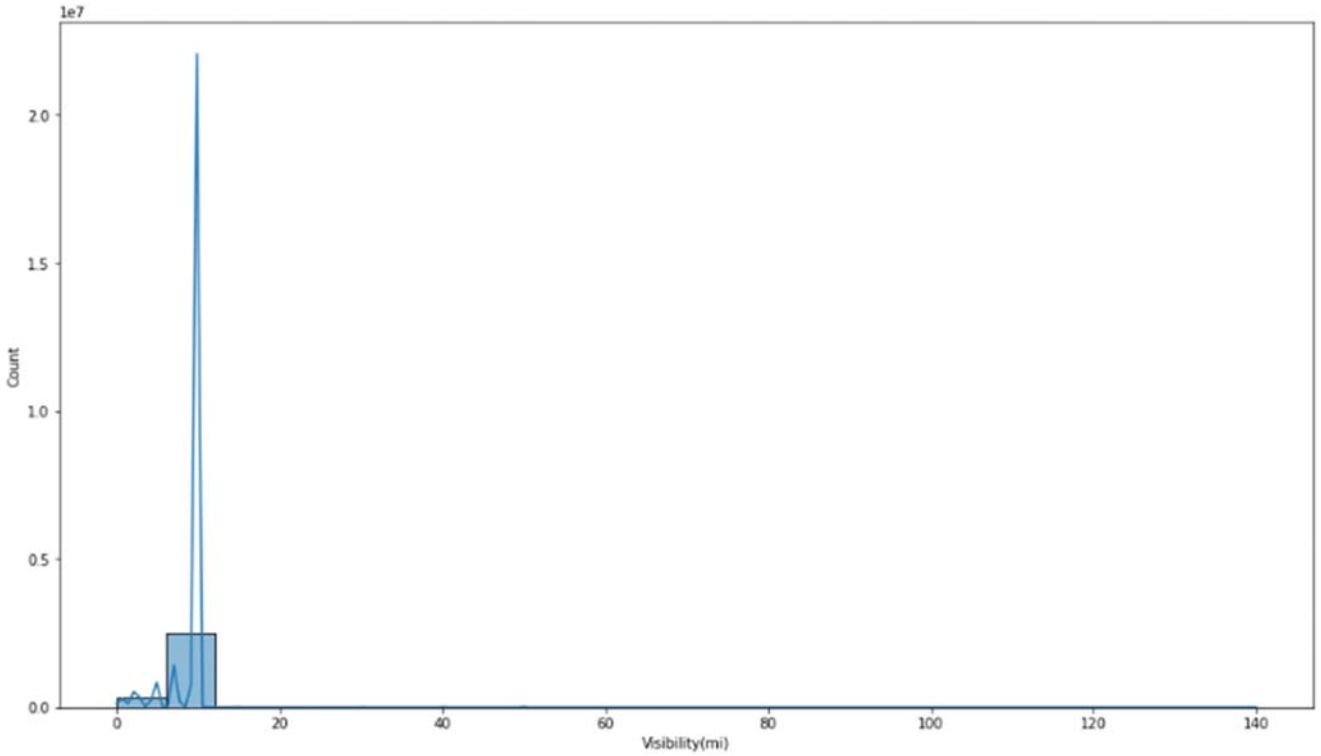


Fig. 10 (b) Histogram of Visibility (mi)

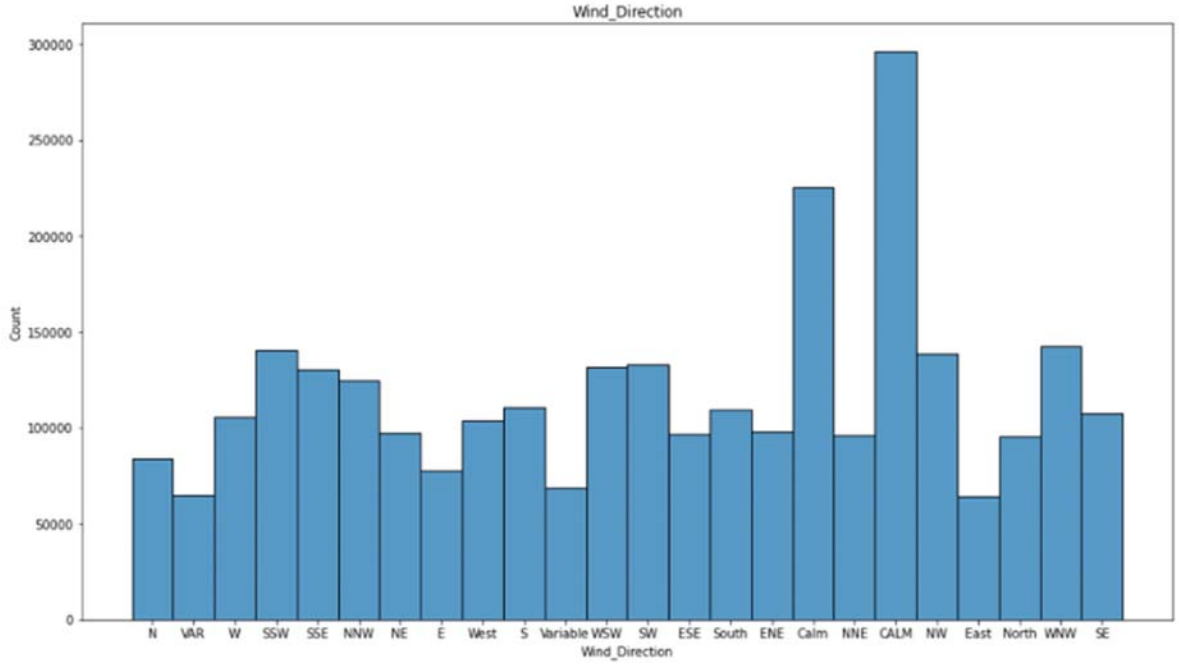


Fig. 11 (a) Histogram wind direction

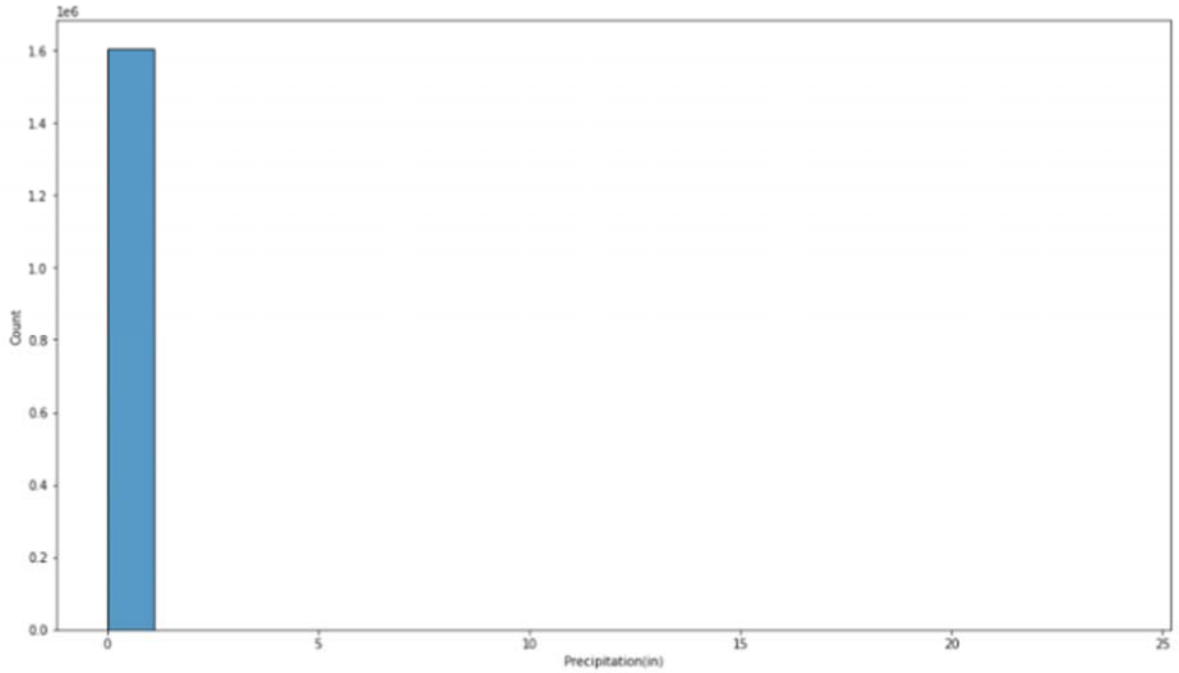


Fig. 11 (b) Histogram of precipitation (in)

Amenity represents if there is a nearby amenity center or rehabilitation provider during a crash. By analyzing this, we can clearly understand the facilities available for crash incidents and how much chance it is for a person to live after a heavy crash. As shown in Table IV, for 2875240 types of crashes, there is no availability of amenities. It exists only for 31370 crash incidents. It indicates very poor infrastructure and public authorities should take this issue seriously and establish opportunities in these crash locations to save many lives. With this kind of infrastructure, the chance of dying in a fatal accident is very high.

Bump indicates whether there is an accident near the Bump. Bump is one of the traffic-calming techniques. If an accident happens near a bump, it indicates overspeed by the vehicle driver. From the above value counts, only 579 accidents are at bump out of 2.9 million records.

Crossing indicates whether accidents happened near the crossing. A crossing is generally where the walking public can

cross a road for a certain period. As shown in Table IV, approximately 8% of the crashes are at crossing only, which indicates a severe problem with overspeeding and danger to the walking public. This issue should be taken seriously.

Give Way represents observed values from 2.9 million crashes, whether there is a crash at Give Way sites or not. As can be seen in Table IV, very few crash incidents happened at Give Way sites. No more inferences can be drawn from here.

Junction represents observed values from 2.9 million crashes, regardless of whether they occurred at Give Way locations. It refers to a point where multiple roads intersect, involving two or more directions. Over 10% of accidents occur exclusively at junctions, indicating the Urban phenomenon, as urban areas contain more junctions than rural areas. This implies that crashes may occur more frequently in urban areas compared to rural ones. A crash near a junction indicates wrong judgments by the vehicle's driver, potentially serving as an immediate cause for the crash.

No Exit contains information on whether there is an accident at the *No Exit* traffic rule. *No Exit* is where government authorities intentionally keep a signal of *No Exit* to control the traffic flow. The cause of the crash at these places indicates a traffic rule violation. Table IV shows that very few crashes happened at *No Exit* signs compared to the remaining crashes of 2.9 million. It can be due to traffic violations.

Railway contains information on whether a railway station is near the accident for other transportation sources. Only 0.87% of the crash sites have railways as an alternative commutation.

Roundabout contains data on whether a crash is near or during a roundabout. As can be seen, very few crashes happened at Roundabout when compared to the remaining crashes of 2.9 million. It can be due to traffic violations.

Station holds the data on whether a metro station near the accident site exists for all 2.9 million records. Approximately 2.1% of crash sites have metro facilities as alternative commutation.

Stop holds the data on whether there is a stop signal near the accident site for all 2.9 million records. Table IV shows that approximately 1.8% of crashes happened at stop signals, indicating overspeed or jerk movements.

Traffic Calming contains whether there is a traffic-calming structure or not during a crash. Traffic calming is a technique used by public authorities to manage the speed and flow of vehicles during commuting hours. If there is a crash nearby, it may indicate an overspeeding effect. There are very few crashes at traffic-calming sites.

Traffic Signal contains information on whether a crash is near a traffic signal. Crash near traffic signals indicates the violation of traffic rules or overspeeding. As presented in Table IV, approximately 17% of the accidents are near traffic signals, which indicates the citizen's behavior in following the rules. A crash can only happen at a traffic signal by overspeeding and traffic violations.

Turning Loop contains information on whether there is an accident near turning loops. The information in Table IV shows almost zero crashes near traffic loops. It indicates a better structure to regulate the traffic flow than junctions and traffic signals.

Table V shows the number of accidents versus some other important parameters.

TABLE V
THE NUMBER OF ACCIDENTS VS SOME EFFECTIVE PARAMETERS

	Sunrise Sunset	Civil Twilight	Nautical Twilight	Astronomical Twilight
Day	1941068	2073629	2212270	2321705
Night	965432	832871	694230	584795

Sunrise Sunset indicates whether it is a day or night according to sunset and sunrise on Earth. Civil Twilight contains information on whether it is a day or night according to Civil Twilight. Civil twilight is when the sun is below 6 degrees of the horizon. Nautical Twilight contains information on whether it is a day or night, according to Nautical Twilight. Nautical twilight is when the sun is not visible, but there is still light. Astronomical Twilight contains information on whether it is a day or night according to Astronomical twilight. Astronomical twilight is when the sun is below 18 degrees of the horizon.

Fig. 12 illustrates the correlation among some variables in the dataset as an Image matrix. Correlation is the relation between two variables. It indicates how two variables are related. It is mainly used in finding patterns in the data visually.

In Fig. 12, every correlation graph between two individual variables is shown as a 9x9 Image matrix where each matrix element is the correlation between one variable vs another. All the diagonal Image elements indicate the correlation between the variable and itself. There are nine quantifiable variables apart from categorical variables. *Start_Lat*, *Start_Lng*, Temperature, Windchill, Humidity, pressure, visibility, wind speed, and precipitation. All nine variables are placed in rows and columns in the same order. The values are obtained from the correlation Image matrix.

- The correlation between *Start_Lat* and *Start_Lng* roughly mirrors the geographical layout of the United States. The relationship between Latitude and temperature, windchill, and humidity is not uniform, suggesting the presence of diverse landscape features. Latitude's correlation with pressure, windspeed, precipitation, and visibility remains relatively consistent at lower values of these variables. However, it tends to diminish after reaching a certain point.
- *Start_Lng* and other variables follow the same correlation values, like latitude. It suggests there are different geographical areas where crashes happen, and they spread over large ecosystems of the lithosphere.
- *Temperature* and *wind chill* exhibit an almost linear relationship, with higher temperatures corresponding to greater wind chills. These two variables are dependent, and their correlation is close to 99%. Additionally, as temperature rises, humidity experiences a very slight decrease, indicating a weak correlation between temperature and humidity. Conversely, temperature shows no significant correlation with Pressure, visibility, Precipitation, and Wind speed at crash sites.

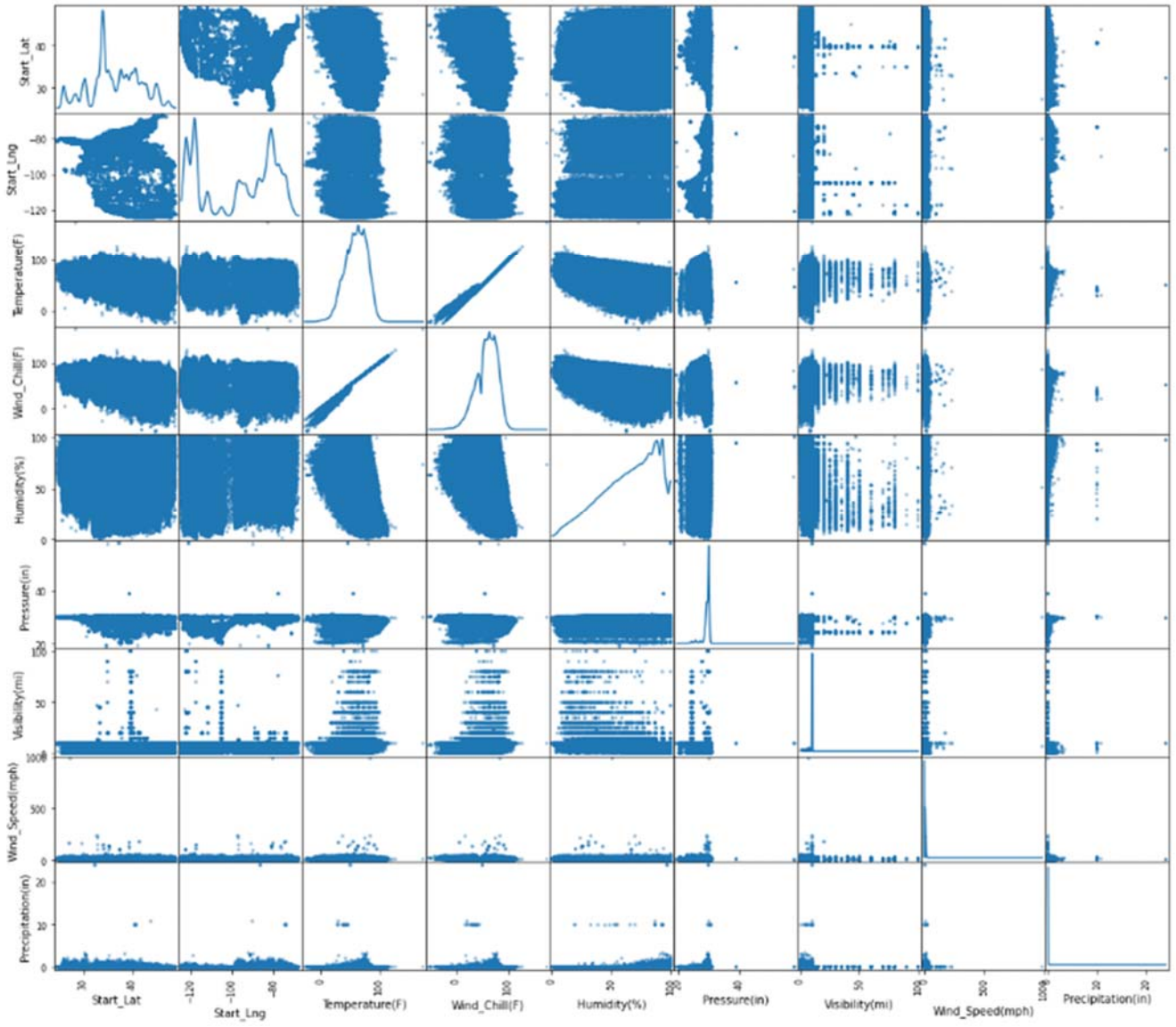


Fig. 12 Correlation visualization

- *Windchill* has a weak correlation with Humidity. The slope is negative. Windchill with pressure, precipitation, wind speed, and visibility have no notable correlation.
- *Humidity* with pressure, precipitation, wind speed, and visibility has no notable correlation.
- *Pressure* has no notable correlation with any variable.
- *Wind speed* has no notable correlation with any variable.
- *Precipitation* and *visibility* have no notable correlations.

Fig. 13 (a) presents the same data as Fig. 12 but in the form of a heatmap. Heatmap makes it visually easy to compare the correlation with other variables. The darker cells indicate lesser correlations and brighter ones indicate higher correlations. All diagonal heat cells are shown as white because the correlation between variables and itself is linear. Fig. 13 (b) shows Covariance among some variables.

Covariance reveals how each variable behaves over other variables. Covariance is defined as the change between variables as one variable vs another. It evaluates the change by either positive or negative and values for change. If change is +ve, it means that as one variable increases, another variable dependent upon it increases. If the change has a -ve sign, it indicates the negative covariance where a change in one variable adversely affects the other variable by decreasing its value. The strength of the covariance is a numerical value that indicates how strong the covariance is. In Fig 13 (b), the variables are arranged in a grid manner with the same number of rows and columns. Every row-column position indicates covariance between one variable and another, arranged in the same order as rows and columns. The darker cells indicate lesser covariances and brighter ones indicate higher covariances. As can be seen, the covariance between most of the variables is the same except for temperature, wind chill, and Humidity. It indicates the covariance high between these variables, which are not independent.

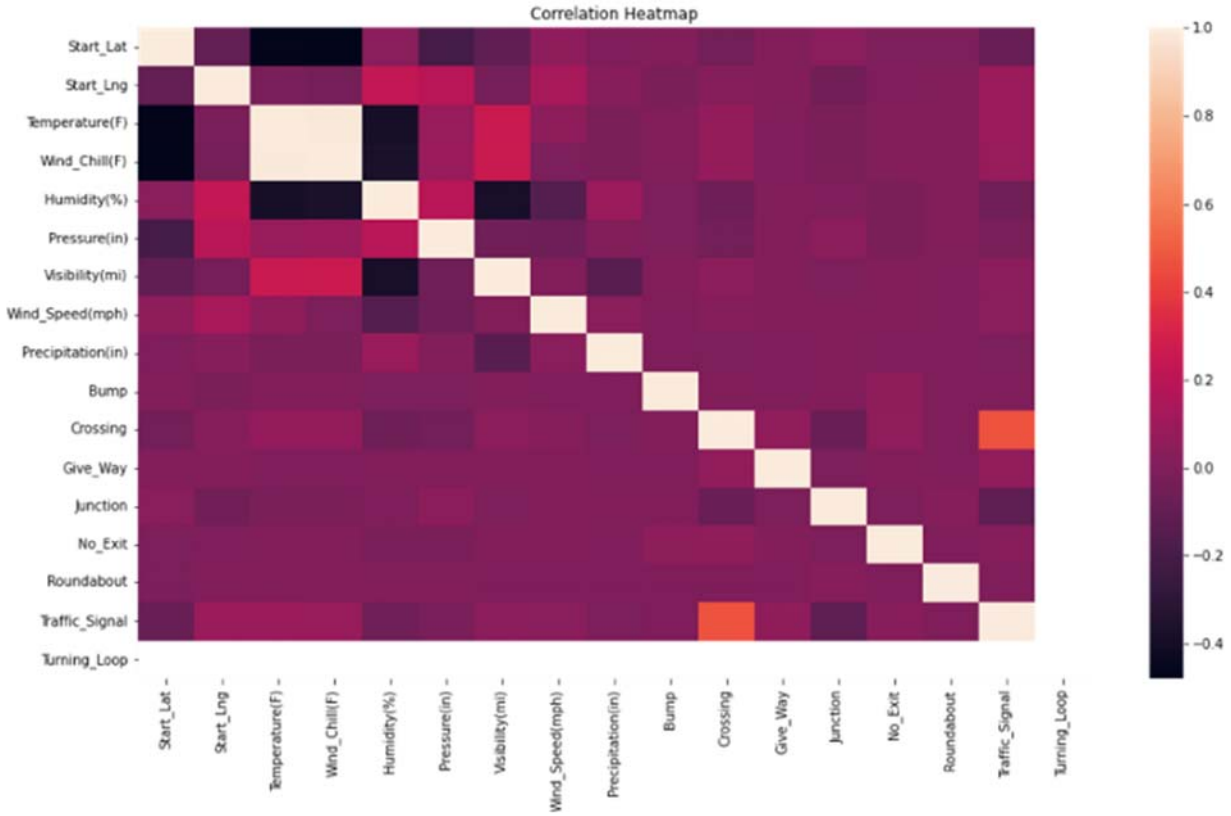


Fig. 13 (a) Heatmap of Correlation among some variables

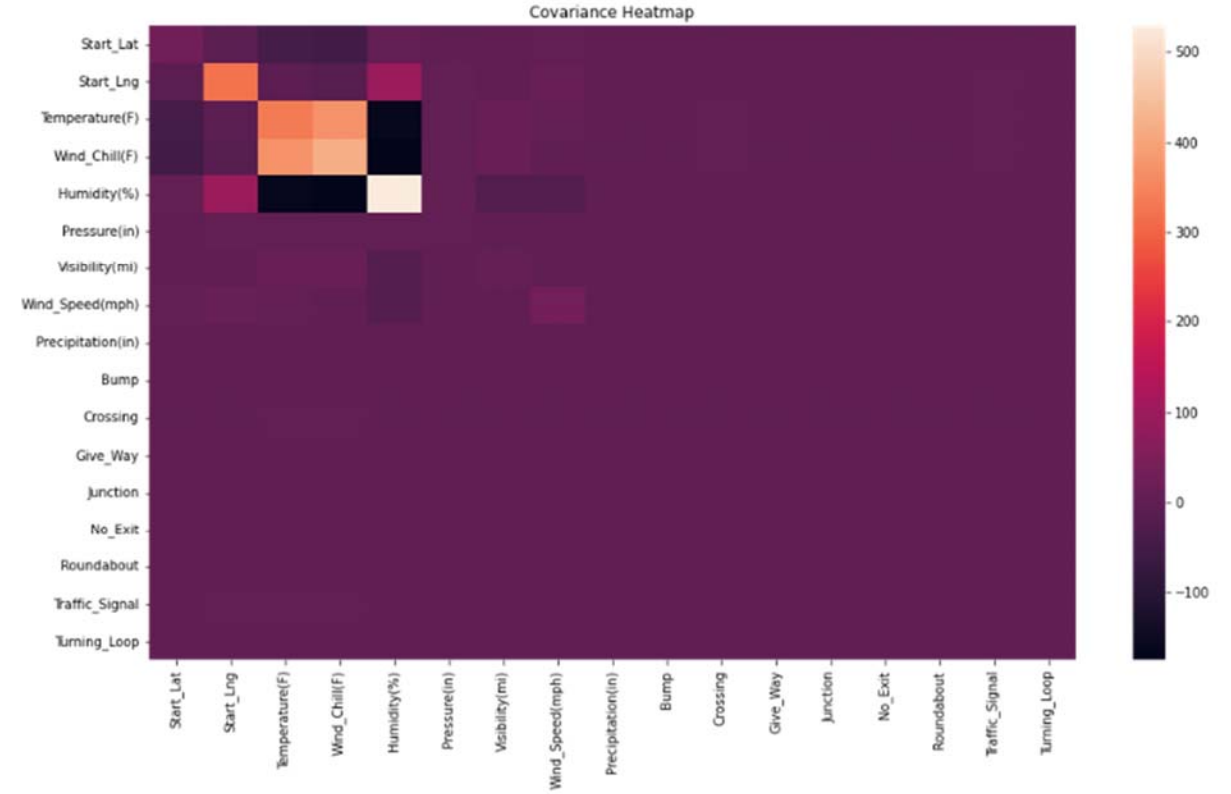


Fig. 13 (b) Heatmap of Covariance among some variables

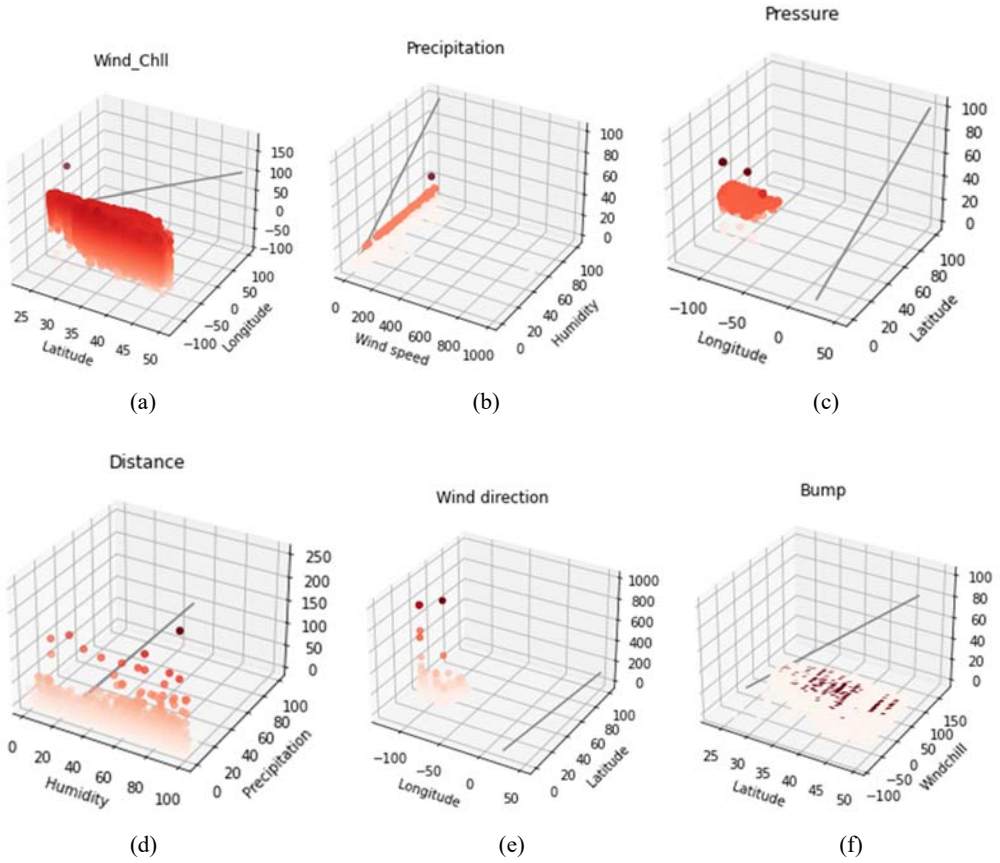


Fig. 14 (a) 3D Representation of Windchill vs Latitude vs Longitude; (b) 3D representation of Wind speed vs Humidity vs Precipitation; (c) 3D Representation of Longitude, Latitude, and pressure; (d) 3D representation of Humidity, Precipitation, and Distance; (e) 3D Representation of latitude, longitude, and Wind Direction; (f) 3D representation of Bump, Latitudes, and Windchill

Some other relationships between variables are in the diagrams represented in Fig. 14. Fig. 14 (a) represents the 3D correlation between Windchill measured in Fahrenheit and the latitude-longitude in the X and Y axes. As latitude and longitude increase, windchill remains constant but gives some blob in between, indicating the underlying geography of provided coordinates and their different types of ecosystems. In Fig. 14 (b), Windspeed and Humidity are plotted in the X and Y axes. The correlation between them and Precipitation in the Z axis shows that Windspeed remained constant, and as Humidity increases, Precipitation remains a weak correlation. In Fig. 14 (c), change in Pressure over certain crash sites is visible for only a small range of values. X and Y axes are Longitude and Latitude. The z-axis took the Pressure Measurement. More accidents occurred around lesser longitude, mid-latitudes, and mid-pressure. In 14 (d), X and Y axes are taken as Humidity and Precipitation, respectively. The Z axis is taken as the Distance of the traffic congestion. Correlation moves towards High values of Precipitation. Humidity and distance have similar kinds of patterns. Precipitation must do something with the distance of congestion. Precipitation generally occurs at 100% humidity. Humidity can vary with temperature and pressure. In Fig. 14 (e), Longitude and Latitude are taken as X and Y axes, respectively, while the Z axis has a Wind direction. Correlation between these three variables indicates that most accidents happen around lesser longitudes, higher wind speeds, and mid-latitudes. In Fig. 14 (f), X and Y axes are taken as Latitude and Wind chill, respectively. While the Z axis has a Bump, it indicates Bump has a lesser correlation with the number of accidents. Mid latitudes and higher wind chills contain more accidents with non-parametric correlation.

B. Results from Multiclass Classification Models

As explained in Section VI, we have used three different multiclass classifiers, including Logistic Regression, Random Forest, and XGBoost, to predict the probability class defined for our problem. Then, we used hyperparameter tuning to determine the right combination of hyperparameters that allows the model to maximize performance. Setting the correct combination of hyperparameters is the only way to extract the maximum performance out of models. The selection of the right combination of hyperparameters is a challenging task. There are two ways to set them.

- Manual hyperparameter tuning: In this method, different combinations of hyperparameters are manually set (and experimented with). This is a tedious process and cannot be practical in cases with many hyperparameters.
- Automated hyperparameter tuning: In this method, optimal hyperparameters are found using some algorithms that automate and optimize the process. This method has been

used in this paper.

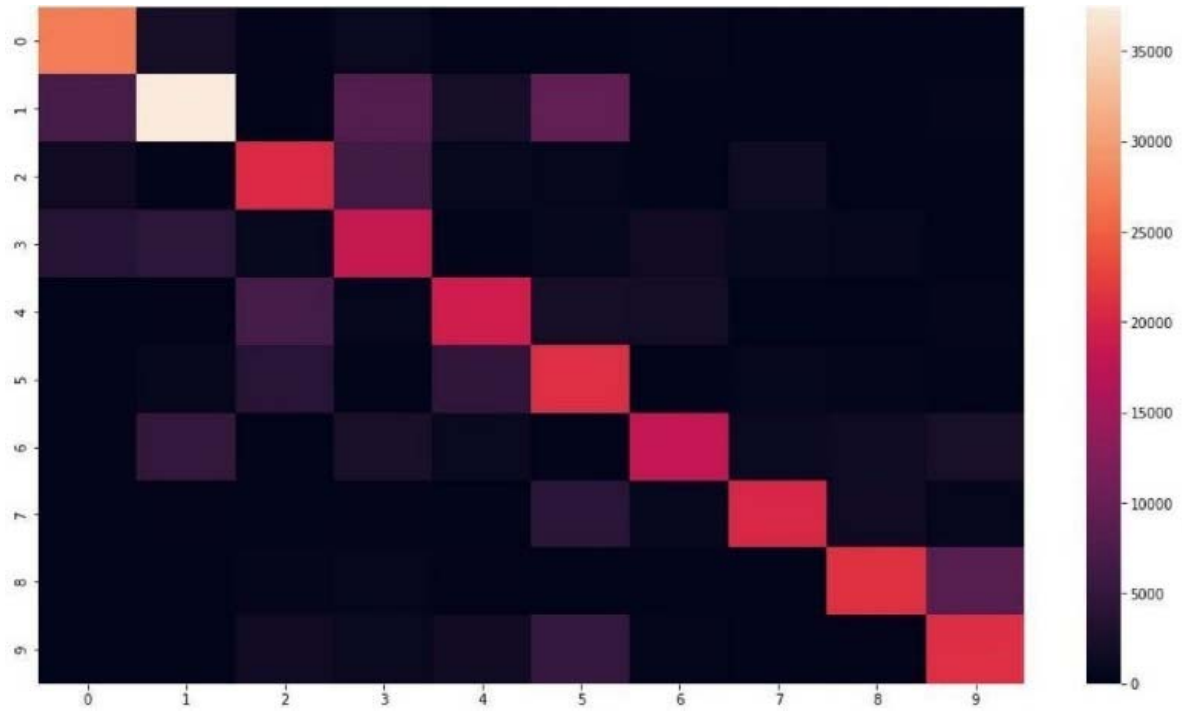
We have reported the numerical results for the abovementioned models with and without hyperparameter tuning.

We first develop multiclass logistic regression to predict the probability class defined for the problem in this paper. The numerical result shows that the overall accuracy is 66.5%, which indicates that we left many variables except Natural and Structural causes. Eliminating spontaneous causes during the prediction class calculation resulted in a decrease in accuracy, signifying the impact on the overall predictive performance. Additionally, it reveals the percentage of predictability for Probability attributed to Natural and Structural causes. Accuracy, at times, may not provide a comprehensive depiction, hence the classification report offers insights into our performance across individual classes.

```
print(classification_report(y_test,Predictions))
```

	precision	recall	f1-score	support
0	0.82	0.91	0.82	31369
1	0.56	0.62	0.54	64549
2	0.66	0.77	0.70	31043
3	0.59	0.70	0.58	31015
4	0.61	0.75	0.62	31431
5	0.67	0.79	0.68	31609
6	0.58	0.68	0.57	31261
7	0.65	0.76	0.67	31398
8	0.68	0.82	0.69	31249
9	0.68	0.71	0.70	31106
micro avg	0.65	0.76	0.66	346030
macro avg	0.65	0.76	0.66	346030
weighted avg	0.65	0.75	0.66	346030
samples avg	0.63	0.73	0.65	346030

(a)



(b)

Fig. 15 (a) Classification Report of Multiclass logistic regression; (b) Classification Report Heatmap

Fig. 15 (a) represents the Multiclass logistic regression classification report for the testing data. It is essential to evaluate accuracy due to potential biases in accuracy curves. The report encompasses a total of 10 classes, where the range 91-100 is denoted as the '0' class, 0-10 as '1', 11-20 as '2', 21-30 as '3', 31-40 as '4', 41-50 as '5', 51-60 as '6', 61-70 as '7', 71-80 as '8', and 81-90 as '9'. These intervals are the classification of which probability the data point belongs to. The accuracy of the model is evaluated in three indicators. Precision indicates performance, Recall measures completeness, and F1 scores assess the model's agility. The values for these metrics are presented as follows:

1. class 91-100 has Precision 0.82, Recall 0.91, f1-score 0.82.
2. class 0-10 has Precision 0.56, Recall 0.62, f1-score 0.54.
3. class 11-20 has Precision 0.66, Recall 0.77, f1-score 0.70.
4. class 21-30 has Precision 0.59, Recall 0.70, f1-score 0.58.
5. class 31-40 has Precision 0.61, Recall 0.75, f1-score 0.62.
6. class 41-50 has Precision 0.67, Recall 0.79, f1-score 0.68.
7. class 51-60 has Precision 0.58, Recall 0.68, f1-score 0.57.
8. class 61-70 has Precision 0.65, Recall 0.76, f1-score 0.67.
9. class 71-80 has Precision 0.68, Recall 0.82, f1-score 0.69.
10. class 81-90 has Precision 0.68, Recall 0.71, f1-score 0.70.

Fig. 15 (b) shows the Heatmap of the Classification report. All ten classes are arranged in rows and columns in the same order. Each meeting point of a row and column is called a heat cell. Heat cells measure the performance. A brighter heat cell concerning other cells indicates that row and column classes perform better than each other. The darker the cell is compared to others, indicating it performed badly in that class. This visualization helps in performing the comparison between different classes. The classification report has the following parameters:

- *F1 score:* F1 score of our model is 0.66 on Average for all ten classes. F1 score indicates the ratio of True Positive to the (True positives + ½ (False positives + False Negatives). It also says how the model is performing in overall situations. *F1 score = TP/(TP+1/2(FP+FN)*
- *Precision:* The precision of our model is 0.655. It indicates what percentage of positives were not considered False Negatives. The formula to calculate Precision is: *Precision = TP/(TP+FP)*
- *Recall:* The recall score of our model is 0.76. It indicates what percentage of True Positives were not labeled as False Negatives. The recall is also called the completeness of the model. The formula to calculate this is: *Recall = TP/(TP+FN)*

Fig. 16 (a) illustrates the accuracy curve of multiclass logistic regression. The machine learning algorithm ran 15,000 iterations over the training dataset to evaluate and visualize its performance. The X-axis represents the number of Iterations over train data. The y-axis represents the accuracy achieved after each iteration. Accuracy improved slowly over each iteration. After teaching 8000 iterations, the accuracy fluctuates around 60-68 values up to 1500. This is the most an algorithm can perform. To reduce the overfitting, iteration was set to 15000 only. Implementing multinomial regression using the One-vs-Rest (OvR) classifier in scikit-learn has been utilized, with the specific choice of a "loglinear" solver.

Fig. 16 (b) illustrates the loss curve over iterations of multiclass logistic regression. The X axis is taken as Iterations. It indicates the number of iterations over train data the algorithm went for. The Y axis is represented by loss for each iteration over training data the algorithm faced. The graph shows loss was gradually decreasing slowly up to 8000 iterations. After 8000 iterations, the loss hovers around 40-32. The algorithm reached the lowest possible loss around this interval. Maximum iterations were set to 15000. If training was done in more iterations, the algorithm would overfit.

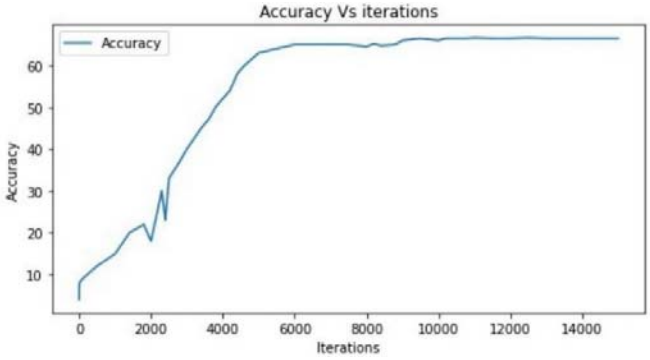


Fig. 16 (a) Accuracy curve over iterations on Data

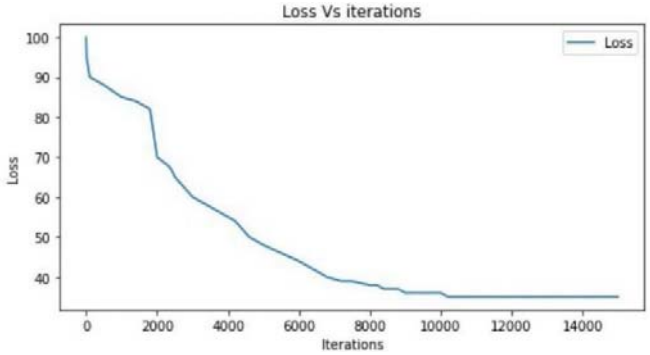


Fig. 16 (b) Loss curve over Iterations till convergence

Table VI represents the accuracy and runtime values for three multiclass classifiers with and without hyperparameter tuning.

TABLE VI
DESCRIPTIVE FACTS OF LOCATION LATITUDE AND LONGITUDE IN DEGREES

	Accuracy (%)	Runtime
Logistic Regression	66.5	1m 53sec
Logistic Regression with hyperparameter tuning	69.3	31m 12sec
Random Forest	70.7	3m 32sec
Random Forest with hyperparameter tuning	73.6	4m 13sec
XGBoost	71.4	71m 12 sec
XGBoost with hyperparameter tuning	75.2	88m 20sec

As shown in Table VI, XGBoost with hyperparameter tuning has provided better accuracy, which is about 75.2%, with a runtime of 88 minutes and 20 seconds. To be realistic, we cannot expect higher accuracy because we had to leave out Spontaneous reasons for the crash, and the probability class values are calculated through random and stochastic processes, which may be biased towards one value. Using other methods of classification is also suffering from the same kind of bias in the dataset. So, it only reveals how much we can predict the probability of a crash due to natural and structural reasons.

VIII. DISCUSSION

This section represents the findings of exploratory data analysis as follows.

- Turning loops are better traffic structures than traffic junctions. They recorded absolutely zero crashes, which indicates that fewer crashes can happen if there is a chance

that we replace all junctions with turning loops, providing that cost-effectiveness.

- Coastal Urban areas like Miami and Florida, along with the other areas surrounding the Great Lakes, have seen many crashes, which indicates the population density and economic activity in that area.
- 17% of the crashes occurred at junctions, while none occurred at turning loops. Traffic signals may have been causing mental pressure for commuters in urban areas.
- Major natural factors for more crashes are Humidity, Wind Chill, Precipitation, and Weather type.
- Minor natural factors for more crashes are Wind speed, Wind direction, Pressure, and Temperature.
- Major structural causes for crashes are Junctions, Traffic signals, and Crossings.
- Minor structural causes for crashes are Bumps, No exit signs, Give way, Roundabouts, and Stop signals.
- Most crash sites have no amenity to save people's lives as early as possible.
- There seems to be an urban-rural gap between crashes. There may be a correlation between traffic volume, population density, and economic activity.
- Approximately 10% of the accidents occurred during the crossing.
- Covariance between Temperature, Pressure, and Humidity is very high. It indicates they may not be very independent.
- Calm and cloudy weather conditions are reported at many crashes. It indicates that the rain level has a neutral effect on the number of crashes.
- California has recorded more accidents, followed by Florida.
- Vermont, Wyoming, North Dakota, and South Dakota recorded very few incidents of crashes.
- The correlation between Visibility, Temperature, and Humidity is very high.
- Correlation between traffic signals and crossing is very high. As it is clear, a major part of the crashes happened in these localities.
- US/Eastern and US/Pacific time zones have seen more crashes. While the US/Mountain region is less.

IX. Contribution of this Paper

Our problem statement aims to predict the likelihood of road crashes by considering both natural and structural factors. This prediction is based on the extensive dataset from the United States government, which includes numerous columns and rows detailing every crash. The technical aspects discussed in this paper are specifically applicable to the United States.

- We categorized the factors contributing to the crash into three independent Variables, including 1) *Natural Reasons* like weather conditions, Humidity, Pressure, Temperature, Visibility, Precipitation, Longitude, Latitude, etc. 2) *Structural reasons* like, Bumps, Traffic signals and Junctions, No exit and Stop signals from authorities, side etc. 3) *Spontaneous Reasons* like vehicle speed, drunk driving, erratic driving behavior, Signal jumps, etc.
- We limited our research to natural and structural reasons due to limited real-time sensor data and no idea of its distribution.
- Dataset contains numerical, Boolean, and categorical variables that were observed during every crash that happened in the United States in all states. Most of the dataset columns are useless for our purpose. After doing some exploratory data analysis, we pointed out 21 columns out of 47 that we need for analysis and prediction through the model.
- To find the probability, we also needed examples of where a crash has never happened, if there is a certain probability of a crash as well as a dataset that does not contain them.
- To find this, we need to generate data that follows the same distribution as the original dataset.
- There are three ways to solve this problem.
  - *Gaussian Process Regression:* This process takes a different approach than normal regression. Instead of calculating only one possible way to create the data using the Bayesian approach, it generates all possible vector spaces that follow the given distribution.
  - *Linear Regression:* This technique is effective for predicting specific values based on dataset vectors. It involves calculating the expected values of each multivariate distribution, creating random vectors from the same distribution, and using linear regression to determine the expected values for these new vectors. These values are then stored for anomaly detection purposes.
  - *Multinomial Expected Value:* Given that the data are multivariate, random data points are generated as vectors following the same multinomial distribution as the original dataset. The expected values for these random points are calculated using the multinomial expected value.
- We selected the multinomial expected value as an anchor to generate new random data to create a train set. It is simple but powerful. After generating new data and expected values, we assigned probabilities close to those of the original distribution. According to Chebyshev's theorem, 60% of the data should fall within two standard deviations from the mean, 80% within three standard deviations, and 99.9% within five standard deviations.
- By using that rule, we created 10 bins of probability by comparing the present distribution with the original distribution.
- The rest of the ten bins are assigned accordingly.
- After the training set was ready, we normalized and scaled the data to ensure accuracy was not skewed.
- We experimented with the dataset by taking 70% of it as a training dataset and used three different multiclass classifiers to predict the probability bin of the data point.
- 30% of data are used as test data for prediction.
- Multiclass XGboost with hyperparameter tuning could achieve 75.2% accuracy. Accuracy suffered because there were no spontaneous reasons.

X. Conclusions and Future Works

This study aims to predict the likelihood of a crash based on natural and structural factors such as weather, temperature,

pressure, wind direction, wind speed, precipitation, humidity, traffic signals, and road features like bumps, crossings, and junctions. The analysis has revealed significant correlations, such as between traffic signals and crossings, which commonly occur together. This demonstrates that exploratory data analysis is an effective method for drawing and quantifying inferences for further use. We applied multiclass logistic regression (OneVsRest), Random Forest, and XGBoost, all with and without hyperparameter tuning, to predict the probability class a crash incident belongs. The results show that XGBoost with hyperparameter tuning gives better accuracy of 75.2% but at a greater computational cost, which is about 88 minutes and 20 seconds. It indicated that we need more and more natural variables to be taken into consideration. The models presented in this paper do not deal with spontaneous causes of accidents like overspeeding, traffic violations, drunk driving, vehicle conditions, etc. Some important measures are identified as follows.

Amenity, rehabilitation, or the help needed at the crash locations, even in the hotspots of accident zones like junctions, crossings, and traffic signals, were also not available in proper proportion. We identified that the only way accidents could occur at junctions is due to human-made spontaneous mistakes like overspeeding and traffic signal violations. 17% of the accidents belong to this category. Crashes near bumps and no exit are also caused by misguided human behaviors of overspeeding. The crashes during Give Way are also misjudgments by humans.

Therefore, calculating probability based on natural structural reasons is only a part of the work. There must be some way to find out spontaneous reasons like overspeeding, drunk driving, and spontaneous reactions to quantify and analyze to find the proper solutions to the menace of these crashes.

In future works, we plan to work on the following ideas.

- To get the proper probability, one must use extensive sensor data and heavy processing to understand the patterns in spontaneous causes like overspeeding and traffic violations. The combination of all these things can perfectly estimate the real-time probability of a crash by any event or cause.
- Data collection is the most important work that needs to be done. Public authorities should make sure that the real-time data are freely available and updated now and then to help academia conduct more research on things that provide valuable solutions. One idea is that there must be a network of traffic and vehicle networks just like the internet in every country to properly manage traffic flow and control the number of crashes. v2X technologies are working towards these. A more systematic approach is needed.
- We plan to use streaming data and Big Data technology for real-time monitoring of road traffic. The road traffic prediction models will also be developed using Spark Streaming.

REFERENCES

[1] B. G. Lee, J. Park, C. C. Pu and W. Chung, "Smartwatch-Based Driver Vigilance Indicator with Kernel-Fuzzy-C-Means-Wavelet Method," in IEEE Sensors Journal, vol. 16, no. 1, pp. 242-253, Jan.1, 2016, doi: 10.1109/JSEN.2015.2475638.

[2] Ashtaiwi, "Intelligent Road Crashes Avoidance System," 2019 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), 2019, pp. 280-286, doi: 10.1109/ICAIIC.2019.8668972.

[3] W. Tong, A. Hussain, W. X. Bo and S. Maharjan, "Artificial Intelligence for Vehicle-to-Everything: A Survey," in IEEE Access, vol. 7, pp. 10823-10843, 2019, doi: 10.1109/ACCESS.2019.2891073.

[4] E. Karlsson and N. Mohammadiha, "A Data-Driven Generative Model for GPS Sensors for Autonomous Driving," 2018 IEEE/ACM 1st International Workshop on Software Engineering for AI in Autonomous Systems (SEFAIAS), 2018, pp. 1-5.

[5] Y. Yoon, C. Kim, J. Lee and K. Yi, "Interaction-Aware Probabilistic Trajectory Prediction of Cut-In Vehicles Using Gaussian Process for Proactive Control of Autonomous Vehicles," in IEEE Access, vol. 9, pp. 63440-63455, 2021, doi: 10.1109/ACCESS.2021.3075677.

[6] X. Bai, L. Sang, J. Khan, H. Kang and Z. Liu, "Initial Thoughts and Application of AI in Geo-hazards Monitoring and Early Warning of Highroad in Beijing," 2019 5th International Conference on Transportation Information and Safety (ICTIS), 2019, pp. 1358-1363, doi: 10.1109/ICTIS.2019.8883743.

[7] Najjar, A., Kaneko, S., & Miyanaga, Y. (2017). Combining Satellite Imagery and Open Data to Map Road Safety. AAAI.

[8] Akin, Darcin & Akbas, Bulent. (2008). A Neural Network (NN) Model to Predict Intersection Crashes Based upon Crash Properties: Driver, Vehicle, and Roadway Surface Characteristics.

[9] Fancello, Gianfranco & Carta, Michele & Fadda, Paolo. (2015). A Decision Support System for Road Safety Analysis. Transportation Research Procedia. 5. 10.1016/j.trpro.2015.01.009.

[10] Silva, P.B., Andrade, M. and Ferreira, S., 2020. Machine learning applied to road safety modeling: a systematic literature review. Journal of Traffic and transportation engineering (English edition).

[11] Tonhauser, M. and Ristvej, J., 2021. Implementation of new technologies to improve safety of road transport. Transportation research procedia, 55, pp.1599-1604.

[12] DfT (2019), Road Safety Statement: Progress Report. A Lifetime of Road Safety, 2019 Road Safety Statement https://www.gov.uk/government/publications/road-safetystatement-2019-a-lifetime-of-road-safety.

[13] Haghani, M., Bliemer, M.C., Farooq, B., Kim, I., Li, Z., Oh, C., Shahhoseini, Z. and MacDougall, H., 2020. https://www.hindawi.com/journals/jat/si/892372/

[14] Li, H., Zhu, M., Graham, D.J. and Ren, G., 2021. Evaluating the speed camera sites selection criteria in the UK. Journal of safety research, 76, pp.90-100.

[15] Hughes, B.P., Anund, A. and Falkmer, T., 2015. System theory and safety models in Swedish, UK, Dutch and Australian road safety strategies. Accident Analysis & Prevention, 74, pp.271-278.

[16] Nogayeva, S., Gooch, J. and Frascione, N., 2020. The forensic investigation of vehicle-pedestrian collisions: a review. Science & Justice.

[17] Damani, J. and Vedagiri, P., 2021. Safety of motorised two wheelers in mixed traffic conditions: literature review of risk factors. Journal of Traffic and transportation engineering (English edition).

[18] Tony Yiu, “Understanding Random Forest”, Towards Data Science,2019

[19] Iyad Lahsen Cherif, Abdesselem Kortebi, “On using eXtreme Gradient Boosting (XGBoost) Machine Learning algorithm for Home Network Traffic Classification”, IEEE Xplore: 13 June 2019

[20] https://towardsdatascience.com/usa-accidents-data-analysis-d130843cde02