University of Tehran

Earth Observation and Geomatics Engineering

Online ISSN: 2588-4360

Homepage: https://eoge.ut.ac.ir/

Spatial Prediction of the Impact of Road Accidents on Traffic Using Machine Learning Algorithms

Mohsen Forouzandeh [1] , ***Bahare Sadat Mousavi*** [2], ***Najmeh Neysani*** [3]✉ , *and* ***Meysam Argany*** [4]

1. *M.Sc. Graduate in Remote Sensing and Geographic Information Systems, University of Tehran, Tehran, Iran. E-mail: Forouzandeh.m@ut.ac.ir*
2. *Ph.D. Student in Remote Sensing and Geographic Information Systems, University of Tehran, Tehran, Iran. E-mail: Bahare.mousavi@ut.ac.ir*
3. *Corresponding author, Associate Professor, Department of Remote Sensing and Geographic Information Systems, Faculty of Geography in University of Tehran, Tehran, Iran. E-mail: nneysani@ut.ac.ir*
4. *Associate Professor, Department of Remote Sensing and Geographic Information Systems, Faculty of Geography in University of Tehran, Tehran, Iran. E-mail: argany@ut.ac.ir*

Article Info

Article type:
Research Article

Article history:
Received 2024-07-06
Received in revised form 2025-01-24
Accepted 2025-02-17
Available online 13 May 2025

Keywords:
Accident severity,
machine learning,
Prediction,
random forest,
decision tree,
naive Bayes,
gradient boosting,
k-nearest neighbor,
voting classification.

ABSTRACT

Predicting accident severity is undeniably one of the most essential aspects of analyzing and managing traffic incidents. Identifying the key factors influencing the severity of road accidents is crucial for mitigating their impact. Analyzing accident events and identifying their patterns enables the prediction and prevention of future accidents. With the vast amount of data available, machine learning has gradually outperformed traditional statistical methods. One promising approach for predicting accident severity is the application of machine learning algorithms. Consequently, machine learning algorithms are increasingly recognized as effective methods for modeling accident severity.

This study aims to implement a framework for predicting accident severity using algorithms such as Random Forest, Decision Tree, Naive Bayes, Gradient Boosting, K-Nearest Neighbors (K-NN), and Voting Classifier. The results of applying these algorithms to accident data for predicting accident severity revealed that the Random Forest algorithm achieved the highest accuracy at 90.8%, while the Naive Bayes algorithm showed the lowest accuracy at 25%. To further enhance the accuracy of accident severity predictions, the Voting Classifier algorithm was employed by combining the two algorithms with the highest accuracy Random Forest and Gradient Boosting while excluding other algorithms due to their lower performance. The final results demonstrated that the Random Forest model alone produced outcomes comparable to the combined model. Consequently, the Random Forest model was selected as the optimal algorithm, achieving an accuracy of 0.91 in predicting accident severity.

Validation of the model's spatial prediction accuracy, along with a comparison of actual and predicted data, revealed a correlation coefficient of 0.99 between the two datasets. This high correlation underscores the model's accuracy and reliability in predicting accident severity.

Cite this article: *Forouzandeh, M., Mousavi, B.S., Neysani, N., & Argany, M. (2025). Spatial Prediction of the Impact of Road Accidents on Traffic using Machine Learning Algorithms. Earth Observation and Geomatics Engineering, Volume 8 (Issue 1), Pages 29-47. http//doi.org/10.22059/eoge.2025.378825.1155*

Publisher: University of Tehran.

DOI: http//doi.org/10.22059/eoge.2025.378825.1155

1. Introduction

Road accidents pose a significant challenge in many countries worldwide, particularly in the United States, where they result in substantial human and financial losses. With the increasing number of vehicles, the complexity of transportation networks, and varying weather conditions, predicting the severity of road accidents and identifying the factors influencing them has become one of the critical challenges in intelligent transportation systems and road safety management. Accident severity, which refers to the extent of damage or injury caused by an accident, is one of the most important parameters in road accident analysis due to its long-term impact on societies and national economies (Bahiru et al., 2018).

Machine learning algorithms are increasingly recognized as effective tools for data prediction and analysis across various fields, including the prediction of road accident severity. These algorithms can process complex datasets, uncover hidden patterns, and aid in improved decision-making and safety evaluations. Common algorithms applied in this domain include Decision Trees, Naive Bayes, Random Forest, XGBoost, K-Nearest Neighbors (K-NN), and ensemble models such as the Voting Classifier (Jamal et al., 2021).

The primary objective of this study is to predict the severity of road accidents in the United States from 2016 to 2023. The research aims to analyze and predict accident severity using datasets containing diverse information such as weather conditions, temperature, humidity, wind, time of day (day or night), geographical locations (cities and roads), and traffic characteristics. By employing various machine learning algorithms, the study seeks to identify the factors that most influence accident severity and provide predictions that can help prevent severe accidents (Bahiru et al., 2018).

The objectives of this study are as follows:

- Predicting Accident Severity: Utilizing machine learning algorithms to predict the severity of road accidents in the United States, considering features such as weather conditions, time of day, geography, and traffic patterns.
- Analyzing Influential Factors on Accident Severity: Identifying and analyzing the features that have the greatest impact on accident severity, including environmental factors, road characteristics, and accident timing.
- Geospatial Modeling of Accident Severity Prediction: Leveraging spatial data to predict accident severity in various geographical regions and identify high-risk areas for accidents.
- Comparing and Evaluating Models: Comparing the accuracy of different machine learning models, including the Voting Classifier and individual algorithms such as Random Forest, XGBoost, Naive Bayes, Decision Tree, and K-NN, in predicting accident severity.
- Providing Practical Recommendations: Based on model results and predictions, offering actionable recommendations to reduce accident severity across different regions and improve road safety policies.

Given the large number of road accidents and their consequences, predicting accident severity is crucial for road safety management, accident prevention planning, and resource allocation to high-risk areas (Gissane, 1965). This research can assist traffic authorities, road safety officials, and other relevant organizations in reducing accidents by identifying hazardous areas and implementing appropriate safety measures to prevent severe accidents and fatalities. Additionally, considering recent advances in machine learning, this study provides innovative and practical solutions for predicting and mitigating the severity of road accidents. It also builds upon previous research efforts that utilized machine-learning algorithms for accident prediction (Chen & Jovanis, 2000).

2. Research Background

Kumar et al. presented an IoT-based vehicle accident detection and classification system.[1] that integrates internal sensors and connected smartphone sensors to detect and report the type of accident. They compared three machine learning models based on NB[2], Gaussian mixture model[3], and decision tree techniques to identify the best ADC model. The results showed that the NB-based ADC model was accurate, with an F1 score of 0.95 (Kumar et al., 2021). Shen and Wei used the extreme gradient boosting algorithm[4] to analyze road transport accident data from seven regions in China. The results indicated that the proposed gradient boosting method had the best modeling performance (Shen & Wei, 2020). Parsa et al. considering the necessity of detecting road accidents as quickly as possible for traffic safety, used extreme gradient boosting to identify the occurrence of accidents using a set of data consisting of traffic, network, demographics, land use, and weather features. The results showed that extreme gradient boosting could detect accidents with an accuracy, detection rate, and false alarm rate of 99%, 79%, and 16%, respectively. Several traffic-related features, especially the speed difference between 5 minutes before and 5 minutes after the accident, had a relatively more significant impact on the occurrence of accidents (Parsa et al., 2020). Elyassami et al. reviewed the causes of accidents and accident severity factors and presented a prediction model for the occurrence and severity of road accidents. They used three machine learning algorithms (decision tree, random forest, and gradient-boosted tree). The findings indicated that inattentiveness to traffic signals and stop signs, road design issues, poor visibility, and adverse weather conditions were the most important variables in the road traffic accident prediction model. Using identified risk factors to create measures that might reduce risks associated with those factors is very important (Elyassami et al., 2021). Ahmed et al. studied single and ensemble machine learning methods, considering various influential factors and their impact on accident severity prediction. The results showed that the random forest outperformed other methods, such as logistic regression, nearest neighbor, naive Bayes, gradient boosting, and adaptive boosting. Logistic regression, nearest neighbor, and naive Bayes performed similarly for binary and multi-class classification. Compared to single methods, ensemble machine learning methods could predict accident severity more accurately, with random forest, gradient boosting, and adaptive boosting being the top three (Ahmed et al., 2021). Manzoor et al. identified important factors correlated with highway accident severity using random forest. The main features affecting accident severity were distance, temperature, wind, humidity, visibility, and wind direction. They presented a set of machine learning and deep learning models combining random forest, and convolutional neural networks called RFCNN to predict road accident severity. The results showed that RFCNN, using 20 important features, improved the decision-making process in predicting accident severity and outperformed other models with an accuracy of 0.991 (Manzoor et al., 2021) . Kushwaha and Abirami used a prediction algorithm to predict the occurrence of road accidents and a classification algorithm to categorize the severity of road accidents. Their comparative study on various machine learning algorithms for road accident data sets showed that random forest, extreme gradient boosting, and naive Bayes algorithms had better results in terms of accuracy (Kushwaha & Abirami, 2021). Karri et al. used support vector machine [5] and nearest neighbor [6] techniques to validate the classification of driving behavior in terms of safe/unsafe stops at signalized intersections. The results showed that the support vector machine technique could infer driver-stopping behavior with a high level of performance (Karri et al., 2021). Zhao and Deng developed an accident prediction model using effective information based on millions of road accident data in the United States, selecting and presenting information from five aspects: traffic, location, weather, points of interest, and time features. They used extreme gradient boosting, LightGBM, CatBoost, stacking, and elastic net algorithms to build a heterogeneous ensemble learning model for predicting accident duration. The results showed that the model had good prediction accuracy and could combine multiple models to provide a degree of importance for influencing factors. The feature importance of the model indicated that time, location, weather, and historical accident statistics were important for accident duration (Zhao & Deng, 2022). Libnao et al. proposed a traffic incident prediction and classification system using the naive Bayes algorithm to predict traffic incidents, which can lead to improved incident management and traffic flow. This system aimed to determine the likelihood of occurrence or non-occurrence of an incident and classify it using traffic data, including location, date, and time. The results showed that this

[1] *ADC*
[2] *Naïve Bayes (NB)*
[3] *GMM*
[4] *XGBoost*
[5] *SVM*
[6] *K-Nearest Neighbor (KNN)*

algorithm could predict and classify incidents with an accuracy of 70.03% (Libnaoa et al., 2023).

Karami and Farajzadeh studied road accidents on the Firoozkooh-Sari route under rainy, snowy, icy, and foggy weather conditions. The results showed the highest risk of accidents exists during adverse weather conditions. With an increase in snowy and icy days, there was a significant increase in accidents in terms of frequency, severity of damages, number of fatalities, and injuries at a 95% confidence level (Karami & Farajzadeh, 2014). Mousavi Fouladi analyzed road accidents on the mountainous and foothill routes from Semnan to Foulad Mahalleh. The results showed that elevations and passes play a fundamental role in causing climatic anomalies and consequently increasing the risk of road accidents in mountainous sections(Mousavi Fooladi, 2011).

Additionally, in favorable weather conditions, the highest frequency of accidents was related to high vehicle speeds in foothill sections and high traffic volume due to weekend travel (Abdolahzadeh Fayegh & Esmailzade, 2013). Abdollahzadeh and Esmailzadeh concluded that biological dimensions, individual and psychological factors, and short-term abilities each impacted the occurrence and severity of road accidents based on the calculated regression coefficients (Abdolahzadeh Fayegh & Esmailzade, 2013). Mirzaei Khezri and Saghayi predicted the severity of fatal road accidents using independent variables such as the cause, type of collision, type of responsible vehicle, accident time, day, and season. They fitted various base classification models such as K-nearest neighbors, naive Bayes, decision tree, and ensemble bagging and boosting classifiers using 2009 fatal accident data. The results showed an increase in the accuracy of the ensemble bagging and boosting classifiers by 0.04, with the ensemble bagging classifier using the decision tree algorithm and the ensemble boosting classifier using the naive Bayes algorithm having high efficiency in predicting the severity of fatal road accidents (Mirzaei Khezri & Saghayei, 2014). Vatanparast et al. studied the role of climatic and human factors in road accidents using fuzzy logic and geographic information systems. The study showed that human factors played a significant role (over 90%) in road accidents. The most dangerous areas for road accidents were identified as the first 5 to 35 kilometers of the entrance and exit of cities using the fuzzy linguistic model. Cloudy and rainy conditions also had the most significant role in accidents in this corridor (Vatanparast et al., 2016). Pourgholami et al. used statistical methods such as means and statistical tests to study accidents. The results showed that most climatic elements, such as cloud cover, relative humidity, wind speed, and precipitation, had a direct role in winter. Still, the temperature only played a role in autumn and had an inverse role in the other three seasons. Relative humidity had an inverse role only in spring, and wind played the least role in spring and winter (Pourgholami et al., 2016).

Vakil Al-Raya and Zarghami studied the prioritization of factors affecting road accidents using the Analytic Network Process (ANP). The results showed that the human factor had the most influence on road accidents, followed by vehicles, roads, and the environment (Vakil Roaya & Zargar, 2017). BabaGholi et al. studied the relationship between accident severity and collision type. They used the multinomial logit model from the choice models to provide a prediction model for accident severity. They compared the results with the binary prediction model using data mining algorithms, including the CART decision tree and MLP-ANN neural network algorithms. The results showed that the best model in terms of prediction accuracy and the ability to provide a prediction formula for each level was the MNL model. The results in the prediction models section showed that the estimated formula could accurately predict the severity of accidents at levels zero and one (Babagoli et al., 2018). Salimi et al. reviewed the performance of the support vector machine algorithm in predicting the severity of road accidents on intercity roads in Iran and determining the influential factors on the severity of motorcycle accidents. The results showed that heavy vehicles and passenger cars were the most important factors in estimating the severity of motorcycle accidents, increasing the probability of injury. For passenger car accidents, factors such as fatigue and drowsiness, pedestrian presence, and left deviation of passenger cars had a direct impact on the severity of passenger car accidents. additionally, in truck accidents, the left deviation of the truck, the driver's age, and increased vehicle speed were the most important factors in increasing the level of injury in this type of accident (Khajesalimi et al., 2018) . Keymanesh and Rahmanian presented a method for identifying and predicting the severity of accidents on intercity roads. By comparing the two models, multivariate analysis[7] and artificial neural network[8] Technique, they stated that model 2 had better results because the total residual value, the difference between the actual observed value and the value predicted by the model, was less in model 2. However, Model 1 seemed to perform better in estimating the number of dangerous accident-prone sections

[7] MVA

[8] ANN

(with many accidents) (Keymanesh & Baradaran Rahmanian, 2021). Tavakoli Kashani et al. studied the factors affecting the severity of intercity road accidents in Zanjan province using two models: a support vector machine and a decision tree. The analysis showed that in the support vector machine model, the type of collision, type of vehicle, and kilometer of the accident, and in the decision tree model, the type of collision, type of responsible vehicle, and kilometer of the accident were the most important factors affecting the severity of accidents in Zanjan province. The decision tree model, with an accuracy of 75%, and the support vector machine model, with an accuracy of 81%, had good performance in predicting the severity of accidents (Tavakoli Kashani et al., 2022).

Koohi and Shabani aimed to identify the factors influencing the severity of intercity road accidents using the multinomial logit[9] Model. Using data from 5 years of intercity accidents in Ilam province, the input variables for the model were selected after evaluating their significance. These variables included young age, high speed, alcohol consumption, head-on collisions, presence of airbags, ejection from the vehicle, wearing seatbelts, very close distance to the vehicle in front, driver's gender, and occurrence of accidents in curves. The log-likelihood ratio and model prediction accuracy at each severity level were used for model validation. The model's accuracy evaluation in predicting accident severity showed that the model provided acceptable results, and except for the variables of curves and close distance, all selected variables contributed to accident severity (Koohi & Shabani, 2023).

Kalantari and Alian used the hierarchical decision tree method to analyze the spatial distribution and probability of road accidents and the factors affecting them, especially the environment and road factors in a geographic information system. They collected data related to road slope, degree of curvature, intersection points, density percentage, climate, and land use around the area and processed it. They mapped the standardized layers using the decision tree and weighted decision tree models. According to the analysis, the study results showed that the most important factor influencing accident occurrence was curvature, with the variables of intersection, climate, density, and slope following in priority (Kalantari & Alyan, 2022).

Nemati et al. predicted the probability of driver fatalities in road accidents in Canada using the binary logit model. The dependent variable in this model was the accident's severity, which is a binary variable (driver fatality and driver injury). The independent variables included types of vehicles, vehicle age, days, time intervals, same direction, opposite direction collisions, accident location, weather conditions, driver's age, and gender. They used 70% of the data for modeling and 30% for model validation. They then used the McFadden test index to evaluate the model's performance. The results showed that the model was a good fit with the data and could predict accident severity changes. Variables such as dry road surface, midnight interval, and vehicle age increased fatalities, while variables like light vehicles, school buses, and same-direction collisions reduced the likelihood of fatal accidents (Nemati et al., 2023).

3. Research Methodology

3.1. Study Area

Based on millions of traffic accident data points in the United States, this study presents a spatial prediction model for accident severity considering weather characteristics. Data from U.S. road accident severity from 2016 to 2023 were used. The United States of America, an independent country in North America, lies between Canada and Mexico, bordered by the Atlantic and Pacific Oceans. Besides the mainland, the U.S. is a federal republic comprising 50 states, one federal district, five autonomous territories, and a collection of scattered islands, covering a total area of 9.8 million square kilometers, making it the third largest country in the world. With a population exceeding 340 million as of 2023, it is the third most populous country globally. The capital is Washington, D.C., and its most populous city is New York City in New York (Parsi, 2023, Parsi, 2023, Worldometer, 2021, Worldometer, 2021). The country shares land borders with Canada to the north and Mexico to the south and maritime borders with Russia off Alaska's northwest coast (Shulski & Wendler, 2007, Klein, 2007). (Figure 1) shows a view of the study area for this research.

3.2. Research Data

The dataset used in this research pertains to road accidents in the United States, comprising approximately 2.2 million records over five years (2016 to 2023). This dataset includes raw information from various road accidents, enriched through processes such as Map Matching and the integration of textual details like weather conditions, days of the week, and traffic points (without

[9] MNL

accidents). Additionally, the data contains critical information such as geographical coordinates, weather conditions, and the precise time of each incident.

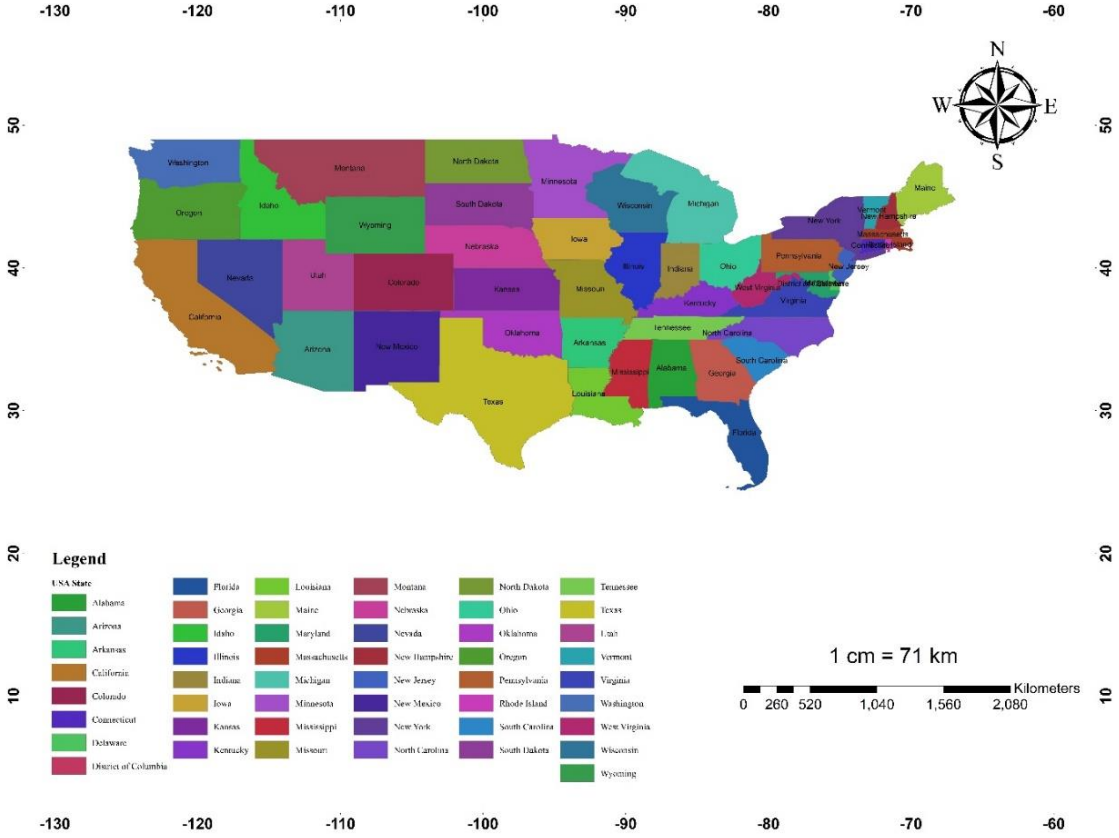


Figure 1. View of the study area (United States)

The features of this dataset include:

• Geographical Coordinates: Latitude and longitude of each accident, used for spatial analyses and predicting accident severity.

• Weather Conditions: Information related to temperature, humidity, wind speed, and precipitation type (rain, snow, or clear).

• Time of Occurrence: Exact timing of the accident (date, time, and day of the week), which can assist in analyzing temporal patterns of accidents.

• Road and Traffic Characteristics: Information such as road conditions, road type (arterial, highway, mountainous), and traffic conditions at the time of the accident, which influence accident severity.

This dataset was selected due to its comprehensive coverage, especially in terms of temporal and geographical aspects, and its relatively complete data, which makes it suitable for analyzing accidents and identifying effective patterns in predicting accident severity.

For data preprocessing, missing values were identified and replaced using the mean or median for each feature. Additionally, the features were normalized to minimize the impact of scale differences on the models and improve prediction accuracy.

3.3. Machine Learning Algorithms for Prediction

Machine learning algorithms utilize statistical methods to analyze data and extract hidden patterns. These algorithms are designed to take input data and make predictions using statistical techniques, with predictions being updated as new data becomes available. Machine learning algorithms are generally categorized into three types: supervised learning, unsupervised learning, and reinforcement learning.

In supervised learning algorithms, models learn from labeled input-output pairs to identify the best possible model. This model is then used to predict outputs for new input data. In unsupervised learning algorithms, the data is unlabeled, and the algorithm seeks to discover patterns in

the data automatically. These algorithms are often used for more complex analyses and uncovering unknown patterns within the data.

In this study, five machine learning algorithms were employed to predict road accident severity: Naive Bayes, Random Forest, Extreme Gradient Boosting (XGBoost), Decision Tree, and K-Nearest Neighbors (K-NN). These algorithms were individually trained and then tested using test data to predict accident severity. Finally, the predictions from these algorithms were combined using the Voting Classifier to maximize prediction accuracy (Karri et al., 2021).

3.3.1. Naive Bayes Algorithm

The Naive Bayes algorithm is one of the simplest machine learning algorithms, relying on Bayes' theorem to calculate the probability of an event given prior knowledge of related conditions. This probabilistic approach assumes independence among the features of the input data, simplifying probability calculations and enhancing computational efficiency. When presented with new data, Naive Bayes calculates the probability of each class based on the observed features and selects the class with the highest probability as the predicted outcome. Despite its simplicity and assumptions, Naive Bayes demonstrates strong performance across a wide range of classification tasks, particularly when the training data is limited compared to the number of features. The Naive Bayes classifier effectively combines a probabilistic model rooted in Bayes' theorem with a decision rule to maximize posterior probability (Elyassami et al., 2021).

3.3.2. Random Forest Algorithm

The Random Forest [10] Algorithm is a classification algorithm introduced by Tin Kam Ho in 1995 and later improved by Leo Breiman in 2001. This algorithm operates by constructing multiple independent decision trees during the training phase and uses randomness in feature selection to create a diverse forest of trees. The Random Forest classifier gathers classifications from each decision tree when classifying a new sample. It combines them using a majority vote or confidence vote strategy to generate a more accurate prediction than any individual tree alone. This principle of ensemble learning enhances predictive accuracy. In the Random Forest model, training samples from random features are used to generate trees. Unlike traditional decision trees, Random Forest classification reduces the risk of overfitting because it contains enough trees to allow each tree to grow fully without pruning. Overall, Random Forest uses multiple uncorrelated classifiers to create an effective ensemble classifier (Elyassami et al., 2021).

3.3.3. Extreme Gradient Boosting (XGBoost)

Extreme Gradient Boosting is another machine learning algorithm for classification and regression problems. It is an advanced version of the Gradient Boosting. [11] An algorithm, unlike a Random Forest, follows a sequential process where each subsequent step improves the accuracy of the previous step. This algorithm uses regression trees as a base learning model. XGBoost includes several optimizations and parallel processing implementations, making it more efficient than traditional Gradient Boosting. It can automatically detect the number of system cores and execute code on all cores to achieve better speed. One important parameter in implementing XGBoost is the regularization parameter, which defines the minimum loss reduction required to perform a split in the tree (Ahmed et al., 2021).

3.3.4. Decision Tree Algorithm

The Decision Tree[12] Algorithm is a predictive modeling approach used for classification and regression tasks, first introduced by Breiman et al. in 1983. This algorithm is mainly used in predictive modeling, statistics, and data mining, among other fields, and consists of leaves and branches (Geyik & Kara, 2020). Branches represent observations about patterns, and leaves represent the target variable of the pattern. The Decision Tree is a supervised algorithm for classification problems involving nodes as leaves. Decision trees use the divide-and-conquer technique to split the problem search space into subsets, with the root and each internal node labeled with a question. The arcs originating from each node represent possible answers to the related question. Each leaf node represents a predicted solution to the problem being examined. The algorithm recursively constructs the tree from top to bottom, selecting the best splitting feature using initial training data. Algorithms differ in how they determine the best feature and its optimal split points. Once this is established, nodes and their arcs are created and added to the tree, and the algorithm continues recursively by adding new subtrees to each branching arc. The algorithm terminates upon

[10] *Random Forest (RF)*
[11] *GBM*
[12] *Decision tree*

reaching certain stopping criteria, and pruning is necessary to avoid overfitting (Elyassami et al., 2021).

3.3.5. K-Nearest Neighbors (K-NN) Algorithm

The K-Nearest Neighbors algorithm is a widely used machine learning algorithm for classification and regression problems based on feature similarity. The K value helps perform operations such as data analysis, distance measurement, clustering, and similarity between datasets, utilizing the Euclidean distance measurement for relevant calculations. Initially, the K value represents the number of neighbors, followed by measuring the Euclidean distance among them (Elyassami et al., 2021).

3.3.6. Voting Classifier Algorithm

The Voting Classifier[13] It is a machine learning model that trains on an ensemble of multiple models and selects an output (class) based on the highest probability of the selected class as the predicted output. This method simply gathers the findings of each classifier submitted to the Voting Classifier and predicts the output class based on the majority vote. The idea is to create a single model that trains with these models and predicts the output based on the combined majority vote for each output class rather than building separate specialized models and finding the accuracy for each. A Voting Classifier employs two types of voting techniques: hard voting and soft voting. In hard voting, the final Prediction is made by majority vote, where the classifier selects the class that repeatedly comes up from the base models. By combining the predictions of various models, the Voting Classifier provides overall better results than other base models (Kumari et al., 2021).

3.4. Correlation Matrix

The Pearson correlation method is a widely used statistical technique for measuring the linear relationship between two variables, with results ranging between -1 and 1. A value of 1 signifies a perfect positive correlation, where an increase in one variable corresponds to a proportional increase in the other. Conversely, a value of -1 indicates a perfect negative correlation, meaning that as one variable increases, the other decreases proportionally. A correlation value of 0 suggests no linear relationship between the variables. This method is particularly effective when the data exhibits a normal distribution and a linear relationship.

Building on this, the correlation matrix is an essential tool in feature engineering and data analysis, as it provides a comprehensive view of the relationships between multiple variables in a dataset. The matrix contains correlation coefficients that quantify the strength and direction of linear relationships between all pairs of features. This tool is especially valuable for several reasons:

1. Identifying Relationships: It helps data scientists quickly identify whether and how variables are related. Positive correlations suggest that as one variable increases, the other tends to increase, while negative correlations indicate an inverse relationship.

2. Feature Selection: In feature engineering, the correlation matrix aids in identifying features with a strong correlation to the target variable. These features are more likely to significantly impact the target and can be prioritized in predictive modeling.

3. Detecting Multicollinearity: The matrix also helps detect multicollinearity, a scenario where independent

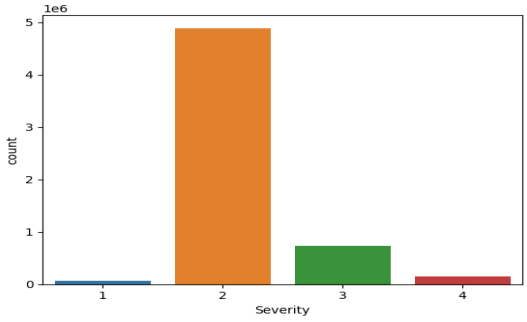


Figure 3. Frequency of observations for the target variable "accident severity"

variables are highly correlated. Multicollinearity can lead to redundancy in features and issues such as unstable coefficient estimates in regression models, which can compromise the interpretability and reliability of the model(Senthilnathan, 2019).

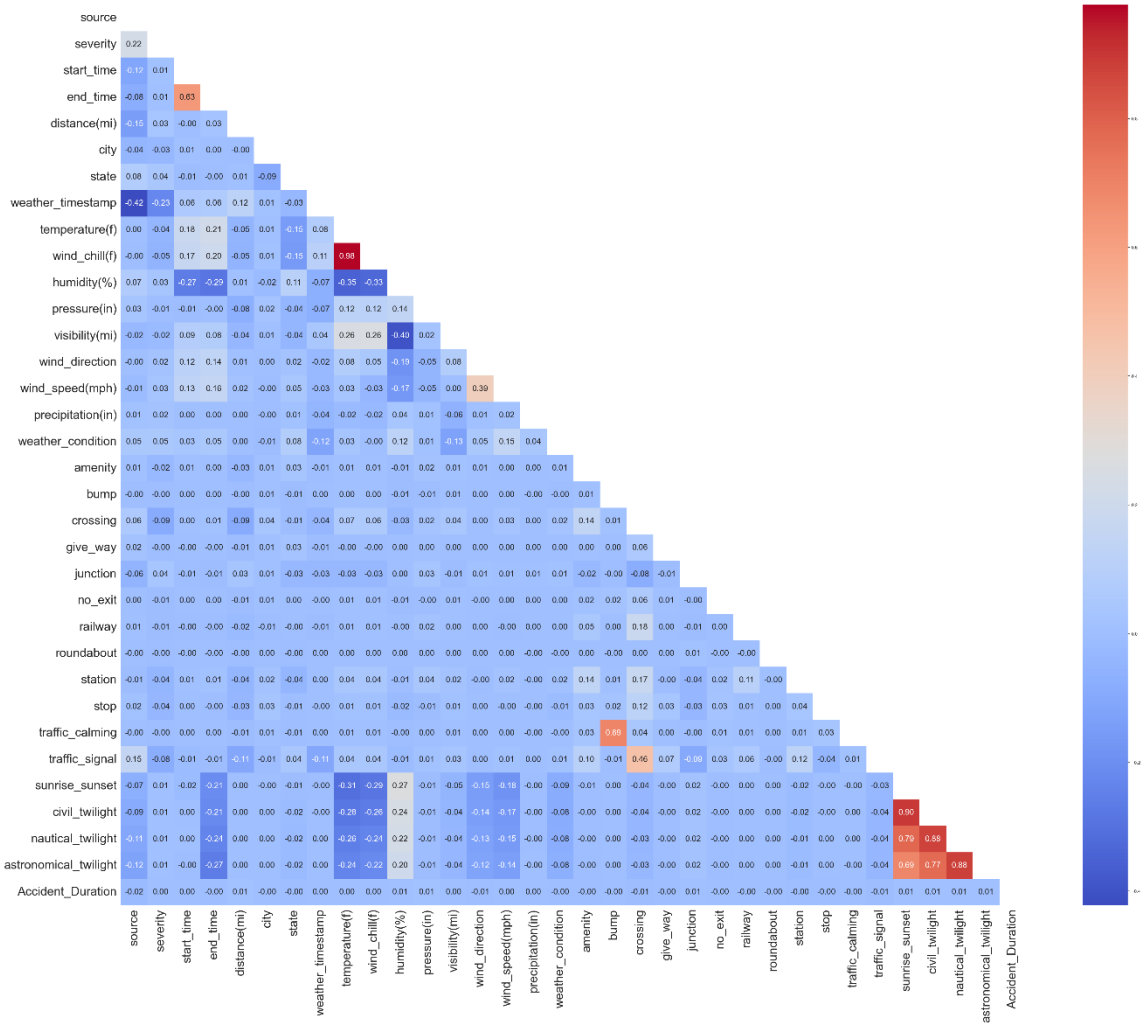
By combining the insights provided by the Pearson correlation method and the correlation matrix, analysts and data scientists can better understand the structure of their data, select relevant features, and improve model performance. These tools are foundational for uncovering linear relationships and ensuring robust statistical and

[13] *Voting Classifier*

machine-learning models(Umer et al., 2020, hang et al., 2014).

3.5. Ensemble Learning

This study used ensemble learning models to predict the severity of road accidents in the United States from 2016 to 2023. After preprocessing and classifying the data, the dataset was randomly split into 70% for training and 30% for testing. The training data were used to train machine learning models to predict accident severity as the target variable. Correlation analyses were then conducted between various influencing factors such as weather conditions (e.g., annual humidity, temperature, wind), days of the week, and the effects of day versus night. Identifying and predicting accident severity based on these factors over seven years is particularly important. In this study, five algorithms Naive Bayes, Random Forest, Extreme Gradient Boosting, Decision Tree [14] , K-Nearest Neighbors, and Voting Classifier[15] were trained for Prediction, and then the test dataset was used to predict accident severity. Subsequently, through a combination of voting techniques, feature weighting, and averaging, an ensemble of machine learning methods achieved an accuracy of 85%. Finally, the geographical Prediction of accident severity for the next year was presented in the results section.



The Voting Classifier used in this research comprises an ensemble of five machine learning algorithms (Naive Bayes, Random Forest, Extreme Gradient Boosting, Decision Tree, K-Nearest Neighbors, and Voting Classifier) that were used to train all machine learning models and predict road

Figure 2. Correlation matrix identifying features and factors influencing accident severity

[14] *Decision Tree*

[15] *Voting Classifier (VC)*

accident severity. The combined Voting Classifier, a collection of individual classifiers, integrates the prediction results of the classifiers and can achieve better outcomes than single classifiers. This study uses the ensemble of five mentioned machine learning algorithms to predict road accident severity. To achieve the objectives of this research, the algorithms were implemented and executed using the Python programming language.

4. Results

The Pearson correlation method is used to measure the linear relationship between two variables, with values ranging from -1 to 1. A value of 1 indicates a perfect positive correlation, meaning that as one variable increases, the other also increases in a perfect manner. Conversely, a value of -1 indicates a perfect negative correlation, meaning that as one variable increases, the other decreases perfectly. When the correlation value is 0, it indicates no linear relationship between the two variables. This method is highly useful for measuring linear relationships and performs best when the data follows a normal distribution and exhibits a linear relationship.

This research plotted a Pearson correlation matrix to identify features and factors influencing accident severity. In this matrix, red colors indicate positive correlations and blue colors indicate negative correlations. The darker the color, the stronger the correlation (Figure 2). Accordingly, a strong positive correlation between wind and temperature indicates a direct relationship between temperature and wind. In contrast, a positive correlation between traffic and speed bumps shows that one cause of traffic is speed bumps, which slow down vehicles and subsequently increase traffic. The positive correlation between traffic and pedestrian crossings also indicated that one cause of traffic is pedestrians, as pedestrian crossings reduce vehicle speed and subsequently increase traffic.

There is a negative correlation between the time of weather observation and its significant impact on traffic, indicating that weather is recorded when there is no traffic. The negative correlation between humidity and temperature showed that wind has an inverse relationship with air temperature, which is an intuitive relationship. Also, the negative correlation between horizontal visibility and air humidity showed that with increased air humidity, horizontal visibility decreases.

Charts were plotted for each year every month to examine the factors influencing accident severity and frequency. Accordingly, accident severity was divided into four classes: one, two, three, and four. The severity of accidents was represented by a number between 1 and 4, where 1 indicates the least impact on traffic (short delay due to the accident), and 4 indicates a significant impact on traffic (long delay due to the accident). Figure 3 shows a bar chart displaying the number or frequency of observations for the levels of the target variable "accident severity." In this chart, the x-axis represents severity levels, and the y-axis represents the number of accidents. The tallest bar, in orange, represents the high number of accidents for severity level 2. Therefore, considering their severity, most accidents relate to severity level 2.

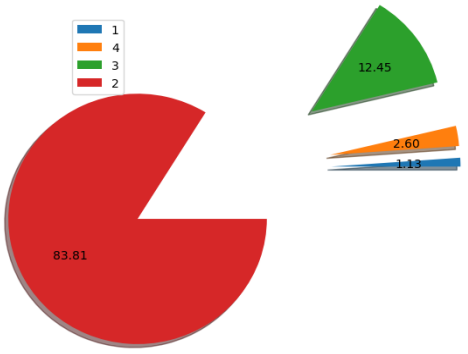


Figure 4. Percentage of severity classes impacting traffic

The pie chart (Figure 4) shows that the highest percentage of the severity impact on traffic is in class 2, accounting for 83.81%, depicted in red. Additionally, the percentage of other classes in the severity impact on traffic is displayed, with the lowest in class 1, shown in blue, at 1.13%.The trend of accidents over 12 months from 2016 to 2021 showed that the number of accidents increased towards the end of the year. Weather conditions seem to lead to more accidents in October, November, and December (Figure 5).

The weekly accident trend from 2016 to 2022 showed a specific pattern, with the highest number occurring on Tuesdays, Wednesdays, and Thursdays and the lowest on Saturdays and Sundays. This indicates that the number of accidents is higher on weekdays (working days) compared to weekends (Figure 6).

The frequency chart (Figure 7) showed that most accidents occurred at latitudes 34 and 41. The peaks represent higher observation rates around specific latitude values. The frequency chart (Figure 8) showed that most accidents occurred at longitudes -93 and -78. The peaks represent higher observation rates around specific longitude values.

xaminations of the number of accidents in each state also showed that the highest number of accidents occurred in the

cities of California.[16], Florida[17], Texas[18], and South Carolina[19], respectively (Figure 9).

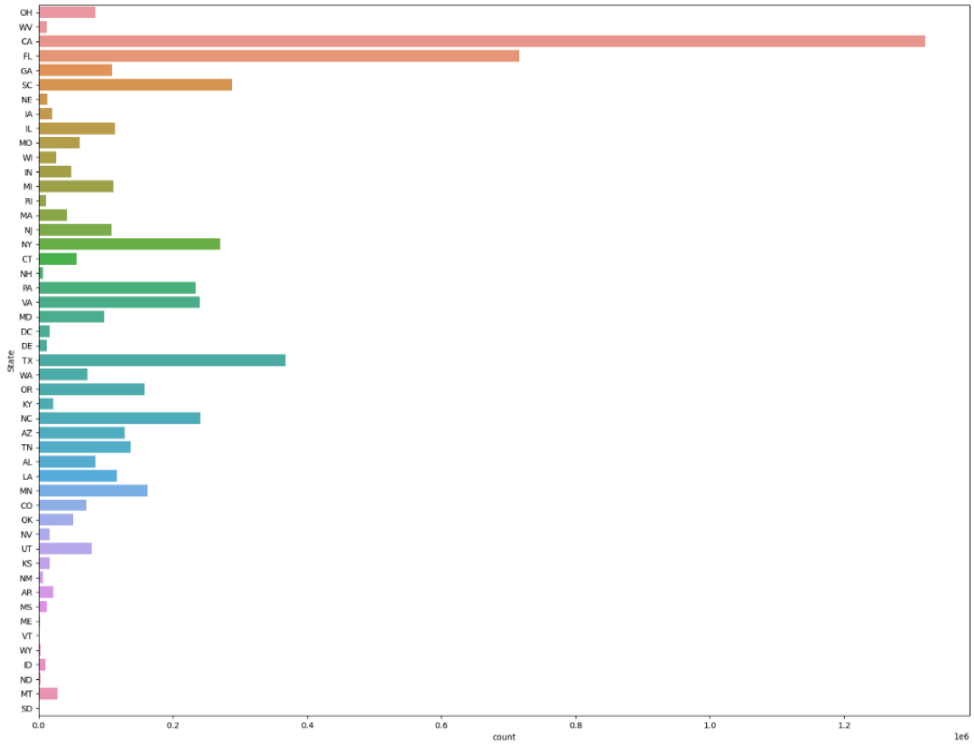


Figure 9. Number of accidents in US cities

An examination of traffic and non-traffic accidents showed that the number of traffic accidents is less than that of non-traffic accidents. This may indicate that reduced vehicle speed in traffic conditions decreases the severity of accidents (Figure 10).

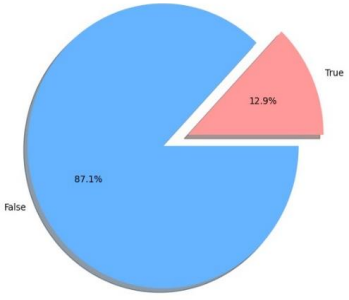
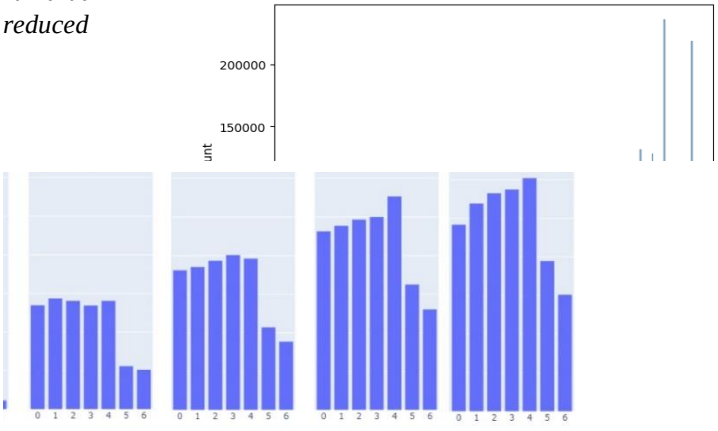


Figure 10. Percentage of traffic and non-traffic accidents

Examining the number of accidents throughout the day showed that the number of accidents at night is 32%, less



gt
cidents over the days of the week from 2016 to 2022

[16] CA
[17] FL
[18] TX
[19] SC

than the number of accidents during the day. Reduced driver visibility at night was not a significant factor in accidents (Figure 11).

Examining the relationship between humidity and the number of accidents, the results showed that with increasing humidity, the number of accidents gradually increases to a maximum at around 90% humidity and then decreases sharply (Figure 12).

Additionally, the impact of humidity over 12-month periods from 2016 to 2022 showed that the number of accidents increases in the later months with increasing humidity during the year. The decrease in temperature and humidity in the later months of the year leads to road slipperiness, increasing the number of accidents (Figure 13).

The results of the trend of the impact of wind on the number of accidents over 12-month periods from 2016 to 2022 showed that the number of accidents increased with increasing wind intensity during the year. Wind affects the control and stability of vehicles, including cars, trucks, trains, airplanes, and ships. Strong winds, especially on highways or bridges, can lead to collisions, deviations, or loss of vehicle control (Figure 14).

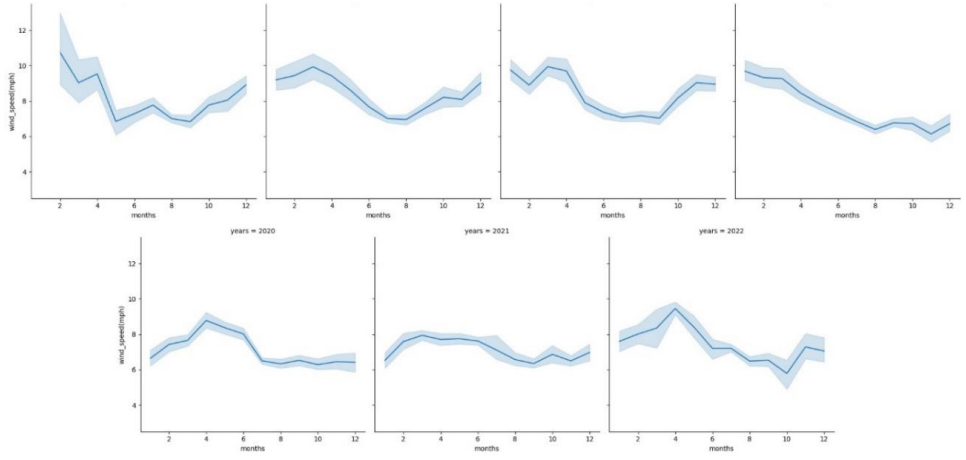


Figure 14. Trend of the impact of wind on the number of accidents from 2016 to 2022

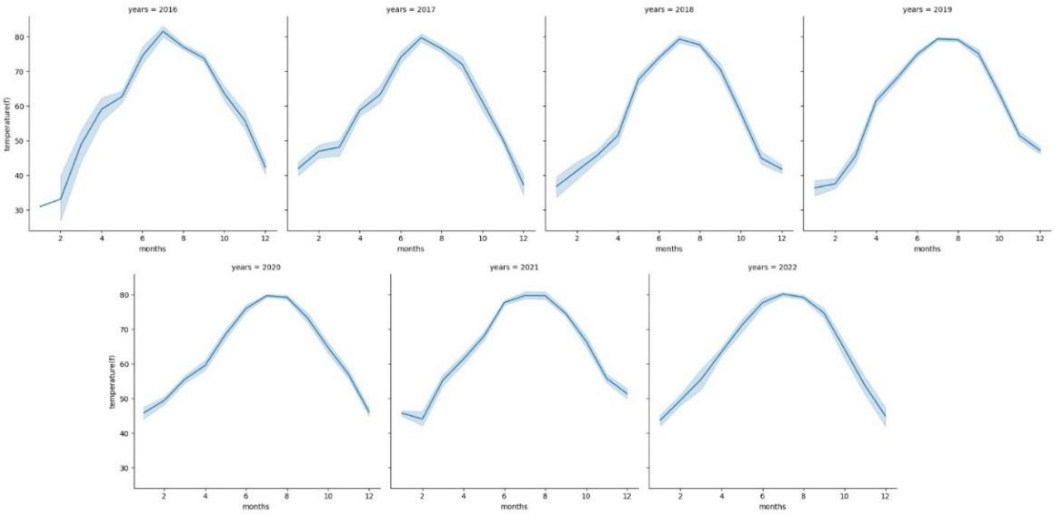


Figure 15. Trend of the impact of temperature on the number of accidents from 2016 to 2022

Table 1. Comparison of machine learning algorithm accuracy for predicting accident severity on traffic impact

Model	Train Score	Test Score	Precision Score	Recall Score	F1 Score	accuracy
Random Forest	*0.99*	*0.91*	*0.90*	*0.91*	*0.90*	*0.91*
Decision Trees	*1.00*	*0.78*	*0.78*	*0.78*	*0.78*	*0.78*
KNN	*0.79*	*0.70*	*0.69*	*0.70*	*0.69*	*0.70*
Naive Bayes	*0.40*	*0.40*	*0.06*	*0.25*	*0.10*	*0.25*
XGBoost	*0.86*	*0.84*	*0.84*	*0.84*	*0.84*	*0.84*

Table 2. Voting algorithm for predicting accident severity on traffic

Model	Train Score	Test Score	Precision Score	Recall Score	F1 Score	accuracy
voting	*0.97*	*0.91*	*0.90*	*0.91*	*0.90*	*0.91*

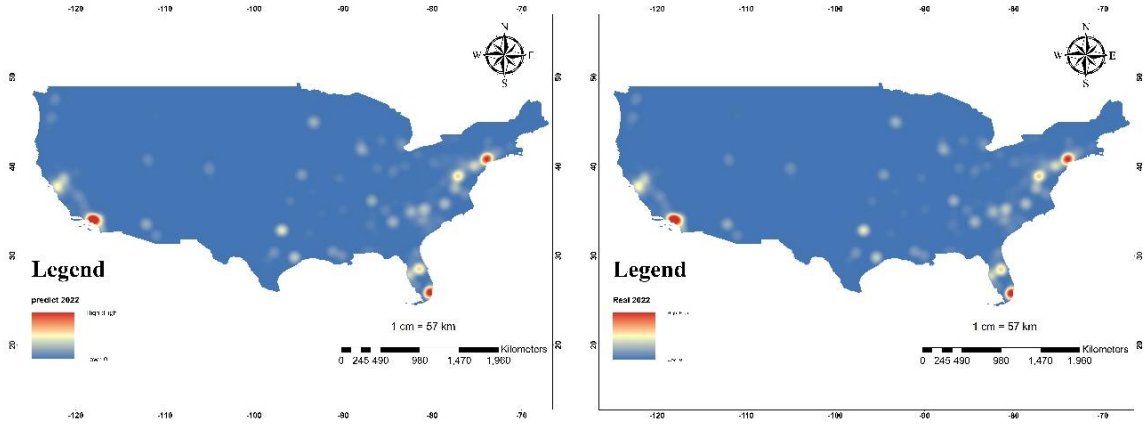


Figure 16. Actual accidents in 2022 - Accident prediction in 2022

The trend of the temperature impact over 12-month periods from 2016 to 2022 showed that the number of accidents increases with increasing temperature during the year. Higher temperatures lead to the susceptibility of asphalt to deformation and road deterioration, which impacts the increase in accidents (Figure 15).

4.1. Algorithm Results

In this study, machine learning algorithms, including Random Forest, Decision Tree, Naive Bayes, Gradient Boosting, K-Nearest Neighbor, and a voting ensemble of Random Forest and Gradient Boosting, were compared as classifiers to predict the severity of road accidents. The results from running these algorithms on accident data, aimed at predicting the impact of accident severity on traffic, showed that Random Forest achieved the highest accuracy of 0.91. At the same time, Naive Bayes had the lowest accuracy of 0.25. Thus, based on the results from Table 1, following Random Forest, Gradient Boosting had the next highest accuracy (0.83), followed by Decision Trees (0.77), and then K-Nearest Neighbor with an accuracy of 0.70 to improve accident severity prediction probability, the Random Forest and Gradient Boosting voting ensemble was chosen due to its higher accuracy compared to other algorithms, which excluded lower-accuracy algorithm validate the spatial accuracy of accident prediction results from the model, real data from 2022 was compared with model-predicted data. Various statistical multivariate analysis methods exist to measure the dependency or relationship between two random variables. The correlation coefficient between two variables indicates their predictability of each other's values. Correlation [20] or covariance [21] The two images were used, resulting in a correlation coefficient of 0.9994, indicating a high correlation between them (Figure 16). Thus, using the

[20] *Covariance*

[21] *Correlation Coefficient*

voting ensemble algorithm with two machine learning algorithms severity prediction improved to an accuracy of 0.91 (Table 2).

5. Discussion

This section discusses the findings from predicting road accident severity using machine learning algorithms in the United States between 2016 and 2023. Additionally, the results are compared with similar studies on predicting accident severity to evaluate the effectiveness of the proposed models.

5.1. Model Accuracy Evaluation

In this study, several machine learning algorithms were trained to predict road accident severity, including Naive Bayes, Random Forest, Extreme Gradient Boosting (XGBoost), Decision Tree, K-Nearest Neighbors (K-NN), and Voting Classifier. The Voting Classifier model, which combines five different models, achieved an accuracy of 85%, delivering satisfactory results compared to individual models. However, further evaluations revealed that Random Forest and XGBoost models, due to their specific features such as resistance to overfitting and the ability to model non-linear relationships, produced results comparable to the combined model (Voting Classifier). Specifically, the Random Forest model alone achieved significant accuracy, closely matching the Voting Classifier's results.

These findings align with previous research, which demonstrated that individual models like Random Forest could perform similarly to ensemble models. For example, Nemati et al. and Kalantari and Alian utilized individual machine-learning models to predict accident severity and, in some cases, obtained comparable results (Kalantari & Alyan, 2022, Nemati et al., 2023).

5.2. Feature Importance in Predicting Accident Severity

In this study, features such as weather conditions (temperature, humidity, wind), time of day (day or night), and geographical location (city, main roads, or secondary roads) were identified as influential factors in accident severity prediction. The models indicated that these features significantly impact the predictions. For instance, weather conditions like high humidity and strong winds were found to increase the likelihood of higher severity accidents.

These findings align with prior studies. Koohi and Shabani, in a similar investigation, examined the impact of weather conditions (e.g., rain and fog) and accident location (e.g., intersections or curves) on predicting accident severity. They found that conditions such as rain or fog significantly increased the likelihood of severe accidents (Koohi & Shabani, 2023). Moreover, Nemati et al. demonstrated in their study that environmental variables, such as road conditions and dry weather, influence accident severity, with severe accidents being more frequent under such conditions (Nemati et al., 2023).

5.3. Impact of Time of Day and Days of the Week

Another critical feature in this study was the time of day and the day of the week. The results showed that more severe accidents occurred during nighttime hours and on specific days of the week (particularly Fridays). These findings are consistent with previous research by Koohi and Shabani and Kalantari and Alian, who noted that nighttime and peak traffic hours significantly influenced accident severity. They concluded that nighttime and weekends (e.g., Fridays) considerably increased the likelihood of severe accidents (Koohi & Shabani, 2023,Kalantari & Alyan, 2022).

5.4. Geography and High-Risk Locations

An essential aspect of this study was the geographical prediction of accident severity. Using geographical coordinates and features such as road type and intersections, predictions for future years identified high-risk areas for accidents. These predictions can assist authorities in implementing appropriate safety measures in hazardous regions.

This type of prediction and geographical analysis has also been utilized by researchers like Kalantari and Alian and Nemati et al. to analyze accidents and identify high-risk areas. They successfully used geographical algorithms, such as spatial analysis, to locate hazardous points on geographical maps (Kalantari & Alyan, 2022, Nemati et al., 2023).In this study, machine learning models, combined with geographical data, predicted accident severity in various locations across the United States.

5.5. Voting Classifier Model

In this research, the Voting Classifier model, which combines multiple machine learning models, achieved an 85% accuracy in predicting accident severity. This model uses hard voting, where the best prediction is based on the majority vote from various models. The results showed that combining multiple models could significantly enhance prediction accuracy. However, after a more detailed evaluation of the results, it was observed that the Random Forest model alone achieved similar accuracy to the Voting Classifier model. Therefore, it is concluded that ensemble models were not necessarily superior to individual models

in this study, as the Random Forest model alone provided comparable results.

5.6. Comparison with Previous Studies

A comparison of the results from this study with previous research indicates that using advanced machine learning algorithms such as XGBoost and Random Forest, along with ensemble methods, can lead to more accurate predictions of accident severity. These findings are consistent with the results of Koohi and Shabani, who used the Multinomial Logit model to predict accident severity and simulated the impact of features such as driver age and alcohol use (Koohi & Shabani, 2023). Similarly, Nemati et al., using the Binary Logit model to predict road accident severity in Canada, found that environmental features and road conditions significantly influenced accident severity (Nemati et al., 2023). These findings align with this study, which examined the effects of weather conditions and geographical location on predicting accident severity.
To evaluate accident severity predictions, five different algorithms Random Forest, Decision Tree, Naive Bayes, Gradient Boosting, and K-Nearest Neighbors were used. To enhance prediction accuracy, a combined Voting Classifier model incorporating Random Forest and Gradient Boosting was applied. After evaluating the results, it was observed that the Random Forest model alone achieved similar accuracy to the combined model. Consequently, the use of the combined model was deemed unnecessary. The Random Forest model, with an accuracy of 91%, was selected as the final model, clearly demonstrating the high efficiency of this algorithm in predicting accident severity.

6. Conclusion

This study investigates road accident severity in the United States from 2016 to 2023 using ensemble learning models to predict the impact of these accidents on traffic and road safety. The findings showed that the Random Forest algorithm achieved the highest accuracy (0.91) in predicting accident severity, while the Naive Bayes algorithm had the lowest accuracy (0.25). Other algorithms, including Gradient Boosting (0.83), Decision Tree (0.77), and K-Nearest Neighbor (0.70), performed relatively well. Ultimately, the Voting Classifier, which combined Random Forest and Gradient Boosting, improved prediction accuracy to 0.91, which was the same as the Random Forest's accuracy.

In the spatial evaluation of the prediction model, comparing the actual data from 2022 with the model's predictions revealed a very high correlation of 0.99947, demonstrating the model's high accuracy in predicting accident severity. This high accuracy provides a valuable tool for forecasting future accidents and informing management decisions regarding road safety. The purpose of using the Voting Classifier in this study was to investigate whether combining multiple algorithms could significantly improve the accuracy of accident severity predictions. However, a detailed comparison of the results showed that the Random Forest model alone produced results similar to the combined model. Therefore, the final choice was based on the Random Forest model, which provided an accuracy of 0.91 in predicting accident severity. These findings suggest that ensemble models in this study did not significantly outperform the Random Forest model, which was able to make accurate predictions on its own.

Additionally, this study identified factors influencing accident severity, including weather conditions, the timing of accidents (e.g., increased accidents in the final months of the year), and geographical locations (e.g., states such as California, Florida, Texas, and South Carolina). The findings revealed that certain variables, such as humidity and wind, play a significant role in accident severity, especially on slippery roads and during adverse weather conditions.
One limitation of this study is the use of data limited to a specific time frame. The research only considered accident data from 2016 to 2023, which may not capture long-term trends or changes over extended periods. Another limitation is the exclusion of other factors, such as vehicle type, driver condition (e.g., fatigue, drug or alcohol use), and social or psychological factors. Moreover, environmental data such as road quality and urban infrastructure, which significantly influence accident occurrence, were not included in this study. Therefore, incorporating these factors in future research could make prediction models more accurate and comprehensive.
It is recommended that future research collect data over longer time periods to better simulate long-term trends and seasonal changes. Additionally, supplementary data on driver conditions, vehicle type, road quality, and other environmental factors should be used to create more accurate and reliable prediction models. Future studies could also focus on developing hybrid models that leverage multiple machine-learning methods simultaneously. These models could capture more complex data features and improve prediction accuracy. Moreover, incorporating additional data, including videos and accident images, and employing more sophisticated algorithms and deep learning models could enhance the precision of analysis and prediction in future research.

Acknowledgements
We acknowledge the use of data from the US Accidents dataset, obtained from

https://smoosavi.org/datasets/us_accidents. We appreciate the efforts of the dataset creators in providing this comprehensive resource for research purposes.

Appendix

The data utilized in this study is sourced from the US Accidents dataset, which includes detailed records of traffic accidents across the United States. This appendix provides an overview of the dataset structure and key variables.

References

Abdolahzadeh Fayegh, F., & Esmailzade, Kh. (2013). An analysis of the role of human factors influential in the occurrence of road accidents from the point of view of the studied drivers: Sardasht-Piranshahr axis in West Azarbaijan province. Relief and Rescue Scientific Research Quarterly, 6(2), 1–12.

Ahmed, S., Hossain, A., & Bhuiyan, M. I. (2021). A comparative study of machine learning algorithms to predict road accident severity. In 20th International Conference on Ubiquitous Computing and Communications (pp. 1–6). IEEE. https://doi.org/10.1109/IUCC-CIT-DSCI-SmartCNS55181.2021.00069

Babagoli, R., Ameli, A. R., Aliasghar, G. T., & Paydar, A. (2018). Presenting a model for predicting the severity of vehicle accidents using accident data (a case study of the Babol-Ganj Afrooz axis). Transportation Research Quarterly, 16(4), 1–14.

Bahiru, T. K., Singh, D. K., & Tessfaw, E. A. (2018). Comparative study on data mining classification algorithms for predicting road traffic accident severity. In 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT) (pp. 1655–1660). IEEE. https://doi.org/10.1109/ICICCT.2018.8473265

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

Breiman, L., Friedman, J. H., Olshen, R., & Stone, C. (1983). Classification and regression trees. Wadsworth. https://doi.org/10.1201/9781315139470

Chen, W.-H., & Jovanis, P. P. (2000). Method for identifying factors contributing to driver-injury severity in traffic crashes. Transportation Research Record: Journal of the Transportation Research Board, 1717(1), 1–9. https://doi.org/10.3141/1717-01

Elyassami, S., Hamid, Y., & Habuza, T. (2021). Road crashes analysis and prediction using gradient boosted and random forest trees. In 2020 6th IEEE Congress on Information Science and Technology (CiSt) (pp. 1–6). IEEE. https://doi.org/10.1109/CiSt49399.2021.9357298

Geyik, B., & Kara, M. (2020). Severity prediction with machine learning methods. In 2020 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA) (pp. 1–6). IEEE. http://dx.doi.org/10.1109/HORA49412.2020.9152601

Gissane, W. (1965). Accidents—A modern epidemic. Journal of the Institute of Health Education, 3(1), 16–18.

Ho, T. K. (1995). Random decision forests. In Proceedings of 3rd International Conference on Document Analysis and Recognition (pp. 278–282). IEEE. https://doi.org/10.1109/ICDAR.1995.598994

Jamal, A., Zahid, M., Rahman, M. T., Al-Ahmadi, H. M., Almoshaogeh, M., Farooq, D., & Ahmad, M. (2021). Injury severity prediction of traffic crashes with ensemble machine learning techniques: A comparative study. International Journal of Injury Control and Safety Promotion, 28(4), 1–20. https://doi.org/10.1080/17457300.2021.1928233

Kalantari, A., & Alyan, S. (2022). Analysis of road accidents with an emphasis on the characteristics of the environment and the road in the spatial information system, a case study: Karaj-Kandavan axis. Human Geography Research, 54(2), 1–15. https://doi.org/10.22059/jhgr.2021.314926.1008216

Karami, Sh., & Farajzadeh, M. (2014). Analysis of road accidents and climate approach using geographic information system: Firouzkoh Sari road. Humanities Teacher's Quarterly, 9(1), 1–10.

Karri, S. L., De Silva, L. C., Lai, D. T. C., & Yong, S. Y. (2021). Classification and prediction of driving behaviour at a traffic intersection using SVM and KNN. SN Computer Science, 2(3), Article 192. https://doi.org/10.1007/s42979-021-00587-8

Keymanesh, M. R., & Baradaran Rahmanian, N. (2021). Predicting the severity of road accidents using artificial neural network and comparing it with multivariate analysis method. Journal of Transportation Engineering, 29(106), 1–10. https://doi.org/10.22034/road.2021.118084

Khajesalimi, M., Khabiri, M. M., & Falahnejad, M. S. (2018). Prediction and investigation of road accident factors using support vector machine algorithm. Journal of Civil and Environmental Engineering, 49(3), 35–43.

Klein, T. (2007). Rhode Island. Marshall Cavendish.

Koohi, M., & Shabani, Sh. (2023). Identifying factors affecting the severity of extra-urban accidents using the multinomial logit (MNL) model of a case study of Ilam province. Road Science Quarterly, 20(2), 45–56. https://doi.org/10.22034/road.2022.107900

Kumar, N., Acharya, D., & Lohani, D. (2021). An IoT based vehicle accident detection and classification system using. IEEE Internet of Things Journal. IEEE University of London. https://doi.org/10.1109/JIOT.2020.3008896

Kumari, S., Kumar, D., & Mittal, M. (2021). An ensemble approach for classification and prediction of diabetes mellitus using a soft voting classifier. International Journal of Cognitive Computing in Engineering, 2, 40–46. https://doi.org/10.1016/j.ijcce.2021.01.001

Kushwaha, M., & Abirami, M. S. (2021). Comparative analysis on the prediction of road accident severity using machine learning algorithms. In International Conference on Micro-Electronics and Telecommunication Engineering (pp. 1–10). Springer. http://dx.doi.org/10.1007/978-981-16-8721-1_26

Labib, M. F., Rifat, A. S., Hossain, M. M., Das, A. K., & Nawrine, F. (2019). Road accident analysis and prediction of accident severity by using machine learning in Bangladesh. In 2019 7th International Conference on Smart Computing & Communications (ICSCC) (pp. 1–6). IEEE. http://dx.doi.org/10.1109/ICSCC.2019.8843640

Libnaoa, M., Misulaa, M., Andresa, C., Mariñasa, J., & Fabregas, A. (2023). Traffic incident prediction and classification system using naïve Bayes algorithm. Procedia Computer Science, 227, 316–325. http://dx.doi.org/10.1016/j.procs.2023.10.530

Malika, S., El Sayed, H., Khana, M. A., & Khana, M. J. (2021). Road accident severity prediction—A comparative analysis of machine learning algorithms. In IEEE Global Conference on Artificial Intelligence and Internet of Things (GCAIoT) (pp. 1–6). IEEE. https://doi.org/10.1109/GCAIoT53516.2021.9693055

Manzoor, M., Umer, M., Sadiq, S., Ishaq, A., Ullah, S., Madni, H. A., & Bisongi, C. (2021). RFCNN: Traffic accident severity prediction based on decision level fusion of machine and deep learning model. IEEE Access, 9, 131654–131664. https://doi.org/10.1109/ACCESS.2021.3113055

Mirzaei Khezri, S., & Saghayei, A. (2014). Forecasting the severity of fatal road accidents in Iran using basic and collective classification models. Rahor Quarterly, 12(32), 1–10.

Mousavi Fooladi, S. R. (2011). Investigating road accidents in foothills and mountain axes, a case study: Semnan - Foulad Mahaleh road. Semnan Police Science Quarterly, 2(6), 1–15.

Nemati, S. A., Amin, R., & Khodaei, A. (2023). Analyzing and predicting the severity of road accidents using the binary logit model: A case study of traffic accident data in Canada in 2019. Omran and Project Magazine, 31, 1–10.

Parsa, A. B., Movahedi, A., Taghipour, H., Derrible, S., & Mohammadian, A. (2020). Toward safer highways, application of XGBoost and SHAP for real-time accident detection and feature analysis. Accident Analysis & Prevention, 136, 105405. https://doi.org/10.1016/j.aap.2019.105405

Parsi, Euronews. (2023, April 19). Population growth: Fertility rate declining, population growth—China, India first. Retrieved March 9, 2025, from https://parsi.euronews.com/2023/04/19/population-growth-fertility-rate-declining-population-growth-china-india-first-r

Pourgholami, M. R., Farajzadeh, M., Gandomkar, A., & Habinzade, A. (2016). Analyzing road accidents with a climate approach and providing a model for preventive traffic police intervention; The case study of the roads in the northwest of the country. Police Science Research Quarterly, 19(2), 1–12.

Sarkar, S., Vinay, S., Raj, R., Maiti, J., & Mitra, P. (2019). Application of optimized machine learning techniques for predicting occupational accidents. Computers & Operations Research, 106, 210–224. https://doi.org/10.1016/j.cor.2019.02.011

Senthilnathan, S. (2019). Usefulness of correlation analysis. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.3416918

Shen, X., & Wei, S. (2020). Application of XGBoost for Hazardous Material Road Transport Accident Severity Analysis. IEEE Access, 8, 206806–206819. https://doi.org/10.1109/ACCESS.2020.3037922

Shulski, M., & Wendler, G. (2007). The climate of Alaska. University of Alaska Press.

Tavakoli Kashani, A., Amirifar, S., Madghalchi, A., Mohammadi, A., & Jazonghi, M. (2022). Prediction of the severity of road accidents by machine learning methods - A case study of Zanjan province. In 19th International Conference on Transportation and Traffic Engineering (pp. 1–10).

Umer, M., Sadiq, S., Ishaq, A., Ullah, S., Saher, N., & Madni, H. A. (2020). Comparison analysis of tree-based and ensembled regression algorithms for traffic accident severity prediction. arXiv:2010.14921. https://doi.org/10.48550/arXiv.2010.14921

Vakil Roaya, Y., & Zargar, S. M. (2017). Determining and analyzing the priority of factors affecting the occurrence of road accidents (case study: roads of Semnan province). Journal of Geographical Information System and Remote Sensing in Planning, 9(3), 1–10.

Vatanparast, M., Afshari, A. R., Rezaei Arefi, M., & Noormohamadi, A. M. (2016). Evaluation of the effect of climatic elements and human factors on the occurrence of road accidents using the fuzzy logic of a case-oriented example, Mashhad Qochan. Journal of Geographical Information System and Remote Sensing in Planning, 8(4), 1–15.

World Health Organization. (2019). Global status report on alcohol and health 2018.

Worldometer. (2021). Largest countries in the world by area. Retrieved March 9, 2025, from https://www.worldometers.info

Worldometer. (2021). Population by country (2021). Retrieved March 9, 2025, from https://www.worldometers.info

Zhang, Y., Zhang, H., Cai, J., & Yang, B. (2014). A weighted voting classifier based on differential evolution. Abstract and Applied Analysis, 2014, 1–6. https://doi.org/10.1155/2014/376736

Zhao, Y., & Deng, W. (2022). Prediction in traffic accident duration based on heterogeneous ensemble learning. Applied Artificial Intelligence, 36(1), Article 2018643. https://doi.org/10.1080/08839514.2021.2018643