Proceedings of the 5th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/99/20251788

Machine Learning-based Prediction Analysis of Potential Factors in Traffic Accidents

Jiawei Jiang[1,a,*], Yangyang Miu[2], Dehao Wu[3]

[1]*School of Computer Science, University of Sheffield, Sheffield, United Kingdom*
[2]*School of Information Engineering, Dalian Ocean University, Dalian, China*
[3]*School of Science, Southwest Petroleum University, Chengdu, China*
a. jjiang52@sheffield.ac.uk
**corresponding author*

Abstract: Every year, traffic accidents cause a great deal of death, serious injury, and financial damage, making it a major global problem. Predicting accident severity is crucial for implementing effective traffic management and safety strategies. This study employs Random Forest and Deep Neural Network (DNN) models and so on from machine learning algorithms to predict the traffic accidents severity, identifying related outcomes from a joint dataset collected by the Seattle Department of Transportation as well as UK's Department for Transport. The results showcase that snowy road surfaces, standing water and dusk or insufficient lighting conditions causes respectively an excess of 10.75%, 10.44% and 13.01% above the average number of causalities in traffic accidents. The DNN model achieved the highest accuracy (91.12%), outperforming other methods by 3–10 percentage, especially in severe crash detection performance. In contrast, Random Forest model offers better stability across multiple classification threshold, illustrating better performance on general tasks. By diagnosing the high-risk conditions, traffic authorities can implement tailored interventions, ranging from boosting road maintenance efforts during adverse weather, improving street lighting, and increasing safety measures during peak times, to boost the standard of protection. Future research should investigate the method to optimise the robustness and generalization of DNN model through techniques like threshold adjustment, or adopting ensemble learning mechanisms as well as devising new segments that are more relevant.

Keywords: Traffic Accident Prediction, Deep Neural Networks, Collision Severity.

1. Introduction

Traffic accidents result in numerous fatalities, injuries, and significant economic losses globally every year [1]. The NHTSA's official website reports that approximately 42,514 people died in the United States in 2022 [2]. Beyond human tolls, these incidents also encompass significant environmental effects [2]. Hence, create a prediction model to estimate the severity of traffic collision by investigating various potential factors such as location, weather and visibility. It seeks to predict accident severity with an emphasis on broadening the dimensions of dataset selection to enhance generalization and comprehensive factor consideration.

Early research has demonstrated how traffic flow and weather conditions affect road safety. Theofilatos and Yannis conducted a comprehensive evaluation using statistical and machine learning

DOI: 10.54254/2755-2721/99/20251788

methods, finding that the interaction between traffic flow and weather conditions significantly affects accident risk, especially during peak traffic hours [3]. Moreover, Peng et al. applied this quantitative analysis to the context of visibility. They found that reduced visibility significantly increased accident risk, highlighting the need for targeted safety interventions during such conditions [4].

With the recent rapid advancement of machine learning technologies, methodologies within this field have experienced profound evolution. Mondal et al. employed non-parametric machine learning algorithms, involving Random Forest and Bayesian Additive Regression Trees in their weather-relevant accident prediction model on four years of crash data from Connecticut [5]. Their model demonstrated its superior predictive performance and ability to analyze a wider range of predictors more than the traditional methods. Subsequently, Wang et al. investigated traffic accident hotspots using an extended interest measure and the Apriori algorithm [6]. Although being limited by data, their research was expected to provided further knowledge on the application of machine learning in improving accident predictions and enabled several future work possibilities especially with the comprehensive studies pending that investigate complex traffic environmental conditions. Likewise, Liu applied several learning models, involving decision trees, random forests, and AdaBoost, in scrutinising high-accident zones across three expressways in Guangdong Province [7]. Not solely, Ahmed et al. employed interpretable machine learning algorithms to assess the intensity of road traffic accidents in New Zealand, highlighting influential contributors such as road type and vehicle counts [8].

Although the predictive ability of Liu and Ahmed's models is impressive, the geographical specificity in their dataset chosen may compromise the model's generalizability across diverse settings. This issue was addressed in the study of Yuan's team, where they developed an innovative method called Hetero-ConvLSTM to address this challenge posed by heterogeneous spatial-temporal data in traffic accident prediction [9]. Following this, Kar and Feng developed an intelligent traffic prediction model that integrates multiple factors and significantly improved the accuracy and reliability of traffic flow predictions [10]. Wang further explored this issue in his work, along with corresponding feature selection methods applied in his smart traffic collision prediction model [11].

This paper aims to use the advances in machine learning to improve the predictive accuracy and generalization of traffic collision severity models. In this work, Random Forest, Nearest Neighbours, LightGBM, and several other models are employed in this study to analyse a diverse range of factors contributing to traffic collisions. In this study, we employ Random Forest, Nearest Neighbours, LightGBM, Neural Networks, and several other models to analyze a diverse range of factors contributing to traffic collisions. This comprehensive analysis enhances understanding of how the timing of incidents, locations, weather conditions, and visibility affect the severity and type of accidents. Subsequently, the model training phase utilizes these critical variables to develop optimized predictive models that forecast the likelihood of accidents with greater precision. In summary, this study will not only enrich the academic literature on traffic safety with theoretical and practical knowledge but also provide valuable information for policy makers to promote preventive action strategies that result in reductions of uncertain road accident risks.

2. Methods

2.1. Data source

The traffic accident data utilized in this study are obtained from two representative public datasets: the Seattle Department of Transportation (SDOT) and the Department for Transport, United Kingdom, which was saved in CSV format, and cover a wide array of geographical elements [12-15]. The datasets describe traffic accident cases including date and time, geographical location, type of vehicles involved, severity rating or number of people involved due to the crash as well as

DOI: 10.54254/2755-2721/99/20251788

environmental factors surrounding it so that a plausible comprehensive view can be obtained in various urban and national contexts. This expansive geographic coverage enhances the generalizability and applicability of our findings to varied traffic systems globally. This dataset contains a total of 772,553 records, and eight variables were meticulously chosen on their relevance and potential impact on the frequency and intensity of traffic collisions. These variables are applied in the following data analysis and model training phases. The selected variables and their correspondence in the datasets are shown in Table 1.

Table 1: Dataset Variable Correspondence and Description

SDOT Dataset	UK Dataset	Description
INCDATE	date	The date on which the accident took place.
INCTIME	time	The time on which the accident took place.
SEVERITYCODE	accident_severity	The severity of the accident, 1-slight, 2-serious, 3-fatal
ROADCOND	road_surface_conditions	0- other, 1- dry, 2- wet, 3- snow, 4- ice or frost, 5- standing water, 6- oil or diesel, 7- mud
LIGHTCOND	light_conditions	0- other, 1- daylight, 2- dusk, 3- dawn, 4- darkness/ lights lit, 5- darkness/ lights unlit, 6- darkness/ no lighting, 7- darkness/ lights unknown
WEATHER	weather_conditions	0- other, 1- clear, 2- raining, 3- snowing, 4- fog or mist, 5- high wind, 6- blowing sand/dirt, 7- unknown
JUNCTIONTYPE	junction_detail	0- others, 1- at intersection, 2- mid-block, 3- driveway junction, 4- ramp junction

2.2. Method introduction

2.2.1. Random Forest

The random forest consists of several independently trained decision trees, each trained on a unique subset of the data. The model has 1000 trees, and the prediction results of each tree are:

$$h_1(x), h_2(x), \ldots, h_{1000}(x) \tag{1}$$

The final prediction result of the random forest is the voting result of all trees.

$$\mathrm{Y} = \mathrm{mode}(h_1(x), h_2(x), \ldots, h_{1000}(x)) \tag{2}$$

For each split, the model will use the square root of the total number of features as the maximum number of features. A random number seed is also set to ensure reproducibility of the results.

$$\mathrm{max\_fetures} = \sqrt{\mathrm{p}} \tag{3}$$

2.2.2. Deep Neural Network (DNN)

The DNN utilized in this study incorporates the following components: an input layer, three dense layers, three dropout layers and an output layer.

The input layer accepts normalized data with specific features and then transfers it to the dense layer where they are conducted linear transformations. Three dense layers applied here were assigned

DOI: 10.54254/2755-2721/99/20251788

different numbers of neurons with 256, 128 and 64. The operation in the dense layer can be mathematically demonstrated as:

$$h_t = \text{ReLU}(W_t[h_{t-1}] + b_t) \tag{4}$$

Where h_t denotes the output of layer t , W_t represents the weight matrix, b_t the bias vector, and h_{t-1} the output from the previous layer, and ReLU is the activation function.

To further control model complexity and suppress overfitting, the incorporation of a dropout layer and L2 regularization in each dense layer following activation is considered. The dropout layer function, through random disconnection, reduces dependency on specific neurons, thus bolstering network generalization. L2 regularization, by penalizing the sum of squares of the weights, reduces the magnitude of model weights, enhancing training stability and generalization of the model.

2.2.3. Analytical method

The study calculated the average of several variables and then displayed them in a bar chart. The purpose is to observe the difference in average accident severity under different factors.

For the association between the target variable (SEVERITYCODE) and the other variables, this study used the chi-square test to explore their correlation.

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} \tag{5}$$

O_i is the observed frequency and E_i Expected frequency.

3. Results and discussion

3.1. Factor analysis

Table 2 presents the mean number of individuals and vehicles involved in traffic collisions under 11 different weather conditions. Analysing data, reveals that accidents occurring during snowing conditions involve the highest average number of people and vehicles, 1.421 and 1.857, respectively. This is closely followed by blowing snow, being 1.402 in PERSONCOUNT and 1.822 in VEHICLECOUNT, which may indicate that snow-related weather is associated with a higher severity of accidents. As a suggestion, traffic management authorities should implement targeted interventions on snowing days.

Table 2: Average number of individuals and vehicles involved in different weather conditions

WEATHER	PERSONCOUNT	VEHICLECOUNT
Blowing Sand/Dirt	1.228	1.818
Blowing Snow	1.402	1.822
Clear	1.285	1.852
Fog/Smog/Smoke	1.315	1.728
Other	1.143	1.733
Overcast	1.385	1.651
Partly Cloudy	1.297	1.781
Raining	1.304	1.790

DOI: 10.54254/2755-2721/99/20251788

Table 2: (continued).

Severe Crosswind	1.311	1.783
Sleet/Hail/Freezing Rain	1.335	1.701
Snowing	1.421	1.857

Similarly, based on the data in Table 3, this research finds that understanding water conditions, the average number of people (1.417) involved in traffic accidents is the highest. Additionally, under snow/slush and wet road conditions, both the mean number of individuals and vehicles involved are relatively high, the former is 1.336 and 1.807, and the latter is 1.318 and 1.802, indicating that slippery road surfaces exacerbate the severity of accidents.

Table 3: Average number of individuals and vehicles involved in different road conditions

ROADCOND	PERSONCOUNT	VEHICLECOUNT
Dry	1.169	1.784
Ice	1.287	1.653
Other	1.085	1.732
Sand/Mud/Dirt	1.275	1.535
Snow/Slush	1.336	1.807
Standing Water	1.417	1.668
Wet	1.318	1.802

As Table 4 illustrates, during dusk, the average number of people (1.558) and vehicles (1.851) involved in traffic accidents is the highest. Under dark conditions with no street lights, both the number of individuals (1.473) and vehicles (1.620) involved are also relatively high. This suggests that insufficient lighting increases the severity of accidents, likely due to reduced visibility. In contrast, during daylight, there are the fewest people and cars.

Table 4: The mean quantity of individuals and vehicles involved in various lighting scenarios

LIGHTCOND	PERSONCOUNT	VEHICLECOUNT
Dark - No Street Lights	1.473	1.620
Dark - Street Lights Off	1.389	1.773
Dark - Street Lights On	1.280	1.813
Dark - Unknown Lighting	1.198	1.773
Dawn	1.343	1.765
Daylight	1.272	1.462
Dusk	1.558	1.851
Other	1.207	1.623
Dark - No Street Lights	1.473	1.620

DOI: 10.54254/2755-2721/99/20251788

Figure 1 illustrates the distribution of traffic accident frequencies across different months, weekdays, and hours. The data reveal that the most hazardous period occurs from October to November, with a notable increase in incidents, peaking at approximately 65,000 to 70,000 accidents. Also, accident frequencies are highest on Saturdays, reaching around 120,000, while Mondays have the fewest accidents, with approximately 80,000 incidents. In terms of time of day, the most dangerous hours are between 15:00 and 17:00, with accident counts peaking at 55,000 to 60,000. These findings indicate that traffic risks are concentrated in specific temporal patterns, particularly during late afternoons, weekends, and the autumn months.

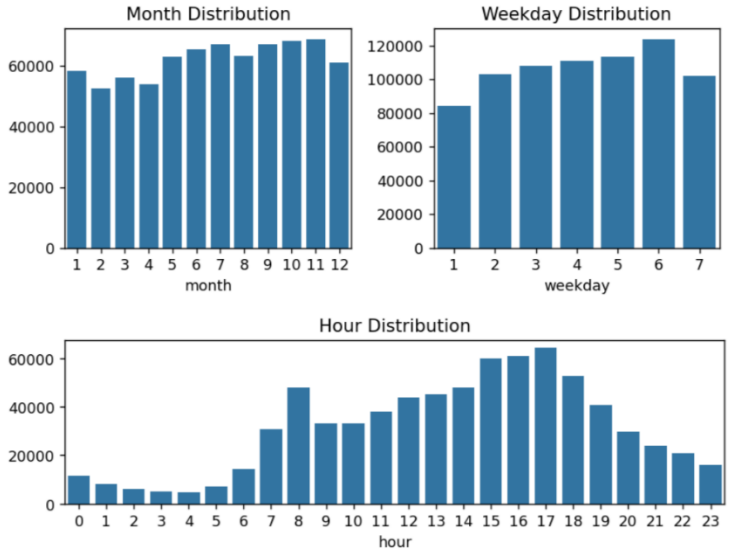

Figure 1: Monthly distribution

3.2. Correlation analysis

To examine the relationship between the categorical variables and the target variable (SEVERITYCODE) in the dataset, the Chi-square test was deployed in this study. Based on the results presented in Table 5, the p-values for all variables are beneath than 0.01, suggesting that the association between each variable and the severity of the accident is statistically notable. Among these variables, JUNCTIONTYPE (type of junction) has the highest Chi-square value, signifying that specific junction types have a significant impact on the severity of accidents. This may be due to the varying complexity of vehicle interactions and the associated risk levels at different types of junctions.

Table 5: Chi-square test results between target variable and variables

Variable	Chi-squared Value	p-value
ROADCOND	258.388	< 0.01
LIGHTCOND	555.695	< 0.01
WEATHER	478.630	< 0.01
JUNCTION TYPE	10925.157	< 0.01

DOI: 10.54254/2755-2721/99/20251788

3.3. Prediction model effect analysis

3.3.1. Result of Common Model

For comparison, four machine learning algorithms: Logistic Regression, K-Nearest Neighbors, Decision Tree, and Random Forest, are employed in this study as well. The Figure 2 displays the accuracy, precision, recall, and F1 score for the prediction models of these algorithms after training on the test set. The Random Forest model performs best with scores of 0.878, 0.877, 0.878, and 0.878 for accuracy, precision, recall, and F1 score, respectively. Figure 2 shows a comparative analysis of the above four machine learning algorithms, evaluating their performance based on accuracy, precision, recall, and F1 score. The Random Forest model outperforms the others, achieving consistently high metrics with values of 0.878 across all four measures. Conversely, K-Nearest Neighbors has the worst precision and recall when tested against each other. This may partially be due to the fact that, as it falls into an algorithm relying on number of neighbors chosen and has proven its weakness in capturing complex decision boundaries mainly with high-dimensional data or a specific imbalance classification.

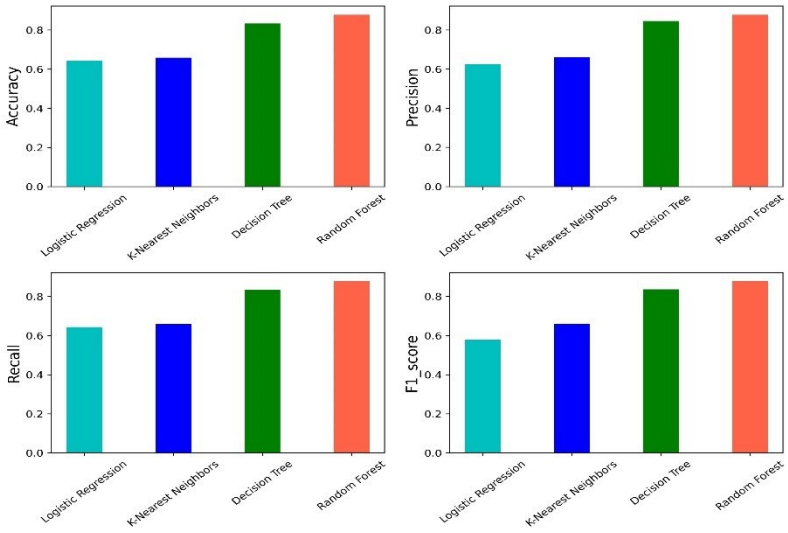

Figure 2: Model effect.

3.3.2. DNN

This DNN model was trained for 30 epochs and early stopping implemented to avoid overfitting. Table 6 summarizes its performance metrics sets compared with the result of Random Forest.

Table 6: Model performance of Random Forest and DNN

Score	Random Forest	DNN
Accuracy	0.8782	0.9112
Precision	0.8778	0.9131
Recall	0.8782	0.9550
F1 score	0.8780	0.9336
AUC	0.8520	0.7590

DOI: 10.54254/2755-2721/99/20251788

By observing these results, it is clear that the DNN generally outperforms Random Forest model in all this metrics (i.e. positive classification ability). A lower Area Under the Curve (AUC) indicates this relief is in exchange for decreased discrimination between positive and negative samples. Given the context of this study, it indicates that the DNN can correctly categorize most severe crashes (an ability desired by traffic management authorities in order to avert high-risk incidents). Moreover, the higher recall rate translates to fewer life-threatening accidents missed. However, the lower AUC compared with Random Forest indicates a higher variability of performance across different thresholds and thus better stability of probability estimation by the DNN. This result suggests that the DNN model has potential applications in scenarios that require focused attention on severe accidents, such as accident warning systems.

4. Conclusion

This study successfully developed predictive models for traffic accident severity by utilizing comprehensive datasets from diverse geographic regions and incorporating multiple influential factors. The analysis demonstrated that environmental conditions such as weather, road surface, and lighting significantly impact accident severity. Factor analysis and chi-square tests confirmed strong correlations between these variables and the severity of accidents.

The analysis revealed that a combination of adverse weather conditions (particularly rain), wet or slippery road surfaces, poor lighting conditions (such as dusk and darkness without street lighting), and certain temporal periods (namely October to November, Saturdays, and between 15:00 to 17:00 hours) is associated with a higher average number of people involved in accidents. These factors collectively contribute to increased accident severity due to reduced visibility, decreased vehicle control, and higher traffic volumes.

For the most performance measures tested (accuracy, precision, recall and F1 score), the DNN model performed better than other machine learning algorithms developed such as Random Forest. This demonstrates a powerful capability of the DNN to capture more complex nonlinear correlations among multiple predictor variables and accident severity. However, the DNN exhibited a lower AUC value compared to the Random Forest model, indicating reduced stability across different classification thresholds. This limitation implies that while the DNN is highly effective at predicting severe accidents, it may be sensitive to threshold settings, potentially affecting its performance in varying operational contexts.

These findings highlight the effectiveness of machine learning methods to improve traffic accident severity prediction models. By accurately identifying the factors that contribute to higher accident severity, traffic authorities would be able to implement targeted interventions at a micro level. For instance, a better quality of road during adverse weather conditions or street illumination in the areas prone to insufficient visibility can lower chances at both high frequency and severity with which accidents occur. Future research should focus on optimizing the DNN model to improve its AUC value and robustness across different thresholds. This could involve techniques such as threshold adjustment, ensemble modelling, and incorporating additional relevant features to enhance model stability and generalization.

Authors Contribution

All the authors contributed equally and their names were listed in alphabetical order.

References

[1] *Flahaut, B. (2004). Impact of infrastructure and local environment on road unsafety. Accident Analysis & Prevention, 36(6), 1055–1066.*

DOI: 10.54254/2755-2721/99/20251788

[2] National Highway Traffic Safety Administration. (2022). Traffic fatalities in 2022. NHTSA. Retrieved September 6, 2024, from https://www.nhtsa.gov/press-releases/2022-traffic-deaths-2023-early-estimates
[3] Theofilatos, A., & Yannis, G. (2014). A review of the effect of traffic and weather characteristics on road safety. Accident Analysis & Prevention, 72, 244–256.
[4] Peng, Y., Abdel-Aty, M., Shi, Q., & Yu, R. (2017). Assessing the impact of reduced visibility on traffic crash risk using microscopic data and surrogate safety measures. Transportation Research Part C: Emerging Technologies, 74, 295–305.
[5] Mondal, A. R., Bhuiyan, M. A. E., & Yang, F. (2020). Advancement of weather-related crash prediction model using nonparametric machine learning algorithms. SN Applied Sciences, 2, Article 1372.
[6] Wang, Y., Sun, Y., & Wang, L. (2021). Analysis of causes of traffic accident hotspots based on improved interest measure and Apriori algorithm. Journal of Zhejiang University: Science Edition, 48(3), 7.
[7] Liu, M. (2022). Research on influence factor analysis and prediction of freeway traffic accidents based on machine learning [Master's thesis, Beijing Jiaotong University]. Beijing Jiaotong University Repository.
[8] Raifman, M. A., & Choma, E. F. (2022). Disparities in activity and traffic fatalities by race/ethnicity. American Journal of Preventive Medicine, 63(2), 160–167.
[9] Yang, Z., Zhang, X., & Yu, T. (2018). Hetero-ConvLSTM: A deep learning approach to traffic accident prediction on heterogeneous spatio-temporal data. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (pp. 984–992). Association for Computing Machinery.
[10] Kar, P., & Feng, S. (2023). Intelligent traffic prediction by combining weather and road traffic condition information: A deep learning-based approach. International Journal of Intelligent Transportation Systems Research, 21, 506–522.
[11] Wang, X. (2024). Analysis of causes and design of predictive models for highway traffic accidents based on machine learning [Doctoral dissertation, Linyi University].
[12] Palay, C. (2023). SDOT collisions all years [Data set]. City of Seattle ArcGIS Online. Retrieved September 6, 2024, from https://data-seattlecitygis.opendata.arcgis.com/maps/SeattleCityGIS::sdot-collisions-all-years-2
[13] Department for Transport. (2023). Road safety data – collisions 2022 [Data set]. Retrieved September 6, 2024, from https://data.dft.gov.uk/road-accidents-safety-data/dft-road-casualty-statistics-collision-2022.csv
[14] Department for Transport. (2022). Road safety data – collisions 2021 [Data set]. Retrieved September 6, 2024, from https://data.dft.gov.uk/road-accidents-safety-data/dft-road-casualty-statistics-collision-2021.csv
[15] Department for Transport. (2021). Road safety data – collisions last 5 years [Data set]. Retrieved September 6, 2024, from https://data.dft.gov.uk/road-accidents-safety-data/dft-road-casualty-statistics-collision-last-5-years.csv