Dean&Francis
ISSN 2959-6157

Understanding Feature Contributions for Traffic Accident Severity Prediction

Wenxin Yan*

School of Computing and Data Science, Xiamen University Malaysia, Sepang, Malaysia
Corresponding author:
swe2309542@xmu.edu.my

Abstract:

Concern around global traffic accidents have increased due to the large toll they take on people and the economy. To investigate factors affecting the result of vehicle accidents in the UK, this study attempts to predict the severity of accidents using a Random Forest model. It investigates difference feature sets including time, context and interaction. Following data preprocessing and class balancing using the Synthetic Minority Oversampling Technique (SMOTE), five feature sets (baseline, temporal, context, interaction and extended) were trained and assessed over 10 iterations. The results show that overall accuracy remains around 0.78 and AUC around 0.60 across all models, indicating stable performance. However, the inclusion of contextual and interaction features slightly improves recall for fatal and serious crashes and reduces the dominance of a few strong predictors, leading to a more balanced feature contribution. These results suggest that richer feature design enhances Random Forest's ability to capture complex crash patterns beyond the majority of slight accidents.

Keywords: Traffic accident severity, Random forest, Feature contribution, UK road accident dataset.

1. Introduction

In recent years, the explosive growth of modern society has led to a sharp surge in the overall number of motorcars on roads. In addition, the overall number of road accidents is increasing, resulting in huge economic and human losses [1]. The World Health Organization (WHO) states that an estimated 1.19 million people are killed each year through vehicle collisions, and tens of millions more are severely injured [2]. Chen et al. indicated that road injuries will cost the global economy $1.8 trillion between 2015-2030. Costs of traffic accidents during this period equate to a 0.12% per annum tax on world gross domestic product [3]. To address this growing challenge, this study investigates how different types of engineered features contribute to predicting traffic accident severity using a Random Forest model.

The forecasting of traffic accident severity has become a major study area in transportation safety analysis. Early studies in the field primarily utilized traditional statistical models to analyze the correlation between the severity of the accident and the contributing factors. These approaches provided

ISSN 2959-6157

interpretable and understandable coefficients that allowed analysts to infer causal relationships, but they employed limiting assumptions in most cases and so were less capable of handling high-dimensional or nonlinear relationships present in complex datasets. In comparison, machine learning (ML) models are proficient at detecting complex interactions among variables, providing higher accuracy of predictions [4]. Among all ML algorithms, Random Forest (RF) has been proposed for traffic accident severity prediction because it excels in robustness, nonlinear relationship handling, and tolerance to noisy or missing data [5]. RF can process different feature types and also learn the importance of each feature. That makes it a suitable algorithm for real-world transport data sets [6]. Various studies have proven its predictability. For example, Adefabi et al. used RF-based importance analysis for the identification of the most important causes of crashes, such as weather and road conditions, and achieved a precision of over 80% [7]. Kaur et al. used RF on highway crash data and found RF to be more stable and interpretable than other ensemble techniques [8]. These results show that Random Forest finds a trade-off between performance, interpretability, and applicability in practical uses to represent the severity of accidents.

Previous research has looked at some of the ways in which the problems of machine learning performance in accident severity forecasting can be improved. Other problems, aside from algorithmic selection, feature engineering and data representation have been understood to be one of the major issues in improving predictive performance and robustness. For example, Alotaibi et al. utilized explainable artificial intelligence (XAI) methods to determine feature importance and found that road geometry, weather, and lighting were among the most significant predictors in some of the classifiers [9]. Similarly, Rifat and Huq demonstrated how incorporating contextual variables such as accident location, vehicle, and time of occurrence improved the prediction of deaths [10]. Even with these advances, feature selection and combinations' performance continues to be study-dependent, actually depends on the nature of the dataset, local variations, and preprocessing techniques. Such differences imply that feature impacts may vary across models or conditions, and therefore there is still more empirical research in some modeling frameworks worthwhile.

Hence, feature-based investigation of the Random Forest model for predicting traffic accident severity is explored in this study, using the UK Road Accident data set that is open to the public to look at how different groups of features, like environmental and road-related variables, contribute to explaining traffic accident severity. The results should help traffic safety planners and analysts by showing them how to prioritize data collection and improve models' performance in smart transportation systems.

2. Methodology

2.1 Data Source and Description

The study uses data from the United Kingdom (UK) Road Accident dataset [11], which is published by the Department for Transport (DfT). The dataset contains police-reported road traffic accidents across Great Britain from 2005 to 2021. which has over 1.5 million accident records in total. Each record describes one accident with details about its location, environmental and road conditions, number of vehicles involved, and number of casualties.

The variable Accident Severity is divided into three categories: fatal, serious, and slight. In the original UK dataset, the distribution is highly imbalanced, consisting of 19,441 fatal cases (1.3%), 204,504 serious cases (13.6%), and 1,280,205 slight cases (85.1%). To reduce computational load while preserving the original proportion, a stratified under-sampling strategy was applied, producing a working subset of approximately 100,000 records. During model training, the Synthetic Minority Oversampling Technique (SMOTE) was applied only to the training portion to oversample the minority fatal and serious categories, ensuring balanced learning without altering the test data distribution.

Some attributes are removed before modeling. Identifiers, post-event fields such as police attendance, and variables with too many missing values are excluded. Continuous variables are filled with their median values, and categorical variables are filled with their mode. This method helps to reduce the influence of extreme values while keeping the overall data distribution consistent.

2.2 Indicator Selection

Feature selection is based on data completeness, relevance, and logical connection to accident conditions. Out of the original 33 variables, 11 are kept for model development, shown in the following Table 1.

Table 1. Description of Variables Used for Random Forest Model Training

Variable	Type	Description
1st_Road_Class	Categorical	5 categories (Motorway, A, B, C, Unclassified).
Road_Type	Categorical	6 categories (Single carriageway, Dual, Roundabout, One-way, Slip road, Other).
Speed_limit	Continuous	20–70 mph; posted speed limit at the accident site.
Junction_Control	Categorical	5 categories (Authorized person, Auto signal, Stop sign, Give way, None).
Road_Surface_Conditions	Categorical	6 categories (Dry, Wet, Snow, Ice, Flood, Other).
Light_Conditions	Categorical	5 categories (Daylight, Darkness–lit, Darkness–unlit, Twilight, Unknown).
Weather_Conditions	Categorical	9 categories (Fine, Raining, Snowing, Fog, Windy, Other).
Urban_or_Rural_Area	Categorical	2 categories (Urban, Rural).
Date	Temporal	Used to extract season.
Time	Temporal	Used to group hours into time periods.
Day_of_Week	Categorical	7 categories (Monday–Sunday).

Categorical features are encoded into numeric values using label encoding. Continuous features remain unscaled because Random Forest is not affected by feature scale. Only pre-accident conditions are used to make the model reflect real situational factors.

2.3 Modeling Approach

A Random Forest (RF) model is used as the baseline classifier. It is chosen because it can handle nonlinear relationships and noisy data. Each tree in the forest is trained with a random subset of samples and features, which helps avoid overfitting and improves generalization.

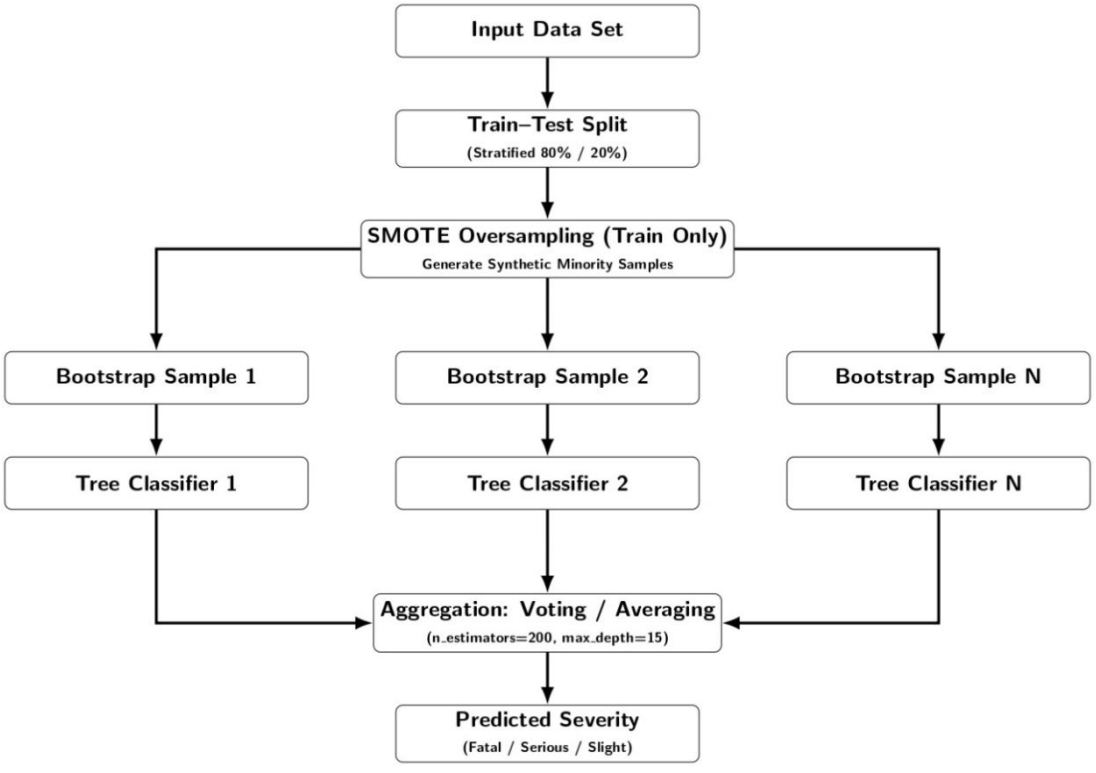

Fig. 1 Structure of the Random Forest model (Picture credit: Original)

As shown in Figure 1, the model uses 200 trees and a maximum depth of 15. The minimum number of samples for splitting and leaf nodes is also limited to control complexity. The class weight is set to "balanced" so that all

ISSN 2959-6157

severity levels have equal importance during training.

Because the dataset is highly unbalanced-slight accidents account for about 85% of all cases-the Synthetic Minority Oversampling Technique (SMOTE) was applied only to the training set. SMOTE creates new synthetic samples for the rare fatal and serious classes based on their nearest neighbors. This method balances the training data and allows the model to learn patterns from all severity levels without changing the original test data.

Model performance is tested through 10 independent runs. In each run, the data is split into 80% for training and 20% for testing. Accuracy, precision, recall, F1-score, and AUC are calculated for every run, and the mean and standard deviation are reported. Feature importance is analyzed using both Mean Decrease in Impurity (MDI) and permutation importance to identify which factors contribute most to prediction accuracy.

3. Result and Discussion

3.1 Baseline Model Results

All experiments were trained on SMOTE-balanced data to ensure fair learning among different severity levels. Each model was repeated ten times using different random seeds, and the average performance was recorded.

Table 2. Baseline Random Forest performance averaged across ten runs

Metric	Mean	Std
Accuracy	0.780	0.002
Precision	0.762	0.001
Recall	0.780	0.002
F1-score	0.767	0.001
AUC	0.604	0.001

As shown in Table 2, the baseline Random Forest achieved an average accuracy of 0.780 and F1-score of 0.767, showing stable results across runs.

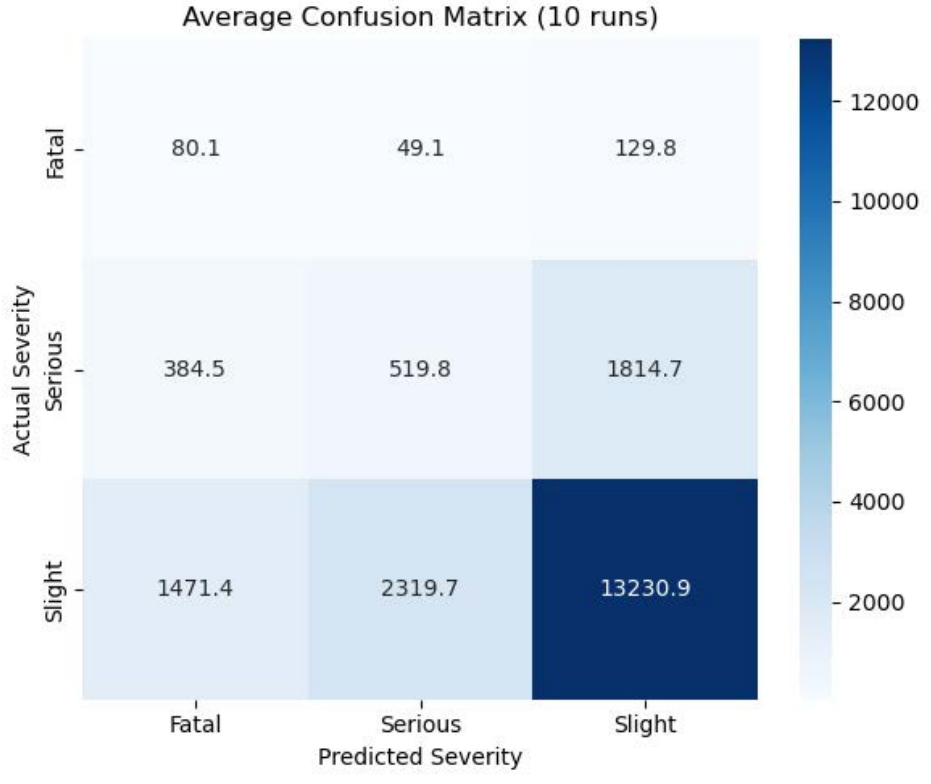

Fig. 2 Average confusion matrix based on test set (Picture credit: Original)

As illustrated in Figure 2, most slight accidents were correctly predicted, while many serious and fatal ones were classified as slight, a common limitation in imbalanced crash data.

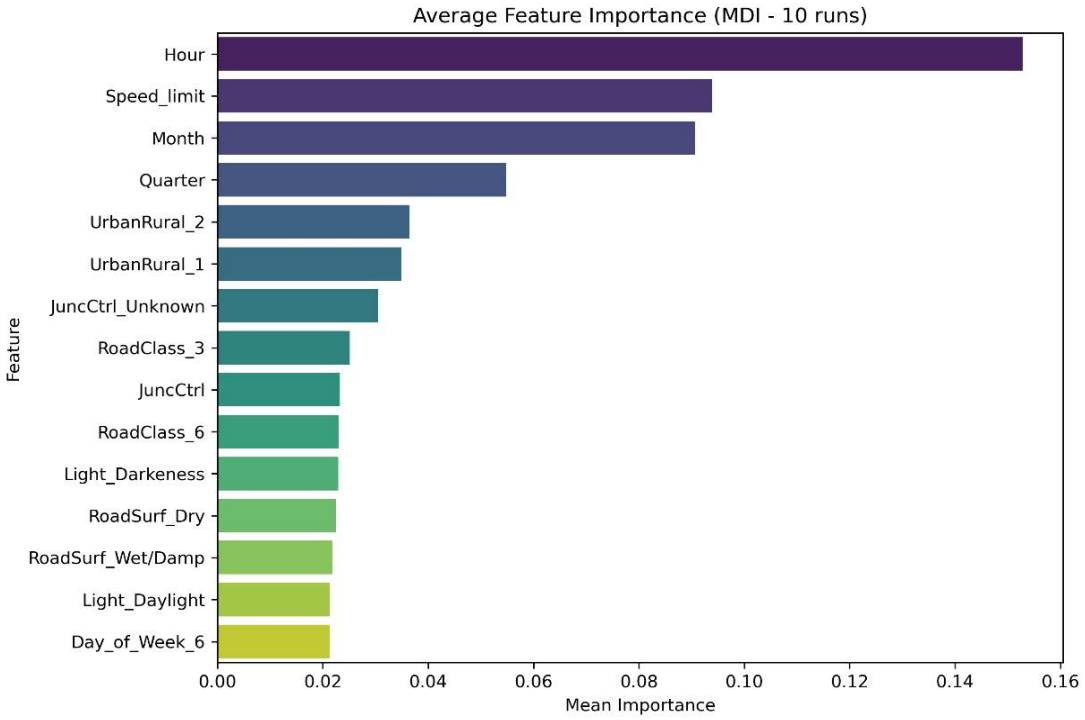

Fig. 3 Fifteen Most Important Features Based on Mean Decrease in Impurity (Picture Credit: Original)

The main predictors in Figure 3 - Hour, Speed Limit, and Month - indicate that both time and road conditions have a strong effect on the severity of the accident, and that other contextual elements like Urban or Rural Area also have a significant impact.

Overall, the baseline model offers significant value as a benchmark, overwhelmingly governed by time and speed factors.

3.2 Comparative Feature Engineering Analysis

To investigate how additional features influence model learning, five datasets were tested under identical parameters: baseline, temporal, context, interaction, and extended. Each represents a different dimension of crash risk.

The temporal features (HourGroup, RushHour, Season, and Weekend) capture when accidents occur, the context features (Speed_Category, Urban_or_Rural_Area) describe where they occur, and the interaction features (e.g., Speed_Urban, Weather × Light) represent how multiple risks combine. The extended set merges all of the features in the above datasets.

Table 3. Comparison of all feature-engineered models

Feature Set	Accuracy	F1 (Weighted)	AUC	Top 10 MDI Cumulative (%)
Baseline	0.780 ± 0.002	0.767 ± 0.001	0.604 ± 0.001	50.3
Temporal	0.790 ± 0.001	0.772 ± 0.001	0.603 ± 0.001	43.9
Context	0.773 ± 0.002	0.765 ± 0.001	0.605 ± 0.001	47.1
Interaction	0.775 ± 0.001	0.766 ± 0.001	0.605 ± 0.001	47.8
Extended	0.784 ± 0.001	0.770 ± 0.001	0.605 ± 0.001	35.9

ISSN 2959-6157

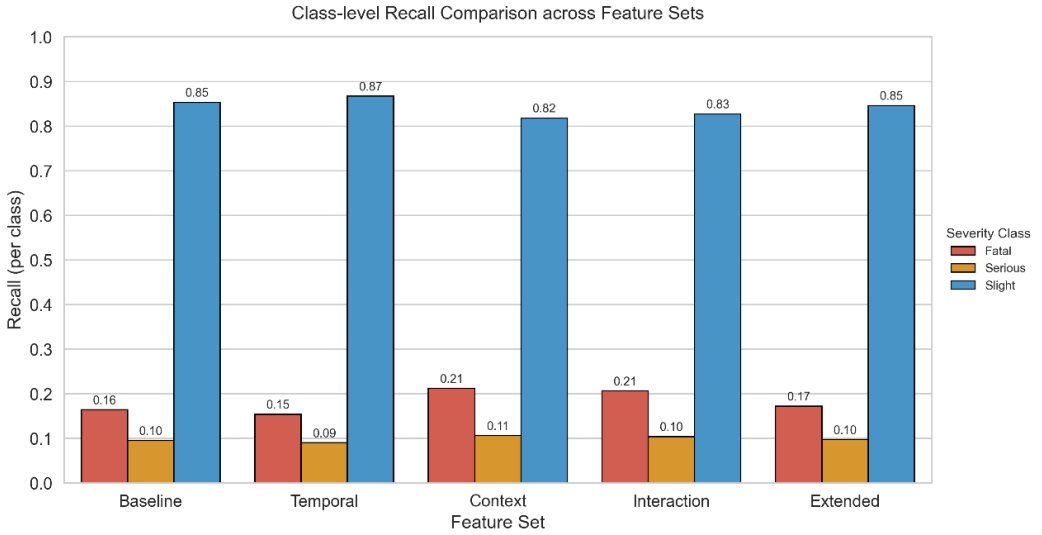

Fig. 4 Comparison of class-level recall for each feature set (Picture credit: Original)

As shown in Table 3, overall accuracy and AUC remain stable at around 0.78 and 0.60, but the cumulative importance of the top ten features dropped from 50.3% in the baseline to 35.9% in the extended model. This means that the expanded models rely on a broader mix of variables rather than a few dominant ones. Figure 4 shows class-level recall, where the baseline performs best on the Slight class but worse on Fatal and Serious. After adding contextual and interaction features, recall for these minority classes improves slightly, indicating that road environment and combined factors provide stronger clues for severe crashes.

Together, the decline in MDI concentration and the change in recall patterns show that the baseline model is mainly sensitive to frequent slight accidents, while additional contextual and interaction variables help uncover the more complex mechanisms behind severe ones.

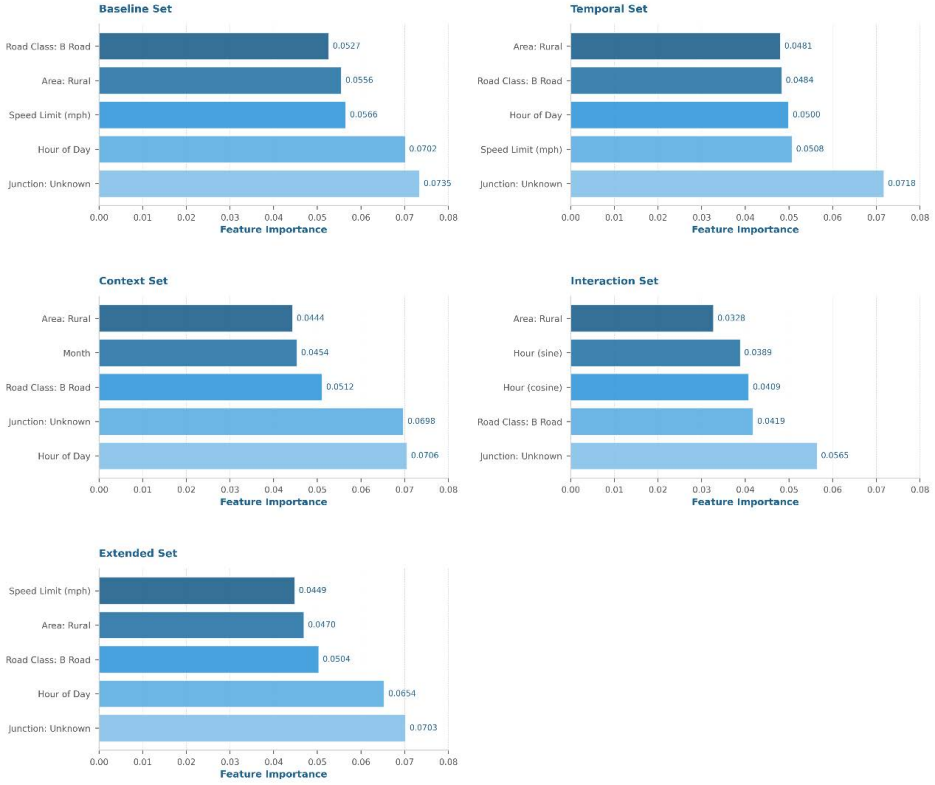

Fig. 5 Comparison of the top 5 feature importance for each feature set (Picture credit: Original)

The feature rankings across the different feature sets shown in Figure 5 reveal several clear and stable patterns. Even though each feature set includes different types of information, the same key variables appear at the top across all of them. These main predictors are junction control, hour of day, speed limit, road class, and urban–rural area. The basic structure and timing of the crash environment feature individual and stable significance even with the addition of some temporal, contextual, and interaction features.

The strongest of all predictors are the junction-related ones. Junction_Control: Unknown is in the top position of every feature set. This suggests that crashes that occur at junctions without any clear control or recorded control are more likely to be severe. Time-of-day effects are also majorly important. Hour is in the top ranks of every feature set, while its cyclic forms Hoursin and Hourcos are included in important features of the interaction set. Road and location attributes like Speed_limit, 1st_Road_Class: B Road, and Urban_or_Rural_Area: Rural also feature quite often. These are consistent with well-established findings in traffic safety: more severe crashes are more common with higher speeds, rural roads, and less controlled junctions. Although interaction features were added to capture combined effects, most of them did not replace the core structural and temporal variables in importance. This shows that the basic physical and operational conditions at the crash site remain the strongest signals in the dataset.

The low standard deviations of accuracy and F1 (below 0.002) show that the results are stable across all experiments. Important features such as Hour and Speed_limit stay at the top in every run, and some engineered features—such as Speed_Urban and the cyclic hour variables—appear in several runs as well. This shows that the new features are actually useful to the random forest and the model can consistently learn to these relationships.

Accident severity is the result of interaction of different factors acting simultaneously. Temporal features answer the question of when are the chances of severe crash more. Contextual features indicate the places where severe crashes are more likely to occur and interaction features describe the different combination of risks that are present. Management of traffic during peak hours and the other findings of the study such as installing better lights at uncontrolled intersections and reduced speed limits on rural roads can be implemented.

4. Conclusion

This study analyzed how various categories of features contribute to the prediction of traffic accident severity using a Random Forest model and the UK Road Accident dataset. The baseline model presented a stable accuracy and F1-scores around 0.78, with results dominated by a few strong predictors such as Hour, Speed_limit, and Junction_Control. After introducing temporal, contextual, interaction, and extended feature sets, overall predictive accuracy remained consistent, but the distribution of feature importance became more balanced, and recall for severe and fatal classes improved slightly. These changes show that richer feature representations enhance the model's ability to capture the complex interplay between temporal, environmental, and road factors rather than relying heavily on single variables.

The analysis reveals that accident severity is influenced by multiple dimensions of risk. Temporal factors determine when crashes occur, speed and road class describe the physical setting, and environmental–structural interactions, such as poor lighting on rural roads or high speeds during peak hours, magnify the likelihood of severe outcomes.

From a practical perspective, this study suggests that transportation agencies can take data centered a wide range of activities from applying stricter regulations on speeding in rural areas with NSP, better lit with areas sight and better management of traffic during peak hours. To address other high risks of rural driving. This study also provided integrating features from the real world into driving sim and machine learning accident predictive pipeline.

Despite these contributions, the scope of this study is still limited in several aspects. First, the current research is based on structured variables involving road, time, and environment, but other important factors such as driver characteristics, vehicle type, and detailed geographic information geographical information are excluded because of the lack of data. These missing factors may also influence accident severity. Second, the current research is limited to a single machine learning algorithm. Therefore, direct comparisons involving other advanced models are not feasible. Moreover, the current research is based on historical data from the UK, and the data is not updated based on recent developments involving the current traffic situation. These limitations indicate that the current research does not fully capture all dimensions of accident risk.

This work can be expanded to include geo-spatial coordinates, vehicle types, and driver demographics to more closely model the behavioral and other regional diversities in user. As smart transport systems and real-time data collection expand, hybrid, feature-rich models may play an increasingly vital role in safety planning and accident prevention.

ISSN 2959-6157

References

[1] Gebru M. K. Road traffic accident: Human security perspective. International Journal of Peace and Development Studies, 2017, 8(2), 15–24.

[2] World Health Organization. Global status report on road safety 2023. Geneva: World Health Organization, 2023.

[3] Chen S., Kuhn M., Prettner K., et al. The global macroeconomic burden of road injuries: Estimates and projections for 166 countries. The Lancet Planetary Health, 2019, 3(9), e390–e398.

[4] Kodepogu K., Manjeti V. B., Siriki A. B. Machine learning for road accident severity prediction. Mechatronics and Intelligent Transportation Systems, 2023, 2, 211–226.

[5] Obasi I. C., Benson C. Evaluating the effectiveness of machine learning techniques in forecasting the severity of traffic accidents. Heliyon, 2023, 9(8), e18812.

[6] Yan M., Shen Y. Traffic accident severity prediction based on random forest. Sustainability, 2022, 14(3), 1729.

[7] Adefabi A., Olisah S., Obunadike C., et al. Predicting accident severity: An analysis of factors affecting accident severity using random forest model. International Journal on Cybernetics and Informatics, 2023, 12(6), 107–121.

[8] Khanum H., Garg A., Faheem M. I. Accident severity prediction modeling for road safety using random forest algorithm: An analysis of Indian highways. F1000Research, 2023, 12, 494.

[9] Alotaibi J. Enhancing traffic accident severity prediction: Feature identification using explainable AI. Vehicles, 2025, 7(2), 38.

[10] Rifat M. A. K., Kabir A., Huq A. S. An explainable machine learning approach to traffic accident fatality prediction. arXiv preprint, 2024.

[11] Ansodariya D. Road accident United Kingdom (UK) dataset. Kaggle, 2020. https://www.kaggle.com/datasets/devansodariya/road-accident-united-kingdom-uk-dataset/data