Road Accident Severity Analysis: Predicting Accident Outcomes and Identifying Hotspots in the United Kingdom Using Machine Learning

Shubhankar Shahade
Department of Computer Science
University of Nottingham
Nottingham, United Kingdom
psxss45@nottingham.ac.uk

Lukshan Sharvaswaran
Department of Computer Science
University of Nottingham
Nottingham, United Kingdom
psyls7@nottingham.ac.uk

Qing Wang
Department of Computer Science
University of Nottingham
Nottingham, United Kingdom
alyqw23@nottingham.ac.uk

***Abstract*- This paper investigates the use of machine learning models to help predict the severity of road traffic accidents and identify popular accident hotspots in the United Kingdom. The dataset has been published by the UK Department of Transport and contains detailed records of personal injury collisions, including environmental conditions, time, location, road type, and severity levels classified as slight, serious, or fatal. Our approach integrates exploratory data analysis (EDA), feature engineering, and supervised machine learning models. The data is preprocessed to handle missing values and encode categorical variables. Some of the challenges involving class imbalance, particularly in underrepresenting fatal incidents are highlighted by the confusion matrix. Models including Logistic Regression, Random Forest, and XGBoost were deployed, along with employing techniques such as SMOTE and class weighting to address class imbalance. For fatal and rare accidents, hyperparameter tuning further improved the performance. Geospatial clustering (DBSCAN) was used to identify accident hotspots with the best performing model being a Tuned LightGBM model with clustering that achieved a macro F1 score of 0.4197 and a fatal recall of 19.3%. The results provide insights into key factors influencing accident outcomes and offer practical recommendations for traffic safety improvements. Future work can incorporate additional data sources and advanced modelling techniques.**

Keywords— Accident severity prediction, Machine Learning, class imbalance, Feature engineering, Geospatial clustering , lightGBM.

I. Introduction

In the UK, road traffic accidents remain a persistent issue with thousands of casualties reported annually. These incidents result not only in the loss of life and serious injuries but also impose significant social and economic burdens on healthcare systems and infrastructure planning. In 2019, 1,752 people died and 25,945 people suffered serious injuries in road traffic incidents in the UK [1]. Generally, the severity of accidents is influenced by factors such as road type, weather conditions, lighting, speed limits, state and/or condition of driver while driving and time of day to name a few. Understanding the severity of accidents—whether slight, serious, or fatal—is critical to informing effective policy decisions, emergency response planning, and preventive road safety measures. While governments and transportation authorities collect extensive data on road accidents, analyzing this information manually to detect patterns or predict outcomes remains a complex challenge. This paper uses the UK Department for Transport's 2023 Road Safety dataset [2], which provides detailed records of personal injury road collisions across Great Britain. The dataset offers an ideal foundation for building classification models to predict accident severity. To understand the structure, quality and limitations of the dataset, an exploratory data analysis is conducted. From the summary statistics conducted on dataset, we see that more than 120000 accidents have been recorded with 30+ original variables. The variables get reduced to 15 upon cleaning as there is a lot of missing data. Columns with more than 95% of rows having missing values have been removed while the rest have been imputed. As observed from Fig. 1 below, the severity distribution is imbalanced with Slight (76%), Serious (22%) and Fatal (1.5%) respectively.

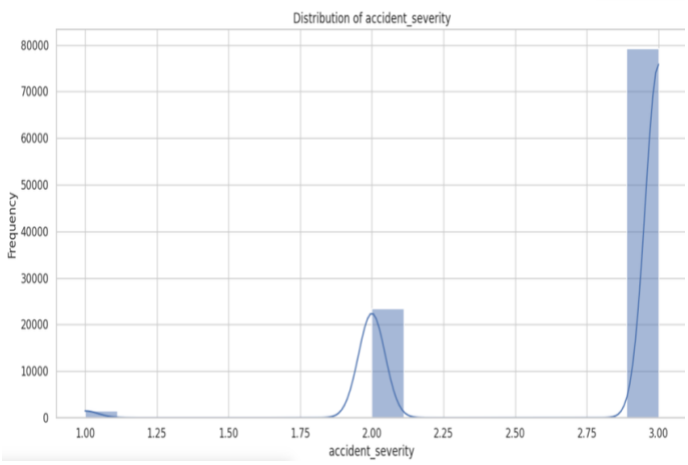

Fig. 1. Distribution of accident severity classes showing class imbalance.

Several other distributions, temporal patterns and correlations are also visualized to get a better idea of the direction we want to head towards. Temporal patterns, such as the day of the week and time of day are known to influence accident likelihood due to variations in traffic volume, driver behavior, and external conditions throughout the week. By visualizing the distribution of accidents across weekdays and weekends through Fig. 2 and Fig. 3 shown

below, we identified noticeable patterns that provide valuable context for feature engineering and risk modelling.

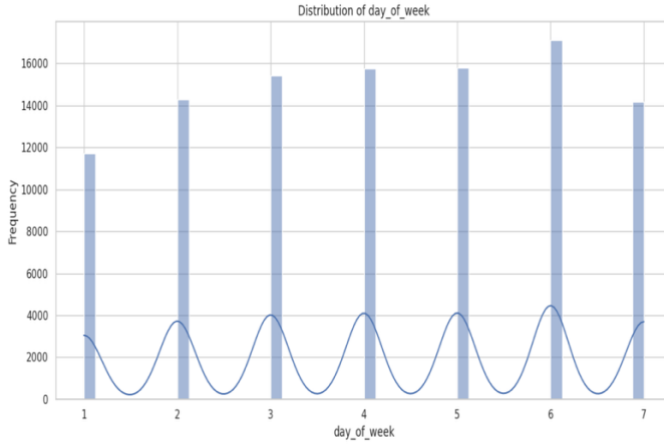

Fig. 2. Distribution of accidents across days of the week.

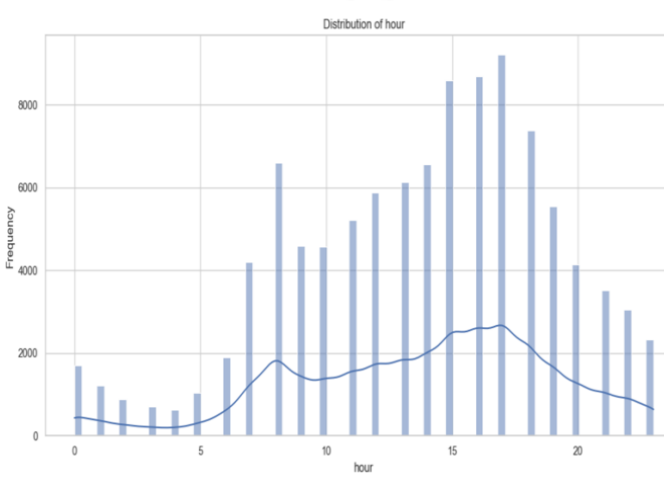

Fig. 3. Distribution of accidents by the hours of the day.

Given the dataset characteristics and real-world relevance, the study is guided by the following research questions:

1. What are the key factors influencing accident severity?
2. Can machine learning models accurately predict accident severity, particularly for rare but critical fatal cases?
3. Can geospatial clustering techniques identify accident hotspots?

Rather than applying models blindly, we take an iterative, data-driven approach — examining data distributions, testing hypotheses, and adapting our strategy in response to results. Each decision we have made, is grounded in reasoning, domain relevance, and a desire to improve real-world safety outcomes. This foundation supports both the development of predictive models and the exploration of spatial patterns related to accident severity.

II. Literature Review

Over the recent past, Machine Learning models have become of increasing interest in helping predict road accident severity and in helping build prediction models that can help reduce road accidents. Often such studies involve using large and complex datasets.

The study performed by Ahmed et al. (2021) [3] was a comprehensive one involving several ML algorithms such as including Logistic Regression, K-Nearest Neighbors (KNN), Naïve Bayes (NB), Random Forest (RF), Extreme Gradient Boosting (XGBoost), and Adaptive Boosting (AdaBoost). Using crash data from New Zealand, the authors applied both binary and multiclass classification frameworks. Their results showed that ensemble models, particularly Random Forest, significantly outperformed single-mode classifiers, achieving 86.64% accuracy in binary classification and 67.67% in multiclass classification. One of the key factors in their model having high prediction accuracy was that their dataset included human factors like driver age and alcohol use and variables such as road conditions and environmental features.

For a comprehensive review of class imbalance techniques, we referred the works of Fernández et al. (2018) [4]. Through their work, they demonstrated that oversampling methods like SMOTE significantly improve model performance in datasets where rare events, such as fatal accidents, are underrepresented. The methodologies chosen in this study have been inspired by their findings in handling class imbalance and improving recall for minority classes.

With time of day being an important factor for crashes, paper published by Daoud et al. (2025) [5] focused specifically on predicting the severity of nighttime crashes using ML models. The research involved the use of seven algorithms, including Random Forest, Artificial Neural Networks (ANN), XGBoost, and Long Short-Term Memory (LSTM) networks. However, they found that Random Forest and ANN models delivered the best performance, with recall rates of approximately 97.6% for severe crashes. the study highlighted how real-time data collection tools could improve future predictive models. The study reiterated the importance of visibility and roadway lighting in avoiding crashes and advocated for how real-time data collection tools could improve future predictive models.

Chen et al. (2016) [6] investigated injury severity patterns in rollover crashes using Support Vector Machine models. Using spatial data, this study incorporated geospatial analysis to identify geospatial patterns that are associated with higher injury severity and to enhance predictive modeling so policymakers are provided with better, actionable insights. Similar to this study, ours also implements DBSCAN clustering to identify accident hotspots is an effective way.

Together, these studies inform and validate the methodology of the present research, underscore the effectiveness of ensemble models and lastly, the value of incorporating diverse, context-aware features. However, they also reveal gaps in addressing class imbalance, feature interpretability, and geospatial analysis — areas this project seeks to advance through advanced sampling methods (SMOTE), feature engineering, and spatial clustering while identifying opportunities for future work, such as incorporating human-centric data and further exploring advanced modeling techniques like deep learning and real-time data integration.

III. METHODOLOGY

This project followed a structured, evidence-based data science pipeline, combining best practices and recommendations from the literature reviewed [3],[4], [5] and [6]. Each methodological choice was driven by the dataset characteristics, the prediction objective, and the need for model interpretability and reproducibility. We structure our analysis as follows:

- Understand the structure, quality, and limitations of the dataset
- Clean and transform the data to prepare it for modelling
- Engineer meaningful features capturing temporal, environmental, and spatial patterns
- Build and evaluate multiple predictive models, accounting for class imbalance
- Explore advanced methods (SMOTE, class weighting, tuning) to improve performance
- Analyse geospatial clustering to identify accident hotspots

A. Data Loading And Initial Exploration

We began by loading the dataset and exploring its structure to understand what we were working with. Table I below shows the sample rows from the original dataset before any processing or cleaning. Initial exploration involved checking the shape, data types, and structure, viewing sample rows and summary statistics, and identifying missing values and columns that may not add value. This helped us anticipate potential issues regarding data quality, encoding quirks and irrelevant fields to name a few. We also flagged columns with only one unique value, high cardinality (e.g., IDs or coordinates), and excessive missingness (>95%). Doing this early prevented wasted effort modelling on features that were uninformative or broken, setting the foundation for a clean, trustworthy dataset.

Dataset shape: Rows = 104258, Columns = 37

:	accident_index	accident_year	accident_reference	location_easting_osgr	locatio
0	2023010419171	2023	010419171	525060.0	
1	2023010419183	2023	010419183	535463.0	
2	2023010419189	2023	010419189	508702.0	
3	2023010419191	2023	010419191	520341.0	
4	2023010419192	2023	010419192	527255.0	

Table I. Sample rows from the original dataset

B. Exploratory Data Analysis (Eda)

We performed distribution analysis of numerical and categorical features to understand their behaviour and give insights into their central tendency, spread, skewness, and potential outliers. This was critical for spotting anomalies and understanding how features might affect modelling — for example, whether scaling or transformation was needed. Highly skewed variables might bias the model, and features with outliers could distort learning. By plotting each feature, we assessed whether any preprocessing was required and gain an intuitive feel for how the data behaved. Correlation matrices (Pearson, Spearman, Kendall) were calculated to detect multicollinearity and inform dimensionality reduction or feature elimination strategies.

C. Data Cleaning

To ensure our dataset was reliable and suitable for modelling, we applied several targeted cleaning steps:

- **Location filtering**: Removed 12 rows lacking essential geospatial coordinates.
- **Low-value columns**: Dropped columns with no variability or high-cardinality identifiers.
- **Placeholder handling**: Replaced placeholder values (-1) with NaN.
- **String normalization**: Standardised categorical string fields by trimming and converting to lowercase.
- **Missing value imputation**: Used median imputation for numeric features to maintain robustness against outliers.

These steps standardize and prepare the dataset for feature engineering and modelling, addressing quality concerns identified in our earlier inspection**.**

D. Feature Engineering And Selection

With a clean dataset, we enriched it by deriving new features aimed at capturing patterns linked to accident severity:

- **Temporal features**: Extracted hour and month; created binary indicators for is_night, is_rush_hour, and is_weekend.
- **Environmental interactions**: Created the road_weather_combo feature by merging road_surface_conditions and weather_conditions.

These features were designed to expose nuanced relationships between external conditions and severity, offering the model more meaningful signals to learn from. To support spatial analysis and mapping, we isolated the subset of accidents with valid latitude and longitude data. This ensured we worked with accurate geolocation coordinates when identifying hotspots and regional trends.

Key steps:

- **Filter Incomplete Locations**: We drop any rows with missing latitude or longitude.
- **Verify Severity Distribution**: We confirm that the filtered geospatial dataset preserves the same severity distribution as the original cleaned dataset.

This geospatial subset enables us to conduct clustering and mapping later, laying the foundation for hotspot detection and spatial correlation analysis.

E. Principal Component Analysis (PCA)

PCA was applied to assess the dimensionality of the data and understand feature variance:

- Selected and standardised all numerical features.
- Visualised cumulative explained variance.
- Exported feature loadings for interpretation.

As seen in Fig. 4 , the first few principal components captured a substantial proportion of the dataset's variance, with the curve steadily increasing and flattening out after around 20 components. This suggested that most of the important variance in the data could be explained using a relatively small number of components, while additional components contribute diminishing returns. We found that geographic features dominated the early components, while features directly tied to accident outcomes appeared in later components. PCA revealed redundancy but was retained for exploratory insight rather than dimensionality reduction, aligning with Fernández et al. [4].

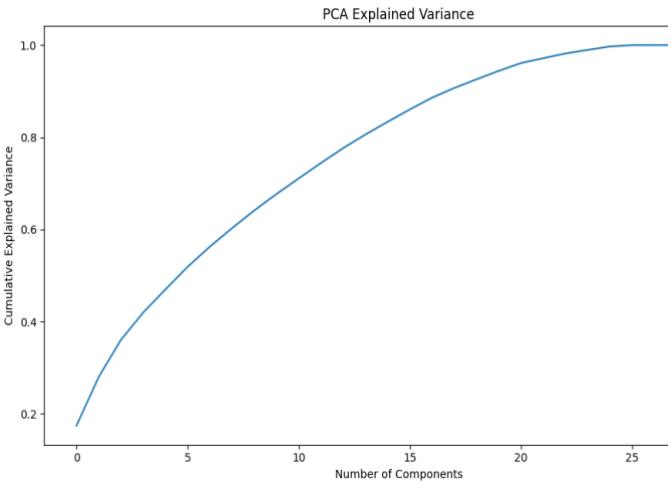

Fig. 4. PCA cumulative explained variance by number of components.

F. Modelling Dataset Preparation

Relevant features were selected by excluding non-informative or high-cardinality identifiers. The target variable (accident_severity) was separated from features. Categorical variables were label encoded, and numerical features scaled using standardization to ensure fair contribution across features, especially for models sensitive to feature magnitude. This ensures our data is numerically compatible with various machine learning models and avoids data leakage or bias from irrelevant columns**.** The dataset was split into a 70:30 stratified train-test split.

G. Class Imbalance Handling

Given the extreme imbalance in the target variable, SMOTE (Synthetic Minority Oversampling Technique) was applied only to the training set to avoid data leakage. This was combined with class weighting in model algorithms to penalize misclassification of minority classes. This choice was supported by best practices discussed in Fernández et al. [4].

H. Model Selection And Training

Initial models trained with enhanced_severity_collision showed artificially inflated performance, revealing data leakage. To ensure our models generalised well and reflected real-world prediction scenarios, we removed enhanced_severity_collision, a post-event feature to eliminate data leakage. This allowed us to evaluate model performance more realistically.

Several algorithms were evaluated:

- **Logistic Regression**: Baseline, interpretable model.
- **Random Forest**: Robust to non-linear relationships.
- **XGBoost**: Utilised class weighting to address imbalance.
- **LightGBM**: Boosting algorithm effective for structured, imbalanced data.

LightGBM emerged as the final model due to its strong recall performance on the minority (fatal) class with minimal trade-off in overall accuracy.

I. Addressing Class Imbalance

After removing the enhanced_severity_collision feature, we observed a significant drop in performance — particularly for fatal and serious accidents. This highlighted a core issue: class imbalance. This imbalance caused models to prioritize accuracy over sensitivity to rare but critical classes. We Applied SMOTE (Synthetic Minority Oversampling Technique) to balance the training set by generating synthetic samples for underrepresented classes and retrained both Logistic Regression and Random Forest on the balanced data compared results with previous models.

To further tackle the severe class imbalance, we moved beyond sampling-based methods and instead leveraged model-based class weighting with XGBoost. XGBoost allowed us to specify weights inversely proportional to class frequencies, which adjusted the

model's loss function to penalize misclassifying rare classes more heavily.

J. Hyperparameter Tuning

Grid Search was applied to Random Forest and LightGBM. Parameters tested included:

- **Random Forest**: n_estimators (50–150), max_depth (6–10), min_samples_split (2–5).
- **LightGBM**: n_estimators (50–150), learning_rate (0.05–0.2), num_leaves (31–50).

Final LightGBM configuration: n_estimators=120, learning_rate=0.1, max_depth=6, num_leaves=31. Grid Search used macro F1 as the scoring metric.

K. Evaluation Metrics

We used accuracy, macro-averaged F1-score, and class-wise recall—particularly fatal recall. Macro F1 was chosen to equally weight performance across classes, avoiding dominance by the majority class [1][2].

L. Geospatial Analysis

DBSCAN clustering was applied to geographic coordinates to identify accident hotspots. Cluster IDs were added as a new feature, one-hot encoded to avoid ordinal bias. Fig. 5 below shoes accident hotspots based on DBSCAN clustering with cluster ID colored differently. This approach followed Chen et al. [4] and enhanced spatial context in predictions. Incorporating cluster information into Random Forest and LightGBM improved fatal accident recall modestly, validating the spatial patterns.

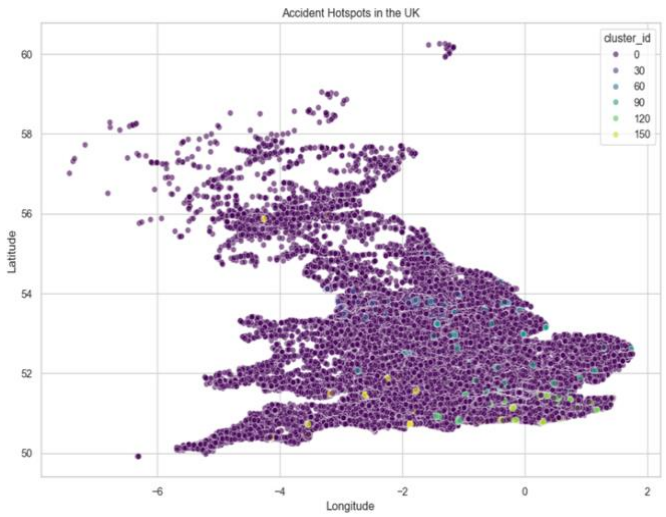

Fig. 5. Accident hotspots in the UK based on DBSCAN clustering.

M. Randomness Control And Reproducibility

Random seeds were set using random.seed(42) and numpy.random.seed(42) for SMOTE, data splitting, and model training. Each model was trained once using a fixed split to ensure reproducibility, with randomness control compensating for the lack of cross-validation.

N. Computing Environment and Infrastructure

All analyses were conducted in Google Colab with Python 3.10. Libraries included pandas, NumPy, scikit-learn, imbalanced-learn, LightGBM, XGBoost, CatBoost, seaborn, matplotlib, and folium (for mapping). The environment allowed for accessible, reproducible, and collaborative analysis.

O. Documentation and Transparency

All code, data processing steps, model outputs, and visualisations were fully documented in a single Jupyter Notebook managed in a GitHub repository to ensure transparency and replicability.

IV. RESULTS

The outcomes of data exploration, modelling, evaluation, and geospatial analysis conducted in the study are presented in this section. Missing values were prominent in environmental and road-related fields. Redundant columns with no variability or excessive cardinality were identified. The initial exploratory data analysis (EDA) also confirmed a severe class imbalance, with 76.06% of accidents classified as Slight, 22.48% as Serious, and only 1.46% as Fatal. Potential Data Quality Issues were found with A small number of rows are missing key geolocation data (e.g. latitude **,** longitude), which could interfere with geospatial analysis. Temporal and Regional Variation in features such as date **,** time **,** and local_authority_* showed enough diversity to warrant engineering into more meaningful insights later. Numerical feature distributions displayed skewness and outliers, while categorical feature distributions exhibited dominant categories but retained enough variability for modelling. Correlation matrices revealed redundancy among spatial features such as longitude, latitude, and location_easting_osgr.

Principal Component Analysis (PCA) revealed that the first few principal components capture a substantial proportion of the dataset's variance, dominated by geographic features like latitude **,** longitude **,** and location_easting_osgr **.** These spatial features explain a large amount of variation and form the basis of early components while outcome-related features like accident_severity **,** number_of_casualties **,** and number_of_vehicles contributed to later components. suggesting they influence more subtle patterns in the data. This analysis revealed a potential disconnect between features with high variance and those that predict severity, reinforcing that high variance does not always equate to predictive power.

After removing the enhanced_severity_collision feature to eliminate data leakage, baseline model performance was assessed. Logistic Regression achieved 48.0% accuracy and a macro F1-score of 0.25, with zero fatal recall. Random Forest performed better, achieving 75.8% accuracy and a macro F1-score of 0.42, but also failed to recall fatal cases.

```
Final model accuracy comparison:
WITH enhanced_severity_collision:
Logistic Regression: 0.8352
Random Forest: 0.9110
WITHOUT enhanced_severity_collision:
Logistic Regression: 0.4804
Random Forest: 0.7575
```

Fig. 7. Model comparisons with and without enhanced severity collision

Applying SMOTE improved class representation in the training data. Logistic Regression improved slightly in recall for fatal accidents but still performed poorly overall. However, Random Forest with SMOTE still failed to recall fatal cases despite achieving a macro F1-score of 0.44. While class balance improved training representation, overfitting and noise from synthetic examples may have limited generalization. This experiment showed that SMOTE helps but is not enough on its own. Class imbalance remains a major challenge for severity prediction. Boosting algorithms were evaluated next. XGBoost achieved comparable overall accuracy to Random Forest (around 76%), however, recall for fatal accidents remained near zero, suggesting that class weighting alone is insufficient to capture rare event patterns. The model still focused heavily on predicting the dominant class ("Slight"), despite the weight adjustments. While promising, this result confirmed that traditional gradient boosting requires further tuning or richer features to detect rare classes like fatal accidents effectively.

To squeeze more performance out of our Random Forest model, we applied Grid Search to tune its hyperparameters. The best model used: n_estimators=100; max_depth=10; min_samples_split=5. Accuracy was like XGBoost (~76%) but recall for Fatal accidents remained at 0%. The model slightly improved F1-score for the "Serious" class compared to untuned Random Forest. Ultimately, this showed that hyperparameter tuning improved model stability but did not overcome the limitations posed by class imbalance or insufficient signal for rare classes.

To explore whether newer boosting-based algorithms could improve performance, we tested LightGBM and CatBoost on the same feature set used for previous experiments. LightGBM (default settings) underperformed, achieving only 56.6% accuracy. It over-predicted the majority "Slight" class and struggled to identify fatal or serious accidents. CatBoost performed better, achieving 76.1% accuracy. However, like earlier models, it still failed to recall any fatal cases and produced a relatively low macro- average F1-score (0.33). Tuned LightGBM showed significant improvement. After optimizing for F1-macro, it reached 62.2% accuracy and a macro F1-score of 0.41. Fatal accident recall increased to 19.3%, making it one of the better models at detecting rare classes. Models and their performance metrics are given below in Table II. While these models offered additional perspectives, they reinforced our earlier insight: the dataset's strong class imbalance severely limits multi-class classification performance. Boosting alone isn't enough — it must be combined with balancing strategies to meaningfully improve recall for rare but critical outcomes like fatal accidents.

Model	Accuracy	Fatal Recall	Macro F1
Random Forest (clustered)	0.6122	0.3457	0.4085
LightGBM (clustered)	0.5664	0.5164	0.3925
Tuned LightGBM (clustered)	0.6228	0.1926	0.4197

Table II. Model performance metrics after incorporating clustering.

DBSCAN clustering identified 167 accident hotspots with the largest cluster (Cluster 0) concentrated in London. Cluster 0 alone accounted for nearly 19,586 accidents, shown in red on the map in Fig. 6 below. Many smaller clusters emerged in other dense regions like Manchester, Birmingham, and along major motorways. However, more than 60% of accidents were labelled as noise, suggesting spatial spread and highlighting the limitations of clustering alone. Incorporating cluster IDs into the feature set improved model performance, particularly for fatal cases. Random Forest with clustering achieved a slightly improved macro F1 and recall on fatal cases LightGBM with clustering showed strong recall on fatal accidents (52%), suggesting sensitivity to spatial context, but with a trade-off in precision Tuned LightGBM recovered overall accuracy while maintaining better class balance, though fatal recall dropped to ~19% Overall performance metrics were similar to non-clustered models, indicating partial redundancy with existing spatial features (e.g. road type, region).

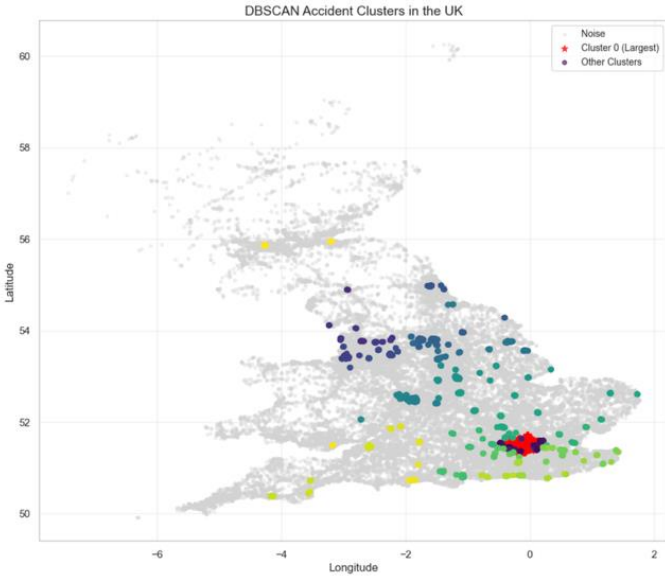

Fig. 6. DBSCAN accident clusters with largest cluster highlighted

V. DISCUSSION

Despite improvements, predicting rare outcomes like fatal accidents remains difficult. In this study we have applied a comprehensive and iterative modelling process that is supported by best practices from the literature we reviewed. Various modelling approaches were rigorously tested and compared to identify the most effective solution. Initial models, including Logistic Regression and Random Forest, demonstrated adequate performance on majority classes but failed to detect minority (fatal) cases, highlighting the impact of class imbalance. To address this class imbalance problem, SMOTE and class weighting were applied. This improved class balance but often introducing overfitting or failing to substantially increase rare class recall. This is consistent with Fernández et al. [4], who emphasized that data imbalance remains a core limitation even when advanced resampling or weighting techniques are applied.

Boosting algorithms, notably LightGBM, provided substantial performance improvements, particularly after hyperparameter tuning. This aligns with findings from Ahmed et al. [3] and Daoud et al. [5], both of whom reported that ensemble and boosting-based models tend to outperform simpler classifiers in severity prediction tasks, especially under imbalanced data scenarios.

Incorporating spatial clustering features further enhanced minority class recall. The LightGBM model highlighted the importance of geospatial context in cases such as ours. LightGBM with clustering achieved a fatal recall of 51.6%. Similar spatial effects were noted by Chen et al. [6], who found that incorporating spatial and contextual features improved the ability of Support Vector Machines to predict injury severity in complex scenarios such as rollover crashes. While adding spatial features did indeed improve sensitivity to fatal cases, it also increased model complexity.

VI. CONCLUSION

This study applied a comprehensive machine learning pipeline to predict the severity of road accidents in the UK using the 2023 Department for Transport (DfT) Road Safety dataset. Through systematic data cleaning, feature engineering, dimensionality analysis (PCA), and iterative modelling, we explored multiple algorithms including Logistic Regression, Random Forest, XGBoost, CatBoost, and LightGBM. We summarised all key results to evaluate which approach performed best not just in terms of overall accuracy, but in their ability to identify rare and critical cases. We compared baseline, SMOTE, and advanced models across accuracy and F1 metrics, conducted class-specific performance analysis (especially for fatal accidents) and lastly, reviewed whether feature tuning and model selection improved performance

Random Forest (without enhanced_severity_collision) offered a strong baseline with 75.8% accuracy, though it failed to recall many fatal cases. SMOTE oversampling improved class balance but did not consistently improve model performance. It slightly boosted Logistic Regression recall for fatalities but introduced instability. XGBoost and CatBoost performed similarly to Random Forest, each with ~76% accuracy, but like others, struggled with minority class prediction. LightGBM (tuned) struck the best balance between accuracy and fatal recall, offering a trade-off between sensitivity and overall performance. Grid Search on Random Forest yielded marginal gains in accuracy but failed to improve macro F1-score.

Random Forest and CatBoost emerged as top performers, particularly in weighted F1 scores. But even these struggled with minority class recall. Feature pruning yielded almost no loss in accuracy, suggesting redundancy and reinforcing the value of PCA and importance ranking in simplifying models. Geospatial clustering (DBSCAN) revealed distinct hotspots, with one dominant cluster around London. However, the high proportion of “noise” points shows that not all risks are spatially concentrated.

Ultimately, no model performed exceptionally well at predicting fatal accidents. We learnt that data imbalance was a serious challenge. Fatal accidents made up only 1.46% of all records, while slight accidents dominated with 76%. This imbalance severely impacted model performance, especially in predicting rare fatal incidents. We also realized Feature leakage matters. The feature enhanced_severity_collision was found to be highly predictive but ultimately represented label leakage, inflating accuracy unrealistically. Removing it exposed the true limitations of our models. Recall for fatal accidents remains poor across all models, including Logistic Regression, Random Forest, XGBoost, LightGBM, and CatBoost. Even with SMOTE and class weighting, most models failed to reliably identify fatal incidents. However, the experimentation confirmed that feature engineering,

class balancing, and tuning are all necessary — but not sufficient — steps when working with real- world, imbalanced datasets.

This project highlighted how iterative model development, critical feature evaluation, and domain knowledge are essential for responsible machine learning. Rather than relying on accuracy alone, we focused on class-specific performance, identified and removed features with data leakage and validated model behaviour through experimentation and visualisations. Despite limitations in predicting fatal accidents, the workflow demonstrates a mature, interpretative approach — blending predictive techniques with thoughtful analysis.

The study emphasizes that ensemble and boosting models enhance performance, but challenges such as class imbalance and rare event prediction require various solutions. We should explore advanced resampling, richer feature engineering, and cost-sensitive or deep learning approaches in future work. The dataset fails to capture the human centered factors such as the driver age, driving experience, ability/ inability to obey traffic laws, mood or state of mind while driving, use of alcohol or substances and other such factors that are known to influence accident severity. Future research should also involve enhancing some more sources of data and aim for data that particularly focuses on human behaviour and the vehicle features and capabilities to help enhance the prediction model. Additionally, integrating external data sources such as traffic volume, vehicle telematics, and real-time weather data could also enrich feature sets. The findings provide a foundation for developing more robust and actionable traffic safety models in the future.

For future research, to better handle class imbalance, we could combine advanced resampling methods such as SMOTE-ENN or ADASYN with cost-sensitive learning. Deep learning models, particularly those leveraging attention mechanisms, might capture complex interactions among variables. Additionally, combining supervised learning with unsupervised techniques, such as anomaly detection or clustering-driven feature generation, presents an exciting avenue to improve rare event prediction. Spatial-temporal models that incorporate both location and time dimensions may also enhance the detection of high-risk patterns.

REFERENCES

[1] Department of Transport, Reported road casualties in Great Britain: annual report 2019 (Page 7), Minster House, September 2020.

[2] Department for Transport, "Road accidents and safety statistics," data.gov.uk, Apr. 2025. [Online]. Available: https://www.data.gov.uk/dataset/cb7ae6f0-4be6-4935-9277-47e5ce24a11f/road-accidents-safety-data (source)

[3] S. Ahmed, M. A. Hossain, M. M. I. Bhuiyan, and S. K. Ray, "A comparative study of machine learning algorithms to predict road accident severity," in *Proc. 20th Int. Conf. Ubiquitous Comput. Commun. (IUCC)*, London, UK, Dec. 2021, pp. 241–248. doi: 10.1109/IUCC-CIT-DSCI-SmartCNS55181.2021.00069

[4] A. Fernández, S. García, M. Galar, R. C. Prati, B. Krawczyk, and F. Herrera, *Learning from Imbalanced Data Sets*. Springer, 2018

[5] R. Daoud, M. Vechione, O. Gurbuz, P. Sundaravadivel, and C. Tian, "Comparison of machine learning models to predict nighttime crash severity: A case study in Tyler, Texas, USA," *Vehicles*, vol. 7, no. 1, pp. 20, Feb. 2025, doi: 10.3390/vehicles7010020.

[6] C. Chen, G. Zhang, M. A. Quddus, Z. Tian, and H. Yue, "Investigating driver injury severity patterns in rollover crashes using support vector machine models," *Accident Analysis & Prevention*, vol. 90, pp. 128–139, 2016.

CREDIT STATEMENT

Term	Contributor
Conceptualization	Shubhankar, Lukshan, Qing
Methodology	Lukshan, Qing
Software	Lukshan, Qing
Validation	Lukshan, Shubhankar
Formal analysis	Qing, Shubhankar
Investigation	Shubhankar, Qing
Resources	Qing, Lukshan
Data Curation	Qing, Lukshan
Writing - Original Draft	Shubhankar, Lukshan
Writing - Review & Editing	Shubhankar, Qing
Visualization	Lukshan, Shubhankar
Supervision	Shubhankar, Lukshan, Qing
Project administration	Shubhankar, Lukshan, Qing