IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Digital Object Identifier

Machine Learning Approaches for News-Based Traffic Accident Analysis in Malta

PAUL SAID, MICHAEL MIZZI, MICHAEL VELLA, ISAAC CUTAJAR
https://github.com/michael-vella/ics5110-assignment

ABSTRACT This study investigates road-traffic incidents in Malta using traditional machine learning techniques applied to unstructured and semi-structured narratives from police press releases and local news articles. A structured dataset is constructed through transparent, rule-based extraction of incident characteristics and is enriched with temporal, geographic, weather and accident characteristics-related features. Exploratory data analysis is conducted to assess data quality, reporting bias, and class imbalance. Given that low-severity incidents are systematically under-represented in news-based sources, we do not model population-level accident severity. Instead, the primary task is formulated as a binary classification problem: predicting whether a reported incident is fatal versus non-fatal. We evaluate interpretable and non-linear baselines, including logistic regression, random forests, support vector machines and k-nearest neighbours, and analyse the relative contribution of feature groups. Results suggest that fatality among reported incidents is more strongly associated with crash context and participant composition than with short-term weather conditions. Ethical considerations related to proxy variables, reporting bias, and deployment risks are discussed.

I. INTRODUCTION

A. CONTEXT AND MOTIVATION

Road traffic accidents remain a significant public safety concern, particularly in geographically constrained regions such as Malta, where dense traffic networks and mixed road usage interact in complex ways [1]–[3]. While official statistics are typically released in aggregated form, detailed contextual information is often communicated through unstructured sources such as police press releases and news articles.

The availability of such textual data presents an opportunity to apply machine learning techniques to support exploratory safety analysis. However, challenges arise from reporting bias, inconsistent terminology, and limited sample sizes. In this context, traditional and interpretable machine learning models are particularly appropriate, as they balance predictive capability with transparency and accountability.

B. RESEARCH QUESTIONS AND OBJECTIVES

Low-severity collisions are less likely to be reported in news and press releases; therefore, the dataset reflects reporting practices rather than the full population of road-traffic accidents. For this reason, the study is framed as a conditional prediction problem: fatality, given that an incident is reported. The project addresses the following research questions:

- **RQ0 (Primary):** Given a reported road-traffic incident, can we predict whether the outcome is fatal versus non-fatal from extracted contextual, temporal, geographic, participant, and weather features?
- **RQ1:** Which feature groups (participant/vehicle composition, road type and region, time, weather) contribute most to fatality prediction among reported incidents?
- **RQ2:** Are there identifiable temporal or geographic contexts (e.g., late evening, weekends/holidays, specific regions or road types) in which the fatality rate among reported incidents is higher?
- **RQ3:** Does incorporating weather variables provide measurable incremental predictive value beyond non-weather features?
- **RQ4:** What are the trade-offs between interpretability and predictive performance across the evaluated models (logistic regression, random forests, support vector machine and k-nearest neighbours)?

C. CONTRIBUTIONS

The main contributions of this work are:

- Construction of a curated, structured dataset of reported road-traffic incidents from unstructured police and news narratives.
- Integration of temporal, geographic, and weather-related context, with feature engineering to reduce sparsity (e.g., road type and regional grouping).
- Formulation and evaluation of a fatality prediction task (fatal vs non-fatal) under class imbalance, including threshold selection and performance analysis.
- Critical discussion of reporting bias, proxy variables, and ethical considerations relevant to potential deployment.

II. RELATED WORK

Research on road traffic accidents involving single or multiple vehicles has frequently focused on regions with comparable demographic and geographic characteristics, particularly island environments and Mediterranean contexts. Across this body of work, the factors most commonly associated with accident severity and fatality can be broadly grouped into driving behaviour (with particular emphasis on speed), road user demographics, road infrastructure, and socioeconomic conditions [1], [4].

Malta is characterised by high population density, limited land area, and a dense road network, features that closely resemble those observed in Greek island regions. Consequently, empirical findings from Greece are often considered transferable to the Maltese context [2], [5]. In particular, research has identified elevated accident risk among foreign drivers and tourists, attributed to unfamiliarity with local road layouts and driving conditions [2]. These observations motivate the selection of comparable explanatory variables in the present study.

Table 1 summarises the principal categories of accident-related features reported in prior literature. These feature groups define the theoretical dimensional space commonly used in accident severity analysis and provide a structured baseline for feature engineering. By grounding feature selection in metrics that have been repeatedly validated in previous studies, this work aims to mitigate limitations arising from the relatively small size of the observed dataset.

While the above studies provide valuable insights, most rely on structured administrative datasets collected by police or national statistics offices. Such datasets typically provide comprehensive coverage of reported accidents but lack rich contextual narratives. In contrast, unstructured sources such as news articles and police press releases remain under-explored in predictive traffic safety research, despite their potential to capture detailed contextual information.

Furthermore, prior work frequently models multi-class injury severity using large-scale datasets. In small-scale or media-derived datasets, however, class imbalance and reporting bias pose significant challenges, as minor incidents are systematically under-represented. This motivates the formulation of the present study as a conditional fatality prediction task among reported incidents, rather than population-level severity modelling.

TABLE 1. Categories of Accident Features Reported in Related Literature

Feature Category	Representative Measures
Injury Metrics	Number of slight, grievous, and fatal injuries; total number of accidents; number of individuals involved; accident and fatality rates per capita; injury type (e.g., thoracic, head, spinal); injury severity scores; accident frequency [3], [4].
Driver Characteristics and Behaviour	Driver age group and gender; vehicle position; vehicle type; use of protective equipment (seat belts, helmets, child restraints); driving experience; self-reported accident history; psychological variables (e.g., driving style, violation intent); speeding-related proxies such as penalty points or sudden braking [1], [5].
Traffic and Exposure Metrics	Annual Average Daily Traffic (AADT); traffic flow; number of vehicles involved; motorisation rate (vehicles per 1000 inhabitants) [4].
Road, Environment, and Location	Area type (urban or non-urban); geographic division (mainland versus island areas); surrounding development intensity; road geometry [2].
Time and Weather Conditions	Temporal indicators (month, day of week, hour); accident timing (day/night, weekday/weekend, tourist season); weather conditions (e.g., rain, fog, storms); night-time lighting conditions [1].

Statistical and mathematical models, such as association rule mining (Apriori algorithm) [2], are often applied in recent research. In our context, we explore ensemble methods such as random forests, which often achieve strong predictive performance but provide limited interpretability, whereas logistic regression remains widely used in safety research due to its transparency and ease of interpretation [6] [7]. Given the exploratory nature of this study and the limited dataset size, we prioritise model-to-model comparison and ranking feature groups in contributing accident fatality.

III. DATA

A. DATA SOURCES

Two accident data sources were provided, local news articles and police press releases. The data was supplied in an unstructured textual format, which is not directly suitable for training or evaluating a machine learning model.

1) Local News Articles

This dataset consisted of 321 observations pre-extracted from two main local news portals. The data is semi-structured with information about incidents that occur on the roads of Malta and Gozo. The records were reported over a time span of reports written between December 2024 and October 2025.

2) Police Press Releases

One hundred and eleven (111), observations were presented in a semi-structured dataset summarising the police press

Importance level	Data Description	Type	Class
Must Have	***Incident Date***	Feature	Date [DD:MM:YYYY]
Must Have	***Incident time***	Feature	{morning, afternoon, night}
Must Have	***Street of incident***	Feature	Text
Must Have	***City***	Feature	City Code
Must Have	***Area type***	Feature	Main road / secondary road
Must Have	***Type of Vehicle***	Feature	unknown, pedestrian, bicycle, motorbike, car, van, bus Notes: 'bicycle' includes e-scooters & 'van' includes heavy vehicles such as trucks
Must Have	***Driver age bracket***	Feature	1: <18, 2: [18 to 24],3: [25 to 49], 4: [50 to 64], 5: 65+, 6: unknown
Must Have	***Driver gender***	Feature	M,F,..
Must Have	***Number of injured persons***	Feature	integer
Must Have	***Severity of injuries***	Label	Fatal, Grievous, not injured, Serious, slight, unknown
Must Have	***Number of vehicles damaged***	Feature	integer
Must Have	***Severity of vehicle damages***	Label	[0,1,2] - (i.e - no, minor, major)
Must Have	***Environment: Weather***	Feature	[Rainy, sunny, windy, ..]
Must Have	***Aspect: holiday/ event / school***	Feature	[no, yes, eve of]
Must Have	***Traffic level @ accident time***	Feature	[0,1,2,3] - (i.e - no info, light, moderate, high)
Nice to Have	***Incident type***	Label	Incident/ Accident/ Report
Nice to Have	***Number of vehicles involved***	Feature	integer
Nice to Have	***Environment: Speed class.***	Feature	[Low-Medium-high]
Nice to Have	***Environment: infrastructure***	Feature	Boolean
Not Important	***Brand/model***	Feature	Text
Not Important	***Driver class***	Feature	Tourist, local, Int. Licence
Not Important	***PPE use infringement***	Feature	Boolean: Lack of use of seatbelts, light, helmets, etc..
Not Important	***Property damage***	Feature	Boolean
Not Important	***Severity of property damage***	Label	[0-5]

FIGURE 1. Targeted features / Labels - Template

release about road traffic incidents in Malta. The data span of the press releases also ranged between December 2024 and October 2025.

As a first step in addressing the data retrieval pipeline of these two datasets a brainstorming session was conducted based on visual screening of the raw data. The basic knowledge acquired from reviews of related works was applied to identify a target features/labels list.

Out of this list a must-have list of data features was identified as shown in Figure 1. This served as a baseline to guide the team in extracting a list of target elements during the preprocessing and feature engineering aspects of the pipeline. Additionally, a list of lower ranked features was also developed and taken into account during the feature engineering aspect. Having this data extraction template led to the initial phase of this project focused on cleaning, deduplicating and curating a structured dataset from the unstructured sources in order to enable the training and evaluation of different machine learning models.

B. PREPROCESSING

Given that the accident data was provided in an unstructured textual format, a multi-stage preprocessing pipeline was required to transform it into a structured dataset suitable for machine learning. The initial preprocessing phase focused on extracting relevant attributes from both datasets. Moreover, some entries did not correspond to traffic accidents and therefore needed to be identified and removed during the data cleaning process.

Although the datasets provided were relatively small and the required features could have been extracted manually, an automated approach was adopted to assess scalability in the event of a larger dataset. A combination of rule-based methods and automated extraction techniques was therefore employed. Regular expressions were used to identify and extract structured information such as accident dates, times, and locations from the textual data. In addition, large language models (LLMs), specifically the GPT `o4-mini` model, were utilised to extract higher-level semantic features that are difficult to reliably obtain using deterministic rules alone.

As part of the LLM-based extraction process, the model was first prompted to determine whether a given entry corresponded to an actual traffic accident. If this condition was not met, no further features were extracted and the entry was marked as a non-accident. Otherwise, the model attempted to extract a predefined set of attributes, including the number of

individuals injured, descriptive characteristics of drivers and vehicles (such as age, gender, and vehicle type), whether the driver involved was a victim, and the severity of the accident. Accident severity was categorised using an ordered set of five classes, ranging from lowest to highest severity: *No Injuries*, *Slight*, *Grievous* (assigned when terms such as "grievous" or similar appeared in the description), *Serious* (assigned when terms such as "serious", "critical" or similar appeared), and *Fatal*. To ensure consistency and reduce post-processing requirements, the model was instructed to output data using a fixed JSON schema with predefined categorical values.

The results extracted using regular expressions were generally accurate for structured attributes such as accident dates. Location extraction proved more challenging using a rule-based regular expression approach, as some articles reported the place of origin of the driver in addition to the accident location, which occasionally led to ambiguity and incorrect assignments. However, the main limitation was to extract semantic information related to the vehicle, driver and severity using rules as no patterns were identified in the source text to perform such extractions.

The features extracted using the LLM were generally reasonable. In most cases, the model could identify accidents from non-accidents and performed well in extracting the location of the accident. However, improved prompt guidance could have further enhanced extraction accuracy. For example, vehicle types were not always explicitly reported in the source text, leading the model to output specific vehicle brands (e.g., *Toyota*) instead of the expected generic category (e.g., *Car*). Furthermore, in some news articles and in four police press releases, multiple accidents were reported within a single entry. This increased the difficulty for both rule-based and LLM-based extraction techniques, as relevant features for multiple distinct accidents were sometimes conflated or incompletely captured.

Due to these limitations, a manual auditing step was introduced to verify and correct the automatically extracted data. This process involved reviewing each extracted entry to resolve ambiguities, correct misclassified attributes, and separate cases where multiple accidents were reported within a single entry. As a result of this manual validation process, a final curated dataset was produced in which each record corresponds to a single accident. Police press releases and news articles were combined into a single dataset, and entries that were not associated with traffic accidents were filtered out. At the end of this stage, each entry in the dataset contained the accident date and time, street and city of occurrence, number of individuals injured, accident severity, and structured details of the vehicles and drivers involved.

During the manual auditing step, it was observed that certain accidents were duplicated, either because they were reported in multiple news articles or appeared in both police press releases and news articles. Entries were identified as duplicates when they shared the same accident date and time and the same location (street and town). To address inconsistencies in street and town naming conventions across different data points, dedicated dimension tables, namely *Dim_Street* and *Dim_Town*, were created to normalise location information. These dimensions were used to standardise variations in spelling, formatting, and abbreviations, and were later joined back to the final version of the dataset so that the cleaned and canonical street and town representations were consistently used. To resolve such duplicates programmatically, a hierarchical selection strategy was applied. When duplicate entries were identified, the record with the highest reported accident severity was retained. If multiple entries shared the same severity level, the entry with the highest reported number of injuries was kept. In cases where both severity and number of injuries were identical, the police press release entry was retained, as this source was generally more structured and contained more detailed information. Any remaining ambiguous cases were flagged for manual review. At the end of this process, the dataset was cleaned of duplicate accident records.

C. FEATURE EXTRACTION

Following the preprocessing stage, additional feature extraction and transformation steps were applied to prepare the dataset for downstream machine learning tasks, as shown in Figure 2. To derive meaningful information from the accident date and time, a categorical temporal feature was created by grouping accidents into time-of-day intervals. Specifically, accidents were classified as occurring in the *early morning* (06:01–08:00), *morning* (08:01–12:00), *afternoon* (12:01–18:00), *evening* (18:01–21:00), *late evening* (21:01–23:00), or *night* (23:01–06:00). This transformation allows temporal patterns to be captured while reducing the granularity of the raw timestamp.

Additional temporal features were also extracted from the accident date information. These included binary indicators capturing whether an accident occurred during a weekend, on a public holiday, during a school holiday period (including the Christmas, Easter, and summer breaks), and whether the date corresponded to a regular school day. These features were included to attempt to capture variations in traffic patterns and road usage associated with changes in daily routines, which may influence accident occurrence and severity.

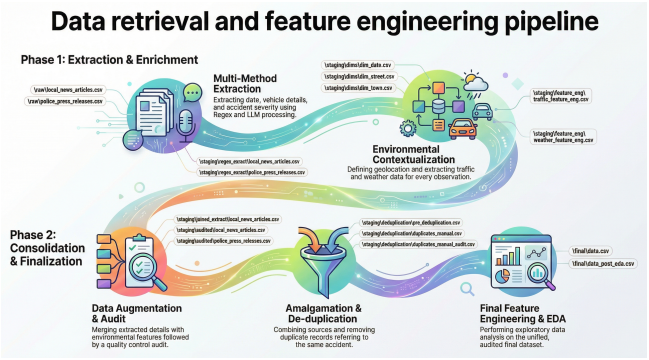

FIGURE 2. High-level overview of the preprocessing and feature engineering pipeline.

Driver and vehicle level information was further enhanced by aggregating individual attributes into accident related features. For each accident, the number of drivers was summarised according to predefined age groups, including under 18, 18–24, 25–49, 50–64, and 65 years or older. In addition, indicators were created to capture whether driver age information was missing for any participant. Driver gender information was similarly aggregated by reporting the number of male drivers, female drivers, and drivers with unknown gender per accident.

Vehicle related features were derived by counting the number of vehicles involved in each accident by type. These included pedestrians, bicycles, motorbikes or motorcycles, cars, vans, buses, and vehicles with unknown classification. By aggregating driver and vehicle information in this manner, the dataset captures both the composition and demographic characteristics of each accident. Unknown or missing values were retained explicitly rather than imputed, as the absence of this information often reflected limitations in the original reporting rather than random missingness. Imputing such values could therefore introduce artificial bias or misleading patterns into the dataset.

Since the extracted street and town values exhibited a high degree of variability, they were not suitable for direct one-hot encoding. To address this, higher-level features were derived from the street information by categorising each street according to its functional road type, namely primary, secondary, tertiary, residential, trunk, or other. This abstraction reduces sparsity while preserving meaningful information about road characteristics that may influence accident outcomes. Road characteristics were gathered and annotated using data from OpenStreetMap[1].

Similarly, town-level information was transformed by grouping towns into broader geographical regions. Each accident was assigned to one of five regions, these being, Central, Gozo, Northern, South Eastern, or Southern Harbour based on official regional classifications (as illustrated in Figure 3). This regional grouping enables spatial patterns to be captured at a coarser but more consistent level, making the features more suitable for machine learning models.

Weather related features were also incorporated into the dataset by extracting weather conditions, temperature, and wind information at the time of each accident using the Open-Meteo API[2]. These features were included to capture environmental factors that may influence driving conditions and accident severity. An attempt was also made to collect traffic related statistics at the time of the accident using the Google Routes API[3]. However, as the Google Routes API does not support historical queries, this information could not be retrieved directly. To address this limitation, traffic conditions were approximated by querying the API for a future date and time at the same location that closely matched the conditions of the original accident.

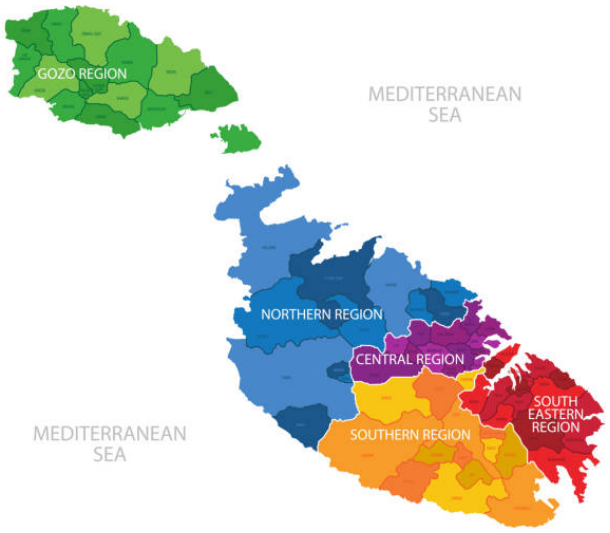

FIGURE 3. Different regions in Malta upon which our regional classifications were based on.

At the final stage of the feature extraction process, a consolidated dataset was produced containing all engineered features derived from temporal, spatial, environmental, driver, vehicle, and accident level information. Table 2 summarises the final set of features used in this study and provides a brief description of each. A total of 219 accident entries remained at this stage.

D. EXPLORATORY DATA ANALYSIS

This section presents an exploratory data analysis (EDA) of the curated accident dataset to gain insight into its underlying characteristics and distributions. The objective of this analysis is to understand the structure of the data, identify patterns and trends, and highlight potential challenges relevant to model development.

As an initial step in the EDA, the distribution of accidents by year was examined, considering that the original data consisted of reports published between December 2024 and October 2025. As shown in Figure 4, the dataset predominantly contains accidents that occurred in 2024 and 2025. A small number of accidents were also recorded in earlier years, with the oldest reported accident being in 2006. This is a result of the data collection process taking place between late 2024 and 2025, during which some news articles reported on older accidents.

Accidents that occurred prior to 2024 (eight entries in total) were considered outliers and subsequently removed from the dataset, reducing the dataset to 211 entries in total. These older records were excluded on the premise that road infrastructure, traffic patterns, and reporting practices may have changed significantly over time, that is; drift in environment, potentially introducing inconsistencies to the analysis. Additionally, the density of the reporting cases was

[1]https://www.openstreetmap.org
[2]https://open-meteo.com/
[3]https://developers.google.com/maps/documentation/routes

TABLE 2. Attributes and features extracted after preprocessing and feature engineering.

Feature Name	Feature Description
Accident datetime	Date and time of accident.
Accident severity	Severity: *No Injury* to *Fatal*.
City (or town)	Town in Malta.
Street	Street name.
Accident time cat.	Time-of-day category.
Drivers < 18	Count of drivers under 18.
Drivers 18–24	Count of drivers aged 18–24.
Drivers 25–49	Count of drivers aged 25–49.
Drivers 50–64	Count of drivers aged 50–64.
Drivers 65+	Count of drivers aged 65+.
Drivers (Unknown)	Count of unknown ages.
Male drivers	Count of male drivers.
Female drivers	Count of female drivers.
Unknown gender	Count of unknown genders.
Total drivers	Total drivers involved.
Pedestrians	Number of pedestrians.
Bicycles	Number of bicycles.
Motorbikes	Number of motorbikes.
Cars	Number of cars.
Vans	Number of vans/heavy vehicles.
Buses	Number of buses.
Unknown vehicles	Unknown vehicle types.
Is weekend	True if Saturday or Sunday.
Is public holiday	True if a public holiday.
Is school holiday	True during school holidays.
Is school day	True on school days.
Street type	Road classification.
Region	Region in Malta.
Max temperature	Max temp on accident day.
Min temperature	Min temp on accident day.
Mean temperature	Mean temp on accident day.
Max wind speed	Max wind speed.
Precipitation	Total daily precipitation.
Is raining	Binary rain indicator.
Traffic level	Estimated traffic level.

not consistent to past years as observations were only gathered in the 2024-2025 period.

Following the exclusion of outliers, exploratory analysis was performed on all features produced during the preprocessing and feature engineering stage, as summarised in Table 2. Initially, analysis focused on accident severity, as at this stage, this was considered the target label for prediction. The distribution of accident severity was visualised using both a bar chart (Figure 5) and a pie chart (Figure 6).

As illustrated in these figures, the majority of reported accidents fall into the higher severity categories. This observation is expected, as minor incidents such as low-impact or bumper-to-bumper collisions are less likely to attract media attention and are therefore under-reported compared to more serious accidents. As a result, the severity distribution exhibits substantial class imbalance, with severe accidents being

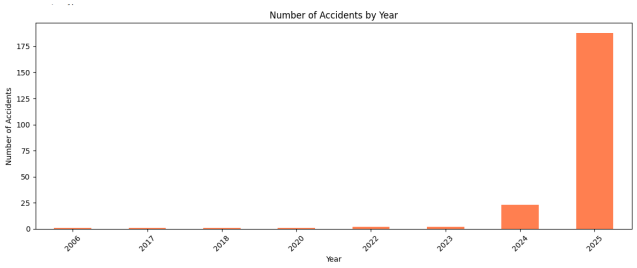

FIGURE 4. Distribution of accidents by year.

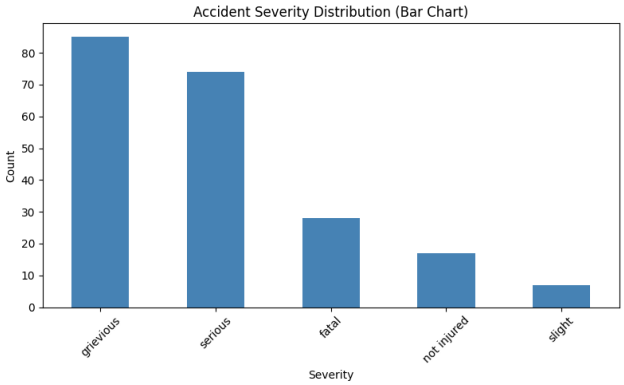

FIGURE 5. Bar chart depicting the distribution of accident severities.

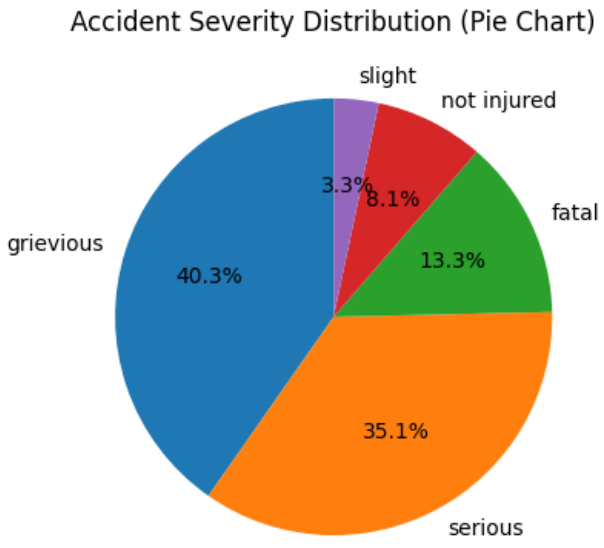

FIGURE 6. Pie chart depicting the distribution of accident severities.

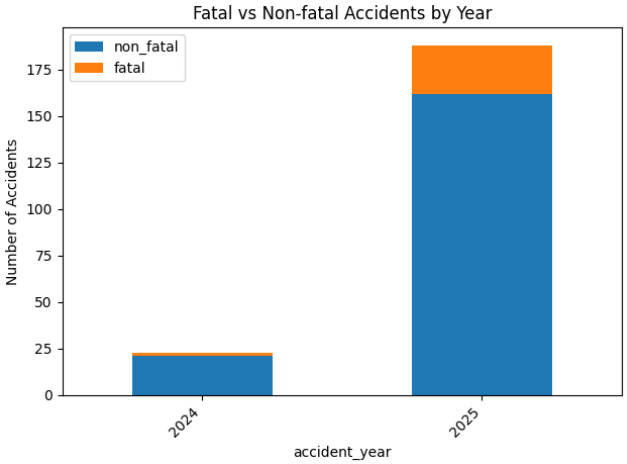

FIGURE 7. Distribution of fatal and non-fatal accidents by year of accident.

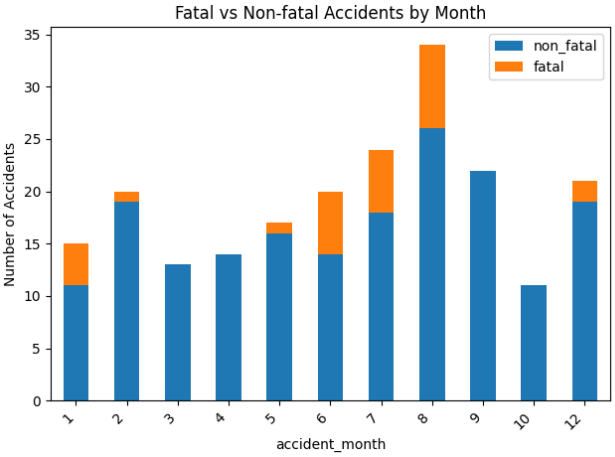

FIGURE 8. Distribution of fatal and non-fatal accidents by month of accident.

overrepresented relative to lower-severity incidents. Given the limited size of the dataset and the uneven distribution across severity classes, multi-class classification was deemed impractical. Consequently, the research question was reformulated to focus on a binary classification task, distinguishing between fatal and non-fatal accidents.

Within the resulting binary formulation, the dataset contains 28 fatal accidents and 183 non-fatal accidents. Although this still represents an imbalanced class distribution, the degree of imbalance is less extreme than in the original multi-class setting and is more suitable for binary classification. As shown in Figures 7 and 8, the majority of fatal accidents in the dataset occurred in 2025, with a noticeable concentration during the summer months of June, July, and August. This seasonal pattern suggests that factors such as increased traffic volume and changes in driving behaviour during the summer period may contribute to higher fatality risk.

Furthermore, as illustrated in Figure 9, accidents occurring during the late evening period (21:01–23:00) exhibit the highest fatality rates, indicating that reduced visibility, fatigue, or risky driving behaviours during late evening hours may play

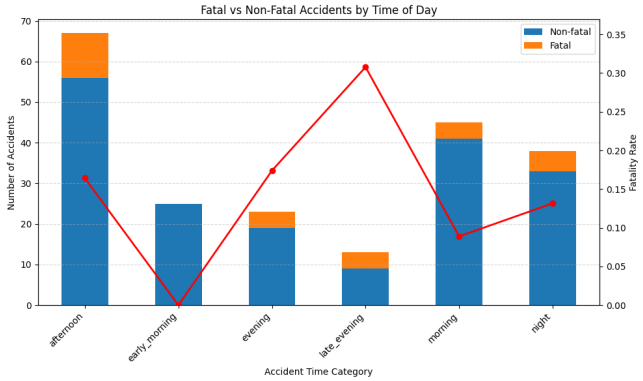

FIGURE 9. Distribution of fatal and non-fatal accidents by accident time category.

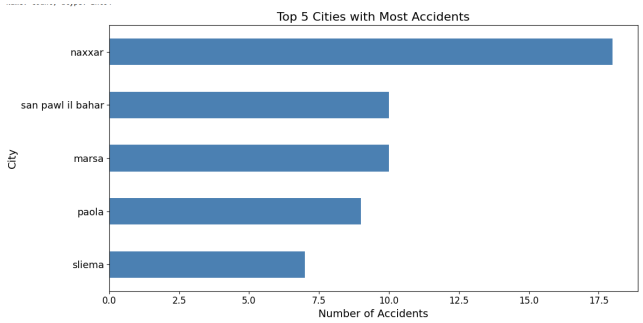

FIGURE 10. Bar chart depicting top 5 cities with most reported accidents.

a significant role in accident severity.

Analysis of accident locations was also performed to identify spatial patterns within the dataset. Figure 10 shows the top 5 towns with the highest number of reported accidents, with Naxxar emerging as the most prominent accident hotspot in terms of total accident count. However, as illustrated in Figure 11, Naxxar does not exhibit the highest number of fatal accidents. Instead, Isla, Valletta, Pembroke, and Floriana each recorded three fatal accidents, indicating that areas with fewer total accidents may still experience higher accident severity.

When analysing accidents by region (see Figure 12), no fatal accidents were reported in Gozo during the study period. In the remaining regions, fatality rates ranged from 8.6% in the Northern Region to 21.7% in the Southern Harbour region,

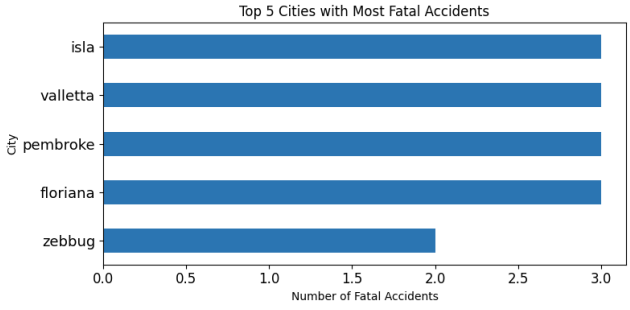

FIGURE 11. Bar chart depicting top 5 cities with most reported fatalities.

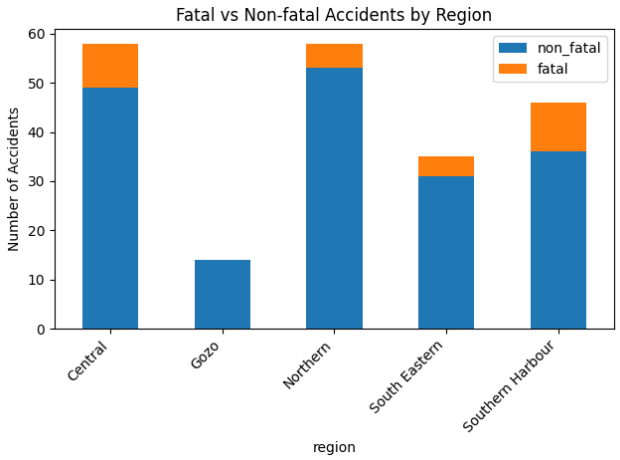

FIGURE 12. Distribution of fatal and non-fatal accidents by region.

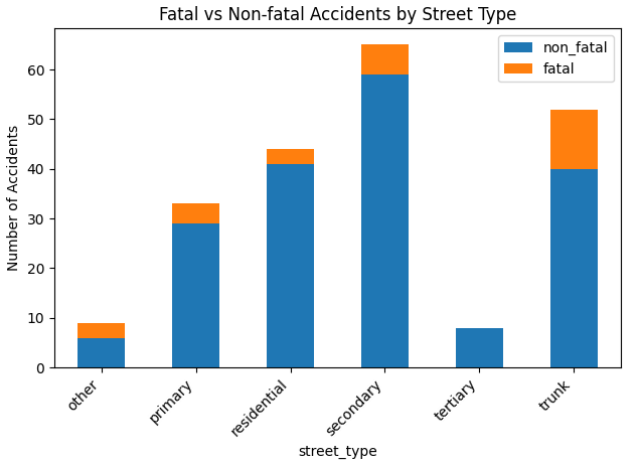

FIGURE 13. Distribution of fatal and non-fatal accidents by street type.

highlighting notable regional variation in accident outcomes.

Similarly, accidents were analysed by street type, as depicted in Figure 13. The street types with the highest fatality rates were the *other* and *trunk* categories, with fatality rates of 33.3% and 23%, respectively. The *other* category represents road types that do not fall into standard functional classifications, such as roundabouts, wharfs, slip roads, and access roads. Accidents occurring on these road types may involve complex traffic interactions or atypical layouts, which can increase the likelihood of severe outcomes.

Trunk roads, which typically support higher traffic speeds and volumes, also exhibit elevated fatality rates. This observation aligns with expectations, as higher-speed environments increase the risk of fatal injuries when collisions occur. In contrast, residential and tertiary roads generally display lower fatality rates, reflecting lower speed limits and reduced traffic intensity. These findings further support the relevance of street type as a key factor in modelling accident severity.

Driver and vehicle level characteristics were also examined as part of the exploratory data analysis. Across all accidents in the dataset, male drivers were involved in the majority of incidents, accounting for 66.3% of recorded drivers, followed by female drivers at 19.1%. The remaining 14.6% correspond to cases where driver gender information was not reported in the source material. This proportion of missing values reflects limitations in media reporting rather than data collection errors.

With respect to age distribution, most drivers involved in accidents were between 25 and 49 years old, representing 44.6% of the dataset. The remaining age groups comprised of drivers under 18 (2.3%), those aged 18 to 24 (12.6%), drivers aged 50 to 64 (16.3%), drivers aged 65 and above (12.6%), and drivers with unknown age (11.7%). This distribution broadly aligns with expected driving demographics, where middle-aged drivers constitute a large proportion of road users.

Analysis of vehicle involvement revealed that cars were the most frequently involved vehicle type, accounting for 51.1% of vehicles in reported accidents, followed by motorbikes at 30.6%. Other vehicle categories, such as bicycles, buses, vans, and pedestrians, constituted a smaller proportion of the dataset.

Analysis was conducted to examine how the number of drivers, driver demographics, and vehicle composition relate to fatal accident outcomes. With respect to driver gender, accidents involving a higher number of male drivers were associated with increased fatality rates. Specifically, accidents involving three male drivers resulted in a fatal outcome in 50% of cases.

Analysis by age group indicated that drivers aged 18 to 24 were the most strongly associated with fatal accidents. When two drivers from this age group were involved in an accident, the probability of a fatal outcome was 50%. In cases where only a single driver aged 18 to 24 was involved, the fatality rate decreased to 22.2%, which remains relatively high compared to other age groups.

Vehicle composition was also found to influence fatality rates. Accidents involving three cars resulted in fatal outcomes in 50% of cases, while those involving two cars exhibited a fatality rate of 23.5%. For motorbike-related accidents, the highest fatality rate was observed when a single motorbike was involved, with a fatality rate of 12.5%, reflecting the increased vulnerability of motorcyclists.

Temperature was also analysed in relation to fatal and non-fatal accidents. As shown in Figure 14, fatal accidents tend to occur more frequently at higher temperatures. This observation is consistent with the seasonal pattern identified earlier (Figure 8), where a higher number of fatal accidents were recorded during the summer months.

Although temperature itself is unlikely to be a direct cause of fatal accidents, it may act as a proxy for other contributing factors, such as increased traffic volumes, longer driving durations, or changes in road usage patterns during warmer periods, for example due to tourism. As such, temperature-related features were retained to capture potential indirect associations with accident severity rather than to imply a causal relationship.

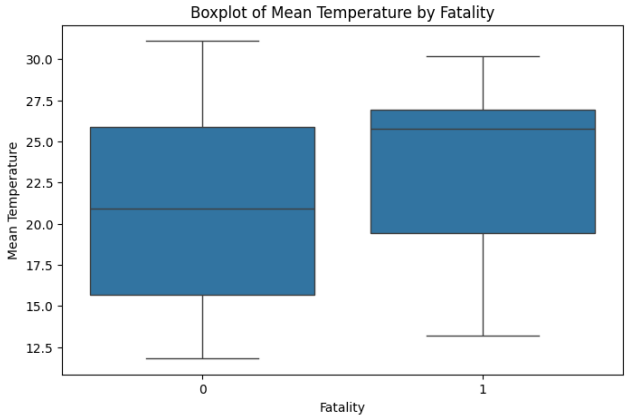

FIGURE 14. Boxplot of mean temperature by fatality.

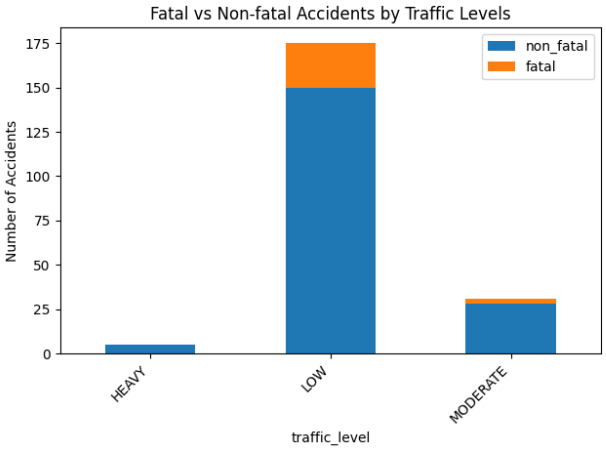

FIGURE 15. Distribution of fatal and non-fatal accidents by traffic level.

Wind speed and precipitation were also analysed in relation to accident fatality. Interestingly, the results suggest that rainfall does not appear to be associated with an increased likelihood of fatal accidents within the dataset. In fact, the majority of accidents, including fatal ones, were observed to occur under non-rainy conditions. One possible explanation for this pattern is that rainfall is relatively infrequent in Malta, and drivers may adopt more cautious driving behaviour during adverse weather conditions.

Traffic level was also analysed using data extracted from the Google Routes API. As described in the feature extraction section, traffic information was approximated by simulating a future date at the same accident time and location, due to the lack of historical traffic data support in the API. As illustrated in Figure 15, the majority of accidents occurred during periods classified as low traffic. Notably, no fatal accidents were observed during periods of heavy traffic.

This pattern is intuitively plausible, as lower traffic density is often associated with higher vehicle speeds, which may increase the likelihood of severe outcomes when accidents occur. However, given the indirect and simulated nature of the traffic data, there is uncertainty regarding the reliability

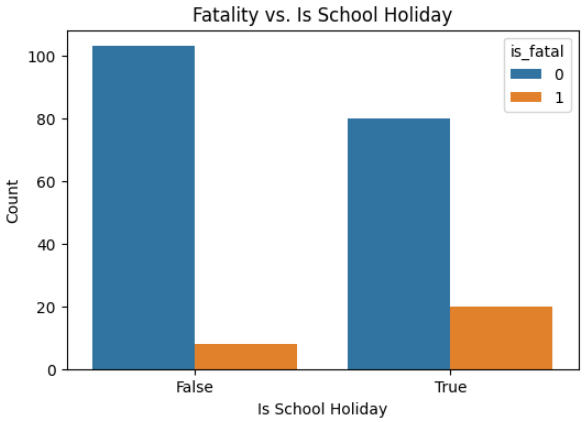

FIGURE 16. Distribution of fatal and non-fatal accidents by school holiday status.

and representativeness of this feature. Consequently, traffic level was excluded from the final feature set to avoid introducing potentially misleading information into the modelling process.

Analysis was also conducted on the special-day features extracted during preprocessing. With respect to overall accident frequency, a larger proportion of accidents occurred on weekdays (72%) compared to weekends (28%). However, no clear difference was observed in fatal versus non-fatal accident rates when comparing weekend and weekday accidents.

Public holidays accounted for a relatively small proportion of accidents, with only 4.3% of recorded accidents occurring on such days. Interestingly, as illustrated in Figure 16, although the majority of accidents occurred outside of school holiday periods (52.6%), accidents that occurred during school holidays (47.4%) exhibited higher fatality rates. It was noted that the school holiday category includes extended breaks such as Christmas, Easter, and the summer recess, during which changes in traffic volume and road usage patterns may influence accident severity.

To conclude the exploratory data analysis, a correlation matrix was computed to examine correlations between the engineered features and the target variable (*is_fatal*). As depicted in Figure 17, the features exhibiting the strongest correlations with fatal accident outcomes were the number of drivers aged between 18 and 24 involved in an accident, whether the accident occurred during a school holiday period, and temperature-related variables.

It is important to note that these correlations indicate association rather than causation. Nevertheless, the results provided useful guidance for feature selection and subsequent modelling decisions. In particular, temperature-related features were consolidated by retaining the mean temperature while dropping the remaining temperature columns to reduce redundancy and multicollinearity.

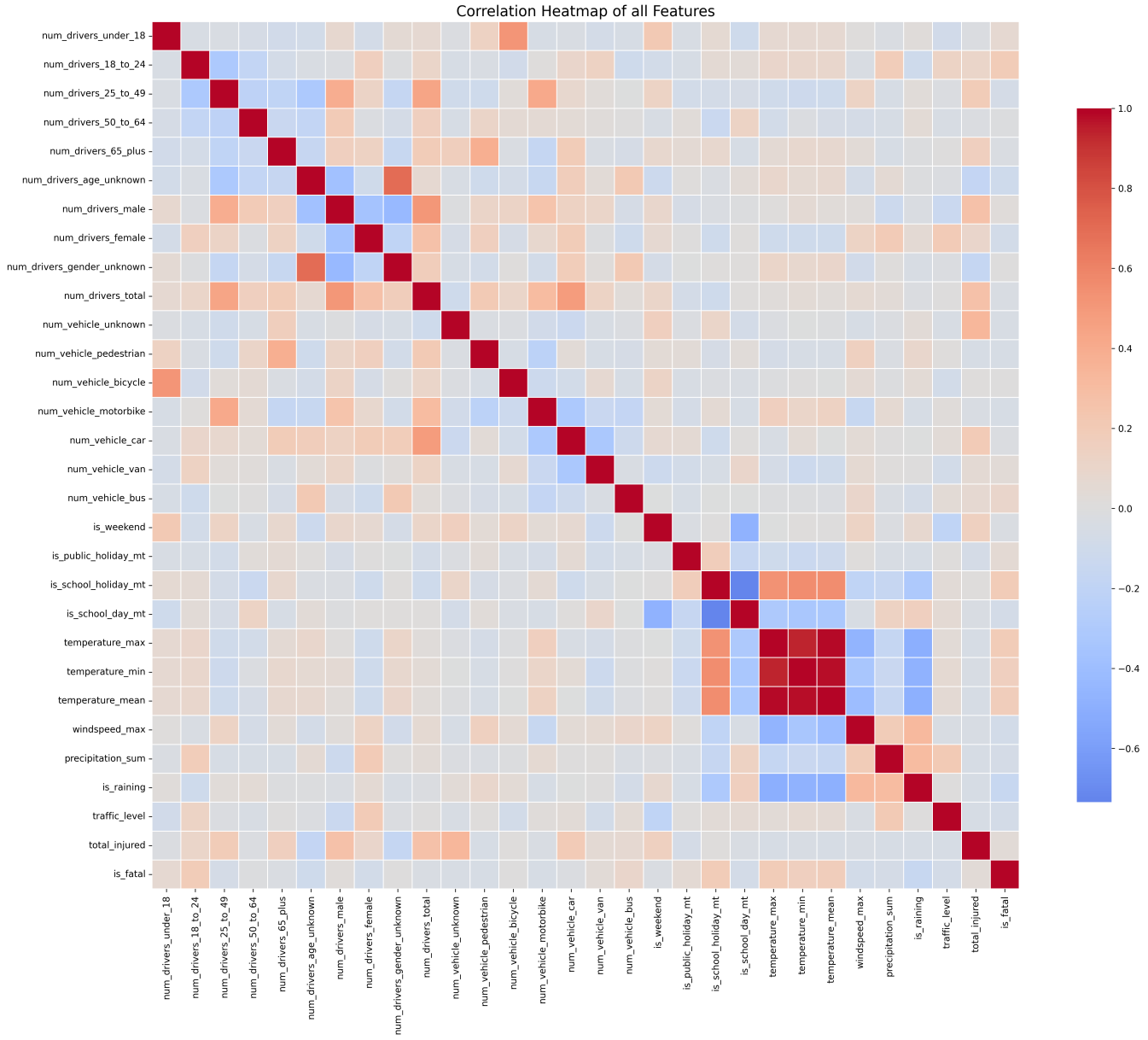

FIGURE 17. Correlation matrix of all extracted features.

E. TASK JUSTIFICATION

Because the dataset is derived from police and news narratives, minor incidents (e.g., low-impact collisions) are under-represented. As a result, modelling multi-class accident severity would conflate injury outcomes with editorial and reporting selection effects. We therefore formulate the primary modelling task as binary classification: predicting whether a reported incident is fatal versus non-fatal. This formulation is aligned with the available labels and supports consistent evaluation under class imbalance, where false negatives (missed fatal incidents) are considered more costly than false positives.

IV. METHODOLOGY

A. LOGISTIC REGRESSION

Logistic regression was selected as the primary baseline model due to its interpretability, robustness on small datasets, and widespread use in traffic safety and injury outcome modelling. Unlike non-linear ensemble models, logistic regression provides directly interpretable coefficients that quantify the direction and magnitude of association between explanatory variables and fatality risk, making it suitable for exploratory and policy-relevant analysis. Results of this analysis are reported in Section V-A1.

1) Preprocessing and Pipeline Design

The model was implemented using a scikit-learn Pipeline architecture to prevent data leakage between preprocessing

and model training. Categorical variables were encoded using one-hot encoding, while numerical features were retained in their original scale due to their limited dynamic range. A stratified train–test split was applied to preserve the proportion of fatal and non-fatal incidents across subsets.

2) Class Imbalance Handling

Given the strong class imbalance in the fatality target variable, class-weighted logistic regression was employed using `class_weight="balanced"`, which penalises misclassification of fatal incidents more heavily than non-fatal incidents. This approach improves model sensitivity to rare fatal outcomes and reduces majority-class bias.

3) Hyperparameter Optimisation

Hyperparameter tuning was performed using grid search with 5-fold stratified cross-validation. The regularisation strength parameter C and penalty type (L1 and L2 regularisation) were explored to control model complexity and reduce overfitting. Model selection prioritised F1-score and recall for fatal incidents, reflecting the safety-critical nature of the prediction task.

4) Threshold Optimisation and Evaluation

Due to the rarity of fatal outcomes, the default probability threshold of 0.5 was found to be suboptimal. A threshold selection procedure was implemented using out-of-fold predicted probabilities to maximise the F1-score, balancing precision and recall. The tuned threshold was subsequently applied to the held-out test set to evaluate generalisation performance.

Model performance was evaluated using Receiver Operating Characteristic Area Under the Curve (ROC–AUC), Precision–Recall AUC (PR–AUC), precision, recall, F1-score, and confusion matrices. ROC–AUC was used to assess ranking performance independent of the classification threshold, while PR–AUC was reported due to the extreme rarity of fatal outcomes. Confusion matrices were generated for both the default and tuned thresholds to visualise the precision–recall trade-off.

5) Coefficient-Based Interpretability Analysis

Logistic regression coefficients were analysed to identify features associated with increased or decreased fatality risk. Positive coefficients indicate features associated with higher fatality likelihood, while negative coefficients suggest protective or lower-risk associations.

The highest-magnitude coefficients corresponded primarily to accident context and participant composition features, including vehicle mix, road type, and driver demographic composition. Weather-related features exhibited comparatively weak coefficients, indicating limited incremental predictive value once temporal, spatial, and contextual factors were accounted for. These findings are consistent with prior literature, which emphasises behavioural and infrastructural determinants over short-term environmental conditions [1].

6) Model Limitations

Logistic regression assumes linear relationships between explanatory variables and the log-odds of fatality, which may not capture complex non-linear interactions present in road-traffic systems. Coefficient estimates may also be unstable due to the limited dataset size and correlated explanatory variables. Nevertheless, the model provides a transparent and auditable baseline for fatality risk modelling and serves as a reference point for comparison with non-linear machine learning models.

B. K- NEAREST NEIGHBOURS (KNN)

KNN was modelled as a classifier decision maker throughout the process focusing on predicting either accident severity class Low (0) or High (1) or accident fatality (Yes/No). The model was developed using a standard Scikit-learn Phyton library [8].

1) Preprocessing and Feature Reduction

The dataset was pre-processed as follows:

- Label encoding of two target variables mentioned above.
- One-hot encoding of categorical features.
- Encoding of the ordinal feature (traffic intensity).
- Convert to numerical data all features including boolean data.

Further exploratory analysis was carried out with visualisation, including a Mutual Information (MI) matrix to identify high dependency between features. Consequently, the following features were omitted from the final evaluation:

- num_drivers_gender_unknown
- accident_month
- accident_day_of_week

2) Algorithm design and k-parameter optimisation

The selected features were grouped (temporal, driver, vehicle, geo-location and weather related), then the dataset was split into training, testing and validation sets and each set normalized separately . The training and testing sets were used to optimize the 'k' hyperparameter and the last validation set was used for model evaluation. The total feature set retained for training consisted in 37 features, 2 boolean labels and a total of 211 observations.

Given the small size of dataset no search optimization techniques were used. While this is a known challenge for KNN, when working with large datasets, this was deemed out of scope. Fifteen (15) different test scenarios were designed and implemented. In training KNN for an optimized 'k' parameter, different features and target labels were assessed. The objective of each of the different scenarios was to maximize the performance of each predictor. In most of the testing cases a 70/20/10 train/test/validate splitting, however a 5-fold stratified testing model was also adopted in some of these experiments.

Note: Euclidean distance was used throughout the training.

3) Experiments' Layout

In each experiment (4.1 to 4.15) the following steps were followed:

- Model training –> ie, in the case of KNN measuring performance for a set of different 'k'-values.
- Determine the optimal 'k'.
- Re-computing model performance with best 'k' using the unexplored dataset.
- Measuring accuracy.
- Plot the Confusion Matrix.
- Compute and summarize key metrics; Precision, Recall, f1- score for each label class.
- Plot ROC and compute the AUC.
- Store the performance metrics in a dictionary for test-to-test comparison.

Note: The last test, 4.15, was performed using a regression method to predict the number of total_injured people in a given accident. This experiment was performed to explore KNN functionality for regression purposes. However, the calculation of the Mean Square Error (MSE) was not used for further evaluation as the experiment was performed only for exploration.

4) Evaluation

'Accuracy' score was repeatedly computed to evaluate each training phase to determine the optimal 'k' parameter. The F-Score, precision and recall for each target label class (low or High) were measured for each test for cross-comparison purposes. Summary of metrics:

1) 'Class 0 Precision'.
2) 'Class 1 Precision'.
3) 'Class 0 Recall'.
4) 'Class 1 Recall'.
5) 'Class 0 F1 Score'.
6) 'Class 1 F1 Score'.
7) 'AUC'.

The ROC and the area under the graph/chart (AUG) metric was also recorded for each experiment.

C. RANDOM FOREST (RF)

Random Forest proved to be a suitable ML model in this project due to its ability to handle mixed feature types, robustness to non-linear relationships, and built-in resistance to over-fitting through ensemble learning. The model naturally accommodates both categorical and numerical features without requiring extensive preprocessing, making it well suited for accident severity prediction where complex interactions between driver demographics, vehicle characteristics, weather conditions, and temporal factors may influence outcomes.

1) Preprocessing and Pipeline Design

The model was implemented using a scikit-learn Pipeline architecture that combines preprocessing and classifier training into a unified workflow. Categorical features (*region*, *street_type*, and *accident_time_category*) underwent one-hot encoding to convert text categories into binary indicators. Numerical features were imputed using median values to handle any potential missing data. Although the current dataset contains no missing values, these imputation steps were included to ensure pipeline robustness and maintain consistency with preprocessing standards across all models in the comparison framework.

2) Class Imbalance Handling

To address the severe class imbalance between fatal and non-fatal accidents, class weighting was applied with a ratio of 1:3 in favour of the minority class representing fatal accidents. This configuration penalises misclassification of fatal cases more heavily during training, thereby improving model sensitivity to rare but critical outcomes. In addition, a stratified 80/20 train-validation split was used to preserve the original class distribution across both subsets.

3) Hyperparameter Optimisation

Hyperparameter tuning was performed using randomised search combined with 5-fold stratified cross-validation. The explored parameter space included:

- Number of estimators: {200, 300, 500}.
- Maximum tree depth: {*None*, 6, 10, 14, 20}.
- Minimum samples required for node splitting and leaf nodes.
- Feature sampling strategies: {*sqrt*, *log2*, 0.3, 0.5}.
- Bootstrap sampling fractions: {0.6, 0.8, 1.0}.
- Class weight configurations.

Model selection prioritised the ROC-AUC as the primary evaluation metric, as it assesses ranking performance independently of the classification threshold. The optimal configuration, consisting of 300 estimators, a maximum tree depth of 20, and class weights {0:1, 1:3}, achieved the highest cross-validated ROC-AUC while maintaining an acceptable train-test gap of at most 0.05, thereby reducing the risk of overfitting.

4) Threshold Optimisation and Evaluation

A custom probability threshold of 0.3 was adopted instead of the default value of 0.5 in order to prioritise recall for fatal accidents. This decision reflects the application context, where false negatives are substantially more costly than false positives. Lowering the threshold increased model sensitivity at the expense of precision, resulting in higher false positive rates but improved detection of critical outcomes.

Model performance was evaluated using multiple metrics, including ROC-AUC, accuracy, precision, recall, and F1-score, computed via 5-fold stratified cross-validation. Confusion matrices were generated for both the default threshold of 0.5 and the custom threshold of 0.3 to visualise classification outcomes. Feature importance analysis was conducted to identify the most influential predictors, providing interpretability and insight into the factors most strongly associated with fatality risk.

D. SUPPORT VECTOR MACHINE (SVM)

A SVM is a supervised learning algorithm widely used for classification and regression tasks. Its primary objective is to identify an optimal decision boundary, also known as a hyperplane, that separates classes while maximising the margin between the closest data points, known as support vectors. Maximising this margin improves generalisation to unseen data.

For linearly separable data, SVM constructs a linear hyperplane in the original feature space. When data is not linearly separable, SVM employs the kernel trick to implicitly project data into a higher-dimensional space, enabling the modelling of non-linear decision boundaries. Common kernels include linear, polynomial, radial basis function (RBF), and sigmoid kernels. The function choice is important as each one forms a different shape of the decision boundary that separates the data points plotted in high dimension.

Model complexity in an SVM is primarily controlled by the regularisation parameter C, which determines the trade-off between maximising the margin and minimising classification error on the training data. A smaller value of C places greater emphasis on margin maximisation, allowing some training samples to be misclassified or lie within the margin. Conversely, a larger value of C penalises misclassification more heavily, forcing the model to fit the training data more closely by reducing margin violations. While this can improve training performance, it increases the risk of overfitting, particularly in the presence of noisy or limited data. Consequently, careful tuning of C is essential to achieve an appropriate balance between bias and variance.

1) Preprocessing and Pipeline Design

Prior to model training, the dataset was prepared through a structured preprocessing pipeline to ensure compatibility with the SVM classifier. Relevant features identified during the exploratory data analysis phase were retained, and the target variable (*is_fatal*) was separated from the input features. Numerical features were standardised using the *StandardScaler* function from *scikit-learn* to prevent features with larger scales from disproportionately influencing the SVM optimisation process. Categorical variables (*region*, *street_type* and *accident_time_category*) were one-hot encoded to allow their inclusion in the model. The dataset was then split into 80% training and 20% testing subsets to enable an unbiased evaluation of model performance.

2) Hyperparameter Optimisation

Hyperparameter optimisation was conducted to identify the most suitable configuration for the SVM model. The parameters explored included multiple values of the regularisation parameter C and a range of kernel functions. The training set was further divided into 75% training and 25% validation splits (which is equivalent to 60/20/20 on the whole dataset). Values of C equal to 0.01, 0.1, 0.25, 0.5, 1.0, 2.0, 5.0, and 10.0 were evaluated in combination with linear, sigmoid, RBF and polynomial kernels. Each hyperparameter combination was trained using a grid search and validated 100 times using different training–validation seeds to reduce the impact of random variation introduced by data splitting. The mean accuracy and recall across these repetitions were then computed and analysed for each C and kernel combination. This procedure enabled a systematic assessment of model sensitivity to varying levels of regularisation and decision boundary complexity. Model selection was guided by performance metrics including accuracy and recall, with particular emphasis placed on recall due to the imbalanced nature of the dataset. Based on these experiments, an SVM using an RBF kernel with $C = 10$ was selected, as it provided an effective balance between overall classification performance and the correct identification of minority class instances.

3) Overall Evaluation

Confusion matrices were used to examine the distribution of correct and incorrect predictions, providing insight into class-specific performance. In addition to overall accuracy, recall was emphasised due to the imbalanced nature of the target variable and the importance of correctly identifying minority class instances. The F1-score was also reported to capture the balance between precision and recall. Furthermore, Receiver Operating Characteristic and Precision–Recall curves were generated to evaluate the model's discriminative ability across different decision thresholds.

4) Model Limitations

SVMs present several limitations that must be considered in the context of this study. SVM performance is highly sensitive to the choice of kernel and hyperparameters, requiring careful tuning to achieve optimal results. In addition, SVMs do not naturally provide probabilistic outputs, which can limit interpretability in safety-critical applications unless additional calibration steps are applied. The model also offers limited transparency, as the learned decision boundary, particularly when using non-linear kernels, does not readily indicate how individual features contribute to predictions. Furthermore, SVMs can struggle with highly imbalanced datasets, as the optimisation objective prioritises margin maximisation rather than class-specific performance, potentially leading to reduced recall for minority classes such as fatal accidents. These limitations highlight the need for careful evaluation metric selection and cautious interpretation of model outputs.

V. RESULTS

A. LOGISTIC REGRESSION

The logistic regression model achieved a test ROC–AUC of 0.635, indicating modest discriminative performance above random chance. Given the rarity of fatal incidents, Precision–Recall AUC was also reported, yielding a PR–AUC of 0.193, reflecting the difficulty of rare-event prediction.

At the default probability threshold of 0.5, the model achieved a recall of 0.83 for fatal incidents, but with low precision (0.24), yielding an F1-score of 0.37. Threshold optimisation using cross-validated F1-score selected a probability

threshold of 0.607. While this improved F1-score in cross-validation, applying the tuned threshold on the held-out test set reduced recall to 0.67 (precision 0.22; F1-score 0.33). This highlights the instability expected under small-sample, imbalanced learning and suggests that threshold selection should be validated carefully on unseen data.

ROC and Precision–Recall curves are shown in Figures 18 and 19, while confusion matrices for both the default and tuned thresholds are presented in Figures 20 and 21.

Compared to non-linear models, logistic regression achieved lower predictive performance but provided transparent coefficient-based explanations, supporting its role as an interpretable baseline model.

1) Coefficient Interpretation

To support interpretability, we analysed the learned logistic regression coefficients. In logistic regression, coefficients represent the change in the log-odds of fatality associated with a one-unit increase in the corresponding feature, conditional on all other variables. Positive coefficients indicate features associated with higher fatality likelihood, while negative coefficients indicate lower fatality likelihood. For categorical variables encoded via one-hot encoding, coefficients are interpreted relative to the omitted reference category.

Table 3 reports the strongest coefficients by absolute magnitude. Higher fatality likelihood was associated with late-evening and afternoon time periods, younger and middle-aged drivers (18–49), school holiday periods, and certain regions (Central and Southern Harbour). These patterns are consistent with exploratory analysis and prior literature, where increased traffic volume, behavioural factors, and higher-speed road environments are linked to severe outcomes.

Exponentiating coefficients yields odds ratios. For example, the coefficient of 1.35 for drivers aged 18–24 corresponds to an approximate 3.86× increase in the odds of fatality, conditional on an incident being reported and all other features being held constant.

Conversely, early-morning and morning periods, accidents occurring in Gozo, and Northern region incidents exhibited negative coefficients, suggesting lower fatality likelihood among reported incidents. This may reflect differences in traffic density, road geometry, and driving behaviour in these contexts.

Several coefficients, such as the negative association for the number of cars and motorbikes involved, should not be interpreted causally. These likely reflect reporting and selection biases in media-based data, where multi-vehicle collisions may be more frequently reported even when injuries are non-fatal, while severe single-vehicle crashes receive disproportionate coverage.

Weather-related variables, including rainfall and temperature, exhibited smaller coefficient magnitudes compared to temporal, spatial, and participant composition features. This suggests limited incremental predictive value once contextual factors are controlled for, consistent with findings in prior traffic safety research.

Coefficient estimates should be interpreted cautiously due to the small dataset size, correlated explanatory variables, and the absence of feature standardisation, which limits direct comparability of coefficient magnitudes across feature types. Nevertheless, the model provides interpretable insights into factors associated with fatal outcomes among reported traffic incidents.

TABLE 3. Top logistic regression coefficients by absolute magnitude (fatal vs. non-fatal among reported incidents).

Feature	Coefficient
accident_time_category_early_morning	-1.579303
region_Gozo	-1.212201
num_vehicle_car	-0.729074
num_vehicle_motorbike	-0.681767
num_drivers_male	-0.662783
num_drivers_female	-0.620547
accident_time_category_morning	-0.587265
num_vehicle_van	-0.436429
region_Northern	-0.363993
is_raining	-0.314290
num_vehicle_pedestrian	-0.224860
is_weekend	-0.221434
temperature_mean	0.200947
num_drivers_under_18	0.245486
is_school_holiday_mt	0.302100
accident_time_category_evening	0.371978
num_drivers_65_plus	0.469983
region_Central	0.470027
region_Southern_Harbour	0.485460
accident_time_category_afternoon	0.513063
accident_time_category_late_evening	0.740952
num_drivers_50_to_64	1.034511
num_drivers_18_to_24	1.347438
num_drivers_25_to_49	1.528350

B. K- NEAREST NEIGHBOURS (KNN) RESULTS

As KNN is a distance-based ML algorithm and relies on finding the best positioning of the predicted value given a n-dimensional feature space, it was not possible to evaluate the results based on weights assigned to each feature. Thus, a set of experiments was designed to identify if/how different features contribute in predicting the severity and fatality of the traffic accidents. In this section, we base our results on the target label predicting if an accident is fatal or not given the set features. It was envisaged that KNN performs best when injected a simple feature set due to curse of dimensionality, a common challenge when exploring KNN. This hypothec aligns to the evaluation of the other models implemented and was deemed the most suitable given the high imbalance of the dataset. Table 4 depicts the most significant test results out of the total fifteen experiments carried out.

The following key parameters were varied during these experiments:

- 4.2 - 'Fatality' classifier using full feature set.
- 4.4 - 'Fatality' classifier using a reduced feature set.
- 4.6 - 'Fatality' classifier based on vehicles involved.
- 4.8 - 'Fatality' classifier based on location/day of accident.
- 4.10 - 'Fatality' classifier based on day and time of the accident.

TABLE 4. KNN Tests' Results - Summary

Test	k	Acc.	C0 Pr.	C1 Pr.	C0 Rc.	C1 Rc.	C0 F1	C1 F1	AUC
T4.2	1	.82	.89	.33	.89	.33	.89	.33	.61
T4.4	7	.86	.86	.00	1.0	.00	.93	.00	.53
T4.6	15	.86	.86	.00	1.0	.00	.93	.00	.39
T4.8	17	.86	.86	.00	1.0	.00	.93	.00	.55
T4.10	5	.86	.86	.00	1.0	.00	.93	.00	.54
T4.12	3	.77	.85	.00	.89	.00	.87	.00	.56
T4.13	7	.86	.88	.50	.97	.17	.92	.25	.66
T4.14	9	.86	.86	.00	1.0	.00	.92	.00	.62

- 4.12 - 'Fatality' classifier based on an aggressively reduced feature set.
- 4.13 - 'Fatality' classifier based on a standardised feature set used to evaluate other ML models during this project.
- 4.14 - 'Fatality' classifier based on PCA transformation of test 4.13 in 5-dimensions.

From the results achieved, the f1 score was reported separately for the two label classes. In this context we interpret the model's performance both when predicting if an accident represents a fatality risk, (class-C1), as well as the opposite hypothesis of predicting the non-fatality scenario, (class-C0). The best f-score achieved to predict fatality was in tests 4.2 and 4.13 respectively. This implies that a 0.33 score was achieved when using all feature set available and 0.25 score was achieved when using the reduced data set as agreed with the research team to adopt as a standard for cross-model evaluation. In terms of precision a score of > 0.85 was consistently achieved when predicting the NULL class, that is, not classifying an incident as fatal. This was best performing when features used when minimal as in tests 4.4, 4.6, 4.8 and 4.10. When interpreting recall, the best results were registered in tests 4.2 and 4.13 with scores of 0.33 and 0.17 respectively. The highest recall was registered when the complete feature set was used. However, both these recall scores are considered to be under-performing. The best AUC score achieved from our testing scheme was of 0.66 in test 4.13, when evaluating the feature set standardised by the research team. The various tests performed to train a KNN classifier did not yield a significant positive result as could be seen in the overall f1 score / AUC table. From the literature review and experimental deduction, we attribute this challenge to two main issues:

- Data scarcity and imbalance. Our data set is very limited and also highly imbalanced.
- There is no clear segregation between data and thus a distance base model will struggle to distinguish between the two classes. In addition, experimentation carried out by changing the 'Euclidian' distance measure to 'Manhattan' did not improve the f1 score for either class.

1) KNN Results' Conclusion

In summary the following key take-outs were deducted from these experiments:

- Reducing the feature set from 37, to 20 improved the performance
- It was observed that scaling/normalising the data resulted in a change in performance
- In all cases, the f1 score for both negative class (non-fatal) and positive class was treated separately. It was decided to judge on the f1 scores separately since the data set is not balanced and contains only few fatality incidents as compared to the complete dataset.
- The results indicated that while precision/recall are 85% for the non-fatal classification the same metrics scored very poor for the positive class. In simple terms the model is classifying any new observations as a non-fatal preferring to predict the majority class of the data set.
- Same interpretation is concluded from the ROC charts where decay falls rapidly from the top right corner and consequently the AUC scored very poor.

C. RANDOM FOREST RESULTS

The Random Forest classifier demonstrated strong predictive performance across multiple evaluation metrics. On the hold-out test set, the model achieved an ROC–AUC of **0.964**, as seen in Figure 18 indicating excellent discriminative ability between fatal and non-fatal traffic accidents. Overall classification accuracy reached **88%**, confirming reliable generalisation to unseen data.

Notably, the model achieved a **recall of 83.3%** for fatal accidents, indicating that the majority of fatal incidents were correctly identified. This characteristic is particularly important in safety-critical contexts, where failing to detect severe outcomes carries substantial consequences. At the same time, the model maintained a high **specificity of 97.2%**, effectively limiting false positive classifications for non-fatal accidents.

Compared to a baseline dummy classifier, the Random Forest model demonstrated a substantial improvement in predictive power, with ROC–AUC increasing from **0.50** to **0.89**, confirming that the model captures meaningful structure beyond random guessing.

1) Feature Importance and Interpretability

An additional advantage of the Random Forest model is its ability to provide **feature importance estimates**. Table 5 shows the top contributing features for predicting accident severity, implemented through scikit-learn RandomForest-Classifier [8]. This is found using the Mean Decrease in Impurity (MDI) method [9], where for each feature, the importance score is computed as the total reduction in node impurity (measured by Gini index for classification tasks) weighted by the probability of reaching each node, aggregated across all decision trees in the ensemble. Introduced by Breiman [10], the importance of feature is calculated as:

$$\text{Importance}(j) = \frac{1}{N_T} \sum_{t=1}^{N_T} \sum_{n \in t} p(n) \cdot \Delta i(n, j) \qquad (1)$$

where N_T is the number of trees, $p(n)$ is the proportion of samples reaching node n, and $\Delta i(n, j)$ is the decrease in impurity at node n if feature j is used for splitting.

TABLE 5. Random Forest feature importance for accident severity prediction.

Feature	Importance
Temperature mean	20%
Number of cars involved	10%
School holiday period	6%
Trunk road location	6%
Number of drivers aged 18–24	5%
Number of male drivers	4%
Southern Harbour region	4%
Number of drivers aged 25–49	3%
Afternoon accidents	3%
Weekend occurrence	3%

The results in table 5 indicate that environmental factors (temperature), vehicle involvement, temporal patterns, and driver demographics are key determinants of accident severity. This helped with the feature selection, ensuring the best feature set to be chosen during model training resulting in the most optimal accuracy possible.

2) Threshold Selection and Trade-offs

As explained in the methodology section, a custom classification threshold of **0.3** was applied to prioritise recall over precision. This resulted in a precision of **62.5%** for fatal accident predictions, implying an increased false positive rate. However, this trade-off is justified given the asymmetric cost of classification errors, where false negatives pose a significantly greater risk than false positives in emergency response scenarios.

3) Overfitting Considerations

Cross-validation analysis indicates a degree of overfitting, with a training ROC–AUC of **1.00** compared to a test ROC–AUC of **0.89**. This performance gap suggests that the model may partially capture dataset-specific patterns, and performance may degrade slightly when applied to completely unseen data.

D. SVM RESULTS

After hyperparameter tuning the SVM on the training and validation sets, the selected model, using an RBF kernel with$C = 10$, was trained on the full 80% training set and evaluated on the remaining 20% held-out test set. Due to the substantial class imbalance in the dataset, with significantly more non-fatal than fatal accidents, model performance was initially compared against the null accuracy baseline. Null accuracy represents the accuracy obtained by always predicting the majority class and, in this case, was 86%.

The SVM achieved an accuracy of 88.4%, representing a modest improvement over the null accuracy baseline. However, given the imbalanced nature of the target variable, accuracy alone is insufficient to fully assess model performance. Therefore, additional evaluation metrics were considered. The model achieved a precision of 66.7%, a recall (sensitivity) of 33.3%, and an F1 score of 44.4%. Furthermore, the ROC–AUC was 89.6% and the Precision–Recall AUC was 65.8%.

These results indicate that while the model is effective at distinguishing between fatal and non-fatal accidents overall, as reflected by the high ROC–AUC, it struggles to identify a large proportion of fatal cases. The low recall suggests that many fatal accidents are misclassified as non-fatal, which is a common challenge when learning from imbalanced datasets with limited positive examples. At the same time, the relatively high precision indicates that when the model does predict a fatal accident, it is often correct. This trade-off highlights the importance of prioritising recall-oriented evaluation metrics in safety-critical applications, where false negatives are generally more costly than false positives.

Due to the use of the RBF kernel, model explainability is limited, as it is not straightforward to determine which features contribute most to the final predictions. The RBF kernel implicitly projects data points into a higher-dimensional feature space, and unlike linear kernels, it does not provide directly interpretable feature weights. As a result, understanding the influence of individual input features is more challenging.

In this study, due to the relatively small dataset, training, validation, and testing times were minimal. However, for larger datasets, SVMs, particularly those using non-linear kernels, can become computationally expensive in terms of both time and memory, as training involves operations on large kernel matrices and numerous matrix multiplications.

VI. ETHICAL CONSIDERATIONS

This study relies on accident narratives extracted from police press releases and news media reports. Such sources reflect editorial and institutional reporting practices rather than a complete or unbiased representation of road-traffic incidents. Minor collisions are systematically under-reported, while severe and fatal incidents receive disproportionate coverage. Consequently, the dataset captures reported fatality risk rather than population-level accident severity, and all results must be interpreted within this conditional framing.

Demographic, geographic, and temporal features may act as proxies for unobserved socioeconomic and infrastructural factors. For example, regional differences may correlate with road quality, enforcement intensity, or population density rather than intrinsic risk. Similarly, driver age and gender variables may reflect exposure patterns or reporting practices rather than causal mechanisms. The inclusion of such variables raises concerns regarding fairness and potential discriminatory inference if models were deployed without contextual safeguards.

Weather and environmental features were incorporated for exploratory purposes, but these variables may also proxy seasonal tourism, traffic volume, or behavioural changes rather than direct causal effects. Additionally, simulated traffic features were excluded from modelling due to concerns about validity and representativeness, highlighting the importance of avoiding synthetic or speculative data in safety-critical systems.

Automated predictive systems for traffic fatality risk could influence policy, resource allocation, or enforcement decisions. Misuse or overinterpretation of such models may lead to biased interventions, disproportionate surveillance of specific regions or demographic groups, or incorrect causal conclusions. Therefore, this work is positioned as an exploratory and analytical tool rather than a deployable decision-support system. Any real-world deployment would require robust validation on representative administrative datasets, continuous monitoring for bias, and governance frameworks ensuring transparency and human oversight.

VII. DISCUSSION

A. MODEL PERFORMANCE COMPARISON

The final notebook evaluates the four ML classifiers through a unified comparison framework designed to ensure methodological rigour and fair assessment. Each model is implemented as a `scikit-learn` Pipeline that encapsulates both preprocessing and classification stages, guaranteeing consistent data transformations across all algorithms.

All four models utilise an identical feature set comprising 20 predictive variables: 3 categorical features (*region, street_type, accident_time_category*) and 17 numerical features covering driver demographics (age groups and gender distributions), vehicle type counts (motorbike, car, van, bicycle, pedestrian), temporal indicators (weekend status, school holidays, time of day), and weather conditions (temperature and rainfall). These features were selected after evaluating the findings of the EDA and performing feature importance analysis at a model level.

The preprocessing pipelines are tailored to each algorithm's requirements:

- **Random Forest (RF):** Receives median-imputed features without scaling, as tree-based methods are scale-invariant.
- **SVM, KNN, and Logistic Regression:** Receive `StandardScaler`-normalised features, as distance-based and gradient-based algorithms are sensitive to feature magnitudes.

To enable direct performance comparison, all models are trained and evaluated on an identical 80/20 stratified train-test split ($random\_state = 42$), ensuring the validation set maintains the same class distribution as the full dataset. Specifically for KNN evaluation the 'k' neighbour parameter was set to '7' from previous analysis, test 4.13 in Table 4. The evaluation framework produces comprehensive outputs including confusion matrices at two classification thresholds (default 0.5 and safety-oriented 0.3), ROC curves with AUC scores, and side-by-side metric comparisons across five key performance indicators:

- ROC-AUC.
- Accuracy.
- Precision.
- Recall.
- F1 score.

This standardised approach ensures that observed performance differences reflect genuine algorithmic capabilities rather than artifacts of inconsistent data handling, feature engineering, or evaluation methodologies.

1) ROC Curve Analysis

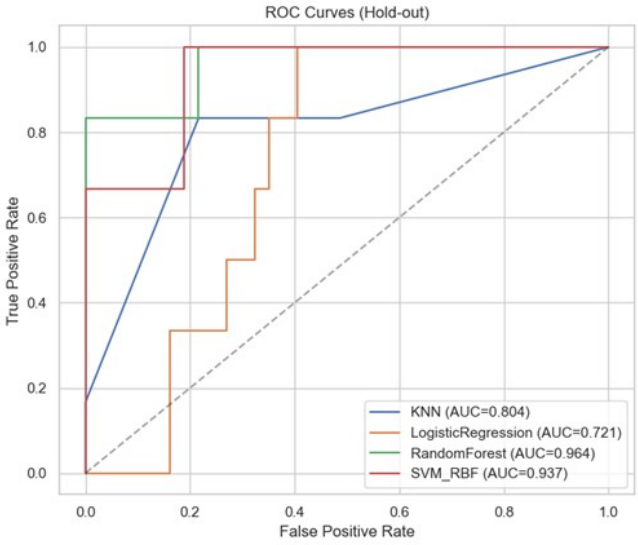

FIGURE 18. ROC Curve of the 4 models

The ROC curve analysis seen in Figure 18 reveals a clear performance hierarchy among the four classifiers. **Random Forest** achieved the highest discriminative power with an $AUC = 0.964$, demonstrating exceptional ability to distinguish between fatal and non-fatal accidents across all classification thresholds. **SVM** (with RBF kernel) followed closely with an $AUC = 0.937$, indicating that its non-linear decision boundary effectively captured complex feature interactions in the accident data.

KNN performed moderately with an $AUC = 0.804$, suggesting that the distance-based approach captured some predictive signal but struggled with the high-dimensional feature space and class imbalance. **Logistic Regression** achieved the lowest $AUC = 0.721$, reflecting the limitations of its linear decision boundary in modeling the non-linear relationships inherent in accident severity prediction.

The substantial gap between Random Forest and Logistic Regression ($\Delta AUC = 0.243$) quantifies the performance gain from ensemble methods and non-linear modeling capabilities, while the minimal difference between Random Forest and SVM ($\Delta AUC = 0.027$) suggests both algorithms successfully captured similar discriminative patterns despite their architectural differences.

2) Hold-Out Test Comparison

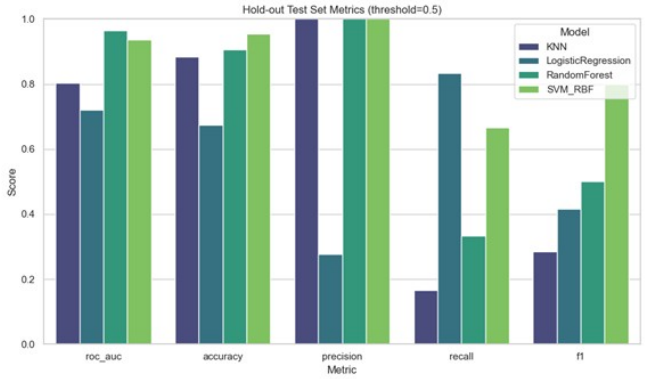

FIGURE 19. Hold-Out Test Set metrics of the 4 models

The hold-out test set metrics at the default 0.5 threshold (see Figure 19) reveal distinct trade-offs across models.

Random Forest and **SVM** achieved the best overall performance. The Random Forest had a 90.7% accuracy, perfect precision (100%), moderate recall (33.3%), and an $F1$ score of 0.50, indicating its predictions are highly reliable but conservative in identifying fatal accidents. **SVM** demonstrated similar characteristics with 95.3% accuracy, perfect precision (100%), better recall (66.7%) and an $F1$ score of 80%, suggesting it successfully identified two-thirds of fatal accidents without generating false alarms.

KNN achieved 88.4% accuracy and perfect precision (100%) but the lowest recall (16.7%) and $F1$ score (0.29), reflecting an extremely conservative classification strategy. Conversely, **Logistic Regression** showed balanced but sub-optimal performance with 67.4% accuracy, 27.8% precision, 83.3% recall, and an $F1$ score of 0.42, demonstrating a tendency to over-predict fatal accidents (13 FP vs. 5 TP).

The precision-recall trade-off is particularly evident when comparing SVM's conservative approach (4 TP, 2 FN, 0 FP) with Logistic Regression's aggressive strategy (5 TP, 1 FN, 13 FP), highlighting the importance of threshold selection based on operational priorities.

3) Confusion Matrix with Default Threshold

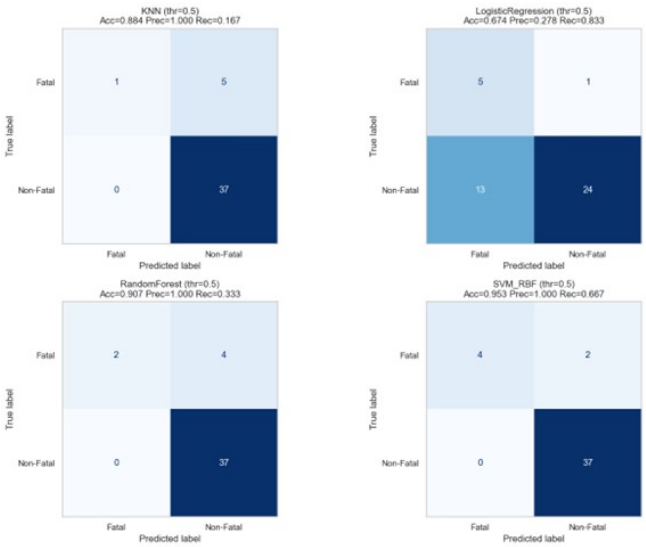

FIGURE 20. Confusion Matrix of the 4 models with Default Threshold

The confusion matrices (see Figure 20) at the 0.5 threshold illustrate each model's error patterns. **Random Forest** correctly classified 37 non-fatal accidents (TN) and 2 fatal accidents (TP) while producing zero false positives ($FP = 0$) and 4 false negatives ($FN = 4$), demonstrating high specificity (100%) but limited sensitivity (33.3%).

SVM achieved an almost identical confusion matrix with 4 TP, 2 FN, 0 FP, and 37 TN, resulting in perfect specificity and improved recall (66.7%). **KNN**'s extreme conservatism is evident in its confusion matrix showing only 1 TP against 5 FN, effectively defaulting to predicting non-fatal outcomes in ambiguous cases.

Conversely, **Logistic Regression**'s aggressive classification yielded 5 TP and 1 FN (83.3% recall) but at the cost of 13 FP, reducing precision to 27.8% and misclassifying over one-third of non-fatal accidents as fatal.

4) Confusion Matrix with 0.3 Threshold

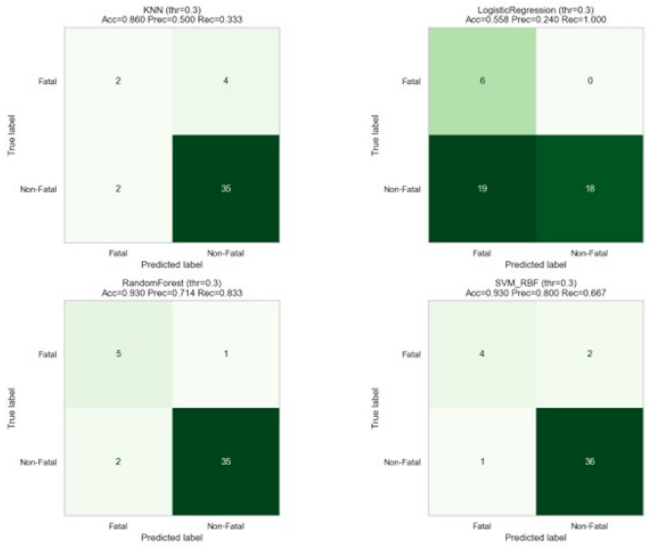

FIGURE 21. Confusion Matrix of the 4 models with 0.3 Threshold

After adjusting the classification threshold to 0.3 (see Figure 21), the precision-recall balance altered dramatically, demonstrating the practical importance of threshold optimisation for safety-critical applications.

Random Forest's recall improved from 33.3% to 83.3% (5 TP vs. 1 FN), identifying five out of six fatal accidents, while precision decreased from 100% to 71.4% due to 2 additional false positives. **SVM** maintained the same previous recall (66.7%) but precision dropped from 100% to 80.0% with 1 false positive.

KNN showed the most dramatic shift; recall increased from 16.7% to 33.3% (2 TP vs. 4 FN) and precision decreased from 100% to 50.0%, though the model remained overly conservative. **Logistic Regression** achieved perfect recall (100%, 6 TP, 0 FN) but at a severe precision cost (24.0%), generating 19 false positives and effectively flagging over half of all accidents as potentially fatal.

These threshold-dependent results underscore that algorithmic improvements alone cannot substitute for thoughtful operating point selection. The choice between the 0.5 and 0.3 thresholds represents a policy decision regarding the acceptable trade-off between false alarms and missed fatal accidents, with the 0.3 threshold prioritising public safety by reducing missed detections at the expense of increased false alarm rates.

B. EVALUATING STUDY OBJECTIVES

This study formulated fatality prediction among reported traffic incidents as a binary classification problem and evaluated multiple machine learning models to address the research questions outlined in Section I-B. The results demonstrate that meaningful patterns can be extracted from unstructured news-derived data, while also highlighting substantial limitations due to data size, bias, and noise.

1) RQ0: Predicting Fatal vs Non-Fatal Outcomes

All evaluated models achieved performance above random chance, confirming that contextual, temporal, spatial, and participant-level features contain predictive signal for fatality among reported incidents. Logistic regression achieved modest discriminative performance (ROC–AUC $\approx$ 0.64), while Random Forest and SVM achieved substantially higher performance, indicating that non-linear interactions and feature combinations play an important role in fatality prediction. KNN exhibited mixed performance, likely due to the high dimensionality and limited dataset size.

2) RQ1: Contribution of Feature Groups

Across models, participant composition (driver age and gender), vehicle composition, road type, and regional indicators were consistently among the most informative features. Logistic regression coefficients and Random Forest feature importance both highlighted the relevance of younger driver age groups, vehicle counts, trunk roads, and specific regions. Weather-related features also showed associations but these should not be interpreted as casual effect.

3) RQ2: Temporal and Geographic Contexts

Exploratory analysis and model outputs identified elevated fatality likelihood during late evening periods, school holiday periods, and in specific regions (e.g., Southern Harbour and Central regions). These findings align with known risk factors such as higher speeds during low-traffic periods and increased exposure during holiday seasons. However, these patterns reflect reported incidents and may be influenced by media coverage practices.

4) RQ3: Incremental Value of Weather Features

Weather variables, including temperature and precipitation, exhibited limited incremental predictive value after controlling for temporal and contextual factors. While temperature correlated with fatality, this likely reflects seasonal exposure patterns rather than a direct causal effect. These findings suggest that weather features contribute marginally compared to behavioural and infrastructural determinants.

5) RQ4: Interpretability vs Predictive Performance

Logistic regression provided transparent and interpretable coefficients, making it suitable for exploratory analysis and policy interpretation. Random Forest and SVM significantly outperformed logistic regression in predictive accuracy but offered less direct interpretability, requiring feature importance measures to explain model behaviour. KNN provided limited interpretability and mixed performance. Overall, the results illustrate the trade-off between interpretability and predictive performance, with ensemble methods capturing complex interactions at the cost of transparency.

6) Limitations

The dataset is small and derived from media sources, leading to selection bias, class imbalance, and unstable coefficient

estimates. Some explanatory variables are correlated, and the lack of feature standardisation limits direct comparison of logistic regression coefficient magnitudes. Simulated traffic data could not be validated and was excluded to avoid misleading conclusions. Therefore, findings should be interpreted as exploratory and hypothesis-generating rather than causal or policy-prescriptive.

VIII. CONCLUSION

This work demonstrates an end-to-end pipeline for transforming unstructured police and media narratives into a structured dataset suitable for machine learning analysis of traffic accidents in Malta. By integrating temporal, geographic, environmental, and participant-level features, the study shows that fatality among reported incidents can be predicted with moderate to high accuracy using traditional machine learning models. Setting up the data pipeline for machine learning was a fundamental part of this research. The journey of structuring the data sets and engineer the features was complex, given the small size of observations, and required various feedback loops to adjust the extraction and validation of the data. Well proven Regex and LLM techniques were employed during this phase however the research team also employed significant effort in manual auditing and reviewing the dataset.

Among the evaluated models, Random Forest and SVM achieved the strongest predictive performance, highlighting the importance of non-linear interactions between features. KNN classification was generally conservative in predicting the non-fatality class and returned low recall. This model struggled to separate the two classes given the relatively large feature-set as compared to the number of observations available. Logistic regression provided interpretable coefficients, revealing associations between fatality and driver demographics, road type, time of day, and regional context, while weather-related variables contributed marginally after controlling for contextual factors. These findings align with established traffic safety literature emphasising behavioural and infrastructural determinants over short-term environmental conditions [1].

The study also underscores critical challenges associated with media-derived datasets, including reporting bias, class imbalance, and limited sample sizes. Future work should incorporate official administrative accident databases, traffic exposure metrics, and longitudinal data to improve robustness and enable causal inference. Advanced techniques such as hierarchical models, survival analysis, and explainable deep learning could further enhance predictive capability and interpretability.

Overall, this research highlights the feasibility and limitations of using unstructured textual sources for traffic safety analytics and provides a transparent framework for exploratory modelling, ethical interpretation, and future data-driven road safety research in Malta and similar contexts.

REFERENCES

[1] I. Perysinakis, A. Spartinou, M.-R. Siligardou, M. Savvides, G. Lianeris, and G. Stamatakis, "Pattern of road traffic injuries in the Rethymnon region, Crete, Greece: A secondary hospital-based study," Rural and Remote Health, vol. 21, no. 4, pp. 1–7, Dec. 2021. [Online]. Available: https://search.informit.org/doi/abs/10.3316/informit.298829925475329

[2] A. Ziakopoulos, E. Michelaraki, D. Nikolaou, K. Folla, and G. Yannis, "Association Rule Mining for Island and Mainland Road Crash Injuries in Greece," Transportation Research Procedia, vol. 72, pp. 163–170, Jan. 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352146523006877

[3] M. Attard, A. S. Bergantino, and M. Intini, "Effects of local urban characteristics and driving behaviour on injuries among pedestrians and cyclists in Malta," Transportation Research Procedia, vol. 82, pp. 81–92, Jan. 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352146524003272

[4] A. Psarras, "Investigation of factors that drive road accidents," Investigation of the determinants of road accidents, 2024, accepted: 2024-03-04T08:44:53Z. [Online]. Available: http://dspace.lib.uom.gr/handle/2159/30172

[5] A. Magri, A. Farrugia, F. Valletta, and S. Grima, "An analysis of the risk factors determining motor insurance premium in a small island state : the case of Malta," 2019, accepted: 2021-04-07T07:51:22Z. [Online]. Available: https://www.um.edu.mt/library/oar/handle/123456789/73114

[6] S. P. Washington, M. G. Karlaftis, and F. L. Mannering, Statistical and Econometric Methods for Transportation Data Analysis. CRC Press, 2011.

[7] H. Mensouri, A. Azmani, and M. Azmani, "Towards an accident severity prediction system with logistic regression," in Proceedings of the International Conference on Advanced Intelligent Systems for Sustainable Development (AI2SD), J. Kacprzyk, M. Ezziyyani, and V. E. Balas, Eds. Cham: Springer, 2023, pp. 396–410.

[8] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.

[9] Scikit-learn Developers, "Feature importance evaluation with a Random Forest," https://scikit-learn.org/stable/auto_examples/ensemble/plot_forest_importances.html, 2024, accessed: 2024-05-22.

[10] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.