Implementing Machine Learning Models for Predicting Road Accident Severity in Northern Ireland

MSc Research Project

Melvin Akash AmbroseDoss
Student ID: x22152601

School of Computing
National College of Ireland

Supervisor: Dr. Anu Sahni

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Melvin Akash AmbroseDoss		
Student ID:	22152601		
Programme:	Msc in Data Analytics	**Year:**	2023 - 2024
Module:	Msc Research Project		
Supervisor:	Dr. Anu Sahni		
Submission Due Date:	14/12/2023		
Project Title:	Implementing Machine Learning Models for Predicting Road Accident Severity in Northern Ireland		
Word Count:	7733	**Page Count**	: 24

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Melvin Akash Ambrose MohanDoss
…

Date: 14th December 2023
…

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Implementing Machine Learning Models for Predicting Road Accident Severity in Northern Ireland

Melvin Akash Ambrose Doss
x22152601

This research project addresses the challenge of enhancing road safety in Northern Ireland by employing advanced machine learning and deep learning techniques to predict road accident severity. By evaluating various algorithms, the project focuses on hyperparameter tuning, a critical step in optimizing the performance of machine learning models. The tuning process is driven by the goal of achieving the highest possible accuracy in predicting accident severity, which is crucial for developing effective model building. The research successfully identifies Random Forest and K-Nearest Neighbors (KNN) as the most effective models, with remarkably high accuracy rates of 98.49% and 99.01% respectively. However, it also uncovers limitations within the artificial Neural Network (ANN) model, indicating a potential area for further refinement. The findings of this study provide valuable insights that can guide policy-making and the design of targeted road safety interventions, potentially saving lives and reducing the frequency and severity of road accidents.

1 Introduction

Road safety is a crucial concern globally, impacting public welfare, economic stability, and societal well-being. Traffic accidents contribute significantly to the loss of human lives, cause severe injuries, and impose a substantial economic burden on healthcare systems and essential infrastructure. Like many regions worldwide, Northern Ireland faces the challenge of addressing the effects of traffic accidents on its communities and road infrastructure. This project seeks to contribute to the reduction of these challenges by employing advanced machine learning models to develop a predictive model for the severity of road accidents in Northern Ireland.

1.1 Significance of this research:

The risk associated with road accidents extends beyond geographic boundaries, affecting people worldwide. The World Health Organization (WHO) estimated that road traffic injuries are a leading cause of death globally, particularly among young people aged 15 to 29 [1]. In addition to the tragic loss of lives, road accidents impose a considerable economic burden on societies, healthcare resources and the overall economic growth. As such, road safety initiatives are critical components of public policy and governance across the globe.

[1] https://www.who.int/health-topics/road-safety#:~:text=More%20than%20half%20of%20all,(18%E2%80%9359%20years).

1.2 The Northern Ireland Context:

Northern Ireland’s unique socio economic and geographical characteristics, faces specific challenges in ensuring road safety. The impact of traffic accidents on the region's infrastructure and the well-being of its residents necessitates a focused and data-driven approach to identify and address risk factors. The goal is to not only react to accidents but to predict and prevent them, creating safer road environments.

1.3 Motivation for the Study:

The motivation behind this research arises from the pressing need to address the rising influence of traffic accidents on public safety and societal well-being in Northern Ireland. The rising toll of fatalities and injuries, coupled with the economic strain on communities, underscores the necessity for proactive and data-driven interventions. The study is driven by the potential of machine learning techniques to forecast traffic accidents accurately and provide insights that can inform targeted preventive measures.

1.4 Research Question:

The research question, "How effective are machine learning and deep learning models, when tuned with optimal hyperparameters, in predicting the severity of road accidents in Northern Ireland?" was derived from Malik et al.'s foundational study, which highlighted the importance of machine learning algorithms in predicting road accident severity.

1.5 Objectives of the Research:

The primary objective of this research is to leverage machine learning algorithms to create a predictive model for the severity of road accidents in Northern Ireland. By analyzing historical accident data, the study aims to uncover trends and patterns related to road accidents, providing a basis for effective preventive measures. The utilization of diverse machine learning algorithms and deep learning algorithms which are logistic regression, random forest, decision tree, k-nearest neighbor and Artificial Neural Network, adds depth and precision to the predictive modelling process.

1.6 Document Structure:

This thesis project explores machine learning applications for road safety in Northern Ireland. It establishes research objectives and motivations. The Literature Review analyzes existing studies, identifying gaps in the papers and discusses what are the positives. The Methodology section outlines the research design, including data collection and model selection. A crucial Design and Implementation chapter details the development and practical application of chosen models, such as Random Forest and K-Nearest Neighbors. The Results and Discussion present model performances and limitations, particularly of the ANN model. Finally, the Conclusion and Future Work summarize key findings and suggest directions for future research in this field.

2 Related Work

The imperatives of enhancing road safety have prompted a growing interest in the implementation of machine learning models to predict accident severity. This pursuit is particularly significant in Northern Ireland, where a customized approach is necessary in areas where the region's road infrastructure and conditions demand it. This section aims to precisely scrutinize key aspects of pertinent research papers, unravelling the methodologies, findings, and implications that contribute to the development of robust predictive models. The study by Malik et.al., (2021) serves as a foundational piece in this emerging field, offering a comprehensive analysis of ML algorithms applied to predict road accident severity. This research aims to enhance the predictive model for road accident severity in the UK by addressing class imbalance with the SMOTE technique and employing a range of machine learning algorithms.

2.1 Machine Learning Models in Road Safety:

This literature review expands upon Malik's work by considering additional studies, technological developments, and methodological considerations relevant to the application of ML in road safety contexts. Machine learning models, such as decision trees, support vector machines (SVMs), and neural networks, have been widely researched for their applicability in various domains, including road safety. The intersection of these models presents a robust analytical framework to tackle the complex and multifaceted issue of road safety. For instance, Amorin et al. (2023) emphasized the utility of ensemble methods that combine multiple ML models to improve predictive performance over any single model. These methods, such as random forests (an ensemble of decision trees) and gradient boosting machines, have shown to be particularly effective in handling the high-dimensional and non-linear nature of road accident data. Moreover, the integration of geographic information system (GIS) technologies with ML models, as detailed by Al-amari et.al. (2021), has facilitated a more nuanced understanding of spatial factors contributing to road accidents. By employing GIS-based decision trees and neural networks, researchers have been able to pinpoint high-risk zones and the contributing environmental factors, thereby enabling targeted interventions. The work by Vanitha et al. (2023) further explores the role of deep learning, a subset of ML that uses structures akin to neural networks, in detecting and analyzing road traffic incidents in real-time. Their research indicates that deep convolutional neural networks (CNNs) can be trained on large datasets of traffic imagery to accurately identify accident-prone situations before they escalate. While these paper demonstrate the potential of integrating GIS with ML, future research could address the limitations in data quality and processing complexities associated with spatial data. Additionally, exploring more robust methods for handling inaccuracies in GIS data could enhance the effectiveness of these models in identifying high-risk zones. This project aims to contribute to the broader knowledge in accident prediction and prevention. The machine learning models and feature selection process helps the overall understanding of factors influencing accidents. The research makes a significant contribution to the existing body of knowledge in areas such as GIS, IoT, spatial data, and CNN. It builds upon the framework established by Malik et al. (2021) in using machine learning for predicting road accident severity, demonstrating the efficiency of such models in real-world scenarios. The integration of GIS, as explored by Alaamri et al. (2021), is particularly noteworthy for its application in understanding the spatial dimensions of road accidents. This is further complemented by the incorporation of IoT technologies, resonating with Ahmed (2021)'s insights into IoT's role in enhancing traffic safety through real-time data analysis. The research also aligns with the advancements in CNN, as discussed by Vanitha et al. (2023), especially in the context of processing traffic imagery for better accident analysis. By utilising these technological approaches, the study not only offers practical solutions for road safety but also paves the way for future research in this domain, suggesting how these advanced tools can be further optimized and integrated for more effective accident prediction and prevention strategies.

Additionally, real-time data processing and predictive analytics are at the forefront of current research. The integration of Internet of Things (IoT) devices with ML models, as discussed in the work of Ahmed (2021), offers real-time data acquisition and analysis capabilities. Such integration has the potential to provide instant predictions and alerts, contributing significantly to accident prevention efforts. Furthermore, ethical considerations and biases inherent in ML models are being scrutinized, as they can impact the fairness and reliability of predictions. Research by Yassin et.al (2020) critically evaluates the potential for machine learning to perpetuate systemic biases in road safety measures, calling for a balanced approach that incorporates ethical guidelines in model development. In the context of Northern Ireland, as the unique road infrastructure and traffic behavior necessitate tailored ML solutions Shaji (2021) have extended this research by incorporating socio-economic data into ML models, thereby offering a comprehensive perspective that includes human factors in predicting accident severity. However, the paper could further explore solutions to the challenges of data privacy, security, and the need for high computational resources in real-time data processing.

2.2 Geographic and Contextual Factors:

The intersection of geography and road safety is a domain of increasing research interest, especially in areas with distinctive terrain and environmental conditions like Northern Ireland. Abduljabbar et al. (2019) laid the groundwork in this field by integrating geographic information systems (GIS) with road accident data, acknowledging the influence of the region's narrow and winding roads on accident severity. This integration provides a multi-dimensional view of accident factors, enabling a more nuanced understanding of how geography influences road safety. Building upon this foundation, Ahmed et al. (2023) introduced context-aware machine learning techniques to further refine predictive models. Their research emphasized the need for models that are responsive not just to the geographic layout but also to temporal and environmental factors such as weather conditions and urbanization levels. These elements are critical as they often have a direct impact on driving behavior and accident risks. While the UK and Northern Ireland share certain similarities, there are also distinct differences in terms of road conditions, climatic factors, administration, and other contextual aspects. This research for Northern Ireland can be useful for Republic of Ireland as Northern Ireland share major similarities with ROI (Ireland). The main reason for considering the Northern Ireland is that there is no official dataset for ROI. The road infrastructure in Northern Ireland may vary from the rest of the UK due to differences in terrain and population density. Some areas might have rural or less-developed roads, which could present different challenges compared to urban or well-maintained roads. The climate in Northern Ireland can be characterized by frequent rainfall and relatively mild temperatures. Weather conditions, including rain and fog, can impact visibility and road surfaces.

Other parts of the UK may experience different weather patterns. For example, regions in Scotland might face colder temperatures and snowfall during winter months, affecting road conditions differently than in Northern Ireland. Road safety regulations and policies specific to Northern Ireland may be distinct from those in the rest of the UK, because the traffic and road policies are governed by the local bodies specific to Northern Ireland. England, Scotland,

and Wales may have separate regulations and policies. That is why these discussion on the integration of IoT devices with ML models for real-time data analysis opens up new and unique avenues for instant predictions and alerts in road safety for different regions. However, their research could be expanded to explore the scalability of these models in different geographic settings and the integration of real-time data sources, such as traffic flow and pedestrian movement, to enhance predictive accuracy.

The importance of context-aware systems is further supported by the work of et Ardakni et. al. (2023), who demonstrated how seasonal variations in weather patterns necessitate dynamic modelling approaches. In their study, machine learning algorithms were trained on weather-related data, revealing that accident patterns in winter differ significantly from those in summer months due to factors like ice, snow, and reduced visibility.

Further research by Rakauskas (2022) extended the scope to urban versus rural distinctions in road safety. Their analysis found that rural roads in Northern Ireland, characterized by less lighting and more severe curvature, present different challenges than urban roads, which are affected by higher traffic volumes and more frequent pedestrian interactions. Rakauskas' analysis of urban versus rural road safety distinctions opens up avenues for research into the development of targeted safety measures for each type of environment. Further studies could focus on the implementation and effectiveness of these targeted measures in reducing accident rates in both urban and rural settings.

The role of human geography in road safety has also been explored. Ziakapolous et al. (2020) investigated the sociodemographic characteristics of areas with high accident rates, linking socioeconomic status and access to public transport with road safety outcomes. Their findings suggest that disadvantaged areas may experience higher accident rates, highlighting the need for targeted safety measures. A novel approach by Dai et al. (2019) integrated social media data to assess public sentiment towards road safety measures. By analyzing geographic-specific discourse, they could gauge community response to infrastructure changes and policy implementations, offering a new avenue for understanding and improving road safety. Dai and colleagues' novel approach of using social media data to assess public sentiment towards road safety measures is intriguing. Future research could expand this to a larger scale and explore the integration of these insights into policymaking and road safety initiatives. Additionally, examining the reliability and representativeness of social media data in this context would be beneficial.

2.3 Data Quality and Challenges:

The integrity of data plays a pivotal role in the field of road safety, especially when employing machine learning (ML) models. Obasi et al. (2023) highlighted the critical nature of this issue, especially within the context of Northern Ireland where road safety data can suffer from incompleteness and inconsistency. Their research advocates for the implementation of advanced data preprocessing techniques to improve data quality before its application in ML models.

Leduc et al. (2008) offer practical insights into this challenge, present several case studies that examines the improvement of road accident data quality. Their collaborative approach, involving government agencies, law enforcement, and researchers, emphasizes the necessity of multi-stakeholder engagement in enhancing the reliability and completeness of datasets.

The literature further expands on this topic, as seen in the work of Arbabzadeh et.al., (2020), who investigate the impact of missing data on the accuracy of predictive models. Their findings underscore the importance of employing sophisticated imputation methods to handle missing values without introducing bias.

Mannering F (2020) address another dimension of data quality, which is the standardization of data formats across different collecting entities. They propose a unified data collection framework that can be employed by various stakeholders, ensuring consistency and facilitating easier data integration for comprehensive analysis.

A study by Ting et.al. (2020) tackles the challenge of outlier detection and the handling of anomalous data points which can skew the results of predictive models. They developed an algorithm specifically designed for road safety data that effectively identifies and processes outliers, enhancing the predictive performance of subsequent models.

The temporal relevance of data is also a critical factor as highlighted by (Ihueze, 2018)who demonstrate that older data may not accurately reflect current road safety conditions. Their research advocates for real-time data collection and processing to maintain the temporal accuracy of ML models.

Moreover, the work by Gudivada, et.al. (2017) discusses the integration of data quality measures into the model training process itself, ensuring that ML algorithms are robust against poor-quality data. They introduce techniques such as data quality-aware feature selection and noise-resistant training algorithms.

To address the complex nature of data quality in road safety, a holistic approach is necessary. As the literature suggests, this includes not only sophisticated technical solutions but also collaborative and standardized practices among data collectors and researchers. The future of this field relies on continual improvement in data pre-processing, collection, and analysis techniques, ensuring that the predictive models developed are both accurate and reliable.

2.4 Evaluation Metrics and Validation Techniques:

Contributing significantly to the methodological underpinnings of the literature, Alagarsamy et al. (2021) conducted analysis of evaluation metrics for road accident severity prediction models. Their study delved into metrics such as accuracy, precision, recall, and F1 score, offering insights into their relative merits in assessing model performance. Extending this discourse, Kumeda et al. (2021) focused on cross-validation techniques, adding depth to the understanding of assessing the performance of integrated machine learning models in road safety. Their work underscored the importance of robust validation processes, acting as a bulwark against overfitting and ensuring the generalizability of models, a critical consideration in the dynamic and complex domain of road safety prediction.

This comprehensive literature review has meticulously explored key research papers at the intersection of integrated machine learning models and road accident severity prediction in Northern Ireland. The nuanced methodologies and findings discussed in these papers offer valuable insights into algorithm selection, the integration of geographic and contextual factors, data quality challenges, and robust evaluation techniques. As the synthesis of this knowledge progresses, it will undoubtedly serve as the bedrock for the development of context-aware predictive models finely attuned to the unique challenges of road safety in Northern Ireland, ultimately contributing to the broader global endeavor of fostering safer road environment.

3 Research Methodology

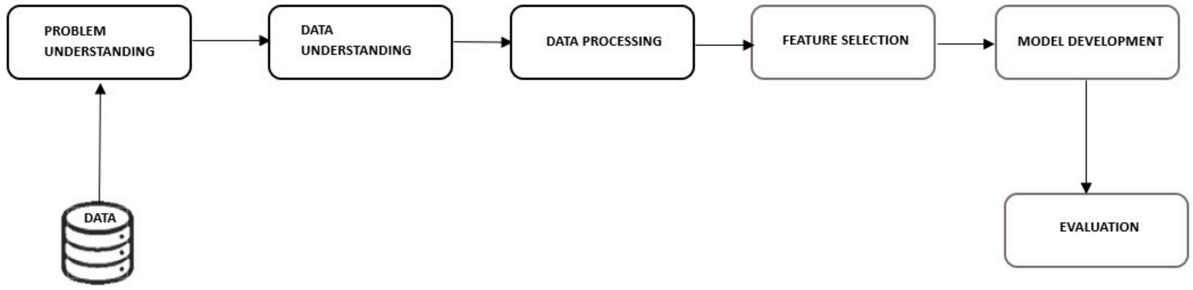


Fig 1: Workflow diagram of the ML model

This research advances the predictive analysis of road accident severity in Northern Ireland, utilizing a range of machine learning techniques. The goal is to create a robust model that can predict accident outcomes with high accuracy, thereby aiding in the development of informed safety measures. This project is based on the hypothesis that a multi-algorithmic approach, through extensive data preprocessing and feature engineering, can significantly improve prediction over current models.

Dataset:

Data collection was conducted systematically, downloaded from official governmental sources[2] to create a database of road accident records in Northern Ireland with a total of nine datasets. This dataset covers a date range spanning from 2020 to 2022, ensuring the inclusion of diverse conditions and a rich set of data points essential for robust analysis.

The resulting dataset comprises nine CSV files, categorically organized into three datasets for each year (2020-2022). These datasets include:

Collision(2020-2022) .csv: This dataset provides detailed information about the circumstances of collisions, including collision severity, the number of vehicles and casualties involved, temporal and geographic details, weather conditions, road types, and the presence of hazards.

Vehicle(2020-2022) .csv: This dataset outlines information related to the vehicles involved, including vehicle type, maneuver at the time of collision, and driver attributes such as age and sex.

Casualty(2020-2022) .csv: This dataset documents details about the casualties resulting from the collisions.

Each of these datasets is substantial, containing 6,000 to 10,000 rows of data. The Collision(2020-2022).csv dataset consists of 25 columns, the Vehicle(2020-2022) .csv dataset

[2] https://admin.opendatani.gov.uk/dataset?organization=police-service-of-northern-ireland&tags=injury+collisions

contains 16 columns, and the Casualty(2020-2022).csv dataset includes 14 columns. This extensive and varied dataset forms the cornerstone of our empirical analysis and predictive modeling efforts in this study.

Data Pre-processing:

A data pre processing pipeline was executed to prepare the dataset for analysis. The initial steps involved the utilization of MySQL Server and Workbench to consolidate collision, vehicle, and casualty data for each year, seamlessly combining them into unified datasets using SQL joins. Subsequently, these combined datasets were imported into Python, where further preprocessing steps were undertaken.

The data pre processing journey commenced with the removal of duplicate records and the systematic handling of missing values, guided by both data distributions and domain expertise. To enable the effective use of categorical variables in machine learning algorithms, they were encoded using techniques such as one-hot encoding and ordinal coding.

A critical aspect of this pre processing was addressing class imbalance within the dataset. To enhance the model's capacity to generalize across different accident severities, the Synthetic Minority Over-sampling Technique (SMOTE) was employed. This technique intelligently balanced the class distribution, ensuring that the predictive models could provide accurate insights into road accident severity. Overall, the combination of SQL-based data consolidation and Python-based pre processing techniques laid a robust foundation for subsequent analysis and predictive modelling.

Feature Engineering:

Feature engineering played a pivotal role in our data preprocessing efforts, serving dual objectives of reducing dimensionality and enhancing predictive power. To achieve these goals, a systematic approach was employed. Correlation analyses were conducted to identify and eliminate redundant features. By assessing the relationships between variables, marking those that exhibited high correlations, and redundant ones were subsequently removed. This step not only streamlined the dataset but also eliminated multicollinearity, which can negatively impact predictive model performance.

Histograms were employed as a tool in this process, aiding in the identification and selection of the most impactful features for accident severity prediction. By visualizing the distribution of each feature and its relationship with the target variable 'a_type' (severity), we could make informed decisions about feature inclusion, ensuring that the selected features had the most significant influence on the predictive power of our models. This iterative and data-driven approach to feature engineering was instrumental in enhancing the overall effectiveness of our predictive models.

Model Development:

The model development phase involved the deployment of four key machine learning algorithms: random forest, decision tree, k-nearest neighbor (KNN), and artificial neural network (ANN.) The algorithms were chosen for its unique strengths: random forest for handling non-linear data, decision trees for their hierarchical decision-making structure, KNN for its effectiveness in capturing the similarity between instances, and for its high-dimensional space classification capability. This diversified approach aimed to leverage the collective strengths of these algorithms, enhancing the robustness and accuracy of the predictive model.

Hyperparameter Optimization:

Hyperparameter optimization was executed through grid search and random search methods, with the goal of fine-tuning the models to achieve optimal performance. Grid search was employed for models with fewer hyperparameters, while random search provided a more efficient alternative for the more complex models with a larger hyperparameter space. The search was constrained within a predefined range of values for each hyperparameter, based on preliminary tests and literature benchmarks.

Model Evaluation:

Model evaluation was conducted using a stratified k-fold cross-validation approach to ensure that each fold was a good representative of the whole. Various performance metrics were employed: accuracy to measure the overall correctness of the model, precision and recall to evaluate the model's performance in predicting high-severity accidents, F1-score as a harmonic mean of precision and recall, and confusion matrix to assess the model's ability to distinguish between classes. These metrics provided a comprehensive understanding of model performance across different aspects.

Validation and Testing:

The validation and testing framework was designed to rigorously assess the model's predictive capabilities on unseen data. A temporal split was used to ensure that the training set consisted of older data and the test set comprised more recent data, reflecting the model's ability to generalize to future conditions. An independent dataset, not used in the training or cross-validation phases, was employed for final testing to provide an unbiased evaluation of the model's performance.

Ethical Considerations:

Throughout the research, ethical considerations were paramount. Data privacy was upheld by anonymizing any personal information within the dataset. Additionally, the research process was designed to avoid any potential biases that could arise from the data or modeling techniques. The research was conducted with a commitment to transparency and accountability, ensuring the results could be trusted and ethically applied to real-world scenarios.

4 Design Specification

4.1 Modelling Techniques:

In the pursuit of road safety enhancement in Northern Ireland, a rigorous analysis of machine learning models is imperative. The selected models, each distinct in approach and capability, collectively contribute to the overarching objective of predicting road accident severity.

4.1.1 Random Forest

Random Forest is an ensemble learning technique that operates by constructing a multitude of decision trees during training and outputting the class that is the mode of the classes (classification) of the individual trees. It was chosen for its ability to handle large datasets with higher dimensionality. It can capture complex, non-linear relationships in the data, which is often the case in accident data. Each tree in a Random Forest gives a classification, and the model chooses the classification having the most votes over all the trees in the forest.Given the complexity and variety of factors influencing road accidents, Random Forest is well-suited for this research as it can provide a more nuanced understanding of the data compared to simpler models. The Random Forest model harnesses the collective power of decision trees to manage the intricacies and varied factors contributing to road accidents. Its ensemble nature allows for a robust analysis that mitigates overfitting, a common issue in singular decision trees. Random Forest model harnesses the collective power of decision trees to manage the intricacies and varied factors contributing to road accidents. Its ensemble nature allows for a robust analysis that mitigates overfitting, a common issue in singular decision trees.

4.1.2 Decision Tree

A Decision Tree is a flowchart-like structure in which each internal node represents a test on an attribute, each branch represents the outcome of the test, and each leaf node represents a class label. The rationale behind selecting a Decision Tree is its ease of interpretation and visualization. It mimics human decision-making processes more closely than other algorithms, making it easier to understand and explain. It works by splitting the data into subsets based on the value of input variables. This splitting is repeated recursively, forming a tree structure. In the context of road accident severity, Decision Trees can be particularly useful in identifying significant factors and their thresholds that lead to different severities of accidents. The Decision Tree stands out for its interpretability, enabling stakeholders with varying technical expertise to understand the model's reasoning. This clarity is crucial in policy formulation, where the rationale for decisions must be transparent and justifiable.

4.1.3 K-Nearest Neighbour (KNN)

KNN is a simple, instance-based learning algorithm that stores all available cases and classifies new cases based on a similarity measure (e.g., distance functions). KNN was chosen for its simplicity and effectiveness in classification tasks where the decision boundary is not well defined. KNN classifies data based on the majority class of its K nearest neighbors. It calculates the distance (such as Euclidean) from the query instance to the labeled instances in the training dataset. This model is suitable for the task as it can effectively classify accident severity by looking at similar past incidents, making it a practical choice for real-world applications. K-Nearest Neighbor (KNN) adds a dimension of simplicity and efficacy, especially when the decision boundary is ambiguous. It's an exemplar of instance-based learning that can leverage the rich historical data on road accidents, using the proximity of data points to predict severity.

4.1.4 Artificial Neural Network (ANN):

In this research, the Artificial Neural Network (ANN) methodology was adopted for its proficiency in handling complex patterns within road accident data. We implemented a multi-layer perceptron architecture, carefully designed with input, hidden, and output layers. The number of neurons in the hidden layers was determined through iterative experimentation to optimize the model’s capacity for feature representation. Activation functions, such as ReLU for hidden layers and softmax for the output layer, were selected to aid in the non-linear transformation of data. To combat overfitting, we utilized dropout and regularization techniques during training. The ANN was trained using a backpropagation algorithm with a stochastic gradient descent optimizer to minimize the loss function effectively. The model’s performance was validated using a split of training and test datasets, ensuring the ANN's generalization to new, unseen data. The Artificial Neural Network (ANN) elevates the model's sophistication, employing a multi-layer perceptron architecture adept at discerning complex, non-linear patterns that simpler models might overlook. The ANN's design includes dropout and regularization to avoid overfitting, ensuring the model's predictive power remains high without becoming overly specialized on the training data. Training the ANN involves backpropagation and stochastic gradient descent, a testament to the model's dynamic learning capabilities and its potential for adaptability in real-world application scenarios. This adaptability is crucial, given the dynamic and often unpredictable nature of road traffic incidents.

4.2 Evaluation Techniques:

To assess the effectiveness of predictive models, an in-depth analysis using various metrics and a confusion matrix was conducted. This approach provided clarity on the accuracy of models in predicting accident severity. Key elements of this evaluation included:

- Precision: Measured the accuracy of positive predictions by the model. High precision indicated fewer false positives.
- Recall: Measured the model's ability to identify all actual positive cases.
- F1 Score: Balanced precision and recall, particularly useful for uneven data distributions.
- Support: Indicated the number of occurrences of each class in the dataset, providing context for other metrics.

The confusion matrix was crucial in the evaluation, displaying the accuracy of model predictions.

These elements helped in calculating precision and recall, offering insight into model performance. Thorough examination of these metrics and the confusion matrix ensured that predictive models were practical and reliable for real-world applications in predicting accident severity.

5 Implementation

5.1 Data Processing Steps

Data Consolidation and Cleaning

The dataset consisted of nine CSV files, organized into three datasets per year (2020-2022). These datasets include "Collision," detailing collision circumstances; "Vehicle," providing information about involved vehicles; and "Casualty," documenting casualty details. Key data points encompass collision severity, vehicle and casualty counts, temporal and geographic information, weather conditions, road types, and hazards. The dataset was extensive, with each CSV file containing 6,000 to 10,000 rows. The initial phase involved merging three years' worth of data (2020-2022) into a single DataFrame, followed by the replacement of blank spaces with NaN values to accurately identify and analyze missing data.

Missing Values Analysis

The assessment of missing values in our dataset revealed a notable pattern, the data was missing completely at random (MCAR). This pattern implies that, theoretically, the missing data could be used in future cases if a larger proportion of this data were present. However, in our specific dataset, the extent of missing data exceeded 90% for certain columns. As a result, a decision was made to address this issue by removing these columns entirely from the dataset. Additionally, there was one column, namely "c_vtype," which exhibited a more modest rate of missing values, accounting for approximately 5% of the data. To use the valuable information contained within this column, a strategy of data imputation was employed. Specifically, the missing values in the "c_vtype" column were imputed using the mean of the available data. This approach allowed us to retain the column's utility while addressing the relatively small proportion of missing values.

Target Variable Examination

The distribution of the target variable 'a_type' (severity) was examined to gain valuable insights into the class balance within the dataset, a crucial aspect for effective model training.

Upon close evaluation, it became evident that the class distribution of severity was highly imbalanced. Specifically, there was a notable disparity in the number of data points for severity levels 2 and 1 (indicating high severity incidents) compared to severity level 3 (representing less severe accidents). The dataset contained a disproportionately low number of records for severity levels 2 and 1, while severity level 3 was more prevalent. This class imbalance presented a significant challenge in training robust predictive models, as the models may lean towards predicting the majority class (severity level 3) and potentially overlook the minority classes (severity levels 2 and 1). Consequently, addressing this class imbalance became a pivotal step in our modeling approach to ensure balanced and accurate predictions for all severity levels.

Data Distribution Visualization

Histograms for all columns were generated to visualize data distributions, enabling an understanding of variable characteristics and the identification of outliers or anomalies.

Data Subset Selection

A subset of necessary variables was carefully curated for in-depth analysis, focusing on features deemed most relevant for predicting accident severity.

Data Export for Visualization

The refined subset was exported to a CSV file to facilitate external visualization and further examination.

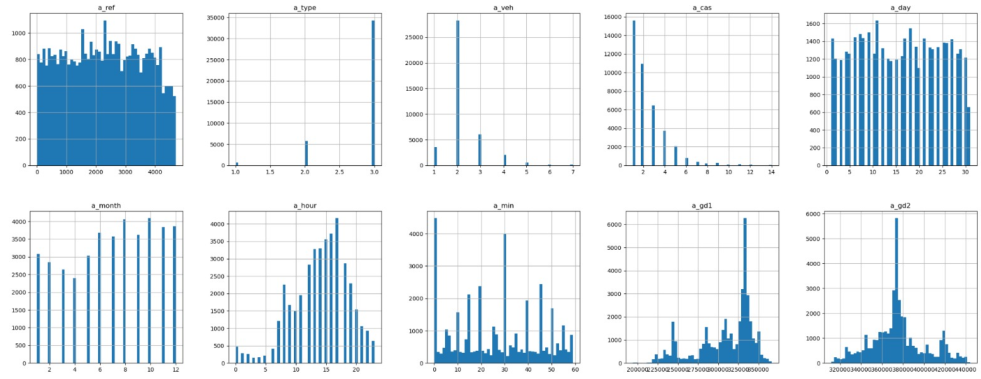
Categorical Data Encoding

The 'a_wkday' column, which consisted the names of the day of the week in which the accident took place, was transformed using ordinal encoding, assigning structured numerical values to weekdays, while the 'a_District' column explained the district in which the accident took place, was processed using one-hot encoding to prepare categorical data for algorithmic analysis.

These methodical data processing steps have prepared the dataset for advanced machine learning applications, ensuring a high degree of data readiness for the subsequent modelling phase.

5.2 Feature Selection by removing columns:

The combined dataset originally comprised 52 columns, which was carefully reduced to 27 critical variables to focus the analysis on the most significant factors contributing to road accidents. This reduction was based on an evaluation of the columns, ensuring the retention of key information while eliminating redundant or less impactful data. The streamlined dataset includes variables such as accident details, vehicle specifics, and casualty information, facilitating a more focused and efficient modelling process.



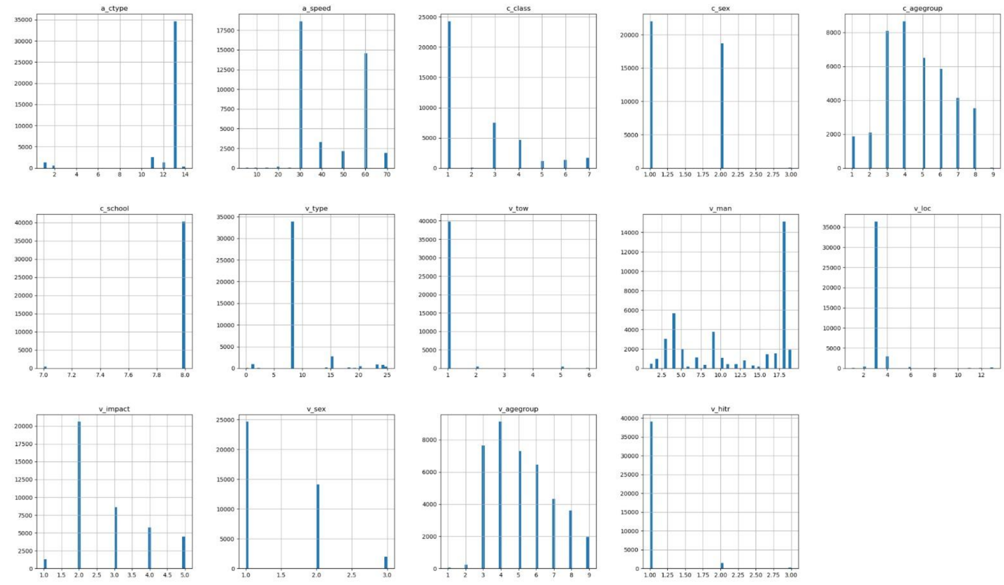


Fig 2: Histogram of all the variables after preprocessing.

5.3 Model Development Phase

In the model development phase of this research, a robust strategy was employed to address the imbalanced dataset, which is crucial for achieving a fair and accurate machine learning model.

Data Balancing with SMOTE

The dataset exhibited a notable imbalance in the target variable 'a_type,' with a substantial overrepresentation of the majority class. Considering the initial dataset where the classes were 34000 vs 600 ,this was the most optimum balance that was got between the two classes.

The class balance produced in the code was an optimised version of the original dataset values because the value with 27000 is fatal injury which is not frequent in Northern Ireland , so increasing this case will lead the machine learning models in making wrong and biased predictions. To rectify this class imbalance issue and create a more equitable learning environment, the Synthetic Minority Over-sampling Technique (SMOTE) was employed on the training data. SMOTE's approach involves generating synthetic samples for the minority classes, effectively augmenting their representation within the dataset.

As a result of the SMOTE analysis, the distribution of the target variable 'a_type' was significantly improved. The previously imbalanced dataset, where the majority class dominated, was transformed into a more balanced dataset. This rebalancing ensured that the machine learning models had a fair and equal representation of all severity levels, enhancing their ability to make accurate predictions across the entire spectrum of accident severity.

Training and Testing Split

The data was split into an 80:20 ratio for training and testing, ensuring that a substantial amount of data was used to train the models while still retaining a significant portion for an unbiased evaluation of the model's performance on unseen data.

Model Training

Random Forest, Decision Tree, K-Nearest Neighbor, and Artificial Neural Network was trained on the combined dataset, which included both the original and the synthetic samples generated by SMOTE and the best performing hyperparameter for each model. This approach aimed to enhance the models' ability to generalize and accurately predict the minority classes.

Model Evaluation

Post-training, each model was evaluated based on accuracy, precision, recall, and F1-score, with their respective confusion matrices visualized. This comprehensive evaluation allowed for the assessment of not only the overall accuracy but also how well each model performed on each individual class, which is particularly important in an imbalanced dataset.

The meticulous approach to model development and training, with an emphasis on addressing data imbalance, sets a precedent for predictive modeling in datasets where minority classes hold significant importance.

6 Evaluation

The refinement of machine learning models through hyperparameter tuning is a pivotal step in predictive modeling, enhancing accuracy and robustness. This section includes the implementation and performance of four distinct models after hyperparameter optimization.

6.1 Experiment 1: Random Forest Tuning

The Random Forest model's hyperparameters were meticulously tuned, leading to an accuracy of 98.49%. The key hyperparameters in this optimization were:

- N estimators: 200 was chosen to increase the model's ability to learn from diverse data aspects.
- Max Depth: None , Min Samples leaf : 1 ,Min samples split : 2.

Accuracy with the best hyperparameters : 0.9849

Table 1: Classification Report for Random Forest with Hyper-parameter tuning:

	Precision	**Recall**	**F1 Score**	**Support**
1	1.00	0.95	0.98	125
2	0.98	0.92	0.95	1193
3	0.98	1.00	0.99	6826
Accuracy			0.98	8144
Macro Average	0.99	0.95	0.97	8144
Weighted Average	0.98	0.98	0.98	8144

```
Confusion Matrix for RandomForestClassifier():
```

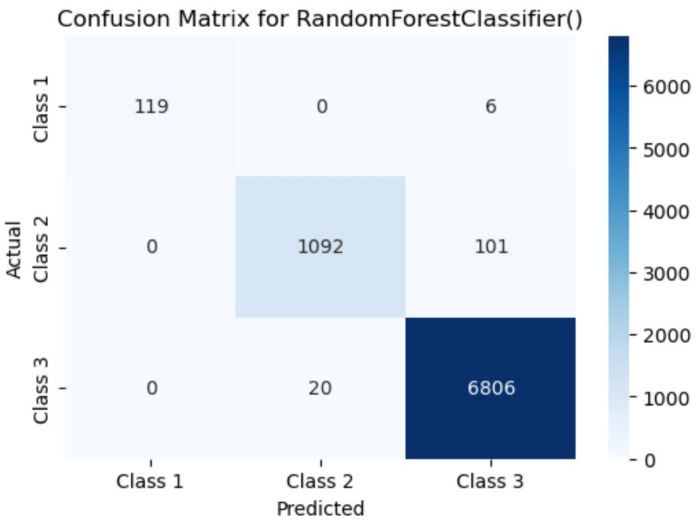


Fig 3: Classification report and confusion matrix for Random Forest Classifier

The precision and recall for each class were notably high, especially for Class 1, demonstrating the model's capability to identify severe accidents reliably.

6.2 Experiment 2: Decision Tree Tuning

Post-tuning, the Decision Tree accuracy stood at 96.97%, with hyperparameters refined as follows:

- Max Depth: Set to 'None', allowing the tree to expand fully where necessary to capture the nuances of the data.
- Min Samples Split: Kept at the default of 2, to begin splitting nodes at the earliest instance of a pattern.
- Min Samples Leaf: Remained at 1 to maintain sensitivity to the training data.

These parameters helped maintain the Decision Tree's interpretability while improving its predictive power.

Accuracy for Decision Tree with Hyper-parameter Tuning : 0.9697
Best Hyper-parameters for Decision Tree : {‘max depth’: None , ‘min_sample_leaf’: 1, ‘min_samples_split’: 2}

Table 2: Classification Report for Decision Tree with Hyper-Parameter Tuning:

	Precision	Recall	F1 Score	Support
1	0.83	0.95	0.88	125
2	0.89	0.93	0.91	1193
3	0.99	0.98	0.98	6826
Accuracy			0.97	8144

Macro Average	0.99	0.95	0.92	8144
Weighted Average	0.98	0.98	0.97	8144

```
Confusion Matrix for DecisionTreeClassifier():
```

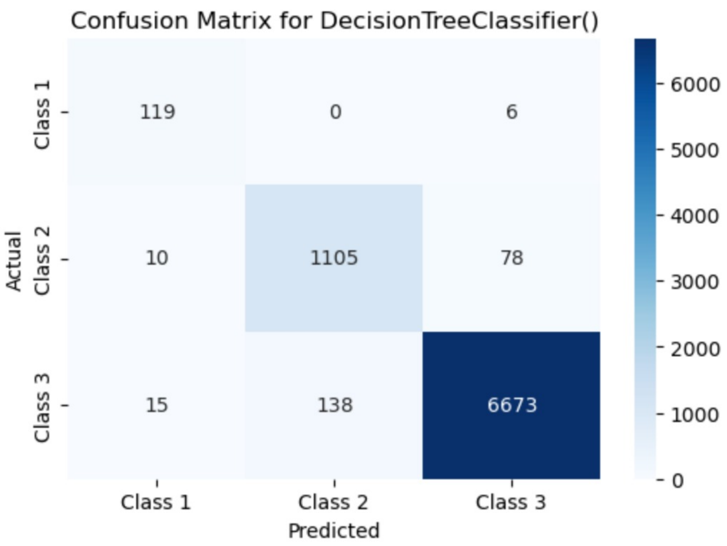


Fig 4: Classification report and confusion matrix for Decision Tree Classifier

6.3 Experiment 3: K-Nearest Neighbors (KNN) Tuning

KNN achieved an impressive accuracy of 99.01% with the following hyperparameters:

- Number of Neighbors: Reduced to 3, the optimal balance for this model to analyze local patterns effectively.
- Distance Metric (p): Set to 1, employing the Manhattan distance which proved more effective than the Euclidean in this context.
- Weights: Utilized 'distance' to give precedence to nearer neighbors, enhancing the prediction accuracy for the target classes.

The adjustments led to substantial improvements in the model's ability to classify road accident severity with high precision and recall.

Accuracy for K-nearest Neighbours with Hyper-parameter Tuning : 0.9901
Best Hyper-parameters for K-nearest Neighbours: {‘n neighbours: 3 , ‘p’: 1, ‘weights’: ‘distance’}

Table 3: Classification Report for K-nearest Neighbours with Hyper-Parameter Tuning.

	Precision	**Recall**	**F1 Score**	**Support**
1	0.94	0.99	0.96	125

2	0.97	0.97	0.97	1193
3	0.99	0.99	0.99	6826
Accuracy			0.99	8144
Macro Average	0.97	0.98	0.98	8144
Weighted Average	0.99	0.99	0.99	8144

```
Confusion Matrix for KNeighborsClassifier():
```

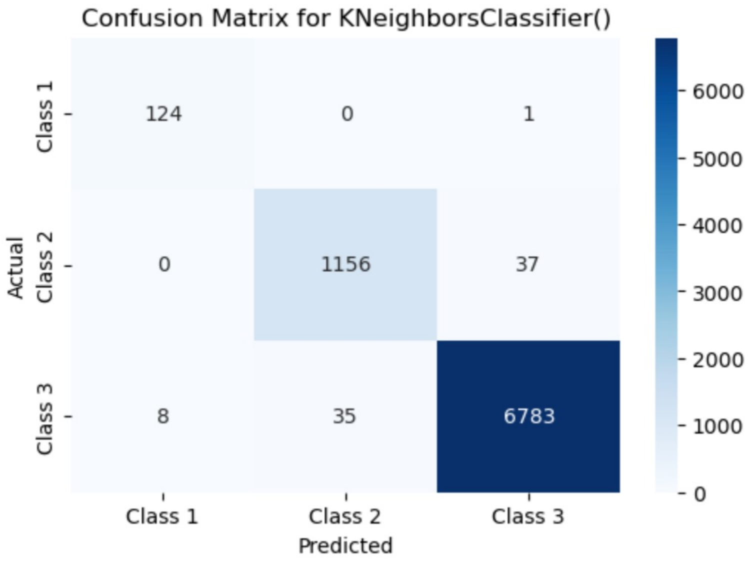


Fig 5: Classification report and confusion matrix for KNN Classifier

6.4 Experiment 4: Artificial Neural Network (ANN) Tuning

The ANN model, after tuning, reported an accuracy of 83.82%. The chosen hyperparameters were:

- Hidden Layer Sizes: Configured to (100, 50), to provide a complex model structure capable of capturing intricate patterns.
- Alpha: Set at 0.001, providing sufficient regularization to tackle potential overfitting issues.

While the ANN's performance was less impressive compared to the other models, these hyperparameters showed promise for further refinement.

Table 4: Classification Report for Artificial Neural Network with Hyper-Parameter Tuning:

	Precision	**Recall**	**F1 Score**	**Support**
1	0.00	0.00	0.00	125

2	0.00	0.00	0.00	1193
3	0.84	1.00	0.91	6826
Accuracy			0.84	8144
Macro Average	0.28	0.33	0.30	8144
Weighted Average	0.70	0.84	0.76	8144

```
Confusion Matrix for KNeighborsClassifier():
```

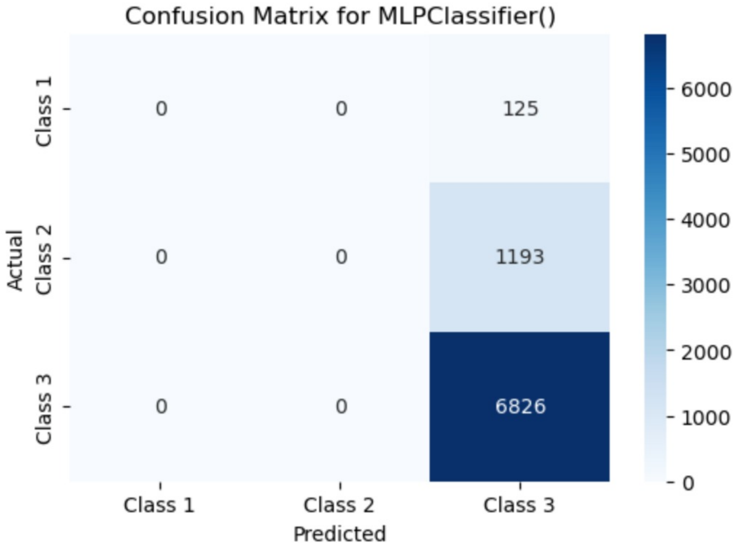


Fig 6: Classification report and confusion matrix for ANN

6.5 Discussion

6.5.1 Comparative Analysis

Table 5: Comparative Analysis of the different model's accuracy

Model	**Accuracy**
Random Forest	0.9849
Decision Tree	0.9697
K - nearest Neighbours	0.9901
Artificial Neural Network	0.8382

Random Forest, with an adjusted accuracy of 98.49%, stood out for its predictive ability, a result of fine tuning hyperparameters such as the number of trees and maximum tree depth. This tuning allowed it to construct a more generalizable model that could handle the dataset provided, enhancing its precision and reducing overfitting.

The Decision Tree demonstrated the value of simplicity and interpretability. Although slightly outperformed by Random Forest, its accuracy of 96.97% indicates a well-fitted model that benefits from optimal complexity control through hyperparameter tuning.

With an accuracy of 99.01%, KNN excelled by finding a delicate balance in hyperparameters - the number of neighbors, the type of weight function, and the distance metric. These parameters were crucial in refining the model's ability to classify data based on the proximity of similar cases, thus capturing localized patterns with high precision.

The ANN model, despite achieving a lower accuracy of 83.82%, highlighted the challenges and potential of neural networks. The model's complexity, managed through hyperparameters like the number of neurons and regularization strength (alpha), pointed to the need for additional feature engineering or more advanced network architectures to improve its prediction accuracy and manage the class imbalance more effectively.

In terms of strategic deployment for road safety applications, the Random Forest and KNN models emerge as the more reliable predictors of accident severity. Their high accuracy and balanced precision-recall provide a strong foundation for developing practical tools that could potentially save lives by predicting and preventing severe accidents.

The ANN's performance, specifically, draws attention to the persistent challenge of model bias towards majority classes, an issue that remains a priority for future research, possibly requiring innovative approaches in data sampling or algorithmic adjustments to achieve equity in model predictions. Moving forward, integrating the strengths of these models could pave the way for a more unbiased predictive system.

6.5.2 Implications for Road Safety

The findings of this analysis carry significant implications for road safety. The predictive accuracy of these models, particularly the KNN, can be instrumental in developing targeted interventions to prevent severe road accidents. The high precision and recall rates for predicting severe accidents mean that safety measures can be more effectively allocated to high-risk areas or conditions.

The enhanced predictability of accident severity also opens avenues for better resource management within emergency response units, enabling them to prioritize areas with potentially higher severity cases. Moreover, the insights gained from models can guide policymakers in understanding the critical factors that lead to severe accidents, helping to shape more informed and effective road safety policies.

Ultimately, the application of these machine learning models could contribute to a significant reduction in both the frequency and severity of road accidents, promoting safer driving conditions and saving lives.

6 Conclusion and Future Work

This thesis has provided a comprehensive evaluation of various machine learning models to predict road accident severity in Northern Ireland. The K-Nearest Neighbors model outperformed others with high accuracy and excellent class balance, followed by the Decision Tree and Random Forest models, which also showed promising results. The ANN, while falling short in this study, presents a valuable area for further research and development.

The implications of this work are significant for the field of road safety, where accurate predictions can inform targeted interventions, policy-making, and emergency response prioritization. The insights gained from this research contribute to the broader goal of reducing traffic-related fatalities and injuries, a critical public health concern.

In conclusion, the future of machine learning in road safety is bright, with many avenues for enhancing model performance and integrating predictive analytics into real-time safety measures. The work completed in this thesis lays a foundation for such advancements, and ongoing research in this area has the potential to bring about substantial improvements in road safety for Northern Ireland and beyond.

The potential for future work in this field of study is vast and promising. Given the moderate performance of the Artificial Neural Network (ANN), subsequent research could explore more complex neural network architectures like Convolutional Neural Networks (CNNs) or Recurrent Neural Networks (RNNs) that may capture spatial and temporal patterns in accident data more effectively. Additionally, the implementation of advanced regularization techniques, such as L1 and L2 regularization, or the use of dropout layers within the ANN, could be investigated to reduce overfitting and improve model generalization.

There is also an opportunity to delve into ensemble methods that combine the predictions of the individual models studied. These ensemble methods, such as stacking or blending, could leverage the strengths of each model to improve overall predictive performance.

Incorporating additional data sources, such as real-time traffic flow or weather conditions, could enhance the models' predictive accuracy. The integration of text data from accident reports using Natural Language Processing (NLP) techniques might also reveal insights into accident causality that are not captured by structured data alone.

Another direction for future research is the exploration of unsupervised learning techniques to detect novel patterns or clusters within the accident data, which could uncover new risk factors or accident types not previously considered.

References

Malik, S., El Sayed, H., Khan, M.A. and Khan, M.J., 2021, December. Road Accident Severity Prediction—A Comparative Analysis of Machine Learning Algorithms. In *2021 IEEE Global Conference on Artificial Intelligence and Internet of Things (GCAIoT)* (pp. 69-74). IEEE.

Amorim, B.D.S.P., Firmino, A.A., Baptista, C.D.S., Júnior, G.B., Paiva, A.C.D. and Júnior, F.E.D.A., 2023. A Machine Learning Approach for Classifying Road Accident Hotspots. *ISPRS International Journal of Geo-Information*, *12*(6), p.227.

Al-Aamri, A.K., Hornby, G., Zhang, L.C., Al-Maniri, A.A. and Padmadas, S.S., 2021. Mapping road traffic crash hotspots using GIS-based methods: A case study of Muscat Governorate in the Sultanate of Oman. *Spatial Statistics*, *42*, p.100458.

Vanitha, R., 2023. Prediction of Road Accidents using Machine Learning Algorithms. *Middle East Journal of Applied Science & Technology (MEJAST)*, *6*(2), pp.64-75.

Ahmed, S., Hossain, M.A., Bhuiyan, M.M.I. and Ray, S.K., 2021, December. A comparative study of machine learning algorithms to predict road accident severity. In *2021 20th International Conference on Ubiquitous Computing and Communications (IUCC/CIT/DSCI/SmartCNS)* (pp. 390-397). IEEE.

Shaji, P., 2021. Good Roads:-Using machine learning to measure how socio and non-socio-economic factors differently influence traffic accident injury rates. *Available at SSRN 3761294*.

Yassin, S.S. and Pooja, 2020. Road accident prediction and model interpretation using a hybrid K-means and random forest algorithm approach. *SN Applied Sciences*, *2*, pp.1-13.

Pourroostaei Ardakani, S., Liang, X., Mengistu, K.T., So, R.S., Wei, X., He, B. and Cheshmehzangi, A., 2023. Road Car Accident Prediction Using a Machine-Learning-Enabled Data Analysis. *Sustainability*, *15*(7), p.5939.

Rakauskas, M.E., Ward, N.J. and Gerberich, S.G., 2009. Identification of differences between rural and urban safety cultures. *Accident Analysis & Prevention*, *41*(5), pp.931-937.

Ziakopoulos, A. and Yannis, G., 2020. A review of spatial approaches in road safety. Accident Analysis & Prevention, 135, p.105323.

Dai, F. and Sujon, M., 2019. Measuring Current Traffic Safety Culture via Social Media Mining. Department of Civil and Environmental Engineering: Washington, DC, USA.

Abduljabbar, R., Dia, H., Liyanage, S. and Bagloee, S.A., 2019. Applications of artificial intelligence in transport: An overview. Sustainability, 11(1), p.189.

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

Ahmed, S., Hossain, M.A., Ray, S.K., Bhuiyan, M.M.I. and Sabuj, S.R., 2023. A study on road accident prediction and contributing factors using explainable machine learning models: analysis and performance. Transportation research interdisciplinary perspectives, 19, p.100814.

Leduc, G., 2008. Road traffic data: Collection methods and applications. *Working Papers on Energy, Transport and Climate Change*, *1*(55), pp.1-55.

Arbabzadeh, N. and Jafari, M., 2017. A data-driven approach for driving safety risk prediction using driver behavior and roadway information data. IEEE transactions on intelligent transportation systems, 19(2), pp.446-460.

Mannering, F., Bhat, C.R., Shankar, V. and Abdel-Aty, M., 2020. Big data, traditional data and the tradeoffs between prediction and causality in highway-safety analysis. Analytic methods in accident research, 25, p.100113.

Gudivada, V., Apon, A. and Ding, J., 2017. Data quality considerations for big data and machine learning: Going beyond data cleaning and transformations. International Journal on Advances in Software, 10(1), pp.1-20.

Ihueze, C.C. and Onwurah, U.O., 2018. Road traffic accidents prediction modelling: An analysis of Anambra State, Nigeria. Accident analysis & prevention, 112, pp.21-29.

Ting, C.Y., Tan, N.Y.Z., Hashim, H.H., Ho, C.C. and Shabadin, A., 2020. Malaysian road accident severity: Variables and predictive models. In Computational Science and Technology: 6th ICCST 2019, Kota Kinabalu, Malaysia, 29-30 August 2019 (pp. 699-708). Springer Singapore.

Alagarsamy, S., Malathi, M., Manonmani, M., Sanathani, T. and Kumar, A.S., 2021, December. Prediction of Road Accidents Using Machine Learning Technique. In 2021 5th International Conference on Electronics, Communication and Aerospace Technology (ICECA) (pp. 1695-1701). IEEE.

Kumeda, B., Fengli, Z., Alwan, G.M., Owusu, F. and Hussain, S., 2021. A hybrid optimization framework for road traffic accident data. International journal of crashworthiness, 26(3), pp.246-257.