Traffic Accident Analysis On Imbalanced Ordinal Variable

Alfred Iyby
School of Computing
Dublin City University
Dublin, Ireland
alfred.iyby2@mail.dcu.ie

Geethu Narippatta
School of Computing
Dublin City University
Dublin, Ireland
geethu.narippatta2@mail.dcu.ie

Priyamol James
School of Computing
Dublin City University
Dublin, Ireland
priyamol.james3@mail.dcu.ie

Sarath Mukundan Adavakkat
School of Computing
Dublin City University
Dublin, Ireland
sarath.mukundanadavakkat2@mail.dcu.ie

Suhara Fathima Shahul Hameed
School of Computing
Dublin City University
Dublin, Ireland
suhara.shahulhameed2@mail.dcu.ie

Abstract—**Traffic accidents are a significant hazard to public safety worldwide, driven by a variety of causes ranging from human conduct to environmental conditions. This work investigates the analysis of traffic accidents, with a specific emphasis on the difficulty of imbalanced ordinal variable. We advocate for the use of data mining techniques to help implement effective preventive policies. We use the Knowledge Discovery in Databases (KDD) methodology to investigate a large dataset of accidents, looking for nuanced patterns and insights. Through this work, we identify critical components leading to accidents using Random Forest and Gradient Boosting models, to facilitate effective preventive strategies, developing a comprehensive understanding of the complex processes that cause such catastrophes. Our findings demonstrate the effectiveness of data mining in accident research, providing useful insights for decision-makers, city planners, and transportation authorities. By utilizing KDD, this study advances accident analysis approaches, highlighting its potential for proactive accident reduction and personalized treatments. Thus, our analysis not only reveals the underlying causes of accidents, but also provides actionable intelligence for stakeholders to implement changes that could save lives and reduce the economic cost of traffic accidents.**

Keywords*— *Accident Analysis, Data Mining, KDD, Severity Prediction

I. Introduction

Traffic accidents are a major global issue, posing hazards to public safety and bearing large economic costs. The multidimensional nature of these accidents, impacted by a variety of factors ranging from human conduct to environmental conditions, emphasizes the importance of thorough investigation and proactive prevention actions. This research dives into the underlying causes of significant traffic accidents, with a special emphasis on the problem given by imbalanced ordinal variable, and seeks to use data mining techniques to improve our understanding and pave the way for effective solutions. The primary goal of this study is to answer the following research question: ”What are the primary contributing factors to serious traffic accidents, particularly when dealing with imbalanced ordinal variable, and how can data mining techniques, such as Random Forest, XGBoost and LightGBM models, be used to develop effective preventive measures?” We use Knowledge Discovery in Databases (KDD) methodology to unravel the complex web of elements that contribute to accidents, offering policymakers, urban planners, and transportation authorities with practical information. This project seeks not only to uncover these elements, but also to investigate the possibilities of data-driven approaches for predicting and preventing future mishaps.

In the following sections, we will look at relevant work and critically evaluate previous works in the subject of accident analysis. By discussing the merits and limits of previous research, we create the framework for our study's methodology. The Data Mining Methodology section describes our technique in depth, covering data collection, preprocessing, the use of Random Forest, XGBoost and LightGBM models, and evaluation criteria. We then describe our evaluation findings, which demonstrate the efficiency of our methodology in identifying important accident factors and offering practical implications for accident prevention methods.

Through this work, we hope to contribute to the improvement of accident analysis methodologies by emphasizing the importance of data mining, namely Random Forest and Gradient Boosting models, in improving road safety, particularly when dealing with skewed ordinal variable. By putting light on the complicated relationships within accident data, we lay the groundwork for evidence-based decision-making and targeted interventions. This study calls for the incorporation of data-driven approaches, into road safety policy, emphasizing the potential to reduce the impact of traffic accidents on society

II. Literature Review

In the study [1], aimed at predicting road accident severity using the road accidents data published by the UK government. The researchers outlined the task of dealing with dataset imbalance and tested multiple sampling techniques to improve the results of model prediction. Furthermore, the authors explored anomaly detection algorithms to improve the classification of severe accidents. The sampling techniques used in the study were SMOTE, NearMiss, and weighted sampling. PCA was used with SMOTE to improve the representation of the features. To treat the severe accidents as anomalies, the researchers used anomaly detection algorithms, such as K-means clustering. The Conclusion reveals the importance of choosing the right sampling algorithm. Although the weighted sampling achieved the best overall accuracy, it was unable to classify severe cases well. The SMOTE with PCA method also showed good promise, but their predictions did not exceed the 50%-accuracy baseline. The NearMiss method did not result in particularly useful outcomes presumably due to there being too few training examples. Anomaly detection methods showed viability for severe accident classification, indicating that a variety of methods need to be tested to find better alternatives for imbalanced datasets.

In order to develop a machine learning predictive model, a study attempted to identify the variables that influence the severity of traffic accidents [2]. Using a dataset of car accidents from 49 states in the US between February 2016 and December 2019, they analyzed the data by handling missing values, normalizing the data, and encoding features. Key factors identified included speed limit, weather, lane number, and time of accident. The Random Forest algorithm proved effective in predicting accident severity with 97.2% accuracy. The study's findings provide valuable insights into factors that contribute to accident severity and can help improve road safety.Machine learning techniques [3] were used in the work to forecast the severity of accidents in Michigan, USA. Several algorithms were used, such as Random Forest, AdaBoost, Logistic Regression, and Naive Bayes. With an accuracy of 75.5%, Random Forest performed better than other methods. Predictive modeling-based preventive solutions for high-risk zones on Michigan roadways were recommended by the study.

The severity of traffic accidents was predicted using machine learning algorithms and exploratory data analysis on historical data in [4]. A 0.83 accuracy rate was attained via Artificial Neural Networks. The study demonstrated how well AI and data analysis can anticipate the severity of accidents.The goal of another study in [5] was to predict crash severity using data on traffic accidents. Algorithms such as Random Forest, Decision Tree, and Bagging performed better than the others; Random Forest obtained the greatest accuracy of 98.80%. The study's conclusions included suggestions for enhancing traffic safety.

Using machine learning methods, a study [6] examined and forecasted traffic incidents in India. K-means clustering and regression techniques were used. The study's objectives were to locate accident hotspots and make recommendations for future improvements to improve forecasts.Using machine learning methods, the authors examined the variables determining the severity of driver injuries in traffic accidents. The study [7] offered a solid approach for assessing and forecasting traffic accidents. A decision tree method [8] was used, to study forecast the death toll from auto accidents in Algeria. The model shown potential for accurately predicting road accident fatalities, with an accuracy of 78.125%.

The study's objective was to evaluate data on traffic accidents in order to pinpoint contributing variables and enhance traffic safety protocols [9]. To improve the accuracy of accident predictions, sophisticated analytical approaches were used. Using data mining tools [10], the study concentrated on predicting the severity of traffic accidents in real time. Improved classification accuracy was demonstrated by using classifiers such as Naïve Bayes, Random Forest, MLP, and AdaBoost.Using boosting algorithms [11], the authors predicted the severity levels of traffic accidents and examined the elements that affect the severity. The study shed light on international risk variables that influence auto accidents. The TRAFFIC ACCIDENTS_2019_LEEDS dataset [12] was used to estimate the severity of traffic accidents. With the best accuracy, the Random Forest classifier highlighted the significance of machine learning for predicting the severity of traffic accidents.

III. Methodology

A. Dataset

The dataset used is a nationwide traffic accident dataset that includes 49 states in the United States. Since February 2016, the data has been collected on a continuous basis via a variety of data providers, including numerous APIs that give streaming traffic events. These APIs broadcast traffic events gathered by a range of groups, including the US and state Departments of Transportation, law enforcement agencies, traffic cameras, and traffic sensors embedded in road networks. Currently, this dataset contains over 1.5 million accident records.

The dataset has several fields that provide extensive information about each accident record. The 'ID' field is a unique identifier for each accident entry. The 'Source' field specifies where the original accident data originated. 'Severity' is a ordinal value ranging from 1 to 4, with 1 indicating the least influence on traffic (resulting in a brief delay) and 4 indicating a major impact (causing a long delay). The 'Start_Time' and 'End_Time' variables show the start and finish timings of the accidents, both in the local timezone. Geographical data contains 'Start_Lat' and 'Start_Lng', which indicate the latitude and longitude GPS coordinates of the accident's starting point, as well as 'End_Lat' and 'End_Lng' for the accident's ending point. 'Distance(mi)' specifies the length of the road affected by the accident in miles.

The 'DateTime' field shows the distribution of accidents over time, from January 14, 2016, to March 31, 2023. The counts vary over time, with peaks and troughs showing shifts in accident rates over time. Finally, a few sample records are shown, each named 'A-1', 'A-2', 'A-3', and so on, with related information such as 'Source', 'Severity', 'Start_Time', 'End_Time', 'Start_Lat', 'Start_Lng', and 'Distance(mi)'. These entries provide insight into the dataset's format and content, highlighting specific accidents and their properties.

Repository : Traffic Accident Analysis

Dataset : US_Accidents_March23_sampled_500k.csv

B. KDD Methodology

The KDD (Knowledge Discovery in Databases) method is a systematic and thorough approach to obtaining useful insights from huge datasets. It is a multi-step process that begins with data selection, which involves identifying relevant datasets from a variety of sources, including databases, data warehouses, and raw sensor data. The quality and relevancy of the data chosen at this step have a significant impact on the overall process results. The selection procedure is critical, since it requires a thorough grasp of the situation at hand to determine which data is required for effective analysis.

After data selection, the following stage is data pre-processing. Raw data frequently contains noise, errors, or missing values, rendering it unsuitable for analysis. Data pre-processing entails cleaning and preparing the data for subsequent analysis. This stage involves resolving missing values, deleting duplicates, and normalizing the data to maintain consistency. Furthermore, data may need to be changed into an appropriate format, such as transforming text data to numerical form using techniques like one-hot encoding.

Data transformation occurs after data has been cleaned and prepared. This stage entails transforming the data into an optimum format for mining. Data transformation may include data aggregation, the creation of new features, or the reduction of the dataset's dimension. Principal Component Analysis (PCA) and feature engineering can be used here to improve the data for mining algorithms. The goal is to develop a dataset that is both informative and easy to analyze.

The KDD method revolves around data mining, which involves applying multiple algorithms to prepared data to uncover patterns, linkages, or trends. There are various data mining approaches, including classification, clustering, regression, and association rule mining. The choice of technique is determined by the nature of the problem and the desired results. For example, classification aims to organize data into predetermined classes based on input characteristics, whereas clustering puts comparable data points together without predefined classes.

After data mining, the identified patterns must be understood and evaluated. This stage is critical in deriving meaningful insights from the data. Interpretation entails comprehending the patterns' meaning and importance in the context of the problem domain. For example, if a data mining algorithm finds a client category that is prone to churn, the interpretation step includes determining why they are churning and what actions may be taken to keep them. Evaluation verifies that the identified patterns are accurate and dependable, frequently use techniques such as cross-validation or testing against a separate dataset to check the model's performance.

The final step in the KDD process is knowledge presentation. The insights gained throughout the data mining and interpretation processes must be properly disseminated to stakeholders. This can be accomplished through reports, infographics, dashboards, or direct interaction with decision-making systems. The goal is to make the insights available and intelligible to those who will utilize them to make educated decisions. Clear communication of the findings is critical for successfully using the knowledge gained via the KDD process.

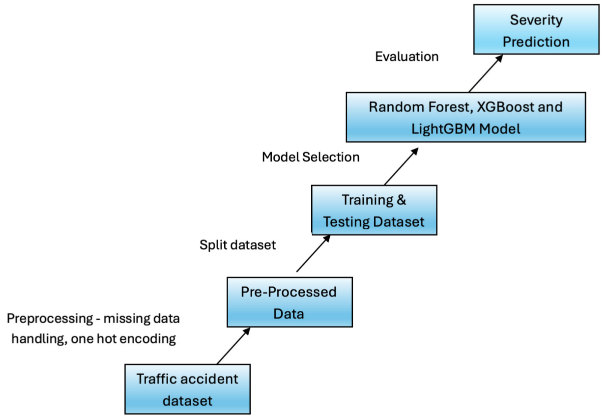
Severity Prediction
Evaluation
Random Forest, XGBoost and LightGBM Model
Model Selection
Training & Testing Dataset
Split dataset
Pre-Processed Data
Preprocessing - missing data handling, one hot encoding
Traffic accident dataset

Fig. 1. Proposed Work

C. Random Forest

When studying traffic accidents with imbalanced ordinal variable like 'Severity', the first critical step is to prepare the data for the Random Forest model by ensuring it is well-formatted and tidy. This data preprocessing step is required to address the imbalanced nature of the 'Severity' variable and assure the Random Forest model's performance.

To address potential missing data points, a SimpleImputer was used to fill in the gaps, with a particular emphasis on temperature values with the mean. Categorical variables such as 'Wind_Direction' and 'Sunrise_Sunset' were converted to numerical format using a ColumnTransformer and one-hot encoding method. To address the issue of imbalanced classes, the Synthetic Minority Over-sampling Technique (SMOTE) was used during the training phase. SMOTE helps to generate synthetic samples for the minority class, which balances the class distribution and improves the model's capacity to generalize to uncommon circumstances.

Following data preprocessing, the dataset is separated into training and testing sets. This separation is critical to our study, particularly given the imbalanced nature of the 'Severity' variable. The training set typically includes 70-80% of the data, with the remaining 20-30% reserved for testing the model's performance.

The Random Forest model is then trained on the prepared training data. Random Forest, being an ensemble learning technique, excels at handling imbalanced datasets and capturing intricate relationships within them. It excels at detecting patterns in parameters such as weather conditions, road types, and vehicle attributes that lead to different levels of accident severity. During training, the Random Forest model learns to categorize accident severity levels using the features provided. The Random Forest framework manages the imbalanced ordinal variable 'Severity', allowing the model to produce accurate predictions at all severity levels.

After training the Random Forest model, the testing set is used to evaluate its performance. Metrics such as accuracy, precision, recall, and F1-score are used to evaluate how well the model predicts accident severity, with an emphasis on its ability to manage the imbalanced ordinal variable 'Severity'.

D. XGBoost

A prediction model for accident severity is developed using XGBoost, a robust gradient boosting technique known for its efficiency and effectiveness in dealing with big datasets. The dataset comprises features such as 'Temperature(F)', 'Wind_Direction', 'Sunrise_Sunset', 'Wind_Speed(mph)', 'Humidity(%)', 'Precipitation(in)', 'Pressure(in)', and 'Visibility(mi)', as well as the goal variable 'Severity', which represents incident severity. To prepare the data for modeling, categorical features are encoded with LabelEncoder, which converts them to a numerical form suitable for model training. The dataset is then divided into training and testing sets to evaluate the model's performance.

These features were chosen because they are often connected with road conditions and have a major impact on accident severity. 'Temperature(F)' describes weather conditions, 'Wind_Direction' and 'Wind_Speed(mph)' describe wind conditions, 'Sunrise_Sunset' indicates visibility based on time of day, and 'Humidity(%)', 'Precipitation(in)', 'Pressure(in)', and 'Visibility(mi)' provide information on humidity levels, precipitation, air pressure, and visibility, respectively. By adding these features, the model can identify complicated patterns and correlations, allowing it to forecast accident severity with greater accuracy. The addition of XGBoost, which is known for its ability to handle complex relationships in data, improves the model's predictive skills, making it an appropriate choice for this purpose.

To address the imbalanced nature of the target variable 'Severity', in which some severity classes may be underrepresented compared to others, we introduce class weights into the XGBoost model. Class weights are inversely proportional to class frequencies, therefore the model prioritizes minority classes during training. This reduces the model's bias towards the majority class and enhances its capacity to generalize across all severity levels. In addition, we use the F1 score as the evaluation metric rather than accuracy since it provides a more thorough assessment by considering both precision and recall, which is especially significant in imbalanced datasets.

The XGBoost model is trained using training data and optimized for the F1 score metric. Early stopping is used to avoid overfitting and determine the ideal amount of boosting rounds. Following training, predictions are made on the test set, and the weighted F1 score is used to assess the model's performance. This methodology ensures that the predictive model is resilient, efficiently handles the dataset's imbalance, and gives correct estimates of accident severity, which is critical for enhancing road safety measures and emergency response systems.

In addition to model training and evaluation, our technique prioritizes interpretability and generalizability. We use XGBoost's feature importance analysis to determine which features have the most significant impact on predicting accident severity. This analysis provides stakeholders with insights into the factors that contribute most to accidents, allowing for more focused actions and policies. Furthermore, we use stratified K-fold cross-validation to guarantee that the model's performance is consistent across different subsets of the data. This strategy reduces the risk of overfitting and gives a more accurate estimate of the model's efficacy when used in real-world circumstances. By integrating these methodologies, our methodology seeks to produce not just a high-performing prediction model but also actionable insights and a dependable framework for analyzing and dealing with accidents.

E. LightGBM

LightGBM (Light Gradient Boosting) is a fast and efficient gradient boosting framework for large-scale data processing. It stands out for its capacity to efficiently manage large datasets with a low memory footprint, because to its novel use of histogram-based methods that optimise both training speed and memory utilisation. Furthermore, LightGBM supports categorical characteristics directly, which reduces the requirement for costly data preprocessing. This capability, together with approaches such as gradient-based one-sided sampling and exclusive feature bundling, allows

As with previous models, data preparation is a key step in the classification of the 'Severity' variable as well as maximising the model's performance and accuracy. The missing data has been handled by employing imputation approaches like mean or median for numerical variables like 'Temperature(F)', 'Visibility(mi)' and most frequent for categorical variables like 'Wind_Direction', 'Sunrise_Sunset' using Python's sklearn module. Unlike other models, one-hot encoding is not necessary since LightGBM can process categorical data directly. Following the completion of all preprocessing stages, the dataset is divided into training and validation sets (90% and 10% of the actual dataset), which are fed into the model. This is an essential step in training and validating your model's performance on untested data.

Then, using various features such as weather conditions ('Temperature(F)', wind direction ('Wind_Direction', 'Sunrise_Sunset') or road conditions ('Bump', 'Crossing', 'Traffic_Signal'), the split dataset is trained on the LightGBM model. Following training, predictions are made using the test dataset and verified using the validation set.

The model's accuracy is determined by using it to predict outcomes on the test dataset. Furthermore, a classification report for the LightGBM predictions is produced. This report contains the F1-score, precision, and recall for each class—critical metrics that provide information on how well the model distinguishes each class. Confusion matrices are finally presented for both the non-normalized and normalised versions of the LightGBM forecasts. These matrices give a detailed picture of the model's performance in categorising various types by graphically representing the accuracy of predictions and highlighting areas where the model may perform poorly.

IV. Experiments and Results

A. Random Forest

The methodology used in this study is intended to precisely identify the underlying causes of significant traffic accidents and provide effective preventive strategies, particularly when dealing with the imbalanced ordinal variable 'Severity'. The selection of important characteristics for projecting accident severity was a deliberate procedure aimed at comprehending the complex aspects that contribute to varied levels of accident severity. Parameters like 'Temperature(F)', 'Wind_Direction', 'Sunrise_Sunset','Wind_Speed(mph)', 'Humidity(%)', 'Precipitation(in)','Pressure(in)','Visibility(mi)' were carefully picked because of their possible influence on accident severity. These variables, among others, play important roles in the frequency and severity of traffic accidents, especially given the uneven nature of the 'Severity' variable. 'Temperature(F)' can affect road conditions and driver behavior; 'Wind_Direction' can affect visibility and driving conditions; and 'Sunrise_Sunset' can alter lighting conditions during accidents.

Model parameterization was critical for achieving optimal performance. GridSearchCV was used to optimize hyperparameters such as 'n_estimators', 'max_depth', 'min_samples_split','min_samples_leaf', and 'class_weight'. This technique sought to strike the optimal balance between model complexity and generalization. Notably, differences in parameterization can have a considerable impact on model performance. For example, higher values of'max_depth' may result in overfitting, whilst lower values may result in underfitting.

The Random Forest Classifier attained an accuracy of 0.73 on the test dataset, demonstrating that it is effective at predicting accident severity. A full categorization report offers information on precision, recall, and F1-score for each severity class. Notably, class 2 (moderate accidents) demonstrated the highest precision and recall, indicating the model's ability to identify these incidents. However, difficulties were encountered in appropriately identifying the least severe (class 1) and severe to most severe incidents (classes 3 and 4).

The results provide substantial insights into the primary causes influencing serious traffic accidents. Temperature, wind direction, and time of day were determined to be important factors. Preventive actions based on these observations could include greater road maintenance during inclement weather, improved traffic control methods during peak accident times, and targeted driver awareness campaigns.

B. XGBoost

The XGBoost model for forecasting accident severity received a weighted F1 score of 0.71. This score measures the model's ability to balance precision and recall across all severity levels, offering a comprehensive assessment of its performance on the test dataset. A weighted F1 score of 0.71 indicates that the model is reasonably effective at predicting accident severity, striking a reasonable compromise between correctly identifying serious accidents and minimizing false alarms.

The weighted F1 score of 0.71 achieved by the XGBoost model demonstrates its ability to balance precision and recall across severity levels. Despite the initial imbalance in the 'Severity' variable, the model's use of class weights and the scale_pos_weight parameter improved its ability to forecast accident severity. This result indicates that the model was successful in addressing the imbalanced nature of the 'Severity' variable, resulting in a respectable level of accuracy in forecasting the severity of accidents. These strategies guaranteed that the model's predictions were not biased toward the majority class, resulting in a more comprehensive and trustworthy predictor of accident severity.

C. LightGBM

The LightGBM model was trained on a large dataset that comprised different road conditions ('Bump', 'Crossing', 'Traffic Signals') and weather conditions ('Temperature(F)', 'Wind Direction', 'Sunrise Sunset', Humidity(%), Pressure(in), Visibility(mi)) as part of the experiment. The goal was to forecast the level of accident severity, which was divided into several categories ranging from minor to large incidents. Raw data was processed directly using LightGBM's effective handling of categorical and continuous data, eliminating the requirement for one-hot encoding or other intensive preprocessing.

The model generated a weighted average F1 score of 0.72 for the features of "Temperature(F)," "Wind_Direction," "Sunrise_Sunset," and "Severity." With an F1-score of 0.84, the model excels in predicting moderate occurrences (Class 1), implying that the most common incidents can be identified and classified accurately. However, it struggles with minor (Class 0) and severe occurrences (Class 3), getting very low F1-scores of 0.04 and 0.10, indicating considerable difficulty in reliably forecasting these less often but critical categories. Severe incidents (Class 2) exhibit poor performance with an F1-score of 0.26.

To reduce overfitting and improve prediction accuracy, the model was adjusted using parameters like num_leaves and min_data_in_leaf. The findings demonstrated LightGBM's ability to identify patterns linking particular weather and road conditions to the severity of accidents.

V. Conclusions

This study advances our understanding of the elements that influence serious traffic accidents and recommends preventive approaches based on data-driven insights. The Random Forest model's accuracy of 73% illustrates its suitability for real-world applications in accident severity prediction. With a weighted F1 score of 0.71 and 0.72, for the XGBoost and LightGBM model demonstrated balanced precision and recall for all severity levels. Both gradient boosting model's provides

Related Work	**Accuracy (%)/ F1-score**
Random Forest	75.5
XG Boost	0.90
LightGBM	0.69
Proposed Model	**Accuracy (%)/ F1-score**
Random Forest	73
XG Boost	0.71
LightGBM	0.72

TABLE I

COMPARISON OF RELATED RESEARCH AND PROPOSED MODEL ACCURACY

a strong predictor of accident severity by mitigating class imbalance. The findings emphasize the significance of meteorological conditions and time-related variables in accident occurrence. However, the model's performance on the least and most severe accidents could be enhanced with more data and feature engineering. Furthermore, the paper emphasizes the need to address imbalanced datasets with approaches such as SMOTE to improve model performance. Further developments could include investigating additional features such as road conditions, visibility, and traffic density. These variables may provide additional information for predicting accident severity. Adjusting the model's hyperparameters and implementing sophisticated techniques such as ensemble learning, or deep learning architectures may boost performance.

REFERENCES

[1] M. Iveta, A. Radovan, and B. Mihaljević, "Prediction of traffic accidents severity based on machine learning and multiclass classification model," in *2021 44th International Convention on Information, Communication and Electronic Technology (MIPRO)*, 2021, pp. 1701–1705.

[2] R. Vijithasena and W. Herath, "Data visualization and machine learning approach for analyzing severity of road accidents," in *2022 International Conference for Advancement in Technology (ICONAT)*, 2022, pp. 1–6.

[3] R. E. AlMamlook, K. M. Kwayu, M. R. Alkasisbeh, and A. A. Frefer, "Comparison of machine learning algorithms for predicting traffic accident severity," in *2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT)*, 2019, pp. 272–276.

[4] A. Esswidi, S. Ardchir, A. Daif, and M. Azouazi, "Severity prediction for traffic road accidents," *Journal of Theoretical and Applied Information Technology*, vol. 101, no. 8, 2023.

[5] S. Malik, H. El Sayed, M. A. Khan, and M. J. Khan, "Road accident severity prediction — a comparative analysis of machine learning algorithms," in *2021 IEEE Global Conference on Artificial Intelligence and Internet of Things (GCAIoT)*, 2021, pp. 69–74.

[6] C. P and S. M, "Road accident prediction and classification using machine learning," in *2022 IEEE 2nd Mysore Sub Section International Conference (MysuruCon)*, 2022, pp. 1–8.

[7] A. Venkat, M. Gokulnath, I. Thomas, and D. Ranjani, "Machine learning based analysis for road accident prediction," *International Journal of Emerging Technology and Innovative Engineering*, vol. 6, p. 31, 2020.

[8] O. Nedjmedine and M. Tahar, "Analysis of road accident factors using decision tree algorithm: a case of study algeria," in *2022 5th International Symposium on Informatics and its Applications (ISIA)*, 2022, pp. 1–6.

[9] A. Chand, S. Jayesh, and A. Bhasi, "Road traffic accidents: An overview of data sources, analysis techniques and contributing factors," *Materials Today: Proceedings*, vol. 47, p. 5135–5141, 2021.

[10] X.-L. Xia, B. Nan, and C. Xu, "Real-time traffic accident severity prediction using data mining technologies," in *2017 International Conference on Network and Information Systems for Computers (ICNISC)*, 2017, pp. 242–245.

[11] M. Akallouch, K. Fardousse, A. Bouhoute, and I. Berrada, "Exploring the risk factors influencing the road accident severity: Prediction with explanation," in *2023 International Wireless Communications and Mobile Computing (IWCMC)*, 2023, pp. 763–768.

[12] I. E. Mallahi, A. Dlia, J. Riffi, M. A. Mahraz, and H. Tairi, "Prediction of traffic accidents using random forest model," in *2022 International Conference on Intelligent Systems and Computer Vision (ISCV)*, 2022, pp. 1–7.