Data Analysis and Prediction on Road Accident

Gendhal Vaidya1, Pranjali Muley2, Shraddha Vaidya3, Parikshit Muley4 ,Anurag Rahate5 , Shalaka Deshpande6 , Prajakta Ghyar7 ,Janya Kinker8 , Mansi Patel9

1 Yeshwantrao Chavan College of Engineering/Computer Technology, Nagpur, India

Email: [gendlalvaidya@gmail.com](mailto:gendlalvaidya@gmail.com)

Email: [shraddha.karale1306@gmail.com](mailto:shraddha.karale1306@gmail.com)

2Data Meghe Medical College and Shalinitai Meghe Hospital & Research Center,Nagpur,Maharshtra

Email: [drpranjali11@gmail.com](mailto:drpranjali11@gmail.com)

Email: [drparikshit19@yahoo.com](mailto:drparikshit19@yahoo.com)

**Abstract**

Road accidents pose a significant threat to public safety and infrastructure. In recent years, the application of machine learning techniques in predicting and analyzing road car accidents has gained traction. This survey paper aims to provide a comprehensive overview of the research conducted in this field, focusing on the use of machine learning models for road car accident prediction. By analyzing the document "Road Accident Prediction Using a Machine-Learning-Enabled Data Analysis," we delve into the key findings, methodologies, and implications of using machine learning for road accident prediction. The study utilizes explainable machine learning models to predict and interpret errors in road accident hotspots using various data sets combined. The goal is to aid drivers and government authorities in detecting stressful driving locations and potential accident-prone areas before accidents occur.

**Introduction**

Traffic accidents are a major global concern, causing significant harm to public health, safety, and the economy. To address these issues, research is focused on data analytics and machine learning (ML) to better understand and predict traffic accidents. By analyzing traffic data, the goal is to identify patterns, predict accident risks, and find ways to improve road safety, reduce injuries, and prevent fatalities.

In recent years, various machine learning techniques, ranging from traditional models like decision trees and random forests to more advanced methods like convolutional neural networks (CNNs) and long short-term memory (LSTM) networks, have been used to analyze traffic-related data. These methods are effective in handling large and complex datasets, such as those gathered from sensors, cameras, mobile devices, and traffic authorities. Machine learning helps predict accident severity, frequency, risk, and even the duration of accidents, using factors like weather, traffic flow, road conditions, and time of day.

However, there are challenges like data imbalance, data quality issues, and making the models more interpretable still need to be addressed. With the growing availability of real-time data from sources like GPS, traffic sensors, and even social media, there's a strong push toward improving real-time accident prediction and decision-making systems. Future research will likely focus on better integrating diverse data sources, enhancing model , and exploring new technologies like autonomous vehicles, driver behavioural pattern to improve traffic safety even further.

This research delves into the various machine learning techniques used to predict and analyze traffic accidents, exploring key findings, current trends, and the future direction of the field. It emphasizes the importance of feature engineering, evaluating models, and predicting accident prone areas and visualization for the same that can ultimately lead to safer roads for everyone.

**Literature Review**

Various methods are used to analyse and predict traffic accidents, with both statistical and machine learning approaches being prominent in the literature.Statistical methods, such as logistic regression, ordered probit models, and mixed logit models, are employed to analyse traffic accident data and determine the influence of different factors. However, these models frequently assume linear relationships between variables, which might not always be accurate [1].

Machine learning techniques, including neural networks, support vector machines, and decision trees, are also applied in this field due to their ability to handle complex, multidimensional data. Deep learning models like Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) are used for predicting traffic accident severity with strong predictive capabilities [1].

Specific models such as Lasso, Ridge, Support Vector Regression (SVR), Decision Tree Regression (DTR), Random Forest Regression (RFR), Multilayer Perceptron(MLP) and Autoregressive Moving Average Model (ARMA) have been used in the analysis [2]. The Random Forest algorithm has been shown to be particularly effective in classifying the severity of injuries [3]. A deep learning model based on LSTM, called TARPML, has been developed for predicting traffic accident risk, outperforming other models in terms of prediction errors. Additionally, the Extreme Gradient Boosting (XGBoost) model\* has been successful in predicting errors in road accident hotspots [4].

Several factors are known to influence traffic accidents:

Traffic accident risk is not uniformly distributed, exhibiting temporal patterns and regional spatial correlations [2].

Factors such as roadway conditions, accident characteristics, vehicle type, casualties, and environmental features contribute to the severity of accidents [1].

Driver-related factors, including drinking and driving, seatbelt use, and driver age, significantly influence injury severity [5][1].

Weather conditions, including rain, snow, and fog, can impair the performance of sensors in autonomous vehicles [6].

Specific weather conditions like temperature, humidity, rainfall, wind, air quality, and visibility affect the selection of intercity travel modes [7l].

Spatial variables, such as province and road type, are critical predictors of errors in accident hotspots [4].

Fatigue, alongside weather variables like dew point, is a key factor in accident hotspots [4].

Traffic flow has a non-linear relationship with accident rates, and precipitation is associated with increased accident frequency [7].

Road hazards, such as vehicle loads, objects, previous accidents, or animals on the road also influence accident risk [8].

The location of the accident, the type of intersection, and the type of vehicle should be considered when analysing accidents [8].

Various data and methodologies are used:

Data sources include historical accident records, telematics data, and weather information [2][4].

Data preprocessing, involving cleaning, integration, and data type conversions, including transforming text data into numerical values is essential [2].

Geographic coordinates, like latitude and longitude, are often combined with accident location data [2].

A confusion matrix and kappa values are used to compare model accuracy, with kappa values measuring the agreement between two observers [5].

Feature selection techniques, like permutation importance, help identify the most influential variables for model training [4].

Performance is evaluated using metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Area Under the Curve (AUC) [2][4].

Bayesian methods are used to incorporate sample information and improve model estimation [6].

SHAP values are used to interpret the most influential predictors in machine learning models [4].

Regarding autonomous vehicles and weather:

Adverse weather conditions, including rain, snow, fog, and hail, can negatively affect sensors in autonomous vehicles, such as LIDAR, GPS, cameras, and radar [6].

Rain can reduce signal power and increase interference, posing a particular problem for millimeter-wave radar [6].

**Methodology**

The intelligent surveillance system is designed to detect the presence and activities of individuals within the field of view of a CCTV camera. It offers two primary functionalities: identifying the presence of individuals and detecting their activities, extending beyond basic movements like walking. The system features an initial verification window requiring the user to input their email, password, and the IP address of the CCTV camera. After successful verification, users can log in using only their email and password. Additionally, a dedicated interface is provided for streaming live CCTV footage.  
The proposed methodology for the project is outlined as follows:

Data Gathering and Integration: The dataset focuses on the topic of transport, specifically road safety data, published by the Department for Transport. It includes provisional mid-year unvalidated data on vehicles, casualties, and collisions for the years 2019 to 2023. The total dataset size is 8.66 MB, obtained from Road Safety Data on data.gov.uk.

Data Preprocessing: PySpark's DataFrame API will be utilized for data cleansing and transformation, addressing missing data, managing outliers, and rectifying inconsistencies. Supplementary data manipulation will be performed using Python libraries such as Pandas for advanced data handling and NumPy for numerical computations.

Exploratory Data Analysis (EDA): Descriptive statistics will be generated using PySpark's statistical functions. Visualization will be performed using Python libraries, with Matplotlib and Seaborn for static visualizations and Plotly for dynamic visualizations. Correlation functions from PySpark ML will be employed to uncover relationships among variables.

**Data Analysis:** The data analysis process includes several critical steps:

1. *Feature Selection:* Features will be selected using the SelectKBest class from sklearn, which identifies the top k features based on statistical scores, ensuring the most relevant variables are used for model development.
2. *Data Splitting:* The dataset will be divided into training and testing sets using the train\_test\_split function from sklearn. This ensures a clear distinction between data for training the model and evaluating its performance on unseen data.
3. *Feature Standardization:* Standardization will be applied using StandardScaler, which removes the mean and scales features to unit variance. This step is essential for machine learning algorithms sensitive to feature scales.
4. *Model Initialization and Training:* Four classification models will be initialized and trained:
   * Logistic Regression: A supervised learning algorithm used for classification tasks to predict categorical outcomes.
   * KNN Classifier: Predicts class labels based on the majority class of nearest neighbors in the training data.
   * Random Forest Classifier: An ensemble method that aggregates predictions from multiple decision trees for improved accuracy and robustness.
   * Decision Tree Classifier: Recursively splits the data based on criteria to create a tree structure for classification and regression tasks.

Model Development: Building upon the prepared data, regression and classification models will be developed using PySpark ML, including Linear Regression, Generalized Linear Regression, Random Forest Regression, and Decision Tree Regression.

Prediction and Scenario Analysis: The trained models will be employed to make predictions on new datasets and simulate various scenarios, providing insights into possible outcomes under different conditions.

Results Synthesis and Reporting: Final results will be synthesized using Pandas and Matplotlib to generate summary tables and visualizations. Jupyter Notebooks will provide an interactive and reproducible research report, while Tableau will be utilized for presenting the analysis and visualizations effectively.

**Experimentation and result**

Accident Trends : Analyze temporal trends to determine if accident rates are on the rise or decline over time.

Patterns by Constraints : Examine patterns influenced by factors such as time of day, weather conditions, day of the week, and location.

Model Accuracy : Develop and validate predictive models that forecast accident occurrences using historical data and various constraints.

Predictive Insights : Create models that pinpoint accident hotspots and periods of increased risk with high accuracy.

Risk Classification : Classify accidents into different risk levels based on factors like severity, time, and location.

Hotspot Mapping : Use GIS tools to identify and visually map accident hotspots, highlighting areas with elevated risk.

Spatial Patterns : Investigate spatial patterns to gain insights into how geographical factors affect accident rates.