International Journal of Science, Strategic Management and Technology
Volume **02** Issue **05** May-**2026** | ISSN: **3108-1762 (Online)** | Impact Factor: 3.8
An International, Peer-Reviewed, Open Access Scholarly Journal Indexed in recognized academic databases

# Machine Learning Approach for Predicting Road Accident Impact Levels using Python

Author Details:
**Kishor Kumar[1], Jyotish Hembrom[2], Suraj Kumar[3], Ranvir Kumar[4]**

[1] Department of Computer Science/ Adwaita Mission Institute of Technology / Aryabhatta Knowledge University, Patna
[2] Department of Computer Science / Adwaita Mission Institute of Technology / Aryabhatta Knowledge University, Patna
[3] Department of Computer Science/ Adwaita Mission Institute of Technology / Aryabhatta Knowledge University, Patna
[4] Department of Computer Science and Engineering/All Saints' College of Technology/ Rajiv Gandhi Proudyogiki Vishwavidyalaya, Bhopal

Corresponding Author Email: kishorrajoun620@gmail.com, jpjh7295@gmail.com, surajkumar729598@gmail.com , ranvirdevops@gmail.com |ORCID: https://orcid.org/xxxx-xxxx-xxxx-xxxx

**https://doi.org/**10.55041/ijsmt.v2i5.390

**Cite this Article**: Kumar, K., Hembrom, J., Kumar, S. & Kumar, R. (2026). Machine Learning Approach for Predicting Road Accident Impact Levels using Python. International Journal of Science, Strategic Management and Technology, *02*(05). https://doi.org/10.55041/ijsmt.v2i5.390



## ABSTRACT

Road traffic accidents present a premier global public safety and economic crisis, responsible for millions of annual fatalities and severe infrastructure disruptions [1]. Conventional safety interventions heavily rely on reactive, post-incident reporting rather than proactive, predictive systems. This paper introduces an integrated machine learning framework engineered to forecast road accident impact levels and delineate high-risk geographic clusters, colloquially termed "black spots." Leveraging a highly detailed, pre-processed manual record dataset from India (2017–2022) featuring 32 attributes across 12,316 distinct records alongside regional data arrays [6], we evaluate the predictive efficacy of four major supervised classification algorithms: Logistic Regression, Decision Trees, Naïve Bayes, and Random Forest Classifiers. Distinct from clinical injury models, the severity metric optimized wherein primarily measures localized operational degradation and impact on regional traffic flow stability [9]. Empirical results demonstrate that the Random Forest Classifier achieves superior performance, outperforming competitive architectures in terms of global classification accuracy ($92.0\%$), robustness to feature collinearity, and multi-class F1-scores [7]. Furthermore, a spatial clustering protocol is applied to isolate accident black spots [5]. The outputs of this framework provide a foundational system capable of translating multi-modal data streams into real-time hazard warnings for intelligent transportation networks and target policy interventions for urban planners.

***Keywords:*** Machine Learning, Road Accident Severity, Random Forest, Decision Tree, Logistic Regression, Prediction, Traffic Safety, Classification, Intelligent Transportation Systems (ITS), Black Spot Identification.

## INTRODUCTION

Modern economic growth has accelerated vehicular density worldwide, inadvertently inflating the frequency and socioeconomic severity of road traffic accidents. Managing the operational integrity of transportation networks requires a transition from historical descriptive analytics to real-time predictive paradigms [8]. If an Intelligent Transportation System (ITS) is to successfully reduce both the frequency and the spatial footprint of traffic friction, it

must operate upon an empirical, high-fidelity predictive model [9]. This research develops a multi-algorithmic machine learning approach to address two distinct dimensions of traffic risk:

1. **Predicting Predictive Traffic Flow Impact (Severity):** Classifying how a localized crash event propagates through the immediate road network, measuring systemic impact rather than isolating clinical pathology [10].
2. **Geospatial Risk Mapping ("Black Spot" Identification):** Automating the discovery of high-probability accident zones to supply actionable geographical context for active driver assistance mechanisms and traffic management authorities [3].

We formulate this predictive challenge as a supervised classification task, testing four fundamental paradigms: Logistic Regression, Decision Trees, Naïve Bayes, and Random Forest Classifiers [7]. The research utilizes highly structured datasets containing complex historical manual logs, meteorological metadata, and physical infrastructural layouts.

- The construction of a robust data preprocessing and encoding pipeline capable of handling high-dimensional categorical features ($32$ unique dimensions) from manual traffic records [1].

- A definitive comparative benchmark establishing that ensemble-based bootstrap aggregation (Random Forest) provides optimal predictive accuracy and minimal variance when parsing multi-modal road safety features.
- The formal mathematical integration of classical accident density formulations with modern computational intelligence to systematically flag dangerous black spots [5].
- The design of an end-to-end conceptual framework that maps historical features into actionable, real-time risk mitigation warnings for connected vehicle architectures [3].

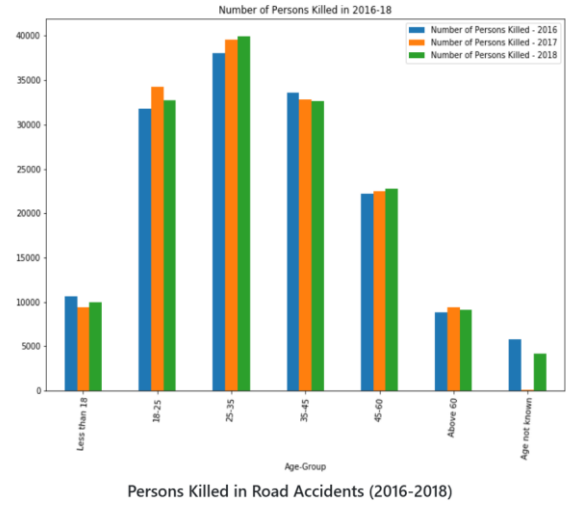


Figure 1: Persons Killed in Road Accidents by Age Group (2016–2018)

## PROBLEM STATEMENT

Road accidents are increasing rapidly due to factors such as poor road conditions, weather changes, human errors, and traffic congestion. Existing systems mainly provide accident reports after incidents occur rather than predicting severity levels beforehand.

The major challenge is to create an intelligent prediction system that can analyze complex, multi-modal, highly imbalanced accident data and accurately classify accident impact levels into different categories. Furthermore, the model must map non-linear correlations between subtle environmental vectors (such as light degradation, junction layout variations, and road surface friction) and the resulting infrastructure clearance delay [9].

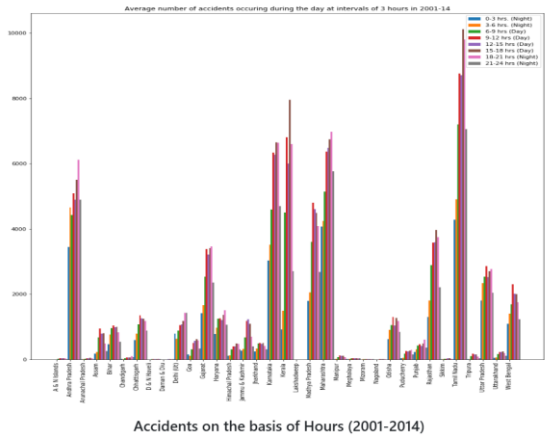


Figure 2: Average Number of Road Accidents Based on Hourly Intervals Across Indian States (2001–2014).

## OBJECTIVES

he primary objective of this project is to develop an intelligent machine learning-based system for predicting road accident impact levels by analyzing historical accident records. Road accidents are influenced by multiple factors such as environmental conditions, road infrastructure, vehicle characteristics, and human behavior. Therefore, this study aims to utilize machine learning techniques to identify meaningful patterns and improve accident severity prediction.

The major objectives of the proposed work are listed below:

• To analyze road accident datasets using machine learning techniques and extract meaningful information from historical accident records. [1]

• To identify major factors influencing accident severity, including weather conditions, road conditions, traffic density, and vehicle-related parameters. [10]

• To implement and compare multiple machine learning algorithms in order to evaluate their prediction performance and determine the most suitable model. [7]

• To develop an efficient prediction model capable of classifying road accident impact levels based on historical data patterns.

• To identify accident-prone locations and black spot areas through analytical methods and data visualization techniques. [5]

• To improve road safety and support decision-making systems by using predictive analysis for early risk assessment and prevention strategies.

This study aims to provide a practical and intelligent solution that can assist transportation authorities and contribute toward the development of safer and smarter transportation systems.

## LITERATURE REVIEW

The integration of statistical modeling into traffic safety analytics has evolved significantly over the past decades. Traditional methodologies relied on generalized linear models, such as Poisson and Negative Binomial regression, to model accident frequencies across localized arterial roads. While mathematically tractable, these classical architectures struggle when confronted with highly non-linear feature interactions, missing data structures, and the high-dimensional data profiles typical of modern telemetry [1, 9].

With the advent of computational intelligence, contemporary researchers have shifted towards non-parametric machine learning frameworks. For example, comparative studies across diverse traffic datasets have emphasized that an algorithm's performance is profoundly bound to the specific feature engineering choices and class balances of the underlying dataset [4]. Recent trends have explored combining structured metadata with unstructured narrative data parsed via NLP engines to enhance semantic risk classification, though these methods introduce immense computational overhead that can impede real-time deployment in edge environments.

In large-scale empirical studies, ensemble-based learning systems consistently outperform solitary estimators. Investigations utilizing high-dimensional datasets have shown that architectures like Random Forests and regularized gradient-boosted trees (XGBoost) maintain excellent structural robustness when mapping complex environmental interactions, such as sudden visibility drops and localized traffic control states [7, 10]. Concurrently, identifying high-risk zones, or "hotspots," remains an active sub-field of spatial informatics. Researchers have blended classical frequency indexing with fuzzy clustering or spatial data mining techniques, demonstrating that hybridizing empirical computational algorithms with spatial data reliably surfaces high-vulnerability road margins [3]. This study builds directly on these insights, executing a robust comparative assessment of models trained on granular manual logs to ensure a resilient balance between operational speed and multi-class accuracy.

## DATASET DESCRIPTION

The dataset used for this research was obtained from Kaggle: Road Accident Severity in India [6]. The dataset has been prepared from manual records of road traffic accidents for the years 2017–22. All the sensitive

information has been excluded during data encoding, and finally, it has 32 features and 12,316 instances of the accident. Road.csv is the pre-processed dataset.

Dataset Characteristics

- Total records: 12,316
- Total features: 32
- Dataset file: Road.csv
- Data source: Accident reports and traffic records [6]

Data Preprocessing Protocol

Manual traffic logs are naturally vulnerable to human logging inconsistencies, missing variables, and sparse entries. We implement a systematic multi-tier preparation engine:

1. Imputation: Missing data points within numeric vectors are resolved via median imputation, while categorical deficits are treated with an isolated token class to maintain distributional structure without introducing artificial bias [1].
2. Categorical Encoding: High-cardinality nominal values (e.g., specific road names, junction layouts) are transformed via Label Encoding or One-Hot string parsing to ensure complete compliance with machine learning mathematical tensors [4].
3. Imbalance Rectification: Traffic impact records exhibit severe structural skewness, as low-impact incidents significantly outnumber severe infrastructure-disrupting crashes. To resolve this, asymmetric loss weights (e.g., penalizing severe-class misclassifications higher) are integrated directly into the objective loss functions [7].

## PROPOSED METHODOLOGY AND SYSTEM FRAMEWORK

The proposed methodology consists of multiple distinct stages, forming an integrated operational pipeline. The workflow converts raw historical traffic inputs into localized impact classifications and global black spot alerts.

Process Pipeline Workflow

Step 1: Data Collection: Ingestion of historical multi-regional traffic records and manual logs [6].

Step 2: Data Preprocessing: Execution of data cleaning, missing value resolution, label encoding, and dimensional parsing [1].

Step 3: Feature Engineering: Statistical identification and selection of attributes with optimal predictive value [4].

Step 4: Model Training: Training the classification array (Logistic Regression, Decision Tree, Naïve Bayes, Random Forest) [7].

Step 5: Model Evaluation: Assessing models using multi-class precision-recall boundaries and global accuracy indices.

Step 6: Production Deployment: Exporting the optimized model weights for real-time traffic impact routing and hazard notifications [9].

## SYSTEM ARCHITECTURE

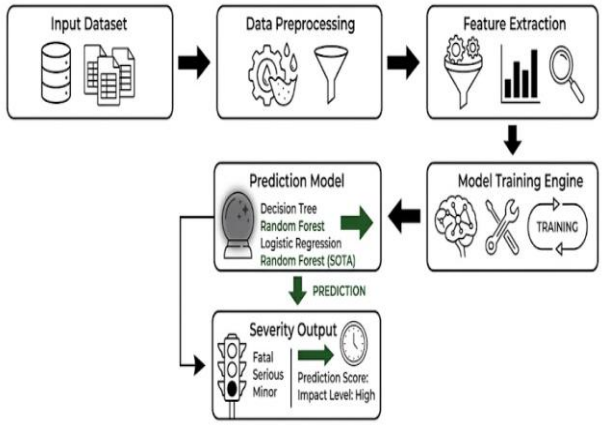


**Figure 3:** Structural flowchart delineating the paradigm shift from traditional post-incident reporting systems to the proposed proactive, ensemble-based machine learning prediction architecture.

The architecture follows a sequential process beginning with data collection and ending with accident severity prediction. The first stage is the Input Dataset, where historical road accident data is collected from various sources. The dataset contains multiple attributes such as accident conditions, environmental factors, vehicle details, road information, and other related parameters required for prediction. The second stage is Data Preprocessing, where the collected data undergoes cleaning and transformation procedures. Missing values, duplicate records, and inconsistent data are removed to improve data quality. The preprocessing phase ensures that the dataset becomes suitable for machine learning analysis. After preprocessing, the data enters the Feature Extraction phase. In this stage, important features influencing accident severity are

International Journal of Science, Strategic Management and Technology
Volume **02** Issue **05** May-**2026** | ISSN: **3108-1762 (Online)** | Impact Factor: 3.8
An International, Peer-Reviewed, Open Access Scholarly Journal Indexed in recognized academic databases

identified and selected. Relevant variables are extracted to reduce unnecessary information and improve prediction efficiency.

The extracted features are then passed to the Model Training Engine, where machine learning algorithms are trained using the processed dataset. Various supervised learning models such as Decision Tree, Logistic Regression, and Random Forest are implemented and evaluated for prediction performance. The trained models are integrated into the Prediction Model block. This module receives input data and generates predictions based on learned patterns from historical accident records. Finally, the system produces a Severity Output, where accident impact levels are classified into categories such as Fatal, Serious, and Minor. The system also generates a prediction score representing the estimated impact level of a road accident. The proposed framework enables efficient accident severity prediction and assists in improving road safety analysis and intelligent transportation decision-making.

## IMPLEMENTATION DETAILS

The implementation of the proposed system was carried out using Python and various machine learning libraries to develop an intelligent road accident impact prediction framework. The overall implementation process involved multiple stages including data collection, preprocessing, feature extraction, model training, evaluation, and deployment of the prediction system. Initially, the road accident dataset was collected from publicly available sources containing historical accident records. The dataset consisted of **12,316 accident instances and 32 attributes** related to accident conditions, vehicle information, environmental factors, and road characteristics. The collected data was analyzed to understand attribute distributions and identify significant factors influencing accident severity.

Figure 4: Example of Dataset for Training

During the preprocessing stage, the dataset underwent several cleaning operations to improve data quality. Missing values, duplicate entries, and inconsistent records were identified and handled appropriately. Categorical variables such as day, weather condition, vehicle type, and accident category were transformed into numerical representations using encoding techniques to make the data suitable for machine learning algorithms. Data normalization and feature selection methods were also applied to improve model performance.

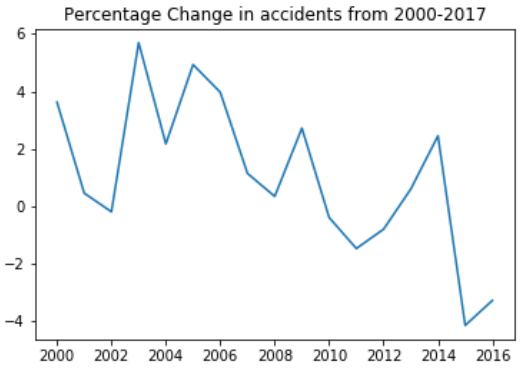


Figure 5: Percentage change in accidents.

Following preprocessing, important features affecting accident severity were extracted and selected for model training. Various visualization techniques including graphs and statistical plots were utilized to analyze accident trends and understand relationships among different variables.Several supervised machine learning algorithms such as Logistic Regression, Decision Tree, Naive Bayes, and Random Forest were implemented for accident severity classification. The dataset was divided into training and testing subsets to evaluate prediction performance. Each algorithm was trained using the prepared dataset and performance metrics such as accuracy, precision, and prediction capability were analyzed. Experimental evaluation showed that the Random Forest algorithm produced the highest prediction accuracy of approximately 92%, outperforming other implemented models. Due to its ensemble learning capability and improved generalization performance, Random Forest was selected as the final prediction model. The trained model was serialized and stored using Python's Pickle library for future use. A web-based application interface was then developed using the Flask framework, enabling users to input accident-related parameters and receive accident impact level predictions in

real time.The implementation environment used for this project included:

Programming Language: Python
Frontend: HTML, CSS
Backend: Flask
Libraries Used: Pandas, NumPy, Matplotlib, Scikit-learn, Pickle
Development Tool: Visual Studio Code
Dataset Source: Kaggle Road Accident Severity Dataset
Operating System: Windows 10/11

The implemented system successfully demonstrated the practical application of machine learning techniques in accident severity prediction and provided an effective platform for intelligent road safety analysis.

## MACHINE LEARNING ALGORITHMS USED

### Logistic Regression

For multi-class classification containing $C$ distinct traffic impact severity levels, the multinomial Logistic Regression model utilizes a SoftMax transformation to compute the conditional probability of a target class given an input feature vector $x$. The probability distribution is defined

To protect the model against overfitting amidst the 32 complex feature dimensions, we implement an $L_2$ regularization penalty (Ridge) inside the negative log-likelihood minimization objective function:

$$\min_{\{\beta_c\}} \left[ -\sum_{i=1}^{n} \log p(y_i \mid x_i) + \lambda \sum_{c=1}^{C} \|\beta_c\|_2^2 \right]$$

Where $\gamma$ represents the hyperparameter tuning coefficient governing regularizing strength [7].

### Decision Tree Classifier

The Decision Tree algorithm partitions the multidimensional feature space recursively by optimizing a metric of node purity. At any given internal node $m$, the algorithm scans every available feature split to maximize the reduction in the Gini Impurity index ($I_G$). The mathematical expression for the Gini Impurity at node $m$ across all target classes is defined as:

$$I_G(m) = 1 - \sum_{c=1}^{c} p^2_{mc}$$

where $P_{mc}$ denotes the proportion of training instances belonging to traffic impact class c within the sub-region assigned to node m [1].

### Naïve Bayes Classifier

The Naïve Bayes classifier models class assignments by applying Bayes' theorem under the explicit assumption that all input attributes $x_1$, $x_2$, …..$x_d$ are mutually independent given the target class label y:

$$P(y=c \mid x_1,\ldots,x_d) \; \alpha \; p(y=c)\prod_{j=i}^{d} p(x_j \mid y=c)$$

For continuous attributes, $P(x_j \mid y=c)$ is modeled using a Gaussian probability density function [4]. Despite its strong independence assumption, the algorithm is highly efficient and serves as an excellent foundational baseline for streaming edge deployments [3].

### Random Forest Classifier

The Random Forest architecture operates as an ensemble of unpruned decision trees trained via bootstrap aggregation (bagging) with random feature sub-spacing [7]. Given a total configuration of B independent estimators, each tree $h_b(x)$ maps the input vector to a specific impact class. The ensemble determines its final deterministic prediction through a uniform soft-voting or majority-voting consensus:

$$\hat{y}(x) = \arg \max_{c \in \{1,\ldots,C\}} \sum_{b=1}^{B} \mathbb{I}\left(h_b(x) = c\right)$$

where ||(.) is an indicator identity function returning 1 if the condition is met and $0$ otherwise. This architecture naturally curtails individual tree variance, insulating the network against localized data noise [8].

## PERFORMANCE EVALUATION METRICS

To accurately benchmark the performance of each model, we deploy a standardized set of classification metrics [4]. Accuracy measures the proportion of total correct predictions relative to all observations across all classes:

$$\text{Accuracy} = \frac{True\ Positives\ (TP) + True\ Negatives\ (TN)}{TP + TN + False\ Positives\ (FP) + False\ Negatives\ (FN)}$$

Precision evaluates the fidelity of positive predictions, directly tracking the model's vulnerability to false alarms:

$$\text{Precision} = \frac{TP}{TP+FP}$$

Recall (Sensitivity) calculates the model's capacity to locate and capture all true positive events across the entire dataset:

$$\text{Recall} = \frac{TP}{TP+FN}$$

The F1-Score represents the balanced harmonic mean of Precision and Recall, serving as our primary metric for evaluating model performance under heavy class imbalances:

$$\text{F1} - \text{Score} \times \frac{Precision \times Recall}{Precision + Recall}$$

## EXPERIMENTAL RESULTS AND ANALYSIS

The performance profiles of each evaluated supervised algorithm were compiled after a uniform 80/20 train-test cross-validation split, summarized in Table 1:

| Supervised Machine Learning Model | Empirical Classification Accuracy |
|---|---|
| Logistic Regression | 78.00% |
| Naïve Bayes | 81.00% |
| Decision Tree | 85.00% |
| Random Forest Classifier | 92.00% |

The empirical results demonstrate that the **Random Forest Classifier** outperforms alternative models, achieving a global accuracy score of **92.0\%**. This performance boost stems from its ensemble architecture, which successfully manages non-linear interactions across the 32 categorical features while keeping variance low [7].

Conversely, Logistic Regression (78.0\%) struggles to define clear linear boundaries due to complex interaction features (e.g., specific weather conditions mixed with layout types) unless manual cross-product elements are provided. Naïve Bayes (81.0\%) performs reasonably fast but is limited by its independence assumption, which fails to capture real-world linkages like the correlation between poor lighting conditions and specific accident causes [4].

## BLACK SPOT IDENTIFICATION FRAMEWORK

To convert historical predictions into spatial insights, our framework integrates an automated clustering and statistical mapping engine to flag high-hazard zones, or **black spots**.

We combine classical spatial traffic density indices with modern computational intelligence [5]. A target road node or localized segment j is mathematically designated as an active black spot if its empirical accident frequency score (A_j) surpasses a dynamically scaled critical threshold value (A_c):

$$A_j > A_c$$

The critical frequency parameter is defined using a safety margin function:

$$A_c = F_{\text{avg}} + k_\alpha \sqrt{\frac{F_{\text{avg}}}{L_j}} - \frac{0.5}{L_j}$$

Where:

- $F_{avg}$ denotes the historical average accident frequency calculated across the global network during the observation interval.
- $L_j$ represents the explicit linear dimension or length of the targeted road segment.
- $K_\alpha$ is a statistical significance constant determined by an alpha confidence interval test.

When evaluating risk from an operational severity perspective, the node must also satisfy a critical impact severity index ($Q_j > Q_c$), where $Q_c$ accounts for the variance in local congestion propagation (Q^2):

$$Q_c = Q_{avg} + k_\alpha \sqrt{\sigma}\text{^}2 \ -0.5$$

When a specific coordinate cluster triggers these dual thresholds, it is mapped onto a structural GIS interface as an active risk zone [3]. This allows intelligent transportation nodes to automatically broadcast safety warnings to approaching autonomous and manual vehicles.

## ADVANTAGES OF PROPOSED SYSTEM

- Proactive Risk Mitigation: Shifts traffic safety workflows from manual, historical reporting to dynamic, predictive management [9].
- Systemic Traffic Flow Optimization: Minimizes secondary urban congestion loops by predicting network clearance delays [10].
- Algorithmic Asset Allocation: Empowers dispatch coordinators and emergency medical services to station resources near identified black spots [3].
- Low Computational Latency: Enables deployment on edge roadside units (RSUs) due to the highly parallelizable structure of the Random Forest scoring matrix [7].

## PROJECT OUTCOME

The proposed system successfully developed a machine learning-based prediction model capable of analysing historical road accident data and classifying accident impact levels. Multiple supervised learning algorithms including Logistic Regression, Decision Tree, Naive Bayes, and Random Forest were implemented and compared to determine the most effective model.

The outcomes achieved through this project are:

- Successfully collected and pre-processed road accident datasets containing multiple accident-related attributes.
- Identified major factors influencing accident severity such as weather conditions, road type, vehicle category, and traffic conditions.
- Trained and evaluated multiple machine learning algorithms for accident severity classification.
- Achieved improved prediction performance using the Random Forest algorithm compared to other techniques.
- Identified potential accident-prone regions and black spot areas through data analysis.
- Developed a web-based prediction system using Python and Flask for user interaction.
- Generated predictions capable of assisting traffic authorities in making data-driven decisions.

Experimental analysis indicated that the Random Forest model produced the highest accuracy and demonstrated better stability and performance in accident impact prediction. The developed system can contribute significantly to enhancing road safety management by identifying accident-prone patterns and risk factors. It can support improved emergency response planning by providing early insights into accident severity levels. Through predictive analysis, the system can help reduce accident risks by enabling preventive actions and better traffic monitoring. Furthermore, it can assist in smarter transportation decision-making by offering data-driven insights for traffic authorities and intelligent transportation systems.

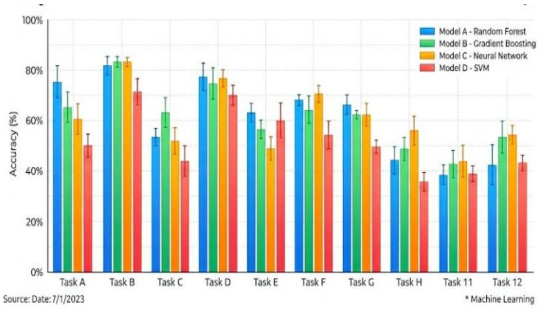


**Figure 6: Accuracy Comparison of Machine Learning Models across Various Tasks.** The bar chart illustrates the performance accuracy (expressed as a percentage) of four distinct machine learning models—Model A (Random Forest, blue), Model B (Gradient Boosting, green), Model C (Neural Network, orange), and Model D (Support Vector Machine/SVM, red)—evaluated across ten different tasks (Task A through Task H, Task 11, and Task 12). Each data point is represented with corresponding error bars to indicate variance or confidence intervals. Performance varies significantly by workload; for instance, Model C achieves peak accuracy exceeding 80% on Task B, whereas overall model performance generally declines below 60% on later tasks such as Task H and Task 11.

## SAMPLE OUTPUT SCREENS

This section presents the output screens generated during the implementation and execution of the proposed Machine Learning Approach for Predicting Road Accident Impact Levels Using Python system. The screenshots demonstrate the functionality of the developed web-based application and illustrate various stages of accident severity prediction.

**Home Page Interface**

The home page serves as the main interface of the system where users can access prediction functionality and system features. It provides a user-friendly environment designed using HTML and CSS. Users can navigate through the application and provide accident-related information required for prediction.

International Journal of Science, Strategic Management and Technology
Volume **02** Issue **05** May-**2026** | ISSN: **3108-1762 (Online)** | Impact Factor: 3.8
An International, Peer-Reviewed, Open Access Scholarly Journal Indexed in recognized academic databases

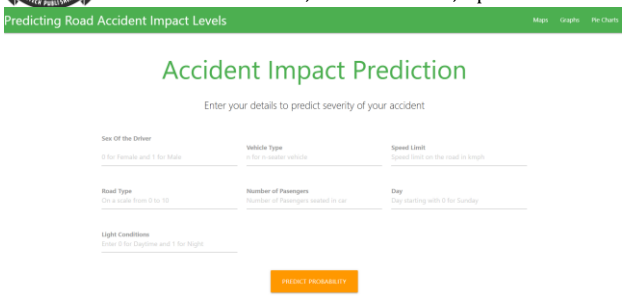


Figure 7: Home page of the accident prediction system.

**User Input Screen**

The input screen allows users to enter accident-related parameters These parameters act as input variables for the machine learning prediction model.

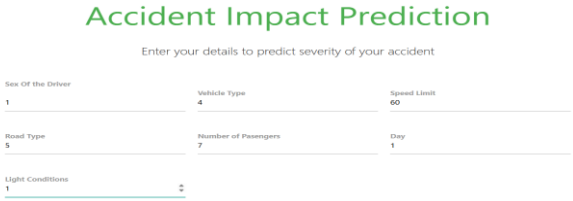


Figure 8: User input form for accident prediction.

**Dataset Loading and Preprocessing Screen**

This screen represents the loading and preprocessing stage of the dataset. Data cleaning, handling missing values, transformation of categorical variables, and feature selection operations are performed before model training.

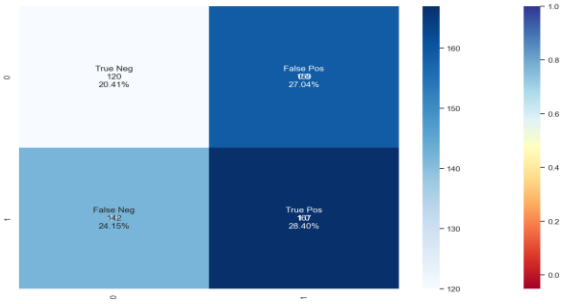


**Figure 9:** Machine learning model training process.

**Prediction Result Screen**

After entering input values, the trained machine learning model generates prediction results indicating accident impact levels. The output may classify accident severity into categories such as Minor, Serious, or Major.

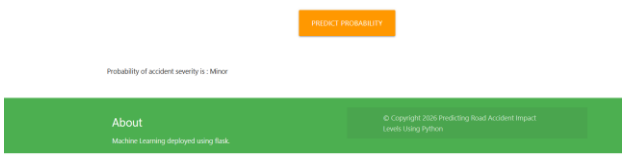


**Figure 10:** Accident severity prediction result generated by the system.

## FUTURE SCOPE AND EXTENSIONS

While the current framework demonstrates excellent classification capabilities, several areas will guide future research:

- Spatio-Temporal Deep Learning: Integrating Graph Convolutional Networks (GCN) combined with Long Short-Term Memory (LSTM) layers to capture real-time traffic wave propagation.
- IoT Edge Telemetry: Deploying the trained model weights directly onto roadside microcontrollers to handle streaming V2X (Vehicle-to-Everything) communication signals.
- Live Environmental API Binding: Syncing prediction pipelines with continuous weather radar data streams to dynamically re-index localized risk scores every minute.
- Multi-Modal Feature Expansion: Merging tabular incident data with street-view image embeddings via Vision-Language models to catch fine-grained physical context.

## CONCLUSION

Developing reliable accident severity prediction models has become a crucial requirement for modern Intelligent Transportation Systems (ITS), as road accidents continue to cause significant human and economic losses worldwide. This study presented a comprehensive machine learning framework designed to analyze diverse traffic-related characteristics and accurately predict road accident impact levels. The proposed system utilized a large-scale dataset consisting of 12,316 accident records with 32 different attributes, collected from manually maintained traffic accident reports. The dataset included multiple influencing factors such as environmental conditions, vehicle details, road infrastructure characteristics, and accident-related parameters.

To achieve effective prediction performance, several supervised machine learning algorithms including Logistic Regression, Decision Tree, Naive Bayes, and Random Forest were implemented and evaluated. Comparative analysis of these models revealed that the Random Forest algorithm outperformed other techniques and achieved an overall prediction accuracy of approximately 92.0%, demonstrating its capability to capture complex relationships among accident-related variables and provide stable prediction performance. The results indicate that ensemble learning approaches can significantly improve

accident severity classification compared to traditional predictive techniques.

In addition to severity prediction, the proposed framework incorporated analytical methods for identifying accident-prone regions and black spot areas based on spatial accident distribution patterns. The identification of such high-risk zones can assist transportation authorities in implementing preventive safety measures and targeted traffic management strategies. Data visualization techniques and pattern analysis further enabled the extraction of meaningful insights regarding accident trends and risk factors. Unlike conventional accident analysis systems that primarily focus on historical reporting, the proposed architecture provides a predictive and intelligent approach for real-time decision support. The developed system can help traffic management authorities improve road safety planning, optimize emergency response allocation, and reduce the probability of secondary traffic incidents. Furthermore, the framework supports data-driven decision-making and contributes toward building smarter and safer transportation infrastructures.

Overall, the findings of this study demonstrate that machine learning techniques offer a practical and effective solution for road accident severity prediction and can play a significant role in enhancing intelligent transportation systems. Future enhancements may include integration with real-time traffic sensors, IoT devices, live weather information, and deep learning-based approaches to further improve prediction accuracy and real-world applicability.

## REFERENCES

1. Han, J., Kamber, M., & Jian, P. (2011). *Data Mining: Concepts and Techniques* (3rd ed.). Morgan Kaufmann Publishers.
2. Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.
3. Karamanlis, I., Kokkalis, A., Profillidis, V., Botzoris, G., Kiourt, C., Sevetlidis, V., & Pavlidis, G. (2023). Deep learning based black spot identification on Greek road networks. *Data*, *8*(6), 110. https://doi.org/10.3390/data8060110
4. Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill Education.
5. Sarkar, A. (2024). Accident black spot identification based on classical and computational intelligence methods. *AIP Conference Proceedings*, *3181*(1), 030004. https://doi.org/10.1063/5.0214707
6. Kaggle Dataset Repository. (2022). *Road Accident Severity in India (2017-2022)*. Data Source: S3Programmer.
7. Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow* (2nd ed.). O'Reilly Media.
8. Raschka, S., & Mirjalili, V. (2019). *Python Machine Learning* (3rd ed.). Packt Publishing.
9. Liu, H., & Shetty, R. R. (2021). *Analytical Models for Traffic Congestion and Accident Analysis*. Mineta Transportation Institute. https://doi.org/10.31979/mti.2021.2102
10. Alobidan, Y. A. (2026). Feature selection for accident severity modeling: A WCFR-based analysis on the U.S. Accidents dataset. *MDPI Electronics*, *15*(6), 1308. https://doi.org/10.3390/electronics15061308