J-INTECH (Journal of Information and Technology)
Accredited Sinta 4 Ministry of Higher Education, Science and Technology Republic of Indonesia SK No. 10/C/C3/DT.05.00/2025
E-ISSN: 2580-720X || P-ISSN: 2303-1425

Traffic Accident Severity Classification System Using Random Forest Algorithm

Ega Muhammad Atsir[1*], Nurmalitasari[2], Aprilisa Arum Sari[3]

[1,2,3]Duta Bangsa University, Faculty of Computer Science, Information Systems, Jl. Bhayangkara No.55, Tipes, Kec. Serengan, Kota Surakarta, Jawa Tengah 57154, Indonesia

Keywords
Classification; Random Forest; System; Traffic Accident

Corresponding Author:
egamuhammadatsir18@gmail.com

Abstract
Traffic accidents pose a major concern in many countries, including Indonesia, causing considerable losses, injuries, and fatalities each year. Properly classifying the severity of these incidents is essential for authorities to establish preventive actions, apply effective countermeasures, and improve overall road safety. Conventional statistical techniques often fall short in capturing the intricate relationships among multiple influencing variables, such as weather, driver experience, vehicle type, number of vehicles, and casualty figures. To address this limitation, this study proposes a machine learning–based classification method using the Random Forest algorithm, known for its robustness in handling complex and high-dimensional data while identifying nonlinear patterns. The model was trained on a traffic accident dataset from Kaggle and incorporated important features, including driver age group, driving experience, type of vehicle, lighting and weather conditions, type of collision, number of vehicles involved, and casualties. The proposed system achieved 81% accuracy, 75% weighted precision, 81% weighted recall, and a weighted F1-score of 77%, demonstrating reliable performance in predicting accident severity levels Slight Injury, Serious Injury, and Fatal Injury. And providing useful insights for data-driven planning in traffic safety management.

1. Introduction

Road traffic incidents remain a significant public safety challenge worldwide, especially in developing nations like Indonesia, where rapid urbanization and increasing vehicle ownership have escalated accident ratesu[1]. Global health data places traffic accidents among the top causes of mortality, resulting in not only numerous deaths and injuries but also severe economic losses, healthcare system strain, and lasting social consequences[2]. Various contributing factors include inadequate infrastructure, poor weather conditions, low vehicle safety standards, and human-related issues such as driver fatigue, inexperience, or negligence[3]. Traditional analytical methods in Indonesia have largely relied on descriptive statistics and simple regression, which are insufficient to detect nonlinear patterns across these multifactorial contributors[4].

In response to the increasing complexity and volume of traffic-related data, the need for advanced analytical techniques has become more critical[5]. Machine learning, particularly the Random Forest algorithm, offers

robust capabilities for handling high-dimensional, heterogeneous data and capturing intricate, nonlinear relationships[6]. As an ensemble method, Random Forest builds multiple decision trees and aggregates their outcomes via majority voting, making it resilient to noise and overfitting[7]. Numerous studies have validated its effectiveness in classifying accident severity[8]. However, most prior research has focused on datasets from developed countries or used limited features, thus failing to capture the full spectrum of factors influencing accident outcomes.

To address this research gap, this study proposes a Random Forest-based classification model trained on a comprehensive traffic accident dataset from Kaggle[9]. The dataset includes relevant features such as driver demographics, vehicle type, environmental conditions, collision details, number of vehicles involved, and casualty counts[10]. The target variable accident severity is categorized into three levels: Slight Injury, Serious Injury, and Fatal Injury. This study aims not only to uncover significant patterns influencing accident severity but also to develop a practical, deployable web-based classification system using Streamlit, ensuring real-time usability for non-technical users and supporting data-driven decision-making in traffic safety management[11].

This system is intended to serve multiple user groups, including traffic police, transportation agencies, and public safety researchers. It provides a practical tool to support rapid decision-making, incident prioritization, and long-term policy development based on classified accident severity. Developed using Streamlit, the system is deployed as a web-based application that can be accessed either locally or publicly, enabling real-time interaction and usability for both technical and non-technical users.

2. Research Method

The data utilized in this research was obtained from a publicly accessible traffic accident dataset hosted on Kaggle.

Table 1. Dataset Variables

Variables	Information
Age_band_of_driver	Categorical grouping based on the driver's age
Driving_experience	Classification of the driver's level of experience behind the wheel
Type_of_vehicle	Classification of the vehicle involved in the incident
Light_conditions	Illumination status at the location during the time of the accident
Weather_conditions	Atmospheric condition present when the accident occurred
Type_of_collision	Nature or category of the collision that took place
Number_of_vehicles_involved	Count of all vehicles participating in the accident
Number_of_casualties	Total number of individuals injured or killed as a result of the crash
Accident_severity	Level of injury severity classified into Slight Injury, Serious Injury, or Fatal Injury categories

This dataset comprises records of traffic accidents obtained from the Addis Ababa Sub-City Police Departments, originally collected for a master's research project. The data spans the years 2017 to 2020 and was derived from manually recorded accident reports. The raw dataset, named RTA Dataset.csv, consists of 12,316 instances and 32 features, though only the relevant attributes were selected for this study, including driver's age band, driving experience, type of vehicle involved, lighting conditions, weather conditions, collision type, number of vehicles involved, number of casualties, and the severity of the accident. The dataset has undergone preprocessing to ensure the removal of sensitive information and to facilitate the application of machine learning algorithms for severity classification. The accident severity itself is categorized into three classes: Slight Injury, Serious Injury, and Fatal Injury. The dataset is publicly available on Kaggle at: https://www.kaggle.com/datasets/saurabhshahane/road-traffic-accidents. The data was chosen due to its relevance and comprehensiveness, making it suitable for developing a reliable classification model[12].

Before building the classification model, the dataset underwent a preprocessing stage to ensure optimal quality for analysis. First, categorical variables such as driver's age band, driving experience, vehicle type, lighting conditions, weather conditions, and collision type were encoded using the Label Encoding technique to convert categorical data into numerical form[13]. Numerical variables, such as the number of vehicles and casualties involved, were then scaled using StandardScaler to standardize the data distribution[14]. This process is essential to guarantee that all features have a balanced influence during the classification phase.

Random Forest was selected as the main classification technique because of its demonstrated effectiveness in managing complex and high-dimensional classification tasks[15]. Random Forest operates as an ensemble learning method that constructs numerous decision trees during the training phase. Each tree produces a classification outcome, and the final decision is derived from aggregating these results through majority voting[16]. This approach greatly minimizes the likelihood of overfitting while improving the model's reliability and predictive precision.

The classification workflow using the Random Forest algorithm in this study involves several essential steps. Initially, the raw data undergoes preprocessing, which includes transforming categorical variables through encoding and applying normalization to numerical attributes for consistency. The dataset, after processing, is partitioned into separate sets for training and testing. During training, the model constructs multiple decision trees by randomly drawing samples of data instances and attributes a method known as bootstrap sampling.

Table 2. Dataset Sharing

Category Data	Percentage
Training	80%
Testing	20%

In this study, the data was split into two subsets, with 80% assigned for model training and 20% reserved for testing. To achieve strong predictive performance, the Random Forest algorithm was trained using 100 separate decision trees.

Random Forest represents a widely applied ensemble technique in machine learning, suitable for solving both classification and regression problems. It constructs numerous decision trees by randomly sampling data and feature subsets during the training phase, which enhances generalization capability and mitigates the risk of overfitting[17]. Each tree in the Random Forest functions on its own, generating predictions using individually sampled training records and chosen features. In classification problems, the outcome is derived from the majority decision across the ensemble, whereas in regression, the result is based on the mean of all predictions.

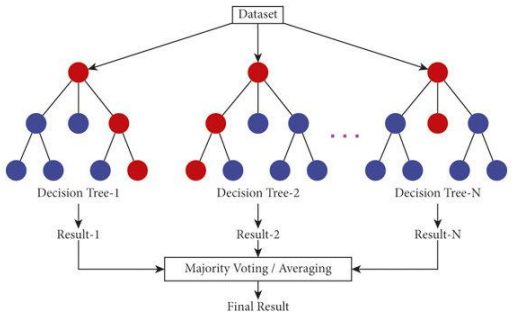

Figure 1. Random Forest Architecture and Voting Mechanism

The Figure 1 illustrates the fundamental structure of the Random Forest algorithm. The process begins with a dataset that is used to build multiple decision trees in parallel. Each tree is trained on a randomly selected

subset of the data and features, allowing for diversity among the trees. Every individual tree generates its own prediction based on learned patterns. These predictions are then aggregated using a majority voting mechanism in classification tasks, or averaging in regression tasks. The final output is the result of this ensemble decision-making process, which improves overall model accuracy and robustness by reducing overfitting and enhancing generalization capability.

The formulas used for each metric are as follows:

a. Accuracy
Accuracy is a widely recognized performance metric that quantifies how many predictions made by the model are correct in relation to the total number of predictions. It is often used as an overall indicator of model performance, particularly in datasets where the class distribution is relatively balanced. The equation used to calculate Accuracy is shown below:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

b. Precision
Precision represents the fraction of true positives among all predicted positive outcomes and is especially important in cases where false alarms carry significant consequences[18]. A higher precision indicates fewer false alarms. The formula for Precision is:

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

c. Recall
Recall often known as sensitivity or the true positive rate—describes the model's capability to correctly identify all true positive cases present in the data[19]. It is especially important when missing actual positive cases can lead to significant consequences. The formula is:

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

d. F1-Score
The F1-score is calculated as the harmonic mean of precision and recall, providing a balanced evaluation metric particularly useful in datasets with class imbalance or when it is important to reduce both false positives and false negatives. A higher F1-score suggests a better balance between these two metrics. The formula is presented as follows:

$$F1\text{-}score = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \quad (4)$$

The final stage of this research involves deploying the evaluated classification model into a functional system that can be directly utilized by end-usersIn this project, the trained model was deployed using Streamlit, a Python-based open-source tool that enables the development of dynamic web applications designed for machine learning use cases[20].

3. Result and Discussions

The dataset used in this study was obtained from Kaggle and contains comprehensive records of road traffic accidents, including categorical and numerical features that are critical for classification.

Table 3. Dataset Snippet

Age_band_of_driver	Driving_experience	Type_of_vehicle	Light_conditions	Weather_conditions	Type_of_collision	Number_of_vehicles_involved	Number_of_casualties	Accident_severity
18-30	1-2yr	Automobile	Daylight	Normal	Collision with roadside-parked vehicles	2	2	Slight Injury
31-50	Above 10yr	Public (> 45 seats)	Daylight	Normal	Vehicle with vehicle collision	2	2	Slight Injury
18-30	2-5yr	Automobile	Daylight	Normal	Vehicle with vehicle collision	1	1	Slight Injury

This dataset includes variables such as driver age group, driving experience, type of vehicle, lighting and weather conditions, type of collision, number of vehicles involved, and the number of casualties. These features were selected based on their relevance to the classification of accident severity.

Before training the model, the dataset underwent several preprocessing steps. The preprocessing stage involved converting categorical attributes into numeric form through Label Encoding, while numerical data was normalized using the StandardScaler approach. These steps are crucial to enable the model to efficiently interpret and learn from the input data.

Table 4. Data Preparation Results

Age_band_of_driver	Driving_experience	Type_of_vehicle	Light_conditions	Weather_conditions	Type_of_collision	Number_of_vehicles_involved	Number_of_casualties	Accident_severity
-0.966718	-1.377891	-1.214155	0.621167	-0.362059	-1.247296	-0.059061	0.448649	Slight Injury
-0.219035	0.487776	0.868879	0.621167	-0.362059	0.570613	-0.059061	0.448649	Slight Injury
-0.966718	-0.756002	2.005080	-1.644466	-0.362059	0.570613	-0.059061	0.448649	Slight Injury

The Table 4 displays several rows of preprocessed input data used in training the classification model. Each feature such as driver age group, driving experience, vehicle type, light and weather conditions, collision type, number of vehicles involved, and number of casualties has been normalized or encoded to prepare it for the machine learning process. The target variable, *Accident_severity*, remains in its original categorical form (e.g., *Slight Injury*) for classification purposes.

Random Forest was selected as the main classification algorithm because of its strength in processing complex and high-dimensional datasets. The training process utilized 80% of the data, while the remaining 20% was reserved for testing the model's performance. A total of 100 decision trees (estimators) were used to optimize

classification performance. This ensemble method aggregates multiple decision tree outputs through majority voting, improving both stability and accuracy.

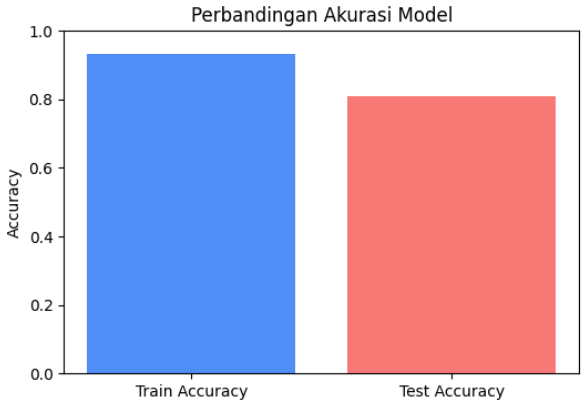

Figure 2. Model Accuracy Comparison

To assess how well the Random Forest model performed, multiple evaluation metrics were utilized, including accuracy, precision, recall, F1-score, and the confusion matrix. The dataset was partitioned into two subsets, with 80% used for training and the remaining 20% reserved for testing. Following data preprocessing and model training, the algorithm achieved a training accuracy of 93.36% and a testing accuracy of 81.05%, as illustrated in Figure 2.

The model's performance was evaluated using a confusion matrix and several classification metrics to assess its ability to correctly predict each category of accident severity. This evaluation aims to measure the accuracy, precision, recall, and F1-score for each class, providing a comprehensive understanding of the model's strengths and weaknesses across both majority and minority classes.

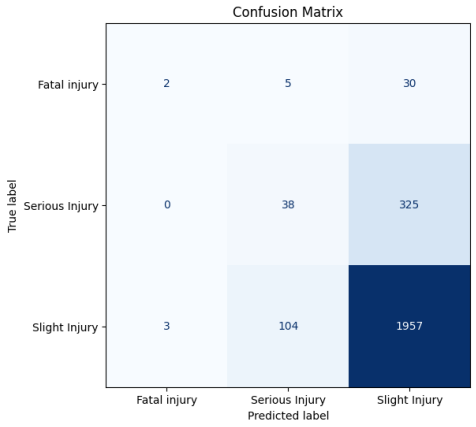

Figure 3. Confusion Matrix of Classification Results

The confusion matrix in Figure 3 shows the distribution of correct and incorrect predictions for each class. The results indicate that the model can generalize effectively, with a reasonable gap between training and testing accuracy, suggesting minimal overfitting. While the model performs strongly for the "Slight Injury" class, it struggles more with minority classes such as "Fatal Injury" and "Serious Injury," reflecting the class imbalance in the dataset. The detailed precision, recall, and F1-score values for each class are presented in Table 5.

Table 5. Classification Metrics per Class

Class	Precision	Recall	F1-Score	Support
Fatal Injury	0.40	0.05	0.10	37
Serious Injury	0.26	0.10	0.15	363
Slight Injury	0.85	0.95	0.89	2064
Accuracy			0.81	2464
Macro Avg	0.50	0.37	0.38	2464
Weighted Avg	0.75	0.81	0.77	2464

From Table 5, the model performs best on predicting "Slight Injury" cases, which dominate the dataset. In contrast, the low F1-scores for "Fatal Injury" and "Serious Injury" indicate poor performance on minority classes, likely due to class imbalance. Future work could apply techniques like SMOTE, class weighting, or resampling-based ensembles to address this issue..

The prototype system was deployed locally using the Streamlit framework for testing and demonstration purposes. However, it can be easily adapted for public access through cloud-based deployment platforms such as Streamlit Cloud or Heroku, enabling broader accessibility for institutions and government agencies.

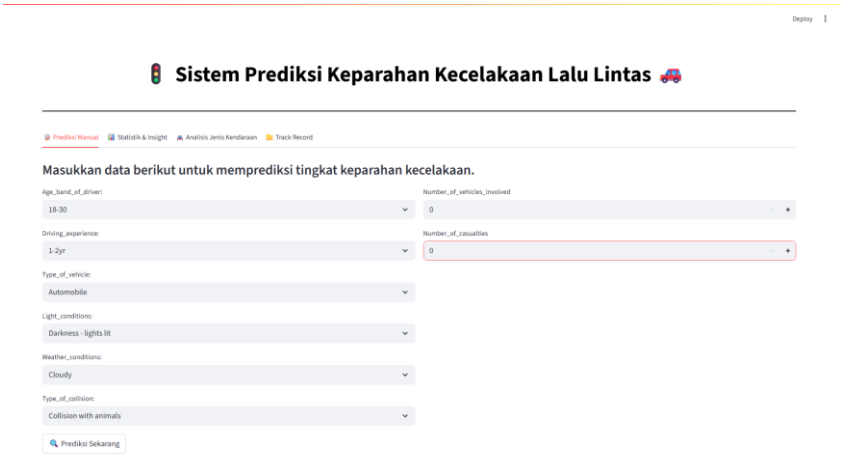

Figure 4. Manual Prediction Interface

In Figure 4, the Manual Prediction interface is shown, where users can input accident-related attributes such as driver age, experience, vehicle type, environmental conditions, and numeric factors. This interface provides a user-friendly form that returns immediate prediction results based on the input parameters.

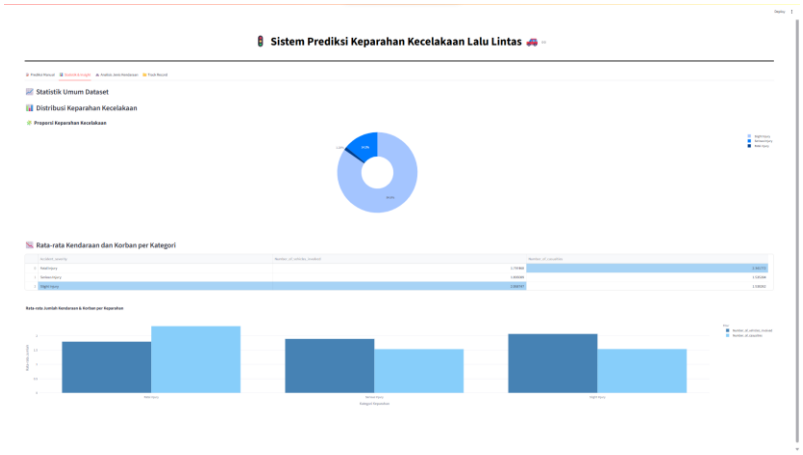

Figure 5. Statistics & Insight interface

Figure 5 illustrates the Statistics & Insight interface, which presents the accident severity distribution using a donut chart and the average number of vehicles involved and casualties per severity level using a grouped bar chart. This visual approach allows users to easily identify the proportion of each accident severity category and compare key averages across different severity levels, enhancing overall understanding of the dataset's general trends.

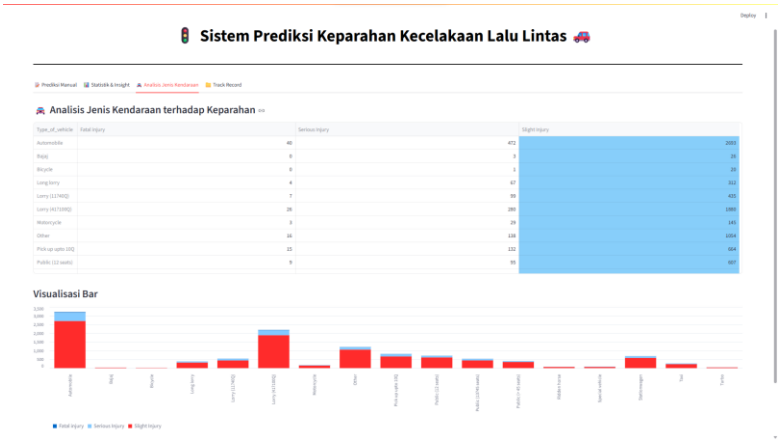

Figure 6. Vehicle Type Analysis Interface

Figure 6 illustrates the Vehicle Type Analysis interface, which displays the distribution of accident severity across various vehicle categories. The interface consists of a highlighted table showing the frequency of each severity level by vehicle type, complemented by a stacked bar chart that visually compares the severity distribution among different vehicles. This layout provides clear insights into which types of vehicles are most commonly involved in severe or minor accidents.

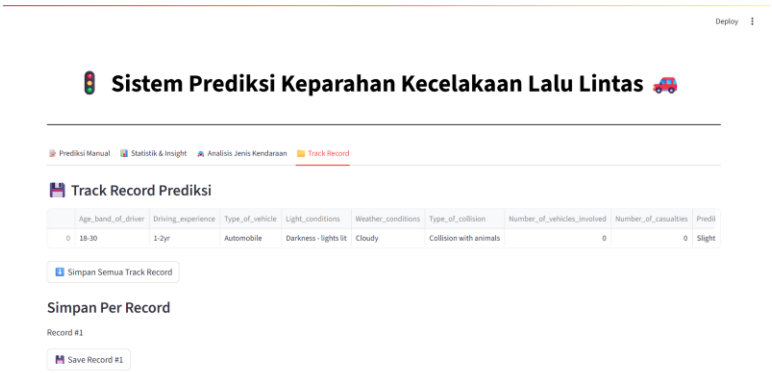

Figure 7. Prediction Track Record Interface

Figure 7 showcases the Prediction Track Record interface, which displays a table of all previously made predictions. Each prediction record includes input details and the resulting prediction, supporting traceability, usability, and accountability within the system.

4. Conclusions and Future Works

This study developed a Random Forest-based classification system to predict traffic accident severity using a dataset from Kaggle. The model incorporated key features such as driver demographics, vehicle type, environmental conditions, and casualty data, and successfully categorized cases into Slight Injury, Serious Injury, and Fatal Injury, achieving a test accuracy of 81%. Deployed using Streamlit, the system allows users to input data and receive interactive predictions along with summary statistics and prediction history. Designed for traffic police, transport authorities, and public safety analysts, the system aims to support incident response and policy planning. While the current model performs well, it does not yet include real-time data such as road infrastructure or live traffic conditions. Future improvements could involve integrating geospatial and temporal data, exploring advanced algorithms like XGBoost, and deploying the system on cloud platforms for broader accessibility via web or mobile interfaces.

5. References

[1] Baiq Sri Susanti, Ristu Haiban Hirzi, Siti Arni Wulandhya, Siti Hariati Hastuti, and Alissa Chintyana, "Analisis Klasifikasi Kecelakaan Lalu Lintas Lombok Timur Berdasarkan Tingkat Keparahan Korban Kecelakaan Menggunakan Metode Support Vector Machine (SVM) dan Bootstrap Aggregating (Bagging)," *J. Eksbar*, vol. 1, no. 1, pp. 1–8, Jul. 2024, doi: 10.29408/eksbar.v1i1.27123.

[2] and U. L. M. B. Rusandi, A. Bisnis, "Penerapan Digitalisasi Camera Analytic Guna Meningkatkan Tingkat Kepatuhan Driver Dan Menurunkan Tingkat Kecelakaan Dump Truck (DT) Di Jalan Hauling Batubara PT . XYZ".

[3] N. S. Kusumastutie, B. Patria, S. Kusrohmaniah, and T. D. Hastjarjo, "A review of accident data for traffic safety studies in Indonesia," *IOP Conf. Ser. Earth Environ. Sci.*, vol. 1294, no. 1, p. 012012, Jan. 2024, doi: 10.1088/1755-1315/1294/1/012012.

[4] H. Yuan and G. Li, "A Survey of Traffic Prediction: from Spatio-Temporal Data to Intelligent Transportation," *Data Sci. Eng.*, vol. 6, no. 1, pp. 63–85, Mar. 2021, doi: 10.1007/s41019-020-00151-z.

[5] F. I. Rahman, L. Lukman, and H. Hildayati, "Sistem Klasifikasi Kerusakan Jalan Metode Machine Learning dengan Algoritma K-Means dan Random Forest," *Arus J. Sains dan Teknol.*, vol. 3, no. 1, pp. 116–126, Jun. 2025, doi: 10.57250/ajst.v3i1.1212.

[6] H. A. Salman, A. Kalakech, and A. Steiti, "Random Forest Algorithm Overview," *Babylonian J. Mach. Learn.*, vol. 2024, pp. 69–79, Jun. 2024, doi: 10.58496/BJML/2024/007.

[7] A. R. Andhika, "Machine Learning Dalam Pengembangan Perangkat Lunak".

[8] Z. F. Olcay, G. Ünkaya, and G. D. Dursun, "The effect of OHS costs on accident severity rate in the construction industry," *Bus. Manag. Stud. An Int. J.*, vol. 9, no. 3, pp. 1076–1087, Sep. 2021, doi:

10.15295/bmij.v9i3.1877.
[9] Mohammad Fokhrul Islam Buian, Ramisha Anan Arde, Md Masum Billah, Amit Debnath, and Iqtiar Md Siddique, "Advanced analytics for predicting traffic collision severity assessment," *World J. Adv. Res. Rev.*, vol. 21, no. 2, pp. 2007–2018, Feb. 2024, doi: 10.30574/wjarr.2024.21.2.0704.
[10] A. N. et Al, "Pendekatan Descriptive Analysis Berbasis Data Untuk Mengevaluasi Kecelakaan Lalu Lintas Di Indonesia".
[11] and M. K. A. Saxena, M. Dhadwal, "Indian Crop Production: Prediction And Model Deployment Using Ml And Streamlit".
[12] A. Azhar, N. M. Ariff, M. A. A. Bakar, and A. Roslan, "Classification of Driver Injury Severity for Accidents Involving Heavy Vehicles with Decision Tree and Random Forest," *Sustainability*, vol. 14, no. 7, p. 4101, Mar. 2022, doi: 10.3390/su14074101.
[13] T. K. Titus and M. Jajuli, "Clustering Data Kecelakaan Lalu Lintas di Kecamatan Cileungsi Menggunakan Metode K-Means," *Gener. J.*, vol. 6, no. 1, pp. 1–12, Jan. 2022, doi: 10.29407/gj.v6i1.16103.
[14] and A. F. A. A. A. Zuhri, R. Kusumawati, M. A. Yaqin, "Development of Academic Community Recommendation System Using Content-Based Filtering at UIN Malang Informatics Engineering Study Program".
[15] H. L. Maharani and S. Zaman, "Implementasi Metode Random Forest untuk Peningkatan Efisiensi Penilaian Status Uji Kelayakan Kendaraan Bermotor di Kota Malang," *J. SAINTEKOM*, vol. 15, no. 1, pp. 106–117, Mar. 2025, doi: 10.33020/saintekom.v15i1.751.
[16] C. AVCI, M. BUDAK, N. YAĞMUR, and F. BALÇIK, "Comparison between random forest and support vector machine algorithms for LULC classification," *Int. J. Eng. Geosci.*, vol. 8, no. 1, pp. 1–10, Feb. 2023, doi: 10.26833/ijeg.987605.
[17] N. S. et Al, "Analysis of the Effectiveness of Traditional and Ensemble Machine Learning Models for Mushroom Classification".
[18] N. Chandrasekhar and S. Peddakrishna, "Enhancing Heart Disease Prediction Accuracy through Machine Learning Techniques and Optimization," *Processes*, vol. 11, no. 4, p. 1210, Apr. 2023, doi: 10.3390/pr11041210.
[19] Jude Chukwura Obi, "A comparative study of several classification metrics and their performances on data," *World J. Adv. Eng. Technol. Sci.*, vol. 8, no. 1, pp. 308–314, Feb. 2023, doi: 10.30574/wjaets.2023.8.1.0054.
[20] H. Dani, "Review on Frameworks Used for Deployment of Machine Learning Model," *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. 10, no. 2, pp. 211–215, Feb. 2022, doi: 10.22214/ijraset.2022.40222.