Factors Influencing Long-Haul Freight Truck Driver Behavior: A Random Forest Analysis in South Sulawesi

Andi Damayanti, Hakzah Hakzah, Misbahuddin, Hamka, Imam Fadly

Civil Engineering Department, University Muhammadiyah, Parepare, INDONESIA

E-mail: AndiDamyanti@gmail.com

| Submitted: August 11, 2025 | Revised: October 09, 2025 | Accepted: December 19, 2025 |

| Published: September 15, 2026 |

ABSTRACT

This study examines factors influencing speeding violations among long-haul freight truck drivers in South Sulawesi using the Random Forest (RF) algorithm. Data were collected from 370 drivers at the UPPKB Datae weigh station in Sidrap Regency, covering variables such as delivery pressure, sleep duration, truck age and size, monthly income, driving experience, daily driving hours, and driver age. Two modeling scenarios were tested: without balancing and with data balancing to address class imbalance. The unbalanced model achieved the highest performance (accuracy = 0.9929; F1-score = 0.9752; AUROC = 0.982), while the balanced model improved minority class detection despite a lower AUROC (0.711). Feature importance analysis revealed sleep duration, truck size, delivery pressure, and driver age as dominant predictors. Biological factors, vehicle characteristics, and operational pressures significantly affected speeding behavior. Policy implications include enforcing limits on working hours, ensuring minimum rest periods, revising incentive structures, providing regular safety training, and employing monitoring technologies such as e-logbooks and GPS. Future research should explore alternative algorithms, advanced balancing techniques, and integration of real-time operational and behavioral data to enhance predictive accuracy.

Keywords: long-haul freight transport, truck driver behavior, random forest, speeding violations and data balancing.

INTRODUCTION

The Road freight transportation constitutes a vital component of national logistics systems, particularly in developing countries such as Indonesia. Trucks serve as the backbone of goods distribution due to their flexibility in reaching diverse regions, including remote areas that are inaccessible to other modes of transportation [1], [2]. However, the growing reliance on trucks for freight transport poses serious challenges to traffic safety. Global reports indicate that more than 20% of severe road traffic accidents involve freight vehicles, with human factors accounting for the dominant share [3], [4]. Human factors, including driver fatigue, economic pressure, and mental health conditions, have long been identified as major determinants of long-haul truck crashes [5], [6]. Truck drivers frequently face extended working hours, delivery deadlines, and insufficient rest opportunities. [7] found that fatigue and poor sleep patterns increase the likelihood of traffic violations by up to 40%. Similarly, [8] emphasized that speeding behavior is often influenced by a combination of psychological stressors and operational pressures.

The Indonesian context presents additional complexities. The widespread practice of Over Dimension and Over Loading (ODOL) further exacerbates drivers' workload. Drivers are often required to undertake long-distance journeys with overloaded trucks across diverse road conditions, ranging from steep gradients to limited road infrastructure [9]. The absence of stringent working hour regulations and weak enforcement mechanisms aggravate this situation, creating heightened safety risks [10]. Furthermore, the work culture within the logistics industry frequently prioritizes meeting delivery deadlines over driver safety, ultimately fostering risky driving behaviors [11], [12]. Although numerous studies have examined truck driver behavior, most prior research has been limited to descriptive approaches or conventional statistical methods, such as logistic regression [13], [14]. Such approaches have limitations in capturing nonlinear relationships and complex interactions among variables. In reality, factors such as sleep duration, driving experience, truck size, and economic pressure often interact in ways that are difficult to model linearly [15].

With advances in data analytics, machine learning methods have increasingly been adopted in transportation studies due to their ability to process multivariate data and detect complex patterns. One algorithm that has proven particularly effective is the Random Forest (RF), which integrates multiple decision trees using bootstrap aggregation (bagging) to enhance accuracy and reduce overfitting risks [16], [17]. RF is recognized for its robustness in handling class imbalance a common issue in social research, including driver behavior studies [18], [19]. Nevertheless, the application of Random Forest in the context of long-haul truck driver behavior in Indonesia remains scarce. Most existing studies in Indonesia focus on operational costs and logistics policies rather than driver behavior and its implications for road safety [1], [20]. Yet, gaining a deeper understanding of the factors influencing driver behavior is essential for formulating evidence-based road safety policies.

Addressing this research gap, the present study aims to analyze the factors influencing the behavior of long-haul freight truck drivers in South Sulawesi using the Random Forest algorithm. This research is expected to contribute to the body of literature on transportation safety in Indonesia and to provide policy recommendations that promote driver welfare and safety through regulated working hours, safety-based incentive systems, and the implementation of telematics technology for monitoring driver compliance.

Truck Driver Behavior and Risk Factors

Driver behavior is a key determinant of road freight transport safety. Long-haul freight truck drivers often face high work pressure, extended travel times, and physically demanding conditions [5]. Factors such as fatigue, insufficient sleep, delivery pressure, and job-related stress have been linked to increased risks of speeding violations and road accidents [7], [21]. In various countries, driver fatigue has been identified as a significant cause of traffic accidents, with studies in Australia and Canada indicating that inadequate rest periods and volume-based incentive systems heighten the likelihood of aggressive driving [6], [11]. Drivers are often compelled to forgo rest breaks to meet strict delivery deadlines, particularly in logistics systems prioritizing time efficiency [22]. Demographic factors such as driver age, driving experience, and educational level also influence driving behavior [3]. While greater experience generally correlates with lower violation rates due to improved decision-making in risky situations, in some cases, extensive experience can foster overconfidence, potentially increasing risk [14].

Operational Conditions and Vehicle Characteristics

Vehicle attributes and operational conditions also affect truck driver behavior. Truck size and age are closely associated with vehicle stability, braking capability, and driver stress levels when navigating challenging terrains [23]. In Indonesia, these challenges are intensified by the persistent practice of Over Dimension and Over Loading (ODOL), forcing drivers to manage heavy payloads on steep gradients and damaged road surfaces [1], [20]. The pressure to deliver goods on time is a major driver of speeding violations. Higher delivery pressure increases the likelihood of risky driving decisions, such as speeding or skipping rest breaks [8].

Machine learning in Driver Behavior Analysis

With increasing data availability and computational capacity, machine learning approaches are increasingly applied to analyze driver behavior and predict accident risks. One prominent algorithm in this domain is the Random Forest (RF), which combines multiple decision trees and produces predictions via majority voting [16]. RF is robust to outliers, capable of handling nonlinear relationships, and effective in datasets with numerous variables and class imbalance issues [18], [23]. In transportation research, RF has been used to predict truck arrival times [23], detect hazardous driving behaviors [19], and evaluate the influence of vehicle factors on driving safety [18]. [16] demonstrated that RF outperforms logistic regression in predicting complex classification outcomes. Similarly, [24] found RF to be highly effective in addressing imbalanced data a common challenge in road safety studies, particularly when minority class instances (e.g., violations) are substantially fewer than majority class cases.

Research Gap

Although Random Forest has been widely applied in international studies, its implementation in Indonesia, particularly in the context of freight truck driver behavior, remains limited. Existing Indonesian research predominantly focuses on vehicle technical aspects, logistics costs, or route characteristics, with limited emphasis on machine learning-based predictive analyses of driver behavior [1], [22]. Consequently, a critical research gap exists for developing a predictive model using Random Forest to accurately explain risky driver behavior, particularly in complex geographical and operational contexts such as South Sulawesi. This study seeks to address this gap and contribute both methodologically and practically to the development of evidence-based freight transportation safety policies in Indonesia.

RESEARCH METHODS

Reasearch Design

This study employed an explanatory quantitative design aimed at analyzing factors influencing the behavior of long-haul freight truck drivers. The researcher was directly involved in both field data collection and data modeling. The predictive model was developed using the Random Forest (RF) algorithm, which is advantageous for handling nonlinear relationships among variables and is robust to multicollinearity and imbalanced data [18], [24]. Furthermore, to evaluate the effect of data balancing techniques on model performance, two approaches were compared: with and without the application of the Balance node.

Population and Sample

The study population comprised long-haul freight truck drivers operating in South Sulawesi Province. Data were collected at the UPPKB Datae weigh station in Sidrap Regency, a primary distribution route for inter-district and inter-provincial freight. A total of 370 truck drivers participated as respondents. Participants were selected using purposive sampling, with inclusion criteria requiring them to be active long-haul drivers, have at least one year of driving experience, and be willing to provide information regarding their working conditions and driving behavior. This sample size was deemed sufficient for multivariate analysis, as recommended by [25].

Research Variables

The study involved one dependent variable and eight independent variables. The dependent variable, driver behavior, was categorized binarily: 1 for drivers who violated speed limits and 0 for those who did not. The independent variables included delivery pressure (X1), daily sleep duration (X2), truck age (X3), truck size (X4), monthly income (X5), driving experience (X6), daily driving hours (X7), and driver age (X8). Variable selection was based on prior studies highlighting personal, economic, and operational factors as key determinants of truck driver behavior [3], [22].

Research Instrument

A structured questionnaire was designed to collect quantitative data on factors influencing driver behavior. The questionnaire comprised four sections. The first covered respondents' socio-demographic characteristics, including age, educational level, marital status, and place of residence. The second addressed job and vehicle characteristics, such as truck type, vehicle age, driving experience, average daily driving hours, and delivery pressure. The third section focused on lifestyle and health aspects, including sleep duration and stimulant consumption. The final section assessed driving behavior, specifically the frequency of speeding violations in the past 30 days. The questionnaire underwent content validity assessment by transportation and road safety experts to ensure relevance and clarity for respondents.

Data Collection Procedure

Data were collected directly by the researcher through face-to-face interviews with truck drivers at the UPPKB Datae weigh station in Sidrap Regency. This approach was chosen to ensure that each question was clearly understood and to obtain valid and in-depth responses from participants. Field surveys were conducted over a two-month period, from May to June 2025. All respondents

pISSN 2302-4240
eISSN 2655-2086
Volume 15, Issue 3, September 2026, pp.0718-0728
DOI: http://dx.doi.org/10.32832/astonjadro.v15i3.21235 http://ejournal.uika-bogor.ac.id/index.php/ASTONJADRO

participated voluntarily and were assured of the confidentiality of their identities and the information provided.

Random Forest Modeling Technique

The modeling in this study employed the Random Forest (RF) algorithm implemented using IBM SPSS Modeler software. Random Forest is an ensemble learning method based on decision trees, which constructs multiple classification trees using the bootstrap aggregation (bagging) technique. Each tree is built from a randomly selected subset of the data with replacement, and the final prediction is determined through a majority voting mechanism. Mathematically, the model prediction can be expressed as:

$$\hat{y} = \text{mode}\,\{h_1(x), h_2(x), \ldots, h_k(x)\}$$

Where $h_k(x)$ represents the prediction of the k^{th} tree for input x, and $\hat{y}$ denotes the final aggregated prediction. The initial stage of modeling included data preprocessing, involving the removal of invalid entries, recoding of categorical variables, scale transformation, and outlier identification. The modeling was performed under two comparative scenarios: without balancing and with balancing. In the unbalanced scenario, the original dataset was used without adjusting class distribution, and the entire dataset was directly applied to model construction. In the balanced scenario, the Balance node was applied to equalize the number of records in each class. Subsequently, the balanced dataset was partitioned using the Partition node into 70% for training and 30% for testing, with a random seed of 1234567 to ensure result replicability. For both scenarios, the RF model was configured with 100 trees and a maximum depth parameter determined through preliminary trials to avoid overfitting while maintaining model stability.

Model Validation

Model validation was conducted to evaluate the performance of the Random Forest algorithm in predicting the behavior of long-haul freight truck drivers. The validation process was carried out separately for two scenarios: the model built without applying any balancing technique and the model incorporating the Balance node. Evaluation was performed by comparing the model's classification results with the actual data using a confusion matrix, which provides the number of correct and incorrect predictions for each class. In addition, the Receiver Operating Characteristic (ROC) curve and the Area Under the Curve (AUROC) value were employed to measure the model's ability to discriminate between drivers who violated speed limits and those who did not. An AUROC value approaching 1 indicates excellent classification performance, while a value near 0.5 suggests that the model performs no better than random guessing [26].

Furthermore, a feature importance analysis was conducted to identify the independent variables that contributed most significantly to the model's classification decisions. This information is essential for understanding which factors most strongly influence driver behavior and for informing the formulation of transportation policy interventions. Comparing the two modeling scenarios was aimed at assessing the effect of the balancing technique on model quality. Previous studies have shown that balancing can enhance the model's sensitivity to the minority class, particularly in datasets with imbalanced distributions [19], [24], [27]. Therefore, this comparative analysis is crucial to ensure that the developed model is not only accurate overall but also equitable in identifying high-risk drivers.

RESULT AND DISCUSSION

Random Forest Modeling Results without Balancing

The initial modeling was conducted without applying any data balancing technique to the target class distribution. Based on the confusion matrix results, the Random Forest model successfully classified the majority of cases with high accuracy. Specifically, 308 out of 309 non-violating drivers ("No" class) were correctly predicted, with only one misclassified as a violator. Conversely, among the 61 drivers who committed violations ("Yes" class), 59 were correctly identified, while two were incorrectly classified as "No." The model's performance evaluation demonstrated excellent results. The test accuracy was recorded at 0.9929, indicating that nearly all cases were classified correctly. The True Positive Rate (TPR) or recall for the violator class was 0.9672, reflecting the model's strong ability to detect violations. The False Positive Rate (FPR) was exceptionally low at 0.0032, showing minimal misclassification of non-violators as violators. Additionally, the F1-score of 0.9752 indicated a well-balanced trade-off between precision and recall. The AUROC value was 0.982, as shown by the ROC curve, which exhibited a steep initial rise, signifying excellent discriminative capability. The ROC curve displayed a sharp classification distribution in the early percentiles, followed by a flattening trend after the 20th percentile. This suggests that while the model excels in distinguishing target classes in the early stages, its performance slightly diminishes beyond a certain threshold. Overall, these results are considered robust and suitable as a baseline model.

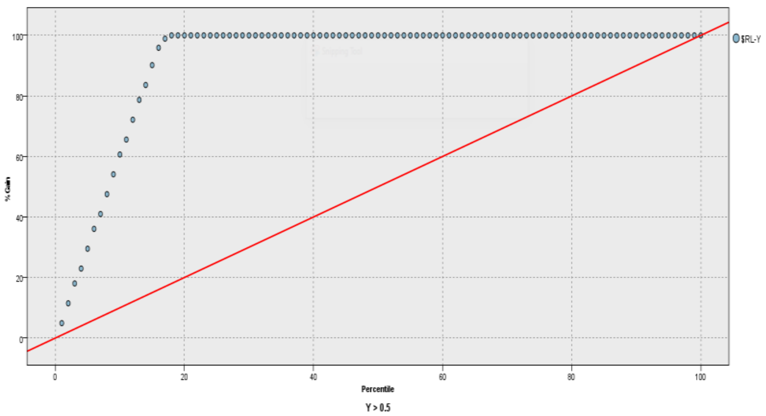

Figure 1. ROC curve of the Random Forest model without data balancing.

The feature importance analysis revealed that the most influential variable in the classification was X2 (daily sleep duration) as F1, followed by X8 (driver age) as F2, and X4 (truck size) as F3. These were followed by X1 (delivery pressure) as F4, X5 (monthly income) as F5, X6 (driving experience) as F6, X7 (daily driving duration) as F7, and X3 (truck age) as F8. These findings indicate that biological factors (sleep duration and age) and vehicle specifications are strongly correlated with the probability of engaging in speeding violations.

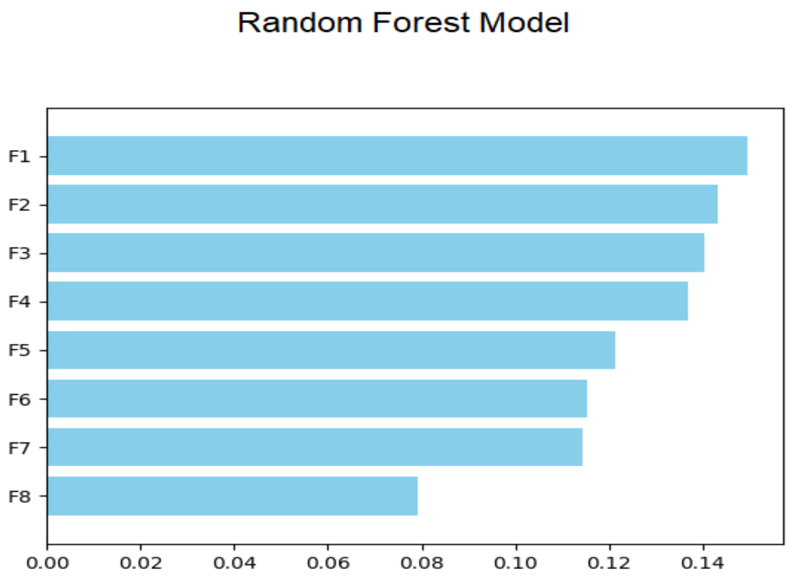

Figure 2. Variable importance ranking (feature importance).

Random Forest Modeling Results with Balancing

The subsequent modeling was performed by applying a balancing technique to the target variable's class distribution. This approach aimed to address the imbalance in the number of observations

pISSN 2302-4240
eISSN 2655-2086
Volume 15, Issue 3, September 2026, pp.0718-0728
DOI: http://dx.doi.org/10.32832/astonjadro.v15i3.21235 http://ejournal.uika-bogor.ac.id/index.php/ASTONJADRO

between violator and non-violator driver classes, enabling the model to learn more proportionally from both categories. The dataset was then split using the Partition node, with 70% allocated for training and 30% for testing. The evaluation results for the test dataset indicated that the Random Forest model with balancing demonstrated good classification performance. The test accuracy was recorded at 0.852, reflecting a high overall classification accuracy. A precision value of 0.923 suggested that the majority of drivers classified as violators were indeed actual violators. The recall (True Positive Rate) of 0.774 indicated that the model correctly detected approximately 77.4% of all violator drivers. The F1-score, recorded at 0.843, represented a balanced trade-off between precision and recall. Meanwhile, the Area Under the Receiver Operating Characteristic Curve (AUROC) was 0.711, indicating that the model's discriminatory ability to differentiate between violator and non-violator classes was within the "good" category.

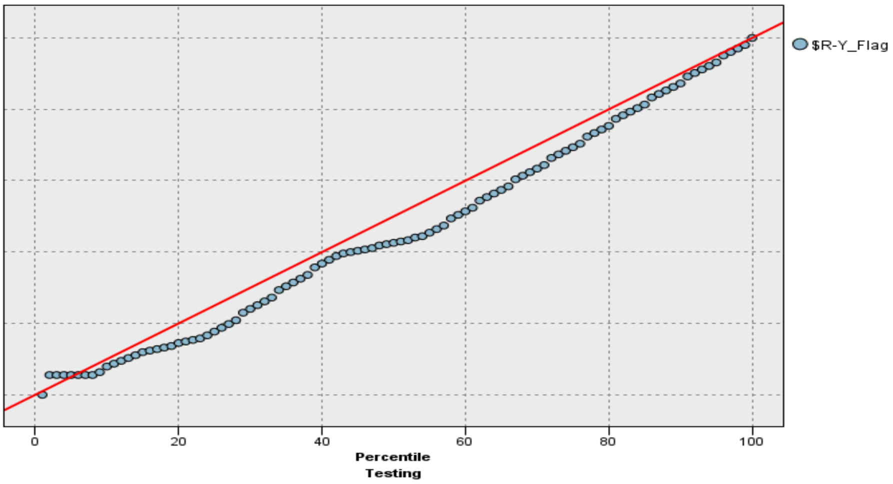

Figure 3. ROC curve of the Random Forest model with data balancing.

The feature importance analysis indicated that the most influential variable in the classification process was X4 (truck size) as F1, followed by X6 (driving experience) as F2, X2 (daily sleep duration) as F3, and X1 (delivery pressure) as F4. Subsequently, X8 (driver age) ranked as F5, X3 (truck age) as F6, X5 (monthly income) as F7, and finally X7 (daily driving duration) as F8, which contributed the least. This ranking suggests that vehicle characteristics and driver experience factors exert a significant influence on the model, whereas truck age and daily driving duration tend to be less dominant in affecting speeding violation behavior.

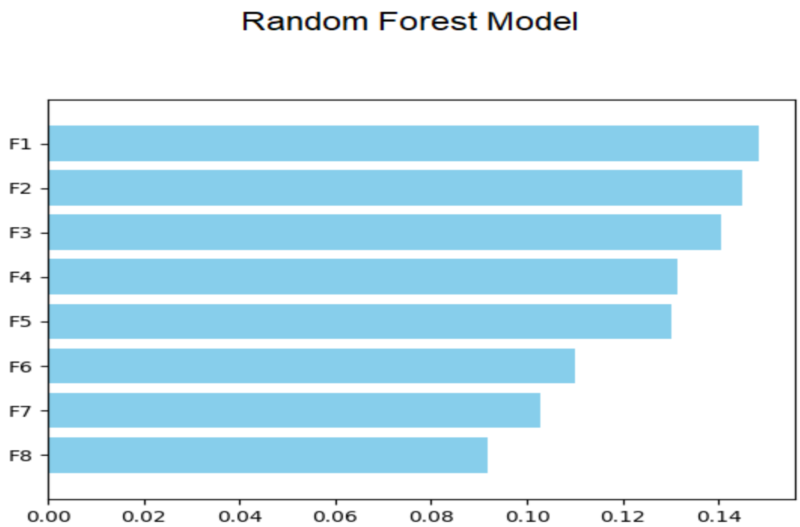

Figure 4. Variable importance ranking (feature importance).

Variable Importance Analysis and Model Comparison

The variable importance analysis aims to identify the dominant factors influencing speeding violations among long-haul freight truck drivers. The Random Forest model facilitates the evaluation of each variable's contribution through feature importance. The ranking of variable importance for both models is presented in Table 1 below.

Table 1. Ranking of variable importance in the random forest model

Rank	Without balancing	With balancing
1	X2 – Daily sleep duration	X4 – Truck size
2	X8 – Driver age	X6 – Driving experience
3	X4 – Truk size	X2 – Daily sleep duration
4	X1 – Delivery pressure	X1 – Delivery pressure
5	X5 – Monthly income	X8 – Driver age
6	X6 – Driving experience	X3 – Truck age
7	X7 – Daily driving duration	X5 – Monthly income
8	X3 – Truck age	X7 – Daily driving duration

The results indicate that variables X2 (daily sleep duration) and X4 (truck size) consistently ranked within the top three in both modeling scenarios. This finding suggests that drivers' physiological conditions and vehicle physical characteristics have a significant influence on speeding propensity. Furthermore, X1 (delivery pressure) appeared in fourth place in both models, demonstrating the consistent effect of operational pressure on driver behavior. However, several notable differences were observed between the two models. In the model without balancing, X8 (driver age) ranked second, indicating that age was an important consideration under the original data distribution. In contrast, in the model with balancing, X6 (driving experience) rose to second place, reflecting that after class distribution was equalized, skill level and driving tenure became more relevant for prediction.

These differences suggest that the balancing technique not only impacts the overall model performance but can also influence the model's perception of the data structure and dominant features. Therefore, the decision to apply or omit balancing should consider the intended application of the model whether it is aimed at general prediction with high accuracy or for detecting and intervening in high-risk driver groups. The quantitative performance comparison between the two models is illustrated in Figure 4 andFigure ***5***. Figure 4 presents the ROC curves for each model, showing that the unbalanced model achieved a True Positive Rate (TPR) of 0.967 with a very low False Positive Rate (FPR), whereas the balanced model achieved a TPR of 0.774. Figure 5 shows the accuracy values (0.993 vs 0.852), F1-scores (0.975 vs 0.843), and AUROC values (0.982 vs 0.711), which generally confirm the numerical superiority of the unbalanced model. However, this advantage should be interpreted in the context of potential bias toward the majority class.

pISSN 2302-4240
eISSN 2655-2086
Volume 15, Issue 3, September 2026, pp.0718-0728
DOI: http://dx.doi.org/10.32832/astonjadro.v15i3.21235 http://ejournal.uika-bogor.ac.id/index.php/ASTONJADRO

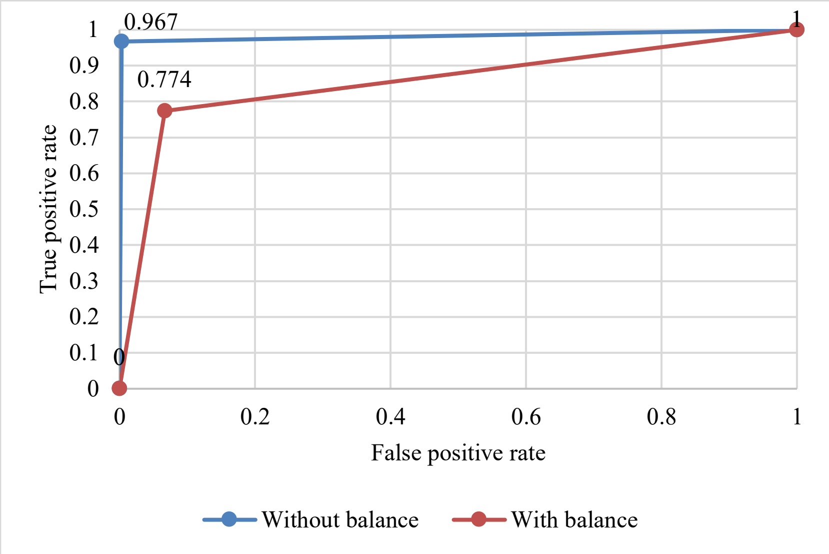

Figure 5. Receiver operating characteristic (ROC) curve.

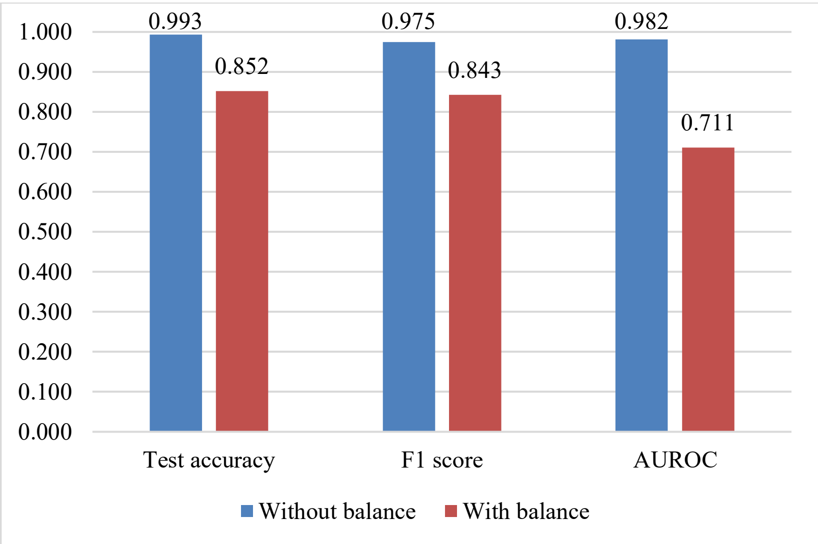

Figure 6. Comparison of accuracy, F1-score, and AUROC.

Limitations and Directions for Future Research

This study makes an important contribution by identifying factors influencing speeding violations among long-haul freight truck drivers using the Random Forest approach. However, several limitations should be addressed in future research. First, the study employed only one modeling method Random Forest without cross-algorithm comparisons such as Logistic Regression, XGBoost, or Support Vector Machines. While Random Forest delivered strong classification results, applying other models could enrich the analysis and provide alternative perspectives on data structure and variable influence. Second, the dataset was derived from weigh station surveys and was cross-sectional in nature, thus not capturing longitudinal changes in driver behavior or situational variations (e.g., nighttime driving, adverse weather, or mountainous routes). Future studies should integrate temporal and spatial data and adopt longitudinal approaches to capture long-term and adaptive behavioral patterns.

Third, several variables relied on self-reported questionnaire data, which are susceptible to perception and social desirability bias. Future research should incorporate field observation, vehicle sensor (CAN bus) data, or GPS records to enable more objective and real-time measurement. Fourth, although data balancing was applied, the technique was limited to internal undersampling/oversampling. Future research should explore algorithm-based balancing methods such as SMOTE (Synthetic Minority Over-sampling Technique) or ADASYN to improve model accuracy with highly imbalanced datasets. Lastly, psychological factors such as stress, aggressiveness, or mental health conditions were not included in the present model, despite their potential significance for driving behavior. Future studies should consider developing hybrid models that integrate technical, operational, and psychological data. Addressing these limitations could lead to more accurate, generalizable, and applicable models to support freight transportation safety policy development in Indonesia.

Policy Implications

The Random Forest model analysis revealed that speeding violations among long-haul freight truck drivers are influenced by a combination of biological, structural, and operational factors. These findings hold important implications for national transportation safety and logistics management policies, particularly in the context of long-distance operations. First, the daily sleep duration variable (X2) consistently emerged as one of the most significant factors, indicating that driver fatigue plays a primary role in speeding tendencies. Therefore, policies regulating maximum daily working hours and minimum rest periods for truck drivers are necessary. GPS-based working hour monitoring systems and e-logbooks could serve as tools to track compliance with rest requirements.

Second, delivery pressure (X1) strongly influences drivers' decisions to exceed speed limits, reflecting operational pressure from management or clients. Logistics companies should review overly stringent delivery schedules and develop incentive systems that prioritize safety and compliance over delivery speed. Third, truck size (X4) and truck age (X3) were found to correlate with speeding behavior. Larger trucks can increase driver fatigue or stress during vehicle handling, while older trucks may compromise performance and safety. Strengthening regulations on technical standards for freight vehicles, including periodic inspections and operational age limits, is therefore essential.

Fourth, driving experience (X6) ranked among the most important variables in the balanced model, indicating that more experienced drivers tend to adapt better and make safer decisions in traffic situations. This finding supports the need for regular training programs and adaptation periods for new drivers before assigning them to long-haul routes. Lastly, driver age (X8) also influenced the model's classification outcomes. Younger drivers may be more prone to risk-taking, whereas older drivers may experience reduced concentration or reaction times. Age-based demographic mapping and workload or route adjustments could thus be an effective strategy for mitigating accident risk. Overall, these results suggest that improving freight transport safety requires interventions beyond technical vehicle regulations, extending to human factors, organizational practices, and monitoring systems. The model's outcomes are expected to support more comprehensive and data-driven decision-making among relevant stakeholders.

CONCLUSION

The Random Forest modeling was conducted under two scenarios: without balancing and with data balancing. In the unbalanced scenario, the model achieved superior performance, with an accuracy of 0.9929, an F1-score of 0.9752, and an AUROC of 0.982, indicating excellent classification capability. The model correctly classified 308 of 309 non-violators and 59 of 61 violators, yielding a True Positive Rate (TPR) of 0.9672 and a False Positive Rate (FPR) of 0.0032. In the balanced scenario, the model demonstrated improved detection of the minority (violator) class, achieving an AUROC of 0.711, an accuracy of 0.852, a precision of 0.923, and a recall of 0.774. This trade-off illustrates the influence of balancing techniques on model sensitivity when dealing with imbalanced data. Feature importance analysis revealed that in the unbalanced model, daily sleep duration (X2), driver age (X8), and truck size (X4) were the top three predictors, while in the balanced model, truck size (X4), driving experience (X6), and daily sleep duration (X2) ranked highest. Across both scenarios, delivery pressure (X1) consistently ranked fourth, confirming its stable impact on speeding behavior. These results highlight the importance of physiological factors, vehicle

pISSN 2302-4240
eISSN 2655-2086
Volume 15, Issue 3, September 2026, pp.0718-0728
DOI: http://dx.doi.org/10.32832/astonjadro.v15i3.21235 http://ejournal.uika-bogor.ac.id/index.php/ASTONJADRO

characteristics, and operational pressures in influencing long-haul truck driver violations, as well as the methodological implications of applying balancing techniques in predictive modeling.

ACKNOWLEDGEMENT

The authors would like to express their sincere gratitude to Mr. Bahar, ST., MT., Head of the Land Transportation Management Center (BPTD) Class II South Sulawesi, Indonesia, and his staff for granting permission, providing support, and offering invaluable assistance during the fieldwork at the weighbridge. Their contributions were essential to the smooth progress and successful completion of this research. The authors also extend their appreciation to the research team and local government authorities for their collaboration, which ensured that all aspects of the project were conducted efficiently and effectively.

REFERENCES

[1] A. Budiharjo, T. Andika, N. Fitriani, R. Rukman, and B. Turasno, “Operational Data Analytics of Over Dimensional and Overloaded Truck in Indonesia,” *RSF Conf. Ser. Eng. Technol.*, vol. 2, no. 2, pp. 88–98, 2022, doi: 10.31098/cset.v2i2.562.

[2] C. Archetti, L. Peirano, and M. G. Speranza, “Optimization in multimodal freight transportation problems: A Survey,” *Eur. J. Oper. Res.*, vol. 299, no. 1, pp. 1–20, 2022, doi: 10.1016/j.ejor.2021.07.031.

[3] D. Meyer *et al.*, “Modelling the relationship of driver license and offense history with fatal and serious injury (FSI) crash involvement,” *J. Safety Res.*, vol. 79, pp. 83–93, 2021, doi: https://doi.org/10.1016/j.jsr.2021.08.008.

[4] B. Zhou *et al.*, “Comparing Factors Affecting Injury Severity of Passenger Car and Truck Drivers,” *IEEE Access*, vol. 8, pp. 153849–153861, 2020, doi: 10.1109/ACCESS.2020.3018183.

[5] A. Williamson, D. A. Lombardi, S. Folkard, J. Stutts, T. K. Courtney, and J. L. Connor, “The link between fatigue and safety,” *Accid. Anal. Prev.*, vol. 43, no. 2, pp. 498–515, 2011, doi: 10.1016/j.aap.2009.11.011.

[6] M. Shattell, Y. Apostolopoulos, S. Sönmez, and M. Griffin, “Occupational Stressors and the Mental Health of Truckers,” *Issues Ment. Health Nurs.*, vol. 31, pp. 561–568, Sep. 2010, doi: 10.3109/01612840.2010.488783.

[7] G. J. Casey, T. Miles-Johnson, and G. J. Stevens, “Heavy vehicle driver fatigue: Observing work and rest behaviours of truck drivers in Australia,” *Transp. Res. Part F Traffic Psychol. Behav.*, vol. 104, no. November 2023, pp. 136–153, 2024, doi: 10.1016/j.trf.2024.05.016.

[8] B. Shandhana Rashmi and S. Marisamynathan, “Factors affecting truck driver behavior on a road safety context: A critical systematic review of the evidence,” *J. Traffic Transp. Eng. (English Ed.*, vol. 10, no. 5, pp. 835–865, 2023, doi: 10.1016/j.jtte.2023.04.006.

[9] H. Hakzah, S. Yusuf, and A. I. Pawelloi, “The Importance of Motorcycle Taxi Transport of Agricultural Products and Operator Income in Indonesia,” *Open Transp. J.*, vol. 16, no. 1, pp. 1–8, 2022, doi: 10.2174/18744478-v16-e2201071.

[10] T. Inkinen and E. Hämäläinen, “Reviewing truck logistics: Solutions for achieving low emission road freight transport,” *Sustain.*, vol. 12, no. 17, pp. 1–11, 2020, doi: 10.3390/SU12176714.

[11] M. K. Lemke, Y. Apostolopoulos, and S. Sönmez, “A novel COVID-19 based truck driver syndemic? Implications for public health, safety, and vital supply chains,” *Am. J. Ind. Med.*, vol. 63, no. 8, pp. 659–662, 2020, doi: 10.1002/ajim.23138.

[12] A. M. Crizzle, M. McLean, and J. Malkin, “Risk factors for depressive symptoms in long-haul truck drivers,” *Int. J. Environ. Res. Public Health*, vol. 17, no. 11, 2020, doi: 10.3390/ijerph17113764.

[13] S. S. Nadler and J. F. Kros, “Truck driver turnover: a logistic regression approach,” in *Contemporary approaches and strategies for applied logistics*, IGI Global Scientific Publishing, 2018, pp. 382–404.

[14] A. P. Afghari *et al.*, “Investigating the effects of sleepiness in truck drivers on their headway: An instrumental variable model with grouped random parameters and heterogeneity in their means,” *Anal. Methods Accid. Res.*, vol. 36, p. 100241, 2022, doi: 10.1016/j.amar.2022.100241.

[15] B. S. Rashmi and S. Marisamynathan, “Investigating the contributory factors influencing speeding behavior among long-haul truck drivers traveling across India: Insights from binary logit and machine learning techniques,” *Int. J. Transp. Sci. Technol.*, vol. 16, pp. 194–211, 2024.

[16] Z. Jin, J. Shang, Q. Zhu, C. Ling, W. Xie, and B. Qiang, “RFRSF: Employee Turnover Prediction Based on Random Forests and Survival Analysis,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 12343 LNCS, pp. 503–515, 2020, doi: 10.1007/978-3-030-62008-0_35.

[17] O. O. Ajayi, A. M. Kurien, K. Djouani, and L. Dieng, “Analysis of Road Roughness and Driver Comfort in ‘Long-Haul’Road Transportation Using Random Forest Approach,” *Sensors (Basel).*, vol. 24, no. 18, p. 6115, 2024.

[18] M. Imani, A. Beikmohammadi, and H. R. Arabnia, “Comprehensive Analysis of Random Forest and XGBoost Performance with SMOTE, ADASYN, and GNUS Under Varying Imbalance Levels,” *Technologies*, vol. 13, no. 3, pp. 1–40, 2025, doi: 10.3390/technologies13030088.

[19] I. K. UMAR, “Machine learning approach for predicting truck drivers’ involvement in an injury accident,” *Sigma*, vol. 43, no. 3, pp. 808–815, 2025.

[20] H. Hakzah, M. I. Ramli, and A. I. Pawelloi, “the Characteristics Operational Cost of Freight Transport in South Sulawesi Province - Indonesia,” *Sinergi*, vol. 25, no. 1, p. 81, 2020, doi: 10.22441/sinergi.2021.1.011.

[21] A. M. Crizzle, R. Toxopeus, and J. Malkin, “Impact of limited rest areas on truck driver crashes in Saskatchewan: A mixed-methods approach,” *BMC Public Health*, vol. 20, no. 1, pp. 1–10, 2020, doi: 10.1186/s12889-020-09120-7.

[22] B. Shandhana Rashmi and S. Marisamynathan, “Investigating the contributory factors influencing speeding behavior among long-haul truck drivers traveling across India: Insights from binary logit and machine learning techniques,” *Int. J. Transp. Sci. Technol.*, no. xxxx, pp. 1–18, 2024, doi: 10.1016/j.ijtst.2024.01.008.

[23] O. O. Ajayi, A. M. Kurien, K. Djouani, and L. Dieng, “Analysis of Road Roughness and Driver Comfort in ‘Long-Haul’ Road Transportation Using Random Forest Approach,” *Sensors*, vol. 24, no. 18, 2024, doi: 10.3390/s24186115.

[24] M. Altalhan, A. Algarni, and M. Turki-Hadj Alouane, “Imbalanced Data Problem in Machine Learning: A Review,” *IEEE Access*, vol. 13, pp. 13686–13699, 2025, doi: 10.1109/ACCESS.2025.3531662.

[25] J. F. Hair, B. J. Babin, R. E. Anderson, and W. C. Black, *Multivariate Data Analysis*. Cengage Learning, 2022. [Online]. Available: https://books.google.co.id/books?id=PONXEAAAQBAJ

[26] T. Fawcett, “An introduction to ROC analysis,” *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, 2006, doi: https://doi.org/10.1016/j.patrec.2005.10.010.

[27] T. Hasanin, T. M. Khoshgoftaar, J. L. Leevy, and R. A. Bauder, “Severely imbalanced Big Data challenges: investigating data sampling approaches,” *J. Big Data*, vol. 6, no. 1, 2019, doi: 10.1186/s40537-019-0274-4.