Available online at www.CivileJournal.org

Civil Engineering Journal

(E-ISSN: 2476-3055; ISSN: 2676-6957)

Vol. 11, No. 11, November, 2025

Predicting Speeding Behavior of Long-Haul Freight Truck Drivers Using Machine Learning Models

Hakzah Hakzah [1*], Andi Damayanti [1], Misbahuddin [1], Abdul Rahman [1]

[1] *Department of Civil Engineering, Universitas Muhammadiyah Parepare, Parepare, Indonesia.*

Received 14 August 2025; Revised 17 October 2025; Accepted 20 October 2025; Published 01 November 2025

Abstract

The behavior of long-haul truck drivers is shaped by the weak enforcement of working-hour rules, tight deadlines, and heavy workloads. Over-dimensioning and overloading practices further increase risks by forcing drivers to handle excessive loads and work for prolonged periods. This study predicts speeding behavior among long-haul freight truck drivers using statistical and machine learning models. Data was collected from 370 respondents at two weigh stations in South Sulawesi, Indonesia, covering eight socio-demographic, economic, and operational predictors. Three models were tested: Binary Logistic Regression (BLR), Random Forest (RF), and Extreme Gradient Boosting (XGBoost). The dataset was balanced and split into 70% training and 30% testing, with performance assessed using accuracy, recall, F1-score, and AUROC. XGBoost delivered the best results, achieving 97.3% accuracy, 93.2% recall, a 96.4% F1-score, and a perfect AUROC of 1.000. RF also showed strong performance with 94.05% accuracy and an AUROC of 0.973, while BLR served as a relevant baseline despite weaker predictions. Key predictors of speeding violations were daily sleep duration, monthly income, and driving experience. This study demonstrates how machine learning can be effectively integrated alongside transportation data under imbalanced conditions, providing evidence-based insights to strengthen freight transport safety.

Keywords: Driver Behavior; Freight Transport; Speeding Violations; Machine Learning; XGBoost; Driver Safety.

1. Introduction

Road freight transportation plays a crucial role in sustaining global logistics, particularly in developing economies where multimodal integration is still limited [1, 2]. Trucks continue to dominate long-distance distribution across varied terrains, yet the sharp growth of heavy vehicle traffic has intensified concerns regarding safety and operational risks. Prior studies have shown that human factors, including fatigue, stress, and economic pressure, remain major contributors to accidents involving freight trucks [3, 4]. Insufficient rest has frequently been linked to fatigue-related incidents in countries such as Australia [5], while payment-based incentives and strict delivery deadlines in North America have been associated with higher tendencies toward speeding and unsafe practices [6, 7]. These findings support theoretical perspectives on driver fatigue and risk-taking behavior, which emphasize that prolonged working hours and financial stress reduce attentiveness, impair reaction times, and increase the likelihood of risky decision-making.

In Indonesia, systemic challenges exacerbate these risks. Weak regulation of driving hours, limited monitoring systems, and the persistent prevalence of over-dimension and overloading practices have been identified as conditions that amplify fatigue-induced behaviors [8, 9]. Drivers also frequently face socio-economic burdens, such as low wages and delivery pressures, which reinforce unsafe choices. According to behavioral economics theory, these structural stressors create conditions in which short-term financial gains are prioritized over long-term safety, thereby elevating the probability of traffic violations.

* Corresponding author: hakzah@umpar.ac.id

http://dx.doi.org/10.28991/CEJ-2025-011-11-015

Recent scholarship has increasingly adopted statistical and machine learning (ML) approaches to analyze risky driving behaviors. Studies have demonstrated, for example, the application of Random Forest to assess road surface quality and driver comfort [10], the use of Binary Logit and tree-based models to evaluate speeding behavior among truck drivers in India [11], the role of tiredness in influencing following distance [12], and the impact of incentive schemes on driver safety performance [13]. Despite these contributions, most research remains limited either to single-factor analysis or to conventional regression methods, which are often inadequate in capturing non-linear relationships and imbalanced data distributions [14, 15]

Emerging research directions highlight the superiority of ensemble methods. Comparative studies of CART, RT, AdaBoost, and GBDT suggest that ensemble approaches outperform single models in classifying unsafe behaviors [16]. Systematic reviews have emphasized the role of monitoring technologies, including telematics and wearable sensors, in reducing risky behaviors [17]. Methodological surveys further advocate hybrid frameworks that combine interpretable statistical baselines with advanced ML techniques to enhance predictive accuracy and policy relevance [18]. Nevertheless, within the Indonesian freight sector particularly in Over Dimension Over Loading-dominated contexts empirical studies integrating statistical and ML approaches remain scarce.

This study aims to address these research gaps by combining Binary Logistic Regression (BLR), Random Forest (RF), and XGBoost to predict speeding behavior among long-haul freight truck drivers in South Sulawesi Province. Unlike previous studies that relied solely on Binary Logit [11] or focused on road surface quality factors in African contexts [10], this research integrates socio-demographic, economic, and operational predictors within a unified ML framework under imbalanced data conditions. Theoretically, it advances fatigue–risk and behavioral economics perspectives by empirically demonstrating how rest duration, income levels, and operational stressors interact to shape risky driving. Methodologically, it contributes to the literature by contrasting classical regression with ensemble learning algorithms, highlighting their performance trade-offs in imbalanced survey datasets [14, 15]. Substantively, the study offers practical contributions to policy by recommending stricter enforcement of rest-hour regulations, the introduction of safety-based incentive systems, and the gradual adoption of telematics in resource-constrained environments. It is therefore one of the first empirical studies in Indonesia to simultaneously evaluate BLR, RF, and XGBoost in the context of long-haul freight trucking safety, providing both methodological innovation and practical implications at national and international levels.

2. Literature Review

2.1. Driver-Related Factors

Driver behavior constitutes one of the most critical determinants of road safety, particularly in the context of long-haul freight transportation. Truck drivers frequently operate under considerable work pressures, extended driving hours, and insufficient rest, all of which are recognized as conditions that can encourage risky behaviors on the road. Prior studies have identified driver fatigue as a major contributor to heavy truck accidents, especially on interstate highways, and have further highlighted how prevailing work practices often compel drivers to continue operating vehicles even when fatigued or drowsy [12, 13, 19-22]. Evidence also shows that sleep disturbances and inadequate rest significantly impair drivers' decision-making capacity, particularly when driving for more than eight hours per day [5]. In developed countries such as the United States and Australia, delivery pressures and volume-based incentive systems have been shown to substantially increase accident risk among truck drivers, as these mechanisms tend to encourage excessive speeding and unsafe driving practices [7]. Similar patterns have also been observed in other regions, where organizational pressures to avoid delivery delays have been directly associated with a higher tendency to exceed speed limits [10].

In Indonesia, the risks appear even more pronounced due to systemic weaknesses in regulatory oversight. The absence of strict monitoring of working and resting hours has been identified as a major factor contributing to chronic fatigue among truck drivers [8]. Additionally, demographic characteristics such as age, driving experience, and educational background play a significant role in shaping driving styles and influencing levels of compliance with traffic regulations [23].

2.2. Vehicle and Operational Environment Factors

In addition to human factors, vehicle characteristics and the operational environment play an equally important role in shaping risky driving behaviors. Overloading and dimensional modifications of heavy vehicles have been linked to conditions that compel drivers to adopt extreme or unsafe driving strategies in order to maintain vehicle control [9]. The combined effect of excessive loads and steep gradients, which are common in regions such as South Sulawesi, has been shown to heighten both physical and mental strain on drivers, thereby increasing the likelihood of unsafe maneuvers [24-26]. Furthermore, environmental conditions such as weather, road surface quality, and traffic density represent additional determinants of driver performance. Evidence suggests that drivers who are accustomed to navigating challenging terrains often develop adaptive strategies that may not be suitable when conditions change, creating new forms of risk [27]. The absence of sufficient technical training further exacerbates this problem, as many drivers are unable to effectively adjust their driving behavior to accommodate environmental variations such as steep inclines, heavy rainfall, or traffic congestion.

2.3. Research Gap and Current Study Position

Previous studies have investigated the factors influencing truck driver behavior through both qualitative and quantitative approaches. Despite these efforts, most research has remained confined to a single analytical method and has not combined statistical and machine learning techniques within the same framework [15, 28-30]. Logistic regression has often been applied to examine relationships such as the impact of work pressure on driving performance, yet this method is constrained in its ability to capture non-linear interactions among variables [10]. In response to these limitations, machine learning algorithms, including Random Forest and XGBoost, have gained increasing attention for their capacity to handle complex factor interactions and produce more reliable predictive models [11, 31]. Random Forest is frequently noted for its classification accuracy, while XGBoost is valued for its balance between predictive performance and interpretability, which makes it useful for practical decision-making.

Building on these developments, the present study seeks to address the gaps within the research by integrating three analytical approaches Binary Logistic Regression, Random Forest, and XGBoost to predict the speeding behavior of long-haul freight truck drivers in South Sulawesi. This integration is designed to deliver a more comprehensive and data-driven understanding of risky driving patterns while also providing evidence-based recommendations for intervention strategies.

3. Material and Methods

3.1. Questionnaire Construction

This study employed a structured, closed-ended questionnaire to collect primary data on the factors influencing the behavior of long-haul freight truck drivers. The design of the questionnaire was guided by the fatigue risk theory, which emphasizes the impact of driver fatigue on risky driving, and the perspective of behavioral economics, which explains how economic pressures may encourage unsafe decision-making.

The questionnaire was divided into four main sections, each corresponding to the study variables. Sociodemographic characteristics included driving experience (X6) and driver's age (X8), both of which prior research links to levels of cautiousness, compliance, and overall driving competence [23]. Operational and vehicle-related characteristics comprised delivery pressure (X1), truck age (X3), truck size (X4), and daily driving duration (X7). These variables reflect workload and vehicle conditions that are often associated with fatigue and aggressive driving in freight operations [8, 9]. Lifestyle and economic conditions were represented by daily sleep duration (X2), which is directly tied to fatigue risk [12], and monthly income (X5), which may signal economic pressures or incentives influencing driver behavior [13].

The behavioral outcome variable (Y) measured risky driving by asking drivers a binary question: "In the past 30 days, have you exceeded the speed limit while driving?" Responses were coded as 0 = No and 1 = Yes. Speeding was chosen as the outcome because it is a common and significant indicator of risky driving, strongly linked to truck-related crash risk [3]. All independent variables (X1–X8) were assessed using a five-point Likert scale. Respondents indicated their level of agreement with each statement, where a score of 5 represented Strongly Agree, 4 indicated Agree, 3 corresponded to Neutral, 2 denoted Disagree, and 1 reflected Strongly Disagree.

The Likert scale was chosen for its effectiveness in capturing drivers' perceptions and attitudes toward both internal and external factors in a comprehensive manner. Accordingly, the questionnaire was designed not only to emphasize technical and operational aspects, but also to integrate sociodemographic, lifestyle, and economic conditions as critical determinants of driver behavior.

3.2. Research Survey Design and Data Collection Procedure

The empirical survey was conducted in South Sulawesi Province, Indonesia, focusing on two official motor vehicle weigh stations that function as primary checkpoints for long-haul freight trucks. The first site was the Maccopa Motor Vehicle Weigh Station in Maros Regency, a strategic gateway connecting the Makassar metropolitan area with inter-district and inter-provincial routes. The second site was the Datae Motor Vehicle Weigh Station in Sidrap Regency, located on one of the busiest freight corridors in the province. Both weigh stations were purposively selected as they represent critical nodes for monitoring freight movement and enforcing regulations related to over-dimension and over-loading practices, while also serving as major hubs for long-haul trucking activity.

Data was collected from 370 respondents, consisting of long-haul freight truck drivers undergoing inspection at these weigh stations. A purposive sampling strategy was applied to ensure that only drivers engaged in inter-regional freight operations were included, excluding those involved in local or short-haul trips. The survey was conducted through structured face-to-face interviews, a method selected to minimize non-response bias and reduce errors associated with self-administered questionnaires. Fieldwork was carried out over a two-month period, from May to June 2025, by trained enumerators who interviewed respondents during both peak and off-peak hours to capture varied operational conditions.

Figure 1 presents the geographical distribution of the study sites, showing the locations of the Maccopa and Datae Motor Vehicle Weigh Stations in South Sulawesi Province. The map highlights the strategic positioning of these weigh stations along national freight corridors, underscoring their importance as representative sites for examining the behavior of long-haul freight truck drivers.

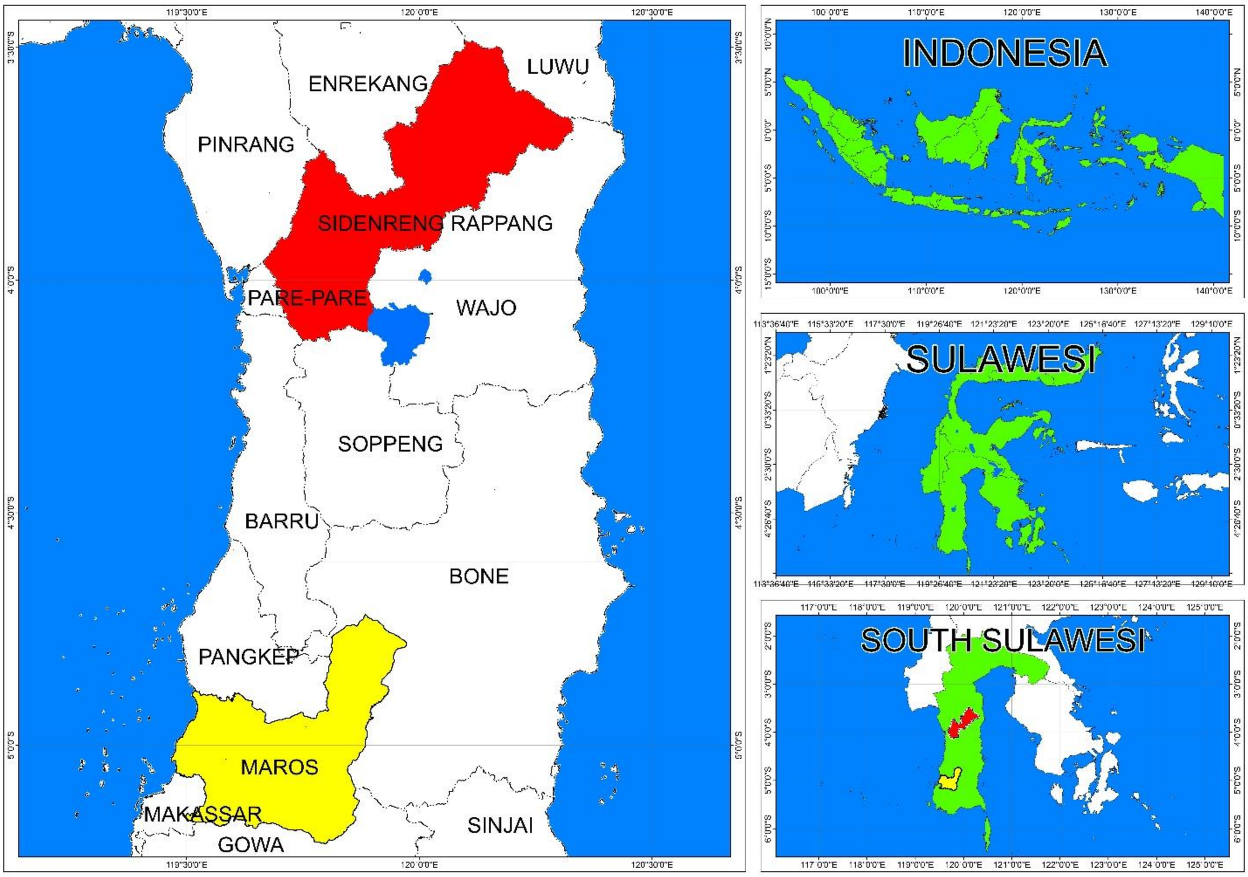

Figure 1. Study area showing the locations of the Maccopa and Datae Motor Vehicle Weigh Stations in South Sulawesi Province, Indonesia

3.3. Modeling Techniques Employed in the Study

Statistical modeling and machine learning algorithms play a crucial role in analyzing and predicting driver behavior, particularly in the context of traffic violations such as speeding. Traditional approaches such as Binary Logistic Regression (BLR) have long been applied to examine the relationship between independent variables and dichotomous outcomes, as well as to identify significant factors influencing driver behavior [10, 32]. However, with the advancement of computational technology and the availability of large-scale datasets, machine learning algorithms such as Random Forest (RF) and Extreme Gradient Boosting Tree (XGBoost) have emerged as increasingly popular alternatives due to their ability to handle non-linearity, variable interactions, and class imbalance [14, 33, 34].

To provide a clear overview of the research process, the methodological workflow of this study was organized into sequential stages, as illustrated in Figure 2. The process began with a literature review and questionnaire design, followed by data collection at two weigh stations in South Sulawesi. The collected data was then preprocessed through balancing and partitioning (70:30 split) before entering the modeling stage. Three models BLR, RF, and XGBoost were developed and subsequently evaluated using confusion matrix, accuracy, recall, F1-score, and AUROC. The results were interpreted through variable importance analysis and comparative evaluation, both of which resulted in the conclusionary statements of the study, including policy recommendations.

Consequently, this study focused on Binary Logistic Regression (BLR), Random Forest (RF), and Extreme Gradient Boosting Tree (XGBoost) to balance predictive accuracy, interpretability, and computational efficiency within imbalanced survey data.

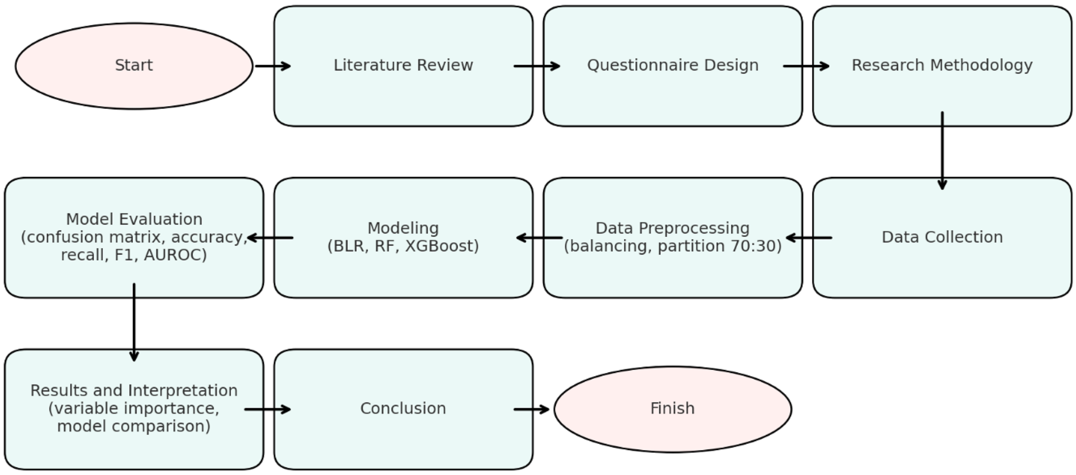

Figure 2. Research workflow summarizing the methodological process from literature review, data collection, preprocessing, modeling, to evaluation

3.3.1. Binary Logistic Regression (BLR)

Binary Logistic Regression (BLR) is one of the most widely used statistical techniques for modeling the probability of a binary outcome, such as whether a speeding violation occurs or not [35, 36]. In this study, BLR was employed to analyze the effects of the predictor variables (X1–X8) on the likelihood that a freight truck driver exceeded the speed limit (Y = 1).

The modeling process began with the transformation of the dependent variable into a binary categorical flag, where responses were coded as Yes (1) for drivers who reported speeding in the past 30 days and No (0) for those who did not. To address the issue of class imbalance, oversampling techniques were applied to ensure adequate representation of both categories [14, 34]. The dataset was subsequently partitioned into 70% training and 30% testing subsets using the Partition node in IBM SPSS Modeler, with a fixed seed (1234567) to ensure replicability [37].

The BLR model was estimated using the Maximum Likelihood method. The probability of a speeding violation was expressed by the logistic function [38]:

$$P(Y = 1|X) = \frac{e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_k X_k}}{1+e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_k X_k}} \tag{1}$$

here $P(Y = 1|X)$ denotes the probability that driver i commits a speeding violation, given the predictor variables X, and β represents the estimated coefficients. A classification threshold of 0.5 was adopted, where probabilities ≥ 0.5 were categorized as violations (Yes = 1) and probabilities < 0.5 as non-violations (No = 0).

Model performance was evaluated using a confusion matrix, sensitivity (True Positive Rate), accuracy, and Area Under the Receiver Operating Characteristic Curve (AUROC). While the BLR model demonstrated limited explanatory power in terms of R^2 values, it remained an essential baseline for comparison with the more advanced ensemble methods applied in this study.

3.3.2. Random Forest (RF)

Random Forest (RF) is an ensemble-based machine learning algorithm that constructs multiple decision trees and aggregates their results to enhance predictive performance while minimizing the risk of overfitting [39]. By combining the outputs of numerous trees through bootstrap aggregation (bagging), RF produces a model that is more stable and robust than a single decision tree, making it highly effective for large-scale and multivariate datasets [34].

In this study, the RF algorithm was applied to classify driver behavior, particularly to predict whether long-haul truck drivers exceeded the speed limit. The model development process comprised four main stages: (1) data balancing using oversampling techniques to address class imbalance, (2) partitioning the dataset into training (70%) and testing (30%) subsets with a fixed seed of 1234567 to ensure reproducibility, (3) training the model using the Random Forest node in IBM SPSS Modeler, where parameters such as the number of trees and maximum depth were configured, and (4) evaluating the model with a confusion matrix and performance metrics, including accuracy, recall, precision, F1-score, and AUROC [14, 40].

Mathematically, the RF prediction for an input x can be expressed as:

$$\hat{Y} = majority\_vote\{h_k(x)\}_{k=1}^{K} \tag{2}$$

where $h_k(x)$ represents the prediction of the k-th tree, and the final classification $\hat{Y}$ is determined by majority voting across all K trees. Within this framework, RF was selected for its ability to capture non-linear interactions among predictor variables while remaining interpretable through variable importance measures, making it particularly relevant for analyzing driver behavior in this study [37].

3.3.3. Extreme Gradient Boosting (XGBoost)

Extreme Gradient Boosting (XGBoost) is an ensemble-based machine learning algorithm that has become one of the most widely adopted methods in modern machine learning practices due to its efficiency, scalability, and high predictive accuracy [41]. The algorithm constructs additive models in a forward stage-wise manner, where new decision trees are iteratively built to correct the residual errors of previously generated trees. This boosting mechanism enables XGBoost to capture complex non-linear relationships and interactions among predictor variables while simultaneously controlling overfitting through regularization [14, 34]. In this study, XGBoost was applied to predict the speeding behavior of long-haul freight truck drivers under conditions of imbalanced survey data. The modeling process began with data preparation, where the dependent variable was encoded into a binary flag (Yes = 1, No = 0) to represent speeding violations. This was followed by data balancing using oversampling techniques to reduce bias arising from class imbalance. Subsequently, the dataset was divided into training (70%) and testing (30%) subsets using a fixed random seed (1234567) to ensure reproducibility of results.

The model was then developed using the XGBoost Tree node in IBM SPSS Modeler, with parameters such as learning rate, maximum tree depth, and the number of boosting rounds configured to achieve optimal performance. Finally, model evaluation was conducted using a confusion matrix and multiple performance metrics, including accuracy, recall, precision, F1-score, and AUROC, consistent with prior studies addressing imbalanced data challenges [15, 40]. Formally, XGBoost seeks to minimize an objective function that combines a differentiable loss function with a regularization term, thereby focusing not only on predictive accuracy but also on controlling model complexity. The objective function can be expressed as:

$$Obj(\theta) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{3}$$

where $l(y_i, \hat{y}_i)$ represents the the loss function (e.g., log-loss for binary classification), f_k denotes the k-th regression tree, and $\Omega(f_k)$ is the regularization term that penalizes model complexity. Within this framework, XGBoost was selected for its proven capability to handle imbalanced datasets and for providing comprehensive measures of variable importance. These advantages make it particularly suitable for analyzing the complex interactions underlying risky driving behaviors among long-haul truck drivers [14, 34].

3.3.4. Model Validation and Hyperparameter Tuning

Validation and evaluation are essential steps in ensuring the reliability and generalizability of machine learning models, particularly when dealing with imbalanced datasets. In this study, the dataset was randomly partitioned into training (70%) and testing (30%) subsets using the Partition node in IBM SPSS Modeler, with a fixed seed of 1234567 to ensure reproducibility. While k-fold cross-validation was not implemented due to computational and practical constraints, this limitation is acknowledged as an area for future research. To evaluate model performance, multiple classification metrics were employed beyond simple accuracy. These included recall (True Positive Rate, TPR), precision (Positive Predictive Value), F1-score (the harmonic mean of precision and recall), and the Area Under the Receiver Operating Characteristic Curve (AUROC). Recall was emphasized as a critical metric since the minority class (speeding violations) carried significant safety implications. AUROC was also adopted to provide a threshold-independent measure of classification performance, capturing the model's ability to discriminate between violators and non-violators [42]. This multi-metric evaluation approach enabled a more comprehensive assessment of each algorithm's strengths and weaknesses. In particular, it allowed for the identification of trade-offs between sensitivity and specificity, which is especially important in transportation safety studies where false negatives (unidentified violators) may result in severe real-world consequences.

4. Results and Discussion

4.1. Respondent Characteristics

The descriptive analysis provides an overview of the socio-demographic, economic, and operational profile of long-haul freight truck drivers included in this study. A total of 370 valid responses were collected at the Maccopa and Datae weigh stations in South Sulawesi Province. Overall, the results suggest that most drivers were middle-aged, relatively experienced, and worked under conditions of modest income, long driving hours, and delivery-related pressure, often with insufficient rest and the use of relatively older vehicles. These characteristics form the contextual foundation for interpreting the predictive modeling results presented in subsequent sections. To illustrate these patterns, the distribution of respondents across key variables is first summarized in Table 1.

Table 1. Distribution of respondent characteristics

Characteristic	Category	Frequency (n)	Percentage (%)
Age (years)	< 30	70	18.9
	31–40	150	40.5
	41–50	110	29.7
	> 50	40	10.9
Driving experience	< 5 years	80	21.6
	6–10 years	120	32.4
	> 10 years	170	46.0
Monthly income (IDR)	< 3,000,000	40	10.8
	3,000,000–4,000,000	90	24.3
	4,000,000–5,000,000	150	40.5
	5,000,000–7,000,000	60	16.2
	> 7,000,000	30	8.2
Daily sleep duration	< 4 hours	80	21.6
	4–6 hours	160	43.2
	7–8 hours	100	27.0
	> 8 hours	30	8.2
Daily driving duration	< 9 hours	90	24.3
	9–12 hours	180	48.6
	> 12 hours	100	27.1
Truck size	2 axles	190	51.4
	3 axles	160	43.2
	≥ 4 axles	20	5.4
Truck age	≤ 5 years	90	24.3
	6–10 years	170	45.9
	> 10 years	110	29.8
Delivery pressure	Never	70	18.9
	Sometimes	130	35.1
	Often	170	46.0

(Note: Frequencies adjusted to match the percentages shown in Figures; total n = 370).

From the information presented in Table 1 and Figure 3, several important patterns can be observed in the distribution of respondents. Delivery pressure emerged as a notable factor, with nearly half of the drivers (46.0%) frequently reporting pressure to meet deadlines. In terms of rest patterns, 43.2% of drivers reported sleeping only 4–6 hours per day, while 21.6% reported fewer than four hours, indicating a high risk of fatigue. Vehicle characteristics further revealed that 45.9% of trucks were 6–10 years old, with the fleet predominantly composed of two-axle (51.4%) and three-axle (43.2%) vehicles, whereas only a small fraction (5.4%) operated trucks with four or more axles. Economic conditions also reflected modest levels, as 40.5% of respondents reported monthly incomes between IDR 4–5 million, while only 8.2% earned above IDR 7 million. Working hours were also extensive, with 48.6% of drivers reporting daily driving durations of 9–12 hours, and more than a quarter (27.1%) exceeding 12 hours per day. Regarding age distribution, the majority of respondents (40.5%) were between 31 and 40 years old, followed by 29.7% between 41 and 50 years, while only 10.9% were over 50 years.

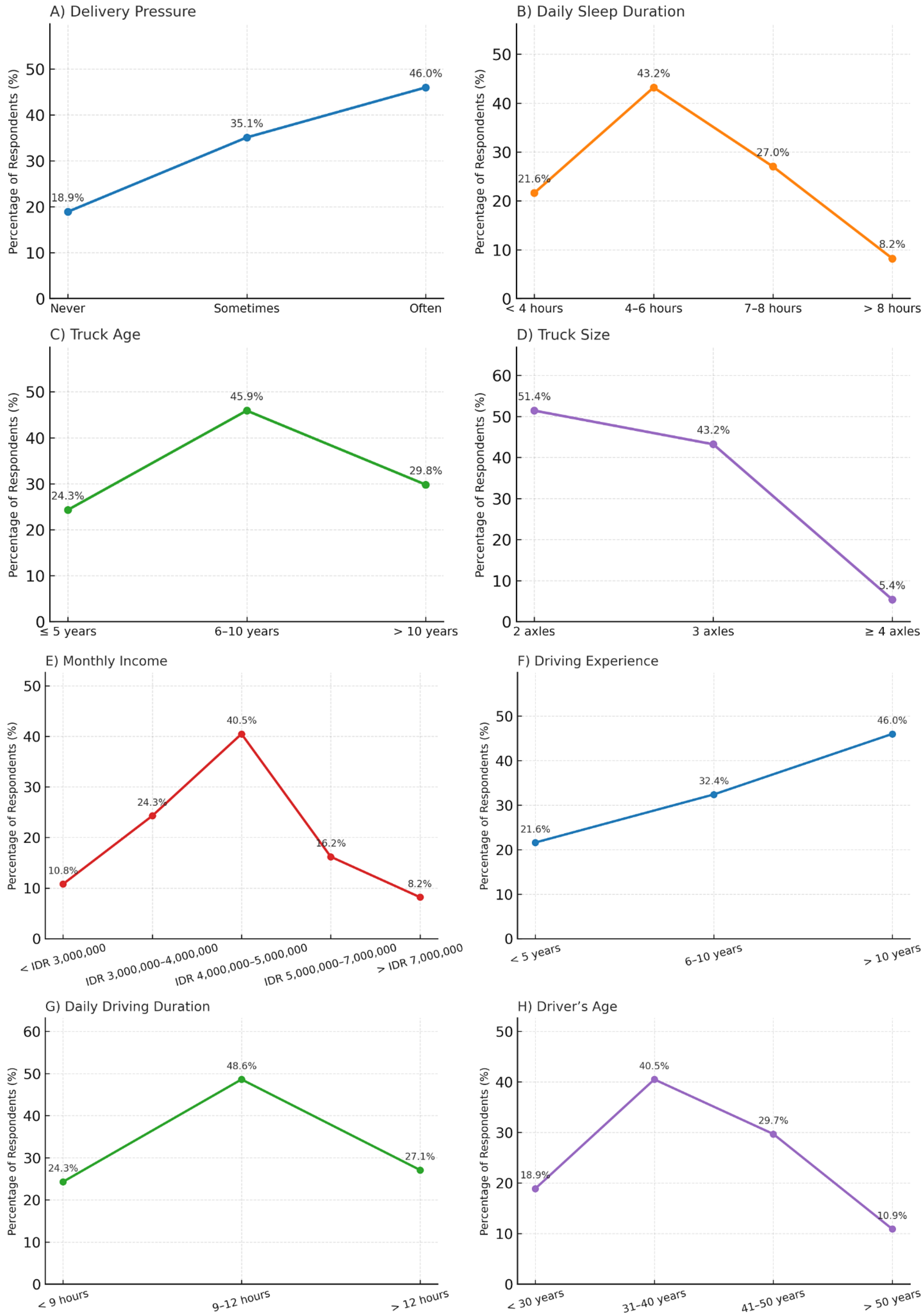

Figure 3. Distribution of respondent characteristics: (A) Delivery pressure, (B) Daily sleep duration, (C) Truck age, (D) Truck size, (E) Monthly income, (F) Driving experience, (G) Daily driving duration, (H) Driver's age

4.2. Results from BLR

The Binary Logistic Regression (BLR) model was applied to examine the relationship between the eight predictor variables (X1–X8) and the likelihood of speeding violations among freight truck drivers. Logistic regression is particularly suitable when the dependent variable is dichotomous (Yes/No), as in this study [35]. In this analysis, all eight predictors were entered simultaneously using the Enter method, ensuring that no variable selection procedure was applied and that each predictor was treated as relevant for constructing the baseline model as shown in Table 2

Table 2. Variables entered/removed

Model	Variables Entered	Variables Removed	Method
1	X8, X3, X2, X6, X1, X5, X7, X4[b]	-	Enter

b. All requested variables entered

The model summary results in Table 3 show that the coefficient of determination (R^2) was 0.020, with an adjusted R^2 of -0.013, indicating that the predictors explained only about 2% of the variance in the dependent variable. This highlights the limited explanatory power of the BLR model, as a negative adjusted R^2 is generally interpreted in the literature as evidence of poor model fit and limited predictive capability [38].

Table 3. Model summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	0.140[a]	0.020	-0.013	0.492

a. Predictors: (Constant), X8, X3, X2, X6, X1, X5, X7, X4

The ANOVA test in Table 4 yielded a significance value of 0.777, which exceeds the 0.05 threshold and indicates that the model is not statistically significant. In practical terms, this suggests that the BLR model makes little meaningful contribution to predicting the dependent variable. Within the literature, such a result where the significance level is greater than 0.05 is generally interpreted as evidence that the model lacks suitability for predictive purposes [43].

Table 4. ANOVA model test

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1.162	8	0.145	0.600	0.777[b]
	Residual	58.051	240	0.242		
	Total	**59.213**	**248**			

b. Predictors: (Constant), X8, X3, X2, X6, X1, X5, X7, X4

The regression coefficient in Table 5 indicated that none of the variables reached statistical significance at the 0.05 level, meaning that no single factor had a meaningful effect on the probability of a driver exceeding the speed limit. For example, variable X2 (daily sleep duration) had a significance value of 0.166, which suggests that its impact on driver behavior was not statistically meaningful. In line with prior methodological discussions, the evaluation of each predictor's statistical significance is essential when interpreting logistic regression models [36].

Table 5. Regression coefficients and significance

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	00.311	0.366	-	0.848	0.397
	X1	0.004	0.041	0.007	0.102	0.919
	X2	-0.062	0.045	-0.089	-10.390	0.166
	X3	-0.005	0.045	-0.007	-0.115	0.909
	X4	0.043	0.048	0.059	0.899	0.369
	X5	0.006	0.042	0.010	0.144	0.885
	X6	0.013	0.043	0.019	0.293	0.770
	X7	0.059	0.046	0.083	10.279	0.202
	X8	-0.033	0.046	-0.047	-0.724	0.470

The predictor importance chart further showed that X2 (daily sleep duration) contributed the most to the model, followed by X8 (driver's age) and X4 (truck size), while X7 (daily driving duration) had the lowest contribution. This visualization offers an overview of the relative weight of each variable in influencing the model outcome. In the literature, such charts are considered useful for exploratory identification of key predictors in classification models [39].

The classification results in Figure 4 showed that the BLR model achieved an accuracy of 0.665, with a recall of 0.598, a precision of 0.556, and an F1-score of 0.576. The AUC value of 0.711 reflects a moderate ability to discriminate between the two classes, suggesting that even with its limited explanatory power (low R^2), the model retained a reasonable capacity for classification [42]. Despite the lack of statistical significance and the minimal explanatory contribution, the BLR model was still considered relevant in this study as a baseline comparator. Its purpose was not to establish strong causal inference, but to provide a benchmark for assessing the advantages of more advanced algorithms such as XGBoost and Random Forest in capturing complex variable interactions and addressing class imbalance. Within the methodological literature, classical statistical models are commonly employed in this way, serving as interpretable baselines for comparison with machine learning approaches [36, 37].

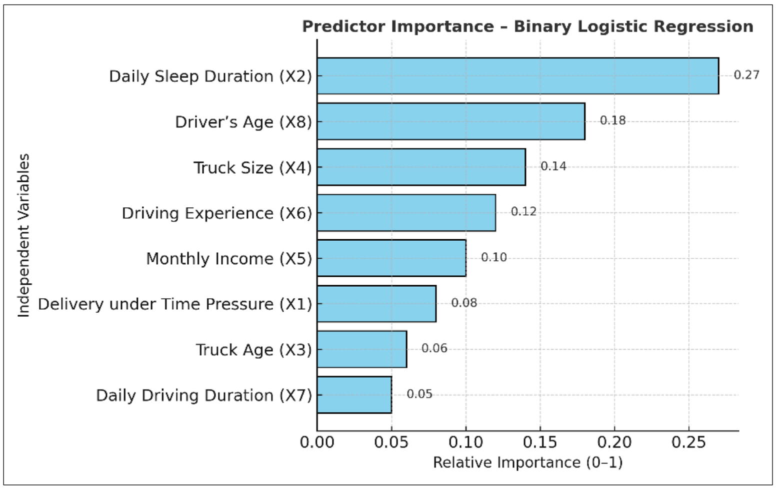

Figure 4. Predictor importance visualization from the BLR

4.3. Results from XGBoost

The feature importance visualization from the XGBoost model, depicted in Figure 5, shows the relative contribution of each predictor in explaining speeding behavior among long-haul freight truck drivers. The algorithm is widely recognized in the literature for its ability to capture non-linear relationships and complex interactions, and it is particularly effective in managing imbalanced datasets [34]. The analysis indicated that X5 (monthly income) emerged as the strongest predictor, with an F-score of 36, suggesting that economic pressure plays a central role in shaping aggressive driving tendencies. This was followed by X6 (driving experience) with a score of 28 and X2 (daily sleep duration) with a score of 24, emphasizing the importance of both fatigue and skill level. In contrast, X4 (truck size) and X1 (delivery under time pressure) showed moderate contributions (23 each), while X8 (driver's age), X7 (daily driving duration), and X3 (truck age) recorded lower scores below 20, reflecting a weaker influence. These results align with previous findings that XGBoost not only improves predictive accuracy but also offers valuable interpretability of the predictor structure [14, 15].

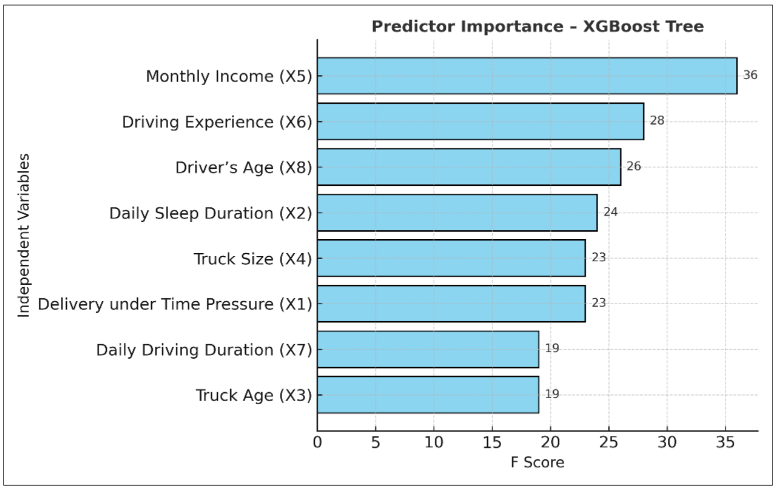

Figure 5. Predictor importance visualization from the XGBoost

The XGBoost model, which was developed using data balancing and a 70:30 train–test partition, achieved a very high classification performance. Testing on the holdout dataset produced an accuracy of 97.3%, a recall (TPR) of 93.2%, a precision of 100%, an F1-score of 96.4%, and an AUROC of 1.000. These results imply that the model was able to identify all speeding violations correctly and demonstrated an excellent capacity to distinguish between violators and non-violators of speed limits. In the literature, XGBoost is consistently recognized as an efficient and robust algorithm for managing complex classification tasks, particularly in contexts involving class imbalance [14, 15, 34]. Despite these outstanding outcomes, the evaluation relied on a single data partition, which may limit generalizability. Prior studies have emphasized that additional procedures, such as cross-validation or testing with external datasets, are necessary to ensure robustness and broader applicability of the model [37, 40].

4.4. Results from RF

The Random Forest model identified X2 (daily sleep duration) as the most influential predictor of speeding violations, followed by X5 (monthly income) and X4 (truck size). Other factors such as X1 (delivery under time pressure), X7 (daily driving duration), and X6 (driving experience) also made notable contributions, although with relatively similar levels of importance. By contrast, X8 (driver's age) showed the lowest importance compared to the other variables. These findings imply that rest quality, economic conditions, and vehicle characteristics play a more significant role in shaping risky driving behaviors. This interpretation aligns with prior studies that have emphasized the influence of both physical and socio-economic determinants on road safety outcomes [15, 44].

The Random Forest model, as shown in Figure 6, developed using data balancing and a 70:30 train–test partition, demonstrated strong classification performance in detecting speeding behavior among truck drivers. Testing with the holdout dataset resulted in an accuracy of 94.05%, a recall (TPR) of 89.8%, a precision of 95.0%, an F1-score of 92.3%, and an AUROC of 0.973. These results indicate that the model was effective in minimizing misclassifications for both violators and non-violators, with the AUROC value confirming its excellent discriminative ability even under complex variable interactions [14, 40]. The robustness of Random Forest stems from its construction of multiple decision trees and the aggregation of their outputs, which enhances stability and reduces the risk of overfitting [37]. Although its performance was slightly below that of XGBoost, the results clearly demonstrate that the use of data balancing significantly improved sensitivity to the minority class. As with the other models, however, the evaluation relied on a single data partition, suggesting that additional validation techniques such as k-fold cross-validation or testing on external datasets would be necessary to confirm robustness and generalizability in broader contexts [15, 34].

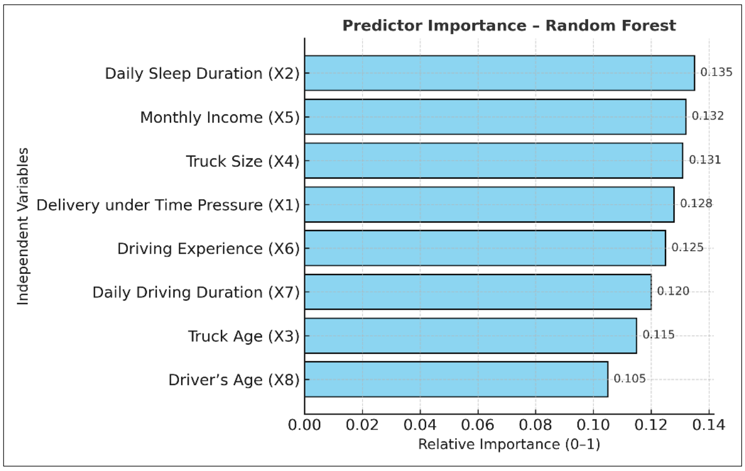

Figure 6. Predictor importance visualization from the RF

4.5. Variable Importance Analysis

The variable importance analysis was carried out to identify the main factors influencing speeding behavior among freight truck drivers. Across the three models BLR, XGBoost, and RF shown in Table 6, the most consistently influential predictors were X2 (daily sleep duration), X5 (monthly income), X6 (driving experience), and X4 (truck size). Within the BLR model, X2 appeared as the most important variable, followed by X8 (driver's age) and X4, suggesting that fatigue and demographic characteristics played a role, although the model overall was not statistically significant. In contrast, the XGBoost Tree ranked X5 as the strongest predictor, followed by X6 and X2, highlighting the importance of economic pressure, driving experience, and rest. Similarly, the Random Forest model confirmed X2 as the dominant predictor, followed by X5 and X4, reinforcing the influence of fatigue and vehicle characteristics. Taken together, the ranking of variable importance across the models was: X2 (daily sleep duration), X5 (monthly income), X6 (driving experience), X4 (truck size), X8 (driver's age), X3 (truck age), X1 (delivery under time pressure), and X7 (daily driving duration). The consistency of results across the ensemble models strengthens the argument that physiological, economic, and experiential factors are central determinants of risky driving. These findings are in line with previous research that emphasized the contribution of work pressure and fatigue to truck-related accidents [5, 6] and further demonstrate that tree-based algorithms such as XGBoost and RF provide advantages in capturing complex interactions among predictor variables [14, 34, 39].

Table 6. Ranking of variables based on importance (combined results from tree models)

Rank	Variable Code	Variable Name	Dominant Models
1	X2	Daily Sleep Duration	RF, BLR
2	X5	Monthly Income	XGBoost, RF
3	X6	Driving Experience	XGBoost
4	X4	Truck Size	RF, BLR
5	X8	Driver's Age	BLR
6	X3	Truck Age	XGBoost
7	X1	Delivery Under Time Pressure	XGBoost
8	X7	Daily Driving Duration	RF

The ROC curve analysis showed that XGBoost delivered the best performance, with a True Positive Rate (TPR) of 0.931 and a False Positive Rate (FPR) of 0.000, followed by Random Forest with a TPR of 0.898 and FPR of 0.031, and BLR with a TPR of 0.776 and FPR of 0.054 shown in Figure 7. These results illustrate the discriminative capacity of the models in identifying speeding violations accurately [14, 15]. Additional evaluation using accuracy, F1-score, and AUROC confirmed these findings: XGBoost achieved an accuracy of 97.3%, an F1-score of 96.4%, and a perfect AUROC of 1.000; RF recorded an accuracy of 94.1%, an F1-score of 92.3%, and an AUROC of 0.973; while BLR produced an accuracy of 87.2%, an F1-score of 80.4%, and an AUROC of 0.865 shown in Figure 8. These outcomes are consistent with prior studies showing that ensemble algorithms such as XGBoost and RF outperform traditional regression-based methods when dealing with complex and imbalanced classification data [34, 37, 40].

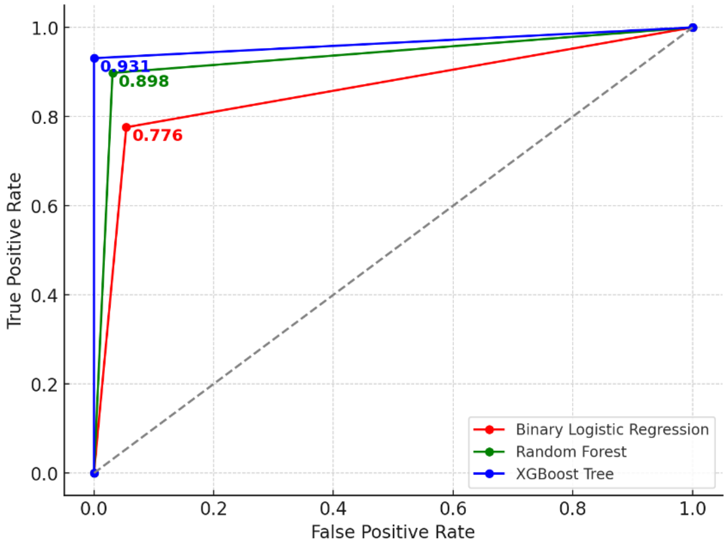

Figure 7. Receiver operating characteristic (ROC) curve

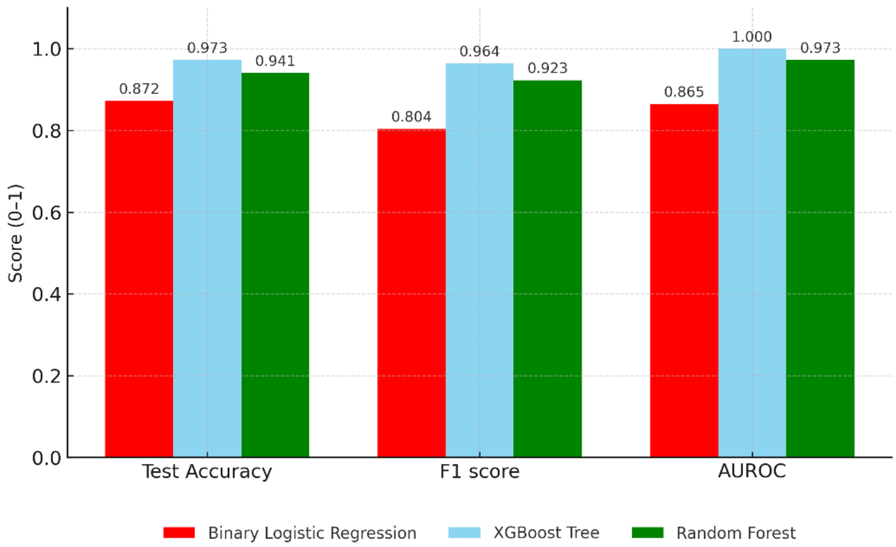

Figure 8. Comparative performance of BLR, RF, and XGBoost models based on test accuracy, F1-score, and AUROC

4.6. Policy Implications

The findings of this study indicate that daily sleep duration and monthly income significantly contribute to the likelihood of truck drivers exceeding speed limits. Accordingly, policies emphasizing the importance of adequate rest for drivers are crucial. Local governments and freight companies should develop safety campaigns highlighting the negative effects of fatigue on driving performance. At the same time, the relatively low monthly income earned by many drivers may encourage them to pursue delivery targets excessively, thereby compromising safety. Hence, reassessment of wage systems, the introduction of safety-based incentives, and the provision of welfare guarantees for drivers are urgent priorities to encourage safer driving behavior.

Furthermore, the results underscore the need for regular traffic safety training, particularly for inexperienced drivers. Such programs may include defensive driving techniques, fatigue management, and awareness of the legal risks associated with speeding violations. On the regulatory side, stricter law enforcement, including the expanded use of technologies such as speed cameras, in-vehicle black boxes, and e-ticketing systems, is recommended to enhance deterrence against violations. Through a combination of educational approaches, economic welfare measures, and stricter regulations, speeding can be significantly reduced, thereby contributing to improved road safety in the freight transportation sector.

5. Conclusion

This study demonstrated that the XGBoost algorithm achieved the highest performance in predicting speeding behavior among long-haul freight truck drivers, with an accuracy of 97.3%, a recall of 93.2%, an F1-score of 96.4%, and a perfect AUROC of 1.000. The Random Forest model also showed strong predictive ability, with an accuracy of 94.05% and an AUROC of 0.973, while Binary Logistic Regression remained relevant as a baseline despite its relatively lower performance. The variable importance analysis consistently identified daily sleep duration, monthly income, and driving experience as the most influential predictors of speeding violations. These results confirm that ensemble learning algorithms provide superior predictive accuracy compared to traditional regression models in analyzing driver behavior.

6. Declarations

6.1. Author Contributions

Conceptualization, H.H.; methodology, H.H. and A.D.; software, H.H. and A.R.; validation, H.H., A.D., M., and A.R.; formal analysis, H.H.; investigation, H.H.; resources, H.H.; data curation, H.H.; writing—original draft preparation, H.H.; writing—review and editing, H.H., A.D., M., and A.R.; visualization, H.H.; project administration, H.H. and A.D.; funding acquisition, H.H. All authors have read and agreed to the published version of the manuscript.

6.2. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

6.3. Funding and Acknowledgements

The authors received no financial support for the research, authorship, and/or publication of this article.

6.4. Acknowledgements

The author would like to express sincere gratitude to Mr. Bahar, Head of the Land Transportation Management Center (BPTD) Class II South Sulawesi, Indonesia, along with his team, for granting the necessary permissions, providing support, and offering valuable assistance during the fieldwork at the weigh stations. Their contributions were essential to the smooth execution and successful completion of this study. The author also extends appreciation to the research team and local government authorities for their cooperation, which ensured that all aspects of the research were conducted efficiently and effectively.

6.5. Conflicts of Interest

The authors declare no conflict of interest.

7. References

[1] Abolhoseini, S., Regehr, J. D., & Mehran, B. (2025). Estimating temporal truck traffic patterns on rural highway systems: A spatially-aware machine learning approach. Journal of Transport Geography, 128, 104345. doi:10.1016/j.jtrangeo.2025.104345.

[2] Basso, F., des Rotours, F., Maldonado, T., Pezoa, R., & Varas, M. (2025). Detecting heavy trucks from mobile phone trajectories using image-based behavioral representations and deep learning models. Scientific Reports, 15(1), 1–14. doi:10.1038/s41598-025-06711-5.

[3] Islam, M., & Mahmud, A. (2025). Unveiling the speeding behavior: Assessing the speeding risks and driver injury severities in single-heavy truck crashes. Safety Science, 187, 106861. doi:10.1016/j.ssci.2025.106861.

[4] Zhong, Y., & Qian, D. (2024). The causal effect of unsafe driving behavior of heavy-duty trucks on accident casualties. Proceedings of the 2024 International Conference on Intelligent Driving and Smart Transportation, 96–101. doi:10.1145/3704657.3704674.

[5] Casey, G. J., Miles-Johnson, T., & Stevens, G. J. (2024). Heavy vehicle driver fatigue: Observing work and rest behaviours of truck drivers in Australia. Transportation Research Part F: Traffic Psychology and Behaviour, 104, 136–153. doi:10.1016/j.trf.2024.05.016.

[6] Crizzle, A. M., McLean, M., & Malkin, J. (2020). Risk factors for depressive symptoms in long-haul truck drivers. International Journal of Environmental Research and Public Health, 17(11), 3764. doi:10.3390/ijerph17113764.

[7] Williamson, A., Lombardi, D. A., Folkard, S., Stutts, J., Courtney, T. K., & Connor, J. L. (2011). The link between fatigue and safety. Accident Analysis and Prevention, 43(2), 498–515. doi:10.1016/j.aap.2009.11.011.

[8] Budiharjo, A., Andika, T., Fitriani, N., Rukman, R., & Turasno, B. (2022). Operational Data Analytics of Over Dimensional and Overloaded Truck in Indonesia. RSF Conference Series: Engineering and Technology, 2(2), 88–98. doi:10.31098/cset.v2i2.562.

[9] Amallia, C. T., & Chymaida, Z. V. (2023). Carrier's Responsibility for Delivery of Goods Due to Overdimension and Overloading in Land Transportation Activities. Jurnal Ilmiah Universitas Batanghari Jambi, 23(2), 1550. doi:10.33087/jiubj.v23i2.3549. (In Indonesian).

[10] Ajayi, O. O., Kurien, A. M., Djouani, K., & Dieng, L. (2024). Analysis of Road Roughness and Driver Comfort in 'Long-Haul' Road Transportation Using Random Forest Approach. Sensors, 24(18), 6115. doi:10.3390/s24186115.

[11] Shandhana Rashmi, B., & Marisamynathan, S. (2024). Investigating the contributory factors influencing speeding behavior among long-haul truck drivers traveling across India: Insights from binary logit and machine learning techniques. International Journal of Transportation Science and Technology, 16, 194–211. doi:10.1016/j.ijtst.2024.01.008.

[12] Afghari, A. P., Papadimitriou, E., Pilkington-Cheney, F., Filtness, A., Brijs, T., Brijs, K., Cuenen, A., De Vos, B., Dirix, H., Ross, V., Wets, G., Lourenço, A., & Rodrigues, L. (2022). Investigating the effects of sleepiness in truck drivers on their headway: An instrumental variable model with grouped random parameters and heterogeneity in their means. Analytic Methods in Accident Research, 36, 100241. doi:10.1016/j.amar.2022.100241.

[13] Meyer, L., Goedhals-Gerber, L. L., & Bod, A. De. (2025). A systematic review of incentive schemes and their implications for truck driver safety performance. Journal of Safety Research, 92, 166–180. doi:10.1016/j.jsr.2024.11.023.

[14] Altalhan, M., Algarni, A., & Turki-Hadj Alouane, M. (2025). Imbalanced Data Problem in Machine Learning: A Review. IEEE Access, 13, 13686–13699. doi:10.1109/ACCESS.2025.3531662.

[15] Wu, O. (2023). Rethinking class imbalance in machine learning. arXiv Preprint, arXiv:2305.03900. doi:10.48550/arXiv.2305.03900.

[16] Niu, Y., Li, Z., & Fan, Y. (2021). Analysis of truck drivers' unsafe driving behaviors using four machine learning methods. International Journal of Industrial Ergonomics, 86, 103192. doi:10.1016/j.ergon.2021.103192.

[17] Fonseca, T., & Ferreira, S. (2025). Monitoring Technologies for Truck Drivers: A Systematic Review of Safety and Driving Behavior. Applied Sciences (Switzerland), 15(12), 6513. doi:10.3390/app15126513.

[18] Mobini Seraji, M. H., Shaffiee Haghshenas, S., Shaffiee Haghshenas, S., Simic, V., Pamucar, D., Guido, G., & Astarita, V. (2025). A state-of-the-art review on machine learning techniques for driving behavior analysis: clustering and classification approaches. Complex & Intelligent Systems, 11(9), 386. doi:10.1007/s40747-025-01988-5.

[19] Crizzle, A. M., Toxopeus, R., & Malkin, J. (2020). Impact of limited rest areas on truck driver crashes in Saskatchewan: A mixed-methods approach. BMC Public Health, 20(1), 1–10. doi:10.1186/s12889-020-09120-7.

[20] Li, T., Chang, B., Sun, Y., Zhou, Y., Zhou, B., & Zhang, L. (2025). Development and application of a generalized fatigue driving monitoring system for open-pit mining vehicles: A case study in a Myanmar copper mine. Results in Engineering, 26, 105206. doi:10.1016/j.rineng.2025.105206.

[21] Li, X., Zhang, W., Li, Q., & Chen, F. (2025). Analysis of driver Lane-Changing behavior under the influence of truck platoons at highway Exits: A driving simulation study. Transportation Research Part F: Traffic Psychology and Behaviour, 113, 174–193. doi:10.1016/j.trf.2025.04.029.

[22] Pradhan, A. K., Lin, B. T. W., Wege, C., & Babel, F. (2024). Effects of Behavior-Based Driver Feedback Systems on the Speeding Violations of Commercial Long-Haul Truck Drivers. Safety, 10(1), 1–11. doi:10.3390/safety10010024.

[23] Afrin, T., & Yodo, N. (2020). A survey of road traffic congestion measures towards a sustainable and resilient transportation system. Sustainability (Switzerland), 12(11), 1–23. doi:10.3390/su12114660.

[24] Guo, Y., Li, L., Sun, Z., Wang, J., & Fu, X. (2025). A Freight Truck Stopping Behavior Prediction Approach Based on Trajectory Dataset. International Journal of Intelligent Transportation Systems Research, 23(1), 475–488. doi:10.1007/s13177-025-00462-3.

[25] Liang, Y., Yuan, H., Wang, Z., Wan, Z., Liu, T., Wu, B., Chen, S., & Tang, X. (2024). Nonlinear effects of traffic statuses and road geometries on highway traffic accident severity: A machine learning approach. PLOS ONE, 19(11), e0314133. doi:10.1371/journal.pone.0314133.

[26] Yang, C., Liu, J., Zhang, Z., Kofi Adanu, E., Penmetsa, P., & Jones, S. (2025). A machine learning approach to understanding the road and traffic environments of crashes involving driver distraction and inattention (DDI) on rural multilane highways. Journal of Safety Research, 92, 14–26. doi:10.1016/j.jsr.2024.11.011.

[27] Sohail, A. M., Teh, Y. W., Khan, N., Khan, Z., & Koubaa, A. (2024). A Comprehensive Review: Analysis of Machine Learning, Deep Learning, and Large Language Model Techniques for Revolutionizing Driver Identification. TechRxiv, Preprinnt, 1-33. doi:10.36227/techrxiv.170974128.87339641/v1.

[28] Alrajhi, A., Roy, K., Kribs, J., & Almalki, S. S. (2025). Automated Detection of Road Defects Using LSTM and Random Forest. IEEE Access, 13, 76857–76867. doi:10.1109/ACCESS.2025.3564172.

[29] Wu, J., Zhang, Z., Wang, Y., Zhang, Y., & Tian, Y. (2024). Method for Identifying Dangerous Driving Behaviors in Heavy-duty Trucks Based on Multi-modal Data. Journal of Transportation Systems Engineering and Information Technology, 24(2), 63–75. doi:10.16097/j.cnki.1009-6744.2024.02.007.

[30] Xue, Z., Xu, H., Cui, H., & Peng, W. (2025). A machine learning-incorporated heuristic column generation algorithm for truck grouping and scheduling problem in trailer-swapping transport mode. Expert Systems with Applications, 278, 127327. doi:10.1016/j.eswa.2025.127327.

[31] Ebrahimi, H., Vosoughi, S., Hosseini, A. F., & Hatami, A. (2024). Comprehensive survey of the truck transportation process: The impact of team driving on health and safety conditions. Heliyon, 10(18), 37880. doi:10.1016/j.heliyon.2024.e37880.

[32] Nadler, S. S., & Kros, J. F. (2018). Truck Driver Turnover. Contemporary Approaches and Strategies for Applied Logistics, 382–404, IGI Global Scientific Publishing, Hershey, United States. doi:10.4018/978-1-5225-5273-4.ch016.

[33] Alotaibi, Y., Nagappan, K., Thanarajan, T., & Rajendran, S. (2025). Optimal deep learning based vehicle detection and classification using chaotic equilibrium optimization algorithm in remote sensing imagery. Scientific Reports, 15(1), 1–16. doi:10.1038/s41598-025-02491-0.

[34] Imani, M., Beikmohammadi, A., & Arabnia, H. R. (2025). Comprehensive Analysis of Random Forest and XGBoost Performance with SMOTE, ADASYN, and GNUS Under Varying Imbalance Levels. Technologies, 13(3), 1–40. doi:10.3390/technologies13030088.

[35] Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2000). Introduction to the logistic regression model. Applied logistic regression, 2, 1-30, John Wiley & Sons, Hoboken, United States. doi:10.1002/9781118548387.ch1.

[36] Menard, S. (2011). Applied Logistic Regression Analysis. SAGE publications, Thousand Oaks, United States. doi:10.4135/9781412983433.

[37] Johnson, J. M., & Khoshgoftaar, T. M. (2019). Survey on deep learning with class imbalance. Journal of Big Data, 6(1), 1-54. doi:10.1186/s40537-019-0192-5.

[38] Field, A. (2024). Discovering statistics using IBM SPSS statistics. SAGE publications, Thousand Oaks, United States.

[39] Jin, Z., Shang, J., Zhu, Q., Ling, C., Xie, W., & Qiang, B. (2020). RFRSF: Employee Turnover Prediction Based on Random Forests and Survival Analysis. Web Information Systems Engineering – WISE 2020. WISE 2020, Lecture Notes in Computer Science, vol 12343, Springer, Cham, Switzerland. doi:10.1007/978-3-030-62008-0_35.

[40] Hasanin, T., Khoshgoftaar, T. M., Leevy, J. L., & Bauder, R. A. (2019). Severely imbalanced Big Data challenges: investigating data sampling approaches. Journal of Big Data, 6(1). doi:10.1186/s40537-019-0274-4.

[41] Chen, T., & Guestrin, C. (2016). XGBoost. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794. doi:10.1145/2939672.2939785.

[42] Fawcett, T. (2006). An introduction to ROC analysis. Pattern Recognition Letters, 27(8), 861-874. doi:10.1016/j.patrec.2005.10.010.

[43] Domínguez-Almendros, S., Benítez-Parejo, N., & Gonzalez-Ramirez, A. R. (2011). Logistic regression models. Allergologia et Immunopathologia, 39(5), 295–305. doi:10.1016/j.aller.2011.05.002.

[44] Meyer, D., Muir, S., Silva, S. S. M., Slikboer, R., McIntyre, A., Imberger, K., & Pyta, V. (2021). Modelling the relationship of driver license and offense history with fatal and serious injury (FSI) crash involvement. Journal of Safety Research, 79, 83–93. doi:10.1016/j.jsr.2021.08.008.