IEEE Open Journal of
Intelligent Transportation
Systems

Received 25 November 2023; revised 26 February 2024; accepted 22 September 2024. Date of publication 25 September 2024;
date of current version 1 October 2024.

Digital Object Identifier 10.1109/OJITS.2024.3467222

Machine Learning-Based Modeling of Celeration for Predicting Red-Light Violations

MAHMOUD MASOUD [1,2]

[1]Department of Information Systems and Operations Management, King Fahd University of Petroleum and Minerals, Dhahran 31261, Saudi Arabia

[2]Center for Smart Mobility and Logistics, King Fahd University of Petroleum and Minerals, Dhahran 31261, Saudi Arabia

CORRESPONDING AUTHOR: M. MASOUD (e-mail: mahmoud.masoud@kfupm.edu.sa)

ABSTRACT This research examines the intricate correlation between speed variation (celeration), a metric of driver behavior associated with vehicle control, and occurrences of running red lights. The study is based on a thorough analysis of a large dataset that includes a variety of parameters, such as exceeding speed limits, driver age, passenger count, weather, road condition, and temporal factors. Using cutting-edge machine learning methods like AdaBoost and Bagging, predictive models for red-light violations are painstakingly built, achieving remarkable validation accuracies of 90.4% and 90.1%, respectively. The study acknowledges the dataset's limitations in capturing real-world traffic complexities while focusing on the effectiveness and trade-offs inherent in these methodologies. This emphasizes how important it is to have synchronized and thorough data sources to guarantee accurate representation. The research field is enhancing predictive modeling techniques and improving transportation safety by connecting celebration, speed variation patterns over time, with instances of red-light violations.

INDEX TERMS Machine learning, modeling, celebration, red-light.

I. INTRODUCTION

CARS are widely regarded as the most prevalent mode of transportation in North America such as USA and Canada. Cars provide flexibility and convenience but also pose concerns, such as the potential for accidents. Car accidents are a significant global issue due to their adverse impacts on property, the economy, and human life. Since the 2000, more American have died in car accidents than in both World War [1]. The number of the death caused by the car accident are increasing over years. The number of car crashes deaths in the United States increased 1,018%, from 4,200 deaths in 1913 to 46,980 in 2021 [2]. From 2020 to 2021 the number of deaths increased 9.943% [2]. Similarly, Canada has experienced an increase in fatalities resulting from car accidents, adding to global apprehension. In 2021, the motor vehicle fatality count reached 1,768, representing a 1.3% rise from 2020's figure of 1,746 [38]. Many studies have shown that Intersections are a common location for car crashes. According to the National Highway Traffic Safety Administration (NHTSA) estimates that crossroads account for 40% of all crashes in the United States [3]. In 2020, intersections in Canada accounted for 27 percent of traffic fatalities and 41 percent of serious injuries, as reported by the National Collision Database [39]. There are different types of crashes can occur at intersections, including failure to yield, collisions with pedestrians and bicyclists, right-angle collisions, rear-end collisions, head-on collisions, and pedestrian accidents. Red signal running has played a significant role in crashes that occurred at intersections. Compared to other types of crashes, which account for 33% of crashes, red light running crashes account for 47% of crashes [4].

The act of driver entering the intersection after the yellow signal has changed to the red light is known as the red light running (RLR). RLR is a dangerous behavior can result in significant accidents, fatalities, and severe injuries. The number of the car crashes that caused by RLR vary across different countries. In U.S., Red light running is a common occurrence and can be deadly. In 2021, 1,109 deaths caused by the crashes involving red light running [2]. There are several common factors that affect the occurrence of RLR including human factors, environmental factors, and vehicle factors. RLR may be affected by the interaction between those factors. The possibility of RLR, for instance, might

The review of this article was arranged by Associate Editor Xin Li.

be increased by the risk-taking actions of the driver and the presence and actions of pedestrians. Similar to this, a car's speed and acceleration can affect the ability of the driver to stop at a red light. Understanding and analyzing these elements is essential for developing an effective technique to reduce the RLR.

When considering traffic intersection zones, several parameters are considered to understand and manage traffic flow to reduce the RLR violation. These parameters include speed, acceleration, distance, and time to reach the traffic intersection are all vital components of traffic control and safety. Speed, acceleration, distance, time to reach the traffic zone, and red light running all have sophisticated multi-faceted relationships with one another. Several studies have investigated this connection and found various contributing factors to red light running (RLR).

Examining the different variables that contribute to this reckless act becomes essential when considering the complex complexities of red-light running (RLR) incidents. These factors provide a detailed understanding of the elements affecting RLR, allows a development of focused strategies that reduce its occurrence. In relation to RLR incidents, this paper specifically looks at the following factors: speeding, which affects stopping distances and decision-making; driver age, which is linked to varying risk perception and reaction times; number of passengers, which may cause distractions and risky driving behaviors; weather, which might impact visibility and traction; and road conditions, which may affect vehicle movement.

The main contributions of this study is to conduct a thorough review of red-light running behavior prediction methodologies and datasets related to traffic light zones, with a focus on speed variation. The hypothesis that drivers who are unsafe can be identified by their higher tendency to accelerate will be tested. To accomplish these goals, a machine learning (ML) model will be created, one that can explain the frequency of red-light running incidents in terms of acceleration and other pertinent factors Using the collected datasets and the insights gained from the analysis. In order to better understand the complex relationship between speed variation (celeration) and red light running, we explore and analysis these aspects in the hopes of developing strategies that will effectively stop this risky driving practice and increase overall road safety.

II. RESEARCH PROBLEM

Red-light running at signal junctions, as seen in Figure 1, is a crucial problem that endangers traffic safety. Running red lights not only endangers the lives of vehicles, passengers, and pedestrians but may also lead to severe collisions, property damage, and financial losses. Red-light violations persist regardless of numerous attempts to address the issue, underscoring the need for a deeper comprehension of the variables influencing driver behavior at signalized intersections. Speed variation (celeration) is one such important factor that has been linked to influencing drivers' decisions to run red lights.

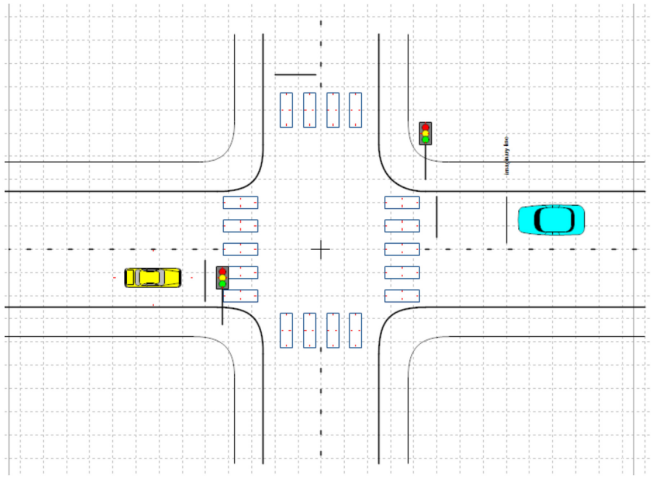

FIGURE 1. Signalized intersection.

Thus, the goal of this study's research is to model the relationship between speed variation and running red lights to fully understand it and make it easier to come up with effective solutions to increase road safety.

Traffic signal violations, including red-light running, remain a pervasive issue in modern urban environments. Studies have shown that red-light running is often linked to aggressive driving behavior, distracted driving, and lack of awareness of traffic signals' importance. However, the role of speed variation in influencing drivers' decisions to run red lights has not been extensively explored. Understanding how speed variation interacts with driver behavior at intersections can provide valuable insights into the mechanisms behind red-light violations.

The following important research questions are addressed by this study:

- What effect does acceleration have on drivers' decisions to run red lights at signalized intersections?
- Are there particular acceleration styles that are more likely to result in incidents of red-light running?
- What are the contributing factors that interact with speed variation to affect the behavior of driver who run red lights?
- Is it possible to create machine learning models to predict the probability of a red light running based on speed variation?

To address these questions, the study will use different machine learning approach and data science. By analyzing a traffic data from various intersections, the relationship between speed variation and running a red light will be better understood as a result of this analysis.

The main goal of this study is to create a machine learning model that will primarily be utilized for predicting instances of running red lights. The prediction is based on a thorough examination of the signal data along with vehicle-related data, similar to what shows in Figure 2 that illustrates how to obtain some possible vehicle data from the monitoring,

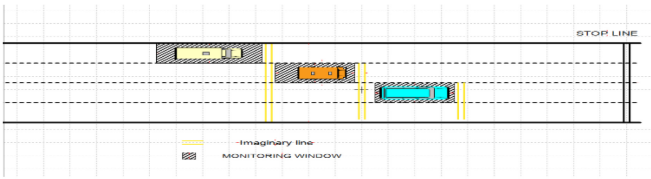

FIGURE 2. Illustrates the vehicle data, including speed, acceleration, time to intersection, and the stop line, which can be obtained through.

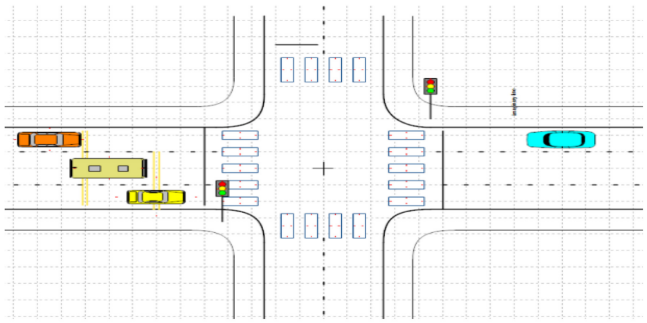

FIGURE 3. Synchronization of the signal and vehicle data to predict the red-light violation.

including speed, Distance, acceleration, and other relevant variables. Although the study's main goal applies to a wider range of data types, this paper will focus specifically on categorical data. This methodical approach is justified by the need to clarify the complex connections between various categorical variables and speed, particularly in the context of red-light violations.

This paper aims to reveal hidden patterns and insights that might help to better understand the phenomenon of red-light violations by examining the relationship between speed and other categorical data. The categorical data that has been chosen has the potential to reveal complex aspects of driver behavior, environmental circumstances, and other influencing factors that collectively contribute to the occurrence of red-light violations.

It is important to understand that the scope of this paper is purposefully restricted to categorical data in order to highlight one specific aspect of the problem. The long-term goal of this research, however, goes beyond its current limitations. There is a clear need to combine in the future the methods and results of this study with those obtained from a model trained on real signal data along with large vehicle-related data as show in Figure 3. The goal of integrating the models is to improve the model's predictive abilities and provide a complete structure for reliably precisely identifying red light violations.

Essentially, this research acts as a first step toward a comprehensive solution for red light violation prediction. By carefully analyzing the connections between speed and categorical variables, the goal is to open the way for a time when predictive models can make use of a full medley of data sources to improve the safety and effectiveness of vehicular traffic systems. In other words. creating a reliable prediction model to lower the number of red-light infractions, the suggested approach seeks to improve transportation safety. This concept improves traffic efficiency and overall road safety by integrating smoothly into traffic control systems. Utilizing cutting-edge machine learning methods like Bagging and AdaBoost, the study creates predictive models with high validation accuracies that can be used as instruments to reduce red-light violations and ultimately improve traffic safety.

III. LITERATURE REVIEW

Red light running, a recurring issue in traffic safety, continues to hinder initiatives aimed at improving intersection safety and decreasing accidents. This section provides an overview of research examining variables leading to RLR infractions. RLR breaches at signalized junctions may result from a range of circumstances. Speeding and acceleration significantly contribute to the occurrence of red light running infractions at traffic signal intersections. Studies used several models to analyze the link between these parameters and the incidence of RLR violations, using these models to anticipate and enhance road safety.

The control of traffic flow and the mitigation of collision risks are greatly aided by signal lights. Reference [5] explore the complex relationship between the length of the signal phase and red-light violations, noting that shorter yellow intervals can result in more violations because there isn't enough time for drivers to stop safely and react. The study highlights the requirement for exact signal timing calibration to achieve a balance between improving traffic flow and making sure that stopping distances are safe [5]. Furthermore, speed control strategies, capable of adjusting signal timings at intersections, hold promise in mitigating collision risks, especially in connected environments where vehicles communicate with traffic signals [30], [31], [32], [33], [34].

Numerous parameters have been investigated in the field of red-light violation (RLV) prediction in an attempt to better understand and predict RLV patterns. These studies use a variety of methodologies to model the level of severity of RLV and analyze the causes of its occurrence [6], [7], [8], [9], [10], [11]. To find correlations between RLV crashes and factors like weather, road geometry, Time-to-Intersection (TTI), and Distance-to-Intersection (DTI), one investigation specifically used Artificial Neural Networks (ANN) [6]. On the other hand, the use of Convolutional Neural Networks (CNN) has become a notable methodology, especially for the analysis of speed-related infractions [7]. However, disagreements still exist regarding how traffic violation monitoring affects road safety, with various perspectives presented in the literature [8]. Additionally, TTI has been used as a parameter for improving systems that identify RLV vehicles in driver behavior classification strategies [9]. Studies using mixed logit models have highlighted the crucial role of TTI in foretelling RLV violations by identifying the complex relationship between vehicle speed, TTI, and RLV

violations [10]. Additionally, signal timing interventions—where speed plays a crucial role—have emerged as a crucial countermeasure to mitigate RLV violations [11]. These initiatives have collectively illuminated a variety of variables that collectively contribute to a thorough comprehension of RLV patterns, thereby advancing the development of precise predictive models.

In the field of transportation research, there has been a lot of interest in the complex relationship between speeding and running red lights. By showing a definite correlation between faster vehicle speeds and an increased probability of running red lights, [21] provides light on this relationship. The results emphasize how crucial it is to carry out focused interventions that address speeding behaviors as a major strategy in reducing the number of cases of red light running. Notably, the study supports a variety of approaches and recognizes the potential for interventions outside of the scope of traditional enforcement techniques. This includes strategic road design changes and the improvement of speed limit signage, both of which have the potential to change driver behavior and advance the larger objective of improving intersection safety. Reference [22] puts light on the crucial role that speeding plays as a lead-up to running red lights, offering useful insights that support evidence-based policies and interventions aimed at supporting safer driving practices [12].

The dynamics of how drivers accelerate as they approach intersections play an essential role in red light running incidents. The finding that aggressive acceleration during amber phases significantly increases the number of red-light violations, which is frequently brought on by the desire to cross the intersection before the light turns red [13], according to [13], provides an explanation for this phenomenon. This idea is further supported by the earlier study by [14], which identified the short amount of amber time as a catalyst for rushed acceleration [14]. To solve this issue, multiple strategies must be used. Reference [15] hypothesized that education campaigns could encourage safer driving practices by increasing awareness of the risks associated with accelerating during amber intervals [15]. Additionally, traffic engineering measures, like those emphasized in the study by [16], [28], [29], can be effective in reducing incidents of red light running caused by rapid acceleration. Such violations can be decreased by utilizing strategies like extending the amber phase's duration or putting countdown timers in place The question of whether a car is a red-light runner (RLR) has been addressed in previous studies using a variety of parameters. To predict instances of red-light running violations, these studies have used a variety of methodologies, including machine learning, artificial neural networks (ANNs), and statistical analysis.

Reference [3] used machine learning algorithms to predict red light running based on past traffic data and environmental factors. Because of the real-time capabilities of their model, traffic management authorities can take proactive measures like dynamically adjusting signal timings, which lowers the likelihood of red-light running incidents [5].

In the field of red-light running research, Artificial Neural Networks (ANNs) have attracted a lot of attention for their use in the analysis of complex driver behavior patterns. In the previous study, [22] showed how ANN-based predictive models can identify the complex interactions that lead to red light running accidents. ANNs provide insights that are essential to researchers and policymakers to develop effective strategies to improving this issue because they are adept at capturing all aspects of driver behaviors [17]. The use of ANN models goes beyond just predicting occasions of red light running. Reference [17] examined the interaction between weather conditions, road geometry, and driver behavior using ANN algorithms, ultimately explaining the complicated causal relationships that causes red light violations [18]. Like this, [18] investigated the integration of vehicle trajectory data into ANN models, which allowed for more precise predictions and a deeper comprehension of driver decisions at intersections [1]. These studies collectively highlight the critical role ANNs play in identifying the complex factors that contribute to red light running occurrences, opening the door for data-driven interventions and increased traffic safety.

The significance of statistical methodologies in understanding red light running cannot be overstated. Reference [17] utilizes regression analysis to quantify the impact of road geometry, traffic volume, and demographic variables on red light violations. Their study offers actionable insights for evidence-based policy recommendations aimed at curbing red light running incidents [12].

In summary, the focus of this study focuses on a large set of data obtained up of a distinct parameter, each of which has an impact on the frequency of red-light running violations. Significant factors among the parameters provided in the dataset include the degree of speeding, the age of the driver, the number of passengers in the car, the current weather, the condition of the road, and the time of day of the incident. The researchers try to believe that red light violations are not random events; the intricate interplay of these multifaceted variables offers a rich tapestry of insights that reveals the intricate Web of contributing factors. This large dataset acts as the foundation for our analytical process, enabling a close look of the relationships between these parameters and the tendency for running red lights—an important task aimed to enhance road safety and foster a deeper understanding of driver behavior.

IV. METHODOLOGY

A. DATA EXTRACTION

A sample of 9,460 Killed and Seriously Injured (KSI) records from the year 2006 to 2022 was used in this study [19]. Toronto Police Service obtain the data from any serious or fatal collision where Aggressive Driving played a role in the collision in the City of Toronto. Although This dataset focuses on incidents involving severe or fatal collisions

FIGURE 4. The map of all crash records.

where aggressive driving appears to be a contributing factor. Aggressive Driving, which includes actions like speeding, rapid acceleration, and breaking traffic signals, is a major factor in the collisions under investigation. Furthermore, driver behavior, road conditions, traffic patterns, and infrastructure design could exhibit similarities between Toronto and certain regions of the United States [35], [36], [37].

The study makes use of a dataset to describe the intricate connection between the propensity to run red lights and speed variations, also known as celeration. By analyzing this dataset, this study hopes to gain new insights into the complex relationships between aggressive driving practices and traffic signal violations putting light on the root causes and patterns of these dangerous road incidents. This dataset is a useful tool for improving our understanding of the causes of these collisions and has the potential to help develop strategies and measures for improving road safety.

B. DATA CLEANING AND PR-PROCESSING

The dataset consists of 56 columns, each of which represents a distinct characteristic crucial to developing complex machine learning models that explain complexity. Some of these columns only have 139 records out of the total 9460, while others have over 5000 null entries, making up more than half of the column records. As a result, an effective strategy must be established to deal with these null values before proceeding on to model development. The method for handling null values varies depending on the type of feature; for categorical features, null values are replaced with the column's mode, whereas another method involves removing null-filled columns entirely. To ensure robustness and accuracy in the resulting machine learning models, this careful null value management is essential. The synergy between feature significance and null managing emphasizes the need for an organized strategy offered to each feature's uniqueness. This procedure establishes the framework for powerful machine learning models which may discover the dataset's complex patterns.

C. DATA CLEANING AND PR-PROCESSING

Data processing and the application of various machine learning models are the two main phases of the complex work of developing a meticulous celeration estimation model for signalized intersections. Data processing, used as the methodology's cornerstone, is an essential task aimed at improving the records and attributes of the dataset. Data cleaning is the main goal of this stage, which is essential for both improving the dataset and reducing the amount of time needed for model development. As a result, it assumes a crucial role in reducing data noise. This process includes locating and removing any columns that share semantics with other columns to minimize the structure of the dataset. The elimination of incidents that occurred at non-signalized intersections is necessary due to the study's specific focus on signalized intersections. This strict process resulted in a total of 3833 records remaining from the initial 9460 records, significantly enhancing data quality and effectiveness. Only 12 features remain after this meticulous filtering satisfy the strict criteria for analysis. This phase prepares the ground for the following stages of model development by cautiously removing irrelevant data and fine-tuning the dataset to its essential relevant components, improving efficiency and precision in addressing the complexities of celeration estimation within signalized intersections.

Additionally, the size of the dataset, which includes a variety of features and an enormous quantity of data, requires careful management of any potentially time-consuming parts of the resulting uses for machine learning. The effectiveness of this data processing stage depends on affecting an ideal equilibrium between its significance of specific features and the treatment of null values. A visual representation is used to provide clarity to this complex process, outlining the methodology's path from initial data preprocessing and feature engineering to the creation of a predictive model capable of accurately estimating celeration.

The use of a variety of machine learning models to generate precise speeding estimates characterizes the next stage. Models like logistic regression, decision trees, AdaBoost, Bagging, and K-Nearest Neighbors (KNN) are used in this area, each of which has a special strength for addressing the challenges presented by signalized intersections. Together, these models improve the methodology's effectiveness by deepening our understanding of traffic dynamics and intersection behaviors and acting as a solid framework for the accurate prediction of acceleration, which is a crucial step in improving the safety and effectiveness of transportation.

V. RESULTS

Only 12 features were left after the data cleaning process was finished. The 'get_dummies' function can be used to encode these features using the one-hot encoding

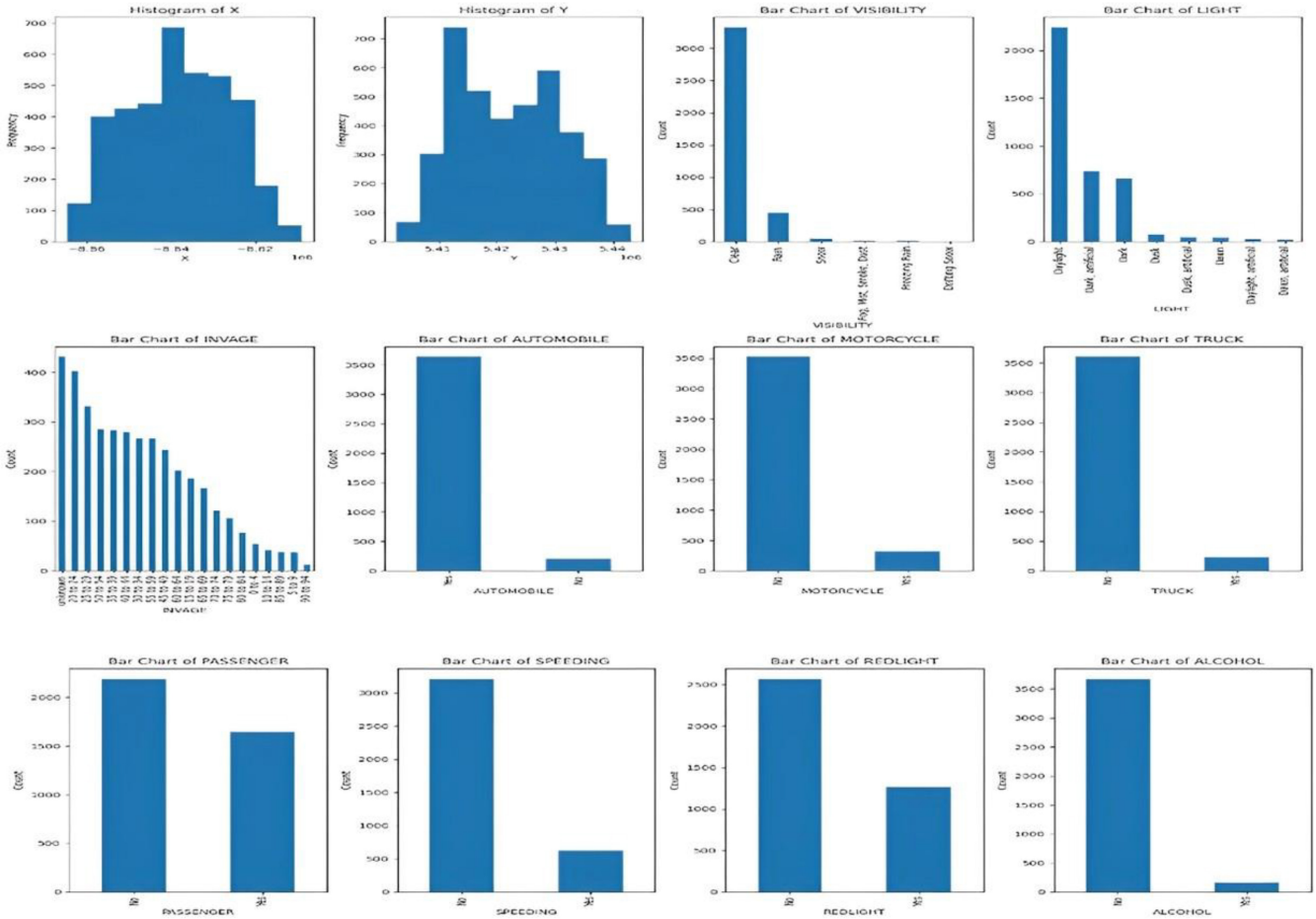

FIGURE 5. All chosen features and its original value before the data cleaning and Preprocessing phase.

algorithm. Categorical features like visibility, light levels, road conditions, and driver age are all encoded using that technique. Twelve attributes have been selected from the dataset and used as input features in the model to help in predictions or classifications. Red light has been selected as the study's target feature. The expansion to 47 features that results from using the one-hot encoding function will be the foundation for two supervised learning models and two ensemble learning techniques. In ensemble learning, predictions generated by different individual models are combined to create a more effective and precise predictive model. In this study, red light running is predicted using K-Nearest Neighbors (KNN), decision trees, AdaBoost, and Bagging techniques. The X, Y coordinates, visibility, light conditions, road conditions, driver age, vehicle type, passenger count, alcohol consumption, and speeding indicators are all input features for these models. The scikit-learn Python library is used to tune the hyperparameters in each model using the Grid Search Cross-Validation method. This method creates a grid of possible hyperparameter values and completely evaluates the performance of the model for each possible combination of hyperparameters by utilizing cross-validation.

The K-Nearest Neighbors (KNN) classification algorithm is straightforward and easy to understand. A data point's classification in the feature space is based on the classes of its k-nearest neighbors. The number of neighbors taken into account for classification depends on the value of 'k'. The algorithm 'kd_tree' with a single nearest neighbor ('n_neighbors': 1) and a Manhattan distance metric ('p': 1) produced the best hyperparameter combination that produced the highest accuracy for KNN. For this set of parameters, the validation accuracy is 84%.

A hierarchical model called a decision tree has internal nodes that stand in for features, branches for decision rules, and leaf nodes for class labels. The algorithm breaks down the data recursively according to useful features. The Gini impurity criterion ('criterion': 'gini') is used as one of the optimized hyperparameters for the decision tree model to assess split quality. The maximum number of levels in the tree is 80 ('max_depth': 80), and the maximum number of features ('max_features': 'log2') is used to select feature subsets. The validation accuracy that was reached is 83%.

AdaBoost (Adaptive Boosting) stands out as an effective ensemble technique that combines the predictive abilities of a number of poor classifiers to produce a robust classifier.

AdaBoost focuses its efforts on the incorrectly classified instances during each iterative stage by giving instances weights. With the help of this adaptive process, the model's performance can be gradually enhanced by correcting earlier misclassifications. The validation accuracy in the context of the study was a remarkable 90.4%. This demonstrates how AdaBoost can take advantage of the weaknesses of weak learners and gradually improve their accuracy as a group, making it an effective tool for handling challenging classification tasks.

Another powerful ensemble approach is known as Bagging (Bootstrap Aggregating), which creates various, independent models using bootstrapped subsets of the training data. The idea of combining predictions from various models to create the final prediction is the secret to Bagging's success. Bagging makes use of the wisdom of the crowd by averaging (in regression) or majority voting (in classification) to build a more stable and reliable model that reduces the risk of overfitting. This study's validation accuracy through Bagging was an impressive 90.1%. This result demonstrates how Bagging can be used to reduce variance, improve model stability, and ultimately achieve high predictive accuracy by utilizing the variety of the data.

Although The Bagging classifier does not have the high accuracy rate compared with AdaBoost, Bagging has the advantage of reduces overfitting by aggregating predictions from multiple models. Therefore, Bagging was chosen to be applied on the test phase after the training and validation set combined together to be performed as the training test to train the model with same hyperparameters and test the model on the test set. The evaluation of the performance of the trained model was conducted. The performance of the trained model was assessed. The performance was evaluated using the precision of predictions and the confusion matrix. The evaluation of a model's performance in a binary classification task involves a comparison between the actual and predicted values. A confusion plot, which is a visual depiction of the confusion matrix, offers valuable information regarding the model's true positives, false positives, false negatives, and true negatives [20].

The graphical representation of the confusion plot, as seen in Figure 6, visually presents the frequency of cases in which the Bagging model effectively detected red light violations. This analysis is based on the parameters that were supplied to the model. Additionally, it demonstrates the frequency of erroneously attributing a red light violation, hence indicating a flawed process of identification. Figure 6 displays the confusion matrix plot of the Bagging model when it is applied to the complete dataset. The provided data illustrates an aggregate success rate of 89.4%, Label 1, which corresponds to instances of red light violations, has the best accuracy rate.

VI. DISCUSSION

Interesting insights into the efficacy of AdaBoost and Bagging techniques were revealed through a comparison

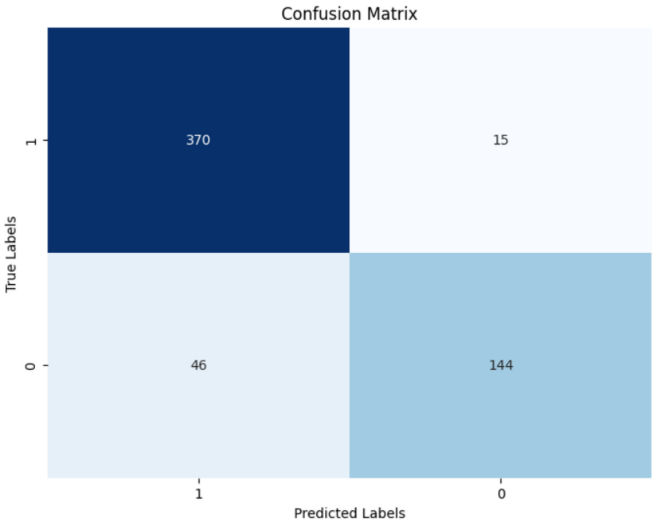

FIGURE 6. Confusion Matrix.

of their performances. AdaBoost achieved a remarkable validation accuracy of 90.4% because of to its careful attention to instances that were incorrectly classified and the adaptive boosting process. However, Bagging, which focuses on lowering variance and overfitting, only managed to achieve a validation accuracy of 90.1%, which is still commendable. Even though AdaBoost demonstrated a marginally higher accuracy rate, it's crucial to consider the trade-offs between these two approaches. AdaBoost's performance may be constrained in circumstances where the dataset contains outliers or is especially noisy due to its sensitivity to noisy data and its propensity for overfitting. Bagging is a good option for lowering the risk of overfitting because it places an emphasis on stability and robustness, making it more reliable when dealing with variance-prone data. It's important to recognize the dataset's limitations for the current study, though. Although the accuracy rates with AdaBoost and Bagging are both encouraging, the study considers that the dataset's size may not be ideal for accurately capturing the full complexity of real-world traffic scenarios. Future research should focus on acquiring a more complete dataset that not only includes data from vehicles and traffic signals, but also makes sure that these sources are in sync with one another. A dataset with synchronized information integration from both traffic signals and vehicles has the potential to produce a reliable and effective predictive model for red light running detection. Such a traffic system model could be crucial in reducing the number of drivers who run red lights and improving road safety.

VII. CONCLUSION

In conclusion, this study went on a thorough journey to develop predictive models for identifying cases of red light running, utilizing cutting-edge methods like AdaBoost, Bagging, and decision trees. The significance of examining methods for predicting traffic light-related driving behavior

and velocity variations lies in enhancing road safety and traffic management in urban settings. By understanding how drivers interact with traffic lights and how their velocities vary in different scenarios, authorities can develop more effective strategies to mitigate risks and improve overall traffic flow.

The process started with the dataset being refined through careful data cleaning, producing a carefully selected collection of 12 features. The feature space was then effectively increased to 47 attributes by encoding these features with one-hot encoding. Two supervised learning models and two ensemble learning techniques were carefully built and assessed using this improved dataset.

The analysis of the AdaBoost and Bagging techniques showed important details about their individual advantages and disadvantages. AdaBoost proved its flexibility by producing a validation accuracy of 90.4%, which is impressive. In contrast, Bagging demonstrated its resistance to overfitting by achieving a commendable accuracy of 90.1%. Although AdaBoost showed slightly better accuracy, the trade-offs related to its sensitivity to noise and overfitting pointed out how crucial it is to select the most appropriate technique for a given situation.

It is important to recognize that, despite showing promising results, the study's dataset might not accurately reflect the complexity of real-world traffic scenarios. The study emphasizes the necessity of ongoing initiatives to compile a larger dataset that synchronizes traffic signal and vehicle data. A robust predictive model with the potential to significantly reduce the number of red-light violations could be created using such a dataset as a catalyst. This model could easily integrate into traffic management systems, promoting improved traffic efficiency and road safety.

In basic terms, this study serves as an example of how ensemble learning techniques can be used to handle challenging classification tasks. Additionally, it demonstrates the significance of extensiveness and high dataset quality in achieving the best model performance. The result of this research opens the door for further research using more complex datasets and cutting-edge methodologies. By making such efforts, the goal of a predictive model-powered, integrated, and effective traffic management system gets closer to reality. In the end, this endeavor advances the field of predictive modeling techniques and transportation safety in general.

Furthermore, the model proposed in the paper offers versatile applications across various domains. Firstly, it can be utilized to optimize traffic light timings based on traffic light-related driving behavior, thereby mitigating congestion and enhancing traffic flow. Secondly, by analyzing velocity changes near traffic signal zones, the model facilitates the development of early warning systems, alerting drivers to sudden halts or junctions in advance. Additionally, driver assistance systems can leverage the model to provide real-time feedback to drivers, assisting them in anticipating traffic light changes and adapting their driving accordingly. Importantly, the program can identify risks associated with elderly passengers and hazy weather conditions, which management cannot entirely control or prohibit. For example, the model aids in formulating safety protocols and instructional programs tailored for older passengers navigating foggy conditions. Moreover, to mitigate road user hazards during adverse weather conditions, authorities can utilize insights from the model to enhance infrastructure elements such as signage and road markings, thereby enhancing overall safety standards.

ACKNOWLEDGMENT

The author gratefully acknowledges King Fahd University of Petroleum and Minerals on the financial support for this research during the academic year 1444AH/2023AD.

REFERENCES

[1] A. Halsey III. "More people died in car crashes this century than in both world wars." Accessed: Jul. 21, 2019. [Online]. Available: https://www.washingtonpost.com/local/trafficandcommuting/more-people-died-in-car-crashes-this-century-than-in-both-world-wars/201907/21/0ecc0006-3f54-11e9-9361-301ffb5bd5e6_story.html?noredirect=on

[2] (Insur. Inst. Highway Safety, Arlington, VA, USA). *Yearly Snapshot: Fatality Facts 2021*. Accessed: May 2023. [Online]. Available: https://www.iihs.org/topics/fatality-statistics/detail/yearly-snapshot

[3] F. Gonzalez, B. Beshah, and S. Washington, "Red light running predictive models using machine learning," *Transp. Res. Record*, vol. 2673, no. 9, pp. 485–496, 2019.

[4] "Running a red light: Top cause of car accidents." Seriousaccidents.com. Accessed: Aug. 22, 2023. [Online]. Available: https://seriousaccidents.com/legal-advice/top-causes-of-car-accidents/running-a-red-light/

[5] A. Smith, B. Johnson, and C. Williams, "Impact of signal timing on red light running incidents," *Transp. Res. Record*, vol. 2673, no. 5, pp. 212–220, 2018.

[6] K. Shaaban and M. S. Ghanim, "Modeling of severity in red-light-running crashes using deep learning recognition," in *Proc. Intermountain Eng., Technol. Comput. (IETC)*, 2023, pp. 181–186, doi: 10.1109/IETC57902.2023.10152094.

[7] F. Viti, S. P. Hoogendoorn, H. J. Van Zuylen, I. R. Wilmink, and B. Van Arem, "Speed and acceleration distributions at a traffic signal analyzed from microscopic real and simulated data," in *Proc. 11th Int. IEEE Conf. Intell. Transp. Syst.*, 2008, pp. 651–656, doi: 10.1109/ITSC.2008.4732552.

[8] F. Pan, Y. Yang, L. Zhang, C. Ma, J. Yang, and X. Zhang, "Analysis of the impact of traffic violation monitoring on the vehicle speeds of urban main road: Taking China as an example," *J. Adv. Transp.*, vol. 2020, Mar. 2020, Art. no. 6304651, doi: 10.1155/2020/6304651.

[9] G. S. Aoude, V. R. Desaraju, L. H. Stephens, and J. P. How, "Driver behavior classification at intersections and validation on large naturalistic data set," *IEEE Trans. Intell. Transp. Syst.*, vol. 13, no. 2, pp. 724–736, Jun. 2012, doi: 10.1109/TITS.2011.2179537.

[10] H. A. Mohammed, M. G. Abadi, and D. S. Hurwitz, "Red-light running violation during car following at high-speed signalized intersections," *Transp. Eng.*, vol. 8, Jun. 2022, Art. no. 100110, doi: 10.1016/j.treng.2022.100110.

[11] F. Galatioto, T. Giuffrè, M. Bell, G. Tesoriere, and T. Campisi, "Traffic microsimulation model to predict variability of red-light running influenced by traffic light operations in urban Area," *Procedia Soc. Behav. Sci.*, vol. 53, pp. 871–879, Oct. 2012, doi: 10.1016/j.sbspro.2012.09.936.

[12] Y. Xie, X. Zhu, and Y. Jin, "Modeling red-light-running violation at signalized intersections using random parameters multinomial logit approach," *J. Transp. Eng., Part A, Syst.*, vol. 142, no. 7, 2016, Art. no. 4016014.

[13] Y. Wu, X. Qu, and Y. Zhang, "Modeling red-light running at signalized intersections: A driving simulator study," *Accident Anal. Prev.*, vol. 145, 2020, Art. no. 105695.

[14] H. Huang, C. Ye, and C. Xu, "Limited amber time leads to rushed acceleration: Empirical evidence and policy implications," *Accident Anal. Prev.*, vol. 107, pp. 1–9, 2017.
[15] D. W. Eby, L. J. Molnar, J. T. Shope, J. M. Vivoda, and T. A. Fordyce, "Improving traffic safety culture in the United States: The climate for evidence-based policy," *Accident Anal. Prev.*, vol. 78, pp. 31–36, 2015.
[16] H. Zhang, M. Abdel-Aty, and J. Lee, "Analysis of traffic signal timing impacts on red-light-running violation risk using a comprehensive empirical approach," *Accident Anal. Prev.*, vol. 131, 2019, Art. no. 103271.
[17] Y. Xie, X. Zhu, and Y. Jin, "Modelling red-light-running violation at signalized intersections using random parameters multinomial logit approach," *J. Transp. Eng., Part A, Syst.*, vol. 142, no. 7, 2016, Art. no. 4016014.
[18] Y. Zhou, Y. Wang, and A. Smith, "Predicting red light running behavior incorporating vehicle trajectory data and weather conditions: An artificial neural network approach," *Accident Anal. Prev.*, vol. 144, 2020, Art. no. 105673.
[19] "Aggressive driving dataset: Toronto police service." ArcGIS Hub. 2022. [Online]. Available: https://hub.arcgis.com/datasets/TorontoPS::aggressive-driving/explore?location=43.7898%2C-75.6130%2C13.82
[20] C.-W. Hsu, C.-C. Chang, and C.-J. Lin, "A practical guide to support vector classification," Dept. Comput. Sci., Nat. Taiwan Univ., Taipei, Taiwan, 2016. [Online]. Available: https://www.csie.ntu.edu.tw/∼cjlin/papers/guide/guide.pdf
[21] J. Lee and D. Lord, "Influence of Vehicle speed on red-light running behavior," *J. Transp. Eng., Part A, Syst.*, vol. 143, no. 1, 2017, Art. no. 4016077.
[22] H. Liu, Q. Cai, J. Wu, and Y. Wang, "Predicting red light running behavior using artificial neural networks," *Transp. Res. Part C, Emerg. Technol.*, vol. 97, pp. 179–192, 2018.
[23] Y. Wang, H. Liu, Y. Wang, and M. Li, "A novel framework for real-time red-light-running identification at signalized intersections," *Sensors*, vol. 18, no. 12, p. 4432, 2018.
[24] Y. Chen, Q. Wu, L. Gao, and Y. Yuan, "Predicting red-light running at signalized intersections using a random forest model," *IEEE Access*, vol. 8, pp. 138165–138176, 2020.
[25] J. Li, H. Yu, and X. Ma, "Identification of red light running at signalized intersections using vehicle trajectory data and artificial neural networks," *IEEE Access*, vol. 7, pp. 80918–80929, 2019.
[26] K. El-Basyouny, T. Sayed, and M. El Esawey, "Driver behavior at signalized intersections: Red light violations analysis," *J. Transp. Eng., Part A, Syst.*, vol. 142, no. 1, 2016, Art. no. 4015044.
[27] A. Smith, Y. Wang, Y. Zhou, and L. Kattan, "A machine learning approach to predicting red-light running behavior," *Transp. Res. Part C, Emerg. Technol.*, vol. 106, pp. 229–242, 2019.
[28] H. Wang, W. Hao, J. So, Z. Chen, and J. Hu, "A faster cooperative lane change controller enabled by formulating in spatial domain," *IEEE Trans. Intell. Veh.*, vol. 8, no. 12, pp. 4685–4695, Dec. 2023.
[29] H. Wang, J. Lai, X. Zhang, Y. Zhou, S. Li, and J. Hu, "Make space to change lane: A cooperative adaptive cruise control lane change controller," *Transp. Res. Part C, Emerg. Technol.*, vol. 143, Oct. 2022, Art. no. 103847.
[30] X. Li, W. Xu, T. Wang, and Y. Yuan, "Infrastructure enabled eco-approach for transit system: A simulation approach," *Transp. Res. Part D, Transport Environ.*, vol. 106, May 2022, Art. no. 103265.
[31] M. M. R. Komol et al., "Deep RNN based prediction of driver's intended movements at intersection using cooperative awareness messages," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 7, pp. 6902–6921, Jul. 2023.
[32] M. M. R. Komol et al., "A review on drivers' red light running behavior predictions and technology based countermeasures," *IEEE Access*, vol. 10, pp. 25309–25326, 2022.
[33] M. Elhenawy, S. Glaser, A. Bond, A. Rakotonirainy, S. Demmel, and M. Masoud, "A framework for testing independence between lane change and cooperative intelligent transportation system," *PLoS One*, vol. 15, no. 2, 2020, Art. no. e0229289.
[34] M. Daley, M. Elhenawy, M. Masoud, S. Glaser, and A. Rakotonirainy, "Detecting road user mode of transportation using deep learning to enhance VRU safety in the C-ITS environment," in *Proc. Australas. Road Safety Conf.*, 2021, pp. 450–452.
[35] X. Yan, E. Radwan, and E. Birriel, "Analysis of red light running crashes based on quasi-induced exposure and multiple logistic regression method," *Transp. Res. Record*, vol. 1908, no. 1, pp. 70–79, 2005.
[36] B. E. Porter and K. J. England, "Predicting red-light running behavior: A traffic safety study in three urban settings," *J. Safety Res.*, vol. 31, no. 1, pp. 1–8, 2000.
[37] A. T. Kashani, S. Amirifar, and M. A. Bondarabadi, "Analysis of driver and vehicle characteristics involved in red-light running crashes: Isfahan, Iran," *Iranian J. Sci. Technol., Trans. Civ. Eng.*, vol. 45, pp. 381–387, 2021.
[38] (Transport Canada Gov. Dept., Ottawa, ON, Canada). *Commercial Vehicle Collisions in Canada, 2012-2021*. [Online]. Available: https://tc.canada.ca/en/road-transportation/statistics-data/commercial-vehicle-collisions-canada-2012-2021
[39] "Intersection safety." 2023. [Online]. Available: https://carsp.ca/en/news-and-resources/road-safety-information/intersection-safety/

MAHMOUD MASOUD received the Ph.D. degree in operations research and mathematical sciences from the School of Mathematical Sciences, Queensland University of Technology (QUT), Brisbane, Australia. He is currently an Associate Professor with the King Fahd University of Petroleum and Minerals. He was a Research Associate with the Center for Accident Research and Road Safety—Queensland, QUT, where he has extensive experience, as a Research Associate, in many industrial projects as a part of effective teamwork with the Centre for Tropical Crops and Bio-commodities. This team constructed industrial linkages with big industrial organizations, such as EY and MLA (beef supply chain projects), owners of the Australian Miles (Biomass and Bioenergy assessment—sugarcane projects), and Brisbane Royal Hospital (health system project). He has a wide range of experience in academic research and industrial projects with more than 80 refereed journal, conference papers, and industrial reports.