A Machine Learning-Based Predictive System for Road Accident Risk Management Using Cloud Architecture

Angel Portal
Faculty of Information Systems Engineering, Universidad Peruana de Ciencias Aplicadas, Lima, Peru
u20201B307@upc.edu.pe

Marlonn Sandoval
Faculty of Information Systems Engineering, Universidad Peruana de Ciencias Aplicadas, Lima, Peru
u202010130@upc.edu.pe

Pedro Castaneda
Faculty of Systems Engineering and Electrical Mechanics, Universidad Nacional Toribio Rodriguez de Mendoza, Amazonas, Peru
pedro.castaneda@untrm.edu.pe

Juan Mansilla-Lopez
Faculty of Information Systems Engineering, Universidad Peruana de Ciencias Aplicadas, Lima, Peru
pcsijman@upc.edu.pe (corresponding author)

Alberto Daniel Garcia-Nunez
Universidad Pontificia Bolivariana, Medellin, Antioquia, Colombia
alberto.garcia@upb.edu.co

Received: 1 October 2025 | Revised: 10 November 2025, 17 November 2025, and 25 November 2025 | Accepted: 26 November 2025

 DOI: https://doi.org/10.48084/etasr.15262

ABSTRACT

This study aimed to address the issue of the high frequency of traffic accidents in Lima Metropolitana by providing a predictive system built using machine learning techniques. A Random Forest (RF) model was trained with a set of historical data on previous accidents, relevant climatic variables, and infrastructure characteristics. The method used involves basic stages of information capture and preprocessing, exhaustive feature engineering to maximize predictive capacity, and implementation in a cloud-based architecture. The results obtained show that the proposed model achieved an accuracy of 50.92%, a recall of 56.97% and an F1 score of 53.77%, demonstrating high efficiency in the detection of geographical areas with a high risk of accident occurrences. Feature importance analysis revealed that variables such as the district (37.2%), the time of the accident (10.8%), and the month (9.5%) are the most significant factors in the predictions, confirming the importance of spatio-temporal patterns in traffic incidents. This proposed system has a scalable and adaptable design that ensures its applicability in urban environments with similar characteristics. Such technologies can significantly contribute to the reduction of accidents and the improvement in traffic management in the capital of Peru.

Keywords-cloud architecture; machine learning; random forest; traffic accident prediction

I. INTRODUCTION

Traffic accidents in Metropolitan Lima are a serious issue with many implications for road safety, public health, and the local economy. The magnitude of this challenge is reflected in the fact that between 2017 and 2022, more than 50% of road accidents throughout the country occurred in the capital, reaching 5,449 accidents in 2022. These events had tragic consequences, resulting in 930 deaths and 7,817 injuries. Among the most reported root causes in the literature are contrary or reckless driving, speeding, and driving under the influence of alcohol or other substances, among other factors that are worsened by poor road infrastructure and road safety

education [1]. These incidents not only put the safety of people at risk, but also increase traffic congestion, reduce transportation efficiency, and lead to high economic and social costs for the city.

Several technological solutions have been proposed to mitigate traffic accidents using big data analysis and Machine Learning (ML) algorithms. For example, in [2], statistics and ML were used to detect areas of risk of accidents in Beijing, while in [3], Deep Neural Networks (DNNs) were employed to improve accuracy in predicting road accidents in South Korea. In Peru, most studies focused on the identification of the factors that cause traffic accidents [4], but the implementation of technological solutions has received lower importance. In addition, most existing solutions remain limited in large urban areas, such as Metropolitan Lima, due to their dependence on specific technological infrastructure and narrow data scope. Many focus only on traffic flow or meteorological factors [5], neglecting the behavioral, socioeconomic, and infrastructure variables essential for scalable and adaptable implementations.

Given these weaknesses, this study presents a new approach for the prediction of traffic accidents in Metropolitan Lima, using multidimensional data and ML algorithms. An ML–based predictive system was designed to identify high-risk geographic areas and temporal patterns of traffic accidents, enabling preventive decision-making and urban safety management. The proposed solution integrates many classic variables (traffic information, weather conditions) with less conventional ones, but with relevant predictive capacity (inferences of driving behavior, socioeconomic elements of risk areas, geolocated data in road infrastructure), to offer more accurate and real-time predictions. In addition, its flexible and scalable design allows it to operate as a prototype that can be replicated in other urban contexts with similar conditions, effectively helping to reduce accidents and optimize traffic management.

Several studies have explored predictive models for traffic accident analysis using ML and Deep Learning (DL) techniques. In [3], deep neural networks achieved high accuracy in accident prediction. In [6], Random Forest (RF) was the best performer in predicting the severity of accidents in the UK, while in [7], ensemble models such as XGBoost were combined with SHAP analysis to explain variable effects. In [8], high accuracy was achieved using RF with imputation on South African data, while in [9], an MDG ensemble model was effective under noisy and inconsistent conditions. These studies highlight the importance of robust models and careful feature selection in accident prediction.

The field of prediction of serious accidents and traffic analysis on motorways has been another research area that has attracted attention. In [10], big data were used with XGBoost and a Bayesian neural network to measure the risk of severe accidents on Turkish highways, achieving an accuracy of 30%. In [11], a risk assessment model used decision trees, taking average speed and weather conditions as variables, thus being able to perform a spatio-temporal risk analysis. In [12], more than one million records from the UK were processed and analyzed with classification models, achieving 85% accuracy and highlighting the power of preprocessing and data encoding. In [13], statistical and supervised models were used to predict accidents on highways, in which the volume of traffic and its deviation were taken as variables, highlighting a statistical correlation. These studies underline the importance of high-speed scenarios, hybrid models, and the exploration of large amounts of data in high-risk road areas.

Previous studies have also investigated how climatic factors and environmental conditions directly affect road safety. The study in [14] quantified the pavement friction threshold below which the risk of an accident increases using regression models with sensor data. The MSGSGCN model [15] was based on graph neural networks to predict the speed of irregular road networks, offering 4% improvement over previous established models. These works show how the environment and climate cannot be neglected in predictive models, since incorporating them can significantly improve accuracy, especially in extreme conditions.

Many advances have also been made in real-time monitoring and automatic detection of road incidents using Artificial Intelligence (AI). In [16], the STVDAE model used spatiotemporal autoencoders, achieving significant detection speed improvements on high-traffic urban avenues. In [17], an automatic system combined classification and prediction of incident duration using RF and Gradient Boosted Trees (GBT), reducing RMSE by up to 45.6% in the prediction of resolution time. The integration of real-time data and the use of optimized models allows for improved incident response, which is crucial for effective smart traffic management.

Several regional studies have also applied similar approaches in Europe and Australia [18-20], confirming the growing interest in adapting ML techniques to local traffic conditions.

II. SYSTEM DESIGN

A. Architecture

The proposed system for the prediction of traffic accidents in Lima, Peru, follows an integrated architecture based on cloud services, real-time data processing, and frontend technologies accessible to the end user. Figure 1 provides a high-level view of how the system works. The architecture includes the following components:

- Frontend: A web application allows users, both private entities and the general public, to access the platform from their browsers. Users can view the prediction results on traffic accident risk areas, generate reports, and query on previous incidents. Access is through an API that connects the frontend to the other components of the system.
- AWS Services (Amazon Web Services): The platform uses AWS Lambda for data processing and Amazon SageMaker to train and run the predictive model. These services allow for scaling the infrastructure without having to worry about servers, adjusting the execution of agile high-performance solutions. Integration with other AWS services, such as Amazon S3 for data storage and API Gateway for communication between modules, allows processing large amounts of data efficiently.

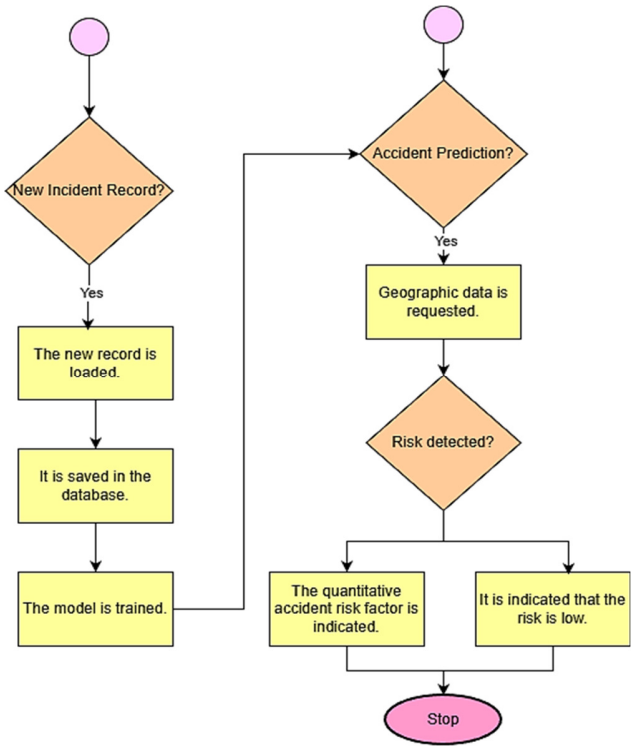

Fig. 1. Flowchart diagram.

- Predictive Model: The predictive model was trained with historical accident records, meteorological data, and infrastructure characteristics of Lima highways. This model is loaded into Amazon SageMaker, where ML algorithms are used to perform traffic accident prediction calculations. The architecture allows the processing of information in real time and sending alerts in a predictive way.
- Data Sources: Data is extracted from various sources, such as public datasets, for example, the records of the National Road Safety Observatory (ONSV), data from the Ministry of Transport and Communications (MTC), and historical meteorological data. This data is recorded and managed through cloud database services, allowing for fast and secure access.

B. Methodology

1) Dataset

Various datasets relevant to the analysis of accident patterns and prevention in Metropolitan Lima were used. One of the main sources was the National Road Safety Observatory (ONSV) dataset (2008–2023) [21], which contains detailed information on fatal accidents, including geographic location, type of event, and consequences. Additional data on participants—both individuals and vehicles in fatal accidents—were incorporated for the period 2021–2023, along with data on road network conditions provided by the Ministry of Transport and Communications (MTC, 2022) [22]. Historical meteorological records for Lima (2000–2014) were also used to assess the impact of weather conditions on accident occurrence [23]. In addition, contextual variables, such as rainfall, holidays, and time of day, were included to enrich the dataset and capture temporal and environmental factors that influence the probability of accidents. The data preparation process was carried out using a systematic cleaning and transformation method. In the first instance, the data was filtered to include only records from Lima, since the model focuses on this geographical region. As part of the data improvement process, additional variables were created from the available temporal information, and holiday dates were included using a list of known dates. A key aspect in data preprocessing was the generation of a balanced dataset by synthetically generating examples of "no accidents." This strategy is essential for the model to correctly learn to recognize the elements that help the occurrence of accidents. To achieve this, an intelligent algorithm was used, which modifies the characteristics of real accident cases by prioritizing the variables with the greatest statistical relevance to ensure the accuracy of the artificially generated cases using the k-Nearest-Neighbors (KNN) algorithm.

The final dataset was divided into three final subsets: training (80%), validation (10%), and testing (10%). The split was performed using a stratified random procedure, ensuring that the proportion of accident and non-accident cases remained consistent across all subsets. The first was used to train the model on the overall patterns, the second to improve it and avoid overfitting, and the third to test the model's performance against new data. Likewise, a cross-validation technique was carried out to ensure a more reliable evaluation and reinforce the prediction functionality of the model. The datasets used in this study are publicly available from official Peruvian institutions. Accident data can be obtained from ONSV [21], road network information from MTC [22], and historical climate data from SENAMHI [23]. Due to data protection regulations, the dataset generated in this study cannot be redistributed.

2) Model

The predictive system for traffic accidents in Metropolitan Lima is based on the RF algorithm, selected because of its ability to handle heterogeneous data naturally and recognize nonlinear relationships between different predictor variables. Its suitability lies in its ability to integrate both categorical and numerical variables, its resistance to overfitting, and its ability to determine the relative importance of each feature in the prediction [24, 25]. The model architecture is divided into two fundamental modules: a data preprocessor and an accident predictor. The former coordinates the vital tasks of loading, cleaning, transforming, and balancing data, while the latter oversees training, comprehensive assessment, and efficient implementation of the predictive model. A chain of sequential and modular methods was implemented to perform data cleaning and transformations. Among the most significant transformations performed, the following should be highlighted:

- Geographic filtering, restricting the analysis to data specific to Lima Metropolitana.
- Comprehensive cleaning of both categorical and numerical variables, addressing missing values and inconsistencies.

- Extraction of relevant temporal characteristics, such as day of the week, month, and period of day.
- Coding of categorical variables using appropriate techniques for inclusion in the model.
- Synthetic generation of "non-accident" cases to balance the dataset and improve predictive capacity.
- Normalization of column names, ensuring compatibility with database systems.

Subsequently, a strict segmentation was defined between the training (80%), validation (10%), and test (10%) sets, applying stratified cross-validation to ensure an appropriate evaluation and avoid overfitting. For the selection of features, the most representative variables were used through an analysis of importance. The variables were: variables related to the area or place of incidence, and variables related to the date and time. These variables were decisive in improving the accuracy of the model during testing. On the other hand, the accident predictor incorporates essential functionalities for the development and validation of the model, including:

- Preparing the datasets, one for training and one for testing.
- Application of advanced balancing techniques, such as SMOTE (Synthetic Minority Over-sampling Technique), to address class imbalance.
- Optimized training of the RF model, using adjusted parameters to maximize its performance.
- Evaluating model performance across multiple metrics.
- Detailed analysis of feature importance, providing valuable information on the most influential factors in accident prediction.
- Serialization of the trained model to facilitate its subsequent implementation and use in the production environment.

The proposed system and its modules were developed using the Python programming language, taking advantage of the efficiency of specialized libraries such as Scikit-learn, which was fundamental in the implementation of the RF model. Pandas was used for efficient data manipulation and analysis, and various visualization tools, such as Matplotlib and Seaborn, were used for a clear and efficient interpretation of the results obtained. To complement the model, a web platform was designed to act as the interface for interaction with the end user. This web application was deployed in the cloud, in particular, AWS. Using AWS cloud services relieved us from secondary tasks such as database management, focusing on the development of a web application that allows data loading, consulting predictions, generating reports, and analyzing statistics from a single platform. The web system contains the following modules:

- Dashboard: Provides interactive visualizations that show the evolution of accidents by period, district, and environmental conditions.
- Data Module: Allows for accessing, filtering, uploading, or exporting historical accident records, with options for debugging and continuous updating.
- Prediction Module: Specific variables (time, weather, type of road, etc.) can be entered to obtain an immediate prediction of the risk of accidents in a given area.
- Reports and Statistics: Offers the ability to download PDF or Excel reports on risk analysis by location or date.
- Configuration and Administration: Manage user access, define roles, update the predictive model, and manage data security.

3) Training

The predictive model was trained using a systematic protocol aimed at optimizing performance. The preprocessed data were initially divided into training (1953 samples) and test (489 samples) sets using a stratified partition that preserved the relationship of accidents and non-accidents in both sets.

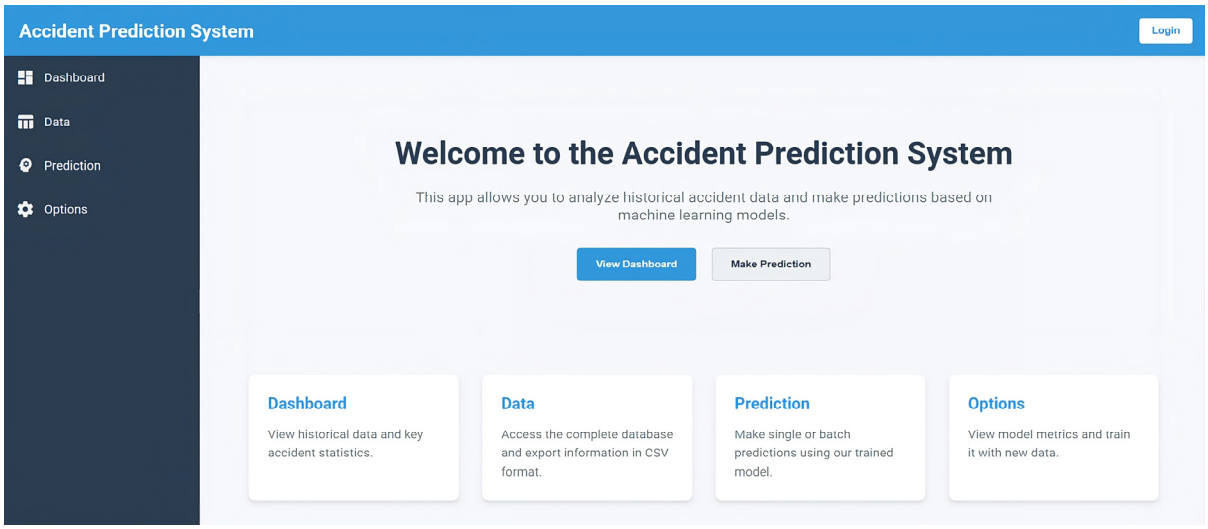

Fig. 2. Interface of the predictive system.

Continuous numerical variables were standardized using the z-score transformation to smooth the convergence of the algorithm. To prevent temporal bias, the chronological distribution of records was maintained in the data splits, preventing seasonal or evolutionary patterns from inducing information leaks between sets. In addition, to reduce the impact of hypothetical local imbalances, SMOTE was applied specifically to the training set, obtaining 977 instances per class. Finally, automatic validations were carried out to ensure the referential integrity of the data, especially for interdependent variables such as the time period of the day or the location-district. This approach ensured an unbiased evaluation of the final performance on the intact test set, simultaneously allowing fine-tuning of the model through the validation set and cross-validation.

The RF algorithm was implemented using the scikit-learn library, complemented by imbalanced-learn for balancing. Hyperparameter optimization followed a two-stage methodology. First, a broad exploration was performed using a coarse-grained grid search that covered a large space of possible configurations, including variations in key parameters such as *n_estimators* (50, 100, 200, 300), *max_depth* (10, 15, 20, None), *min_samples_split* (2, 5, 10, 15), *min_samples_leaf* (1, 2, 4, 5), *max_features* ('sqrt', 'log2', 0.3, 0.5), *bootstrap* (True, False) and *criterion* ('gini', 'entropy'). Once the promising region of the hyperparameter space was identified, the search was refined using fine-grained analysis around the preliminary optimal values. The final selection of hyperparameters was based on the optimization of the weighted F1-score, a metric that provides a balance between precision and accuracy, considering both classes. The resulting configuration comprised *n_estimators*: 200, *max_depth*: 15, *min_samples_split*: 5, *min_samples_leaf*: 2, *max_features*: 'sqrt', *bootstrap*: True, *criterion*: 'gini', *class_weight*: 'balanced', and *random_state*: 42 to ensure reproducibility.

During model training, several monitoring mechanisms were activated to control the process. Performance-based early stopping was established based on the validation set, stopping training when no improvement was seen in 50 consecutive trees. In addition, learning curve visualizations represented the change in error as a function of the size of the training set to diagnose bias-variance regimes. A stability analysis was also performed to test the consistency of predictions and the relevance of features between different cross-validation folds to verify the model's metrics.

After training, the model was serialized and saved as a .pkl file, while its associated metrics were saved in a .json file. These files can be dynamically loaded by the system using specific scripts, which facilitates their reuse at later stages and their inclusion in the prediction flow in production.

4) Evaluation and Statistical Analysis

The classification model was evaluated using multiple complementary metrics, as shown in Table I, to capture different aspects of its performance. Precision measures the proportion of positive predictions that are correct, sensitivity or recall assesses the model's ability to correctly identify all positive cases, and specificity measures how well the model can identify negative cases. The F1-score represents the harmonic mean between accuracy and recall, providing a metric useful when seeking a balance between the correct identification of positive cases and the minimization of false alarms. The area under the ROC curve (AUC-ROC) assesses the model's discriminative capacity based on its performance across all possible decision thresholds. In addition, to avoid randomness problems, a 5-fold stratified cross-validation was used to evaluate the behavior of the model against different partitions of the dataset and obtain more reliable estimates of the real performance of the model.

TABLE I. EVALUATION METRICS

#	Metric	Description	Formula
1	Precision	Proportion of correct predictions among all predictions. Evaluate the overall accuracy of the model.	$\frac{TP + TN}{TP + TN + FP + FN}$
2	Recall	Proportion of true positives among all real positive cases. Important in accident detection.	$\frac{TP}{TP + FN}$
3	F1-score	Harmonic mean between accuracy and recall, useful in unbalanced class scenarios.	$2 \times \frac{Precision * Recall}{Precision + Recall}$
4	Specificity	Proportion of true negatives among all real negative cases. Assesses the ability to avoid false positives.	$\frac{TN}{TN + FN}$

III. RESULTS

Table II presents the quantitative results, which indicate a moderate performance of the model in the test set, with a slight tendency to favor sensitivity (greater ability to identify real accidents) at the expense of precision (higher proportion of false positives).

TABLE II. MODEL EVALUATION METRICS ON THE TEST SET

#	Metric	Value	Standard deviation
1	Precision	0.5092	0.5112
2	Recall	0.5697	0.0510
3	F1-score	0.5377	0.0292
4	Specificity	0.5112	0.0171

The confusion matrix in Figure 3 provides a detailed view of the behavior of the model. The model correctly classified 111 cases as non-accidents (true negatives) and 139 cases as accidents (true positives). In addition, 134 cases were incorrectly classified as accidents while they were not (false positives), and 105 real accidents were not detected by the model (false negatives). This distribution of errors indicates a slight bias towards positive classification, which can be explained by the inherent complexity of predicting the occurrence of accidents based on the available contextual variables. The ROC curve in Figure 4 visualizes the model's discrimination ability across different thresholds. These results are in line with the evaluation on the test set, showing that the model has stable performance across various data subsets. It should be noted that although the model shows some potential in accident prediction, its overall performance, particularly in terms of accuracy and AUC-ROC, still needs improvement.

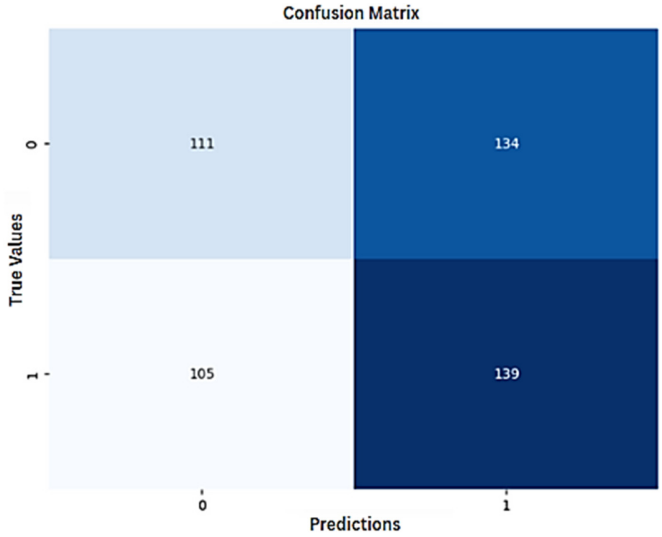

Fig. 3. Confusion matrix.

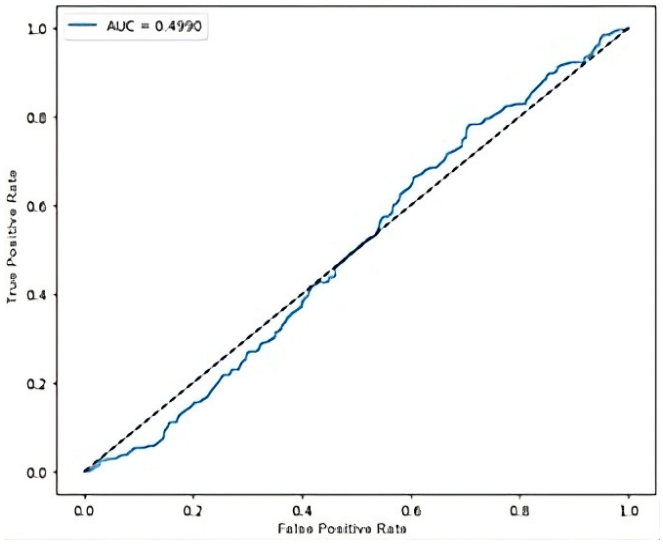

Fig. 4. ROC curve.

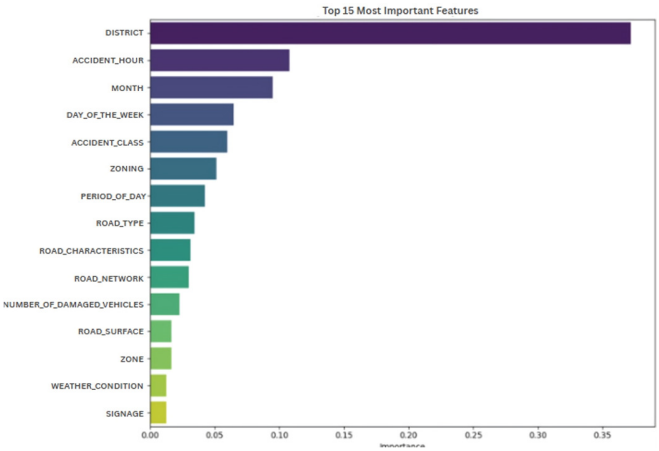

Fig. 5. Most relevant variables.

The feature importance analysis in Figure 5 indicates that the variable "DISTRICT" is by far the most significant predictor (37.2% of relative importance), followed by "TIME OF THE ACCIDENT" (10.8%) and "MONTH" (9.5%). This suggests that the spatiotemporal distribution of accidents shows significant patterns. Temporal context variables (hour, month, day of the week) together account for approximately 31% of the total predictive significance, reflecting the role of temporal patterns in the occurrence of accidents.

To better understand the behavior of the model under different classification criteria, Figure 6 shows a systematic analysis of the impact of the decision threshold on the main metrics. The threshold analysis reveals a characteristic pattern in which accuracy and recall present inversely proportional behaviors as the threshold varies. With low thresholds (≤0.3), the model maximizes recall (98.3-100%) at the expense of modest accuracy (~50%). As the threshold increases, the accuracy tends to improve initially, peaking at around 0.5, and then deteriorating significantly with thresholds above 0.6. The F1 score, which represents the balance between accuracy and recall, reaches its maximum value (~0.66) with thresholds between 0.1 and 0.3, suggesting that a lower threshold setting could optimize the balance between the two metrics. This finding has relevant practical implications, as it allows the model to be adjusted according to the specific requirements of the application: privileging the detection of all possible accidents (high recall) or minimizing false alarms (high accuracy).

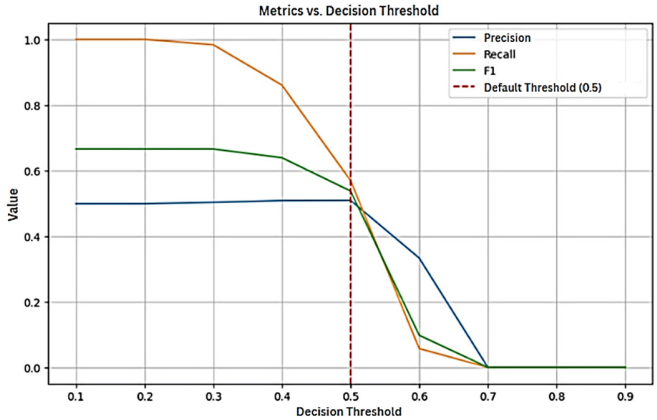

Fig. 6. Metric vs decision threshold.

IV. DISCUSSION

A strong influence of geographical and temporal factors on the occurrence of accidents can be determined, which is consistent with previous findings, such as in [11], which also addressed the importance of spatiotemporal factors in accident risk. Although in the context of highways, the study in [13] also supports the existence of a statistical correlation between traffic volume and its respective temporal deviation with the occurrence of accidents, which supports the existence of temporal patterns in this study. On the other hand, after evaluating the results, a moderate performance can be evidenced, which highlights the potential and, at the same time,

the limitations of the prediction model in a complex urban reality such as Metropolitan Lima.

Comparing the proposed model's results with the literature, important differences in performance can be observed. In [3], 94% accuracy was achieved using DNN, while in [8], 97% accuracy was achieved using RF on South African data. However, such comparisons should take into account discrepancies in geographical contexts, the size of the datasets, and the complexity of the variables. In studies with more comparable performance, similarities with a study for the UK [6] can be noted. The study in [6] noted that RF was the best classifier, although it did not report the overall accuracy. Similar results were achieved in [7], with an accuracy of 73.63% using XGBoost and SHAP explanations, which suggests that the complexity of the accident prediction problem lies in the inherent difficulty of achieving high performance regardless of the proposed algorithm.

Disparities in performance can be attributed to differences in methodology and context. First, the quality and availability of the data are markedly different throughout the studies. For example, while the study in [12] processed more than one million traffic accident records from the UK, achieving an accuracy of 85%, this work used a smaller dataset (with 2422 final samples), which may partly explain its lower performance. Secondly, the social, economic, and cultural context of the city of Lima has its own characteristics that may not be well reflected in models whose training or studies are carried out in countries with more orderly and classic road infrastructures. An important difference with respect to similar research lies in the differences presented by the Lima context. Although, as noted in [16, 17], complementing the model with real-time data could improve predictive capacity, this requires a data collection infrastructure that currently does not exist in Metropolitan Lima, which represents a challenge for future developments.

In practical terms, the proposed predictive system can be used as a decision-support tool for public institutions responsible for mobility and road safety. The system's predictions could help identify critical accident hotspots and high-risk time intervals, allowing authorities to plan evidence-based interventions such as traffic-flow adjustments, infrastructure improvements, or targeted driver-awareness programs. By integrating these predictive insights into operational decision-making, the model goes beyond theoretical analysis and directly contributes to proactive road management. In the long term, this system could be integrated into intelligent transportation platforms at the municipal level, providing continuous feedback and helping to reduce accident rates in dense urban environments such as Metropolitan Lima.

This research advances the current state of knowledge by bridging the gap between high-performance predictive modeling and its contextual implementation in developing urban areas. Previous works have demonstrated accuracy improvements through algorithmic optimization, yet few have translated these models into scalable architectures that can be operationalized by public institutions. The integration of cloud services, multidimensional data, and interpretable ML models presented here extends previous approaches by offering a cost-effective, replicable framework that can be continuously updated with new local information. In this way, this study contributes not only to predictive accuracy but also to the democratization of AI-driven solutions for road safety management.

Therefore, although the model developed has limitations compared to studies in developed countries, the results are promising considering the data constraints and the specific context of Metropolitan Lima. The identified spatiotemporal patterns provide solid foundations for the design of targeted public policies, and the architecture developed sets a benchmark for the application of AI in the prediction and prevention of accidents.

V. CONCLUSION

This study demonstrates the feasibility of building a predictive system for traffic accidents in urban environments such as Metropolitan Lima, using ML and cloud technologies to increase road safety. The proposed model, which is based on the RF algorithm, helped identify fundamental patterns in accidents, first highlighting the role of time and location. Although the model presents moderate results, it has laid relevant foundations in applying AI to the prediction and prevention of incidents. The existence of different challenges of a technical and methodological nature has been highlighted, which can be specified in the need to seek new sources of information (traffic flow in real time, driver information, updated status of the infrastructure), and/or take into account certain qualitative or subjective information that can become fundamental for the prediction of accidents. To improve results, future work should investigate other ML algorithms, superior feature engineering, richer spatiotemporal analysis, or, for example, segmentation by type of accident.

In conclusion, this work can be considered an important first step in AI methods applied to the prediction and prevention of traffic accidents in Metropolitan Lima. Prediction of events such as traffic accidents requires a holistic, multidisciplinary, and multivariate vision, which combines data analysis with knowledge of the context for the prevention of events and incidents such as accidents.

ACKNOWLEDGMENTS

The authors are grateful to the Dirección de Investigación of the Universidad Peruana de Ciencias Aplicadas (UPC) for the support provided for this research work through the UPC-EXPOST-2025-2 incentive.

REFERENCES

[1] "Informe de adjuntía No. 022-2022-DP/AMASPPI," *Defensoria del Pueblo - Perú*. https://www.defensoria.gob.pe/informes/informe-de-adjuntia-n-022-2022-dp-amasppi/.

[2] H. Yang, X. Zhao, S. Luan, and S. Chai, "A traffic dynamic operation risk assessment method using driving behaviors and traffic flow Data: An empirical analysis," *Expert Systems with Applications*, vol. 249, Sept. 2024, Art. no. 123619, https://doi.org/10.1016/j.eswa.2024.123619.

[3] Z. Jin, B. Noh, Z. Jin, and B. Noh, "From Prediction to Prevention: Leveraging Deep Learning in Traffic Accident Prediction Systems," *Electronics*, vol. 12, no. 20, Oct. 2023, https://doi.org/10.3390/electronics12204335.

[4] J. C. Miranda, "Factores que influyen en los accidentes de tránsito ocasionados por el transporte público terrestre en Villa El Salvador, 2021," B.S. Thesis, Universidad Autonoma del Peru, 2023.

[5] Y. Yang, K. Wang, Z. Yuan, and D. Liu, "Predicting Freeway Traffic Crash Severity Using XGBoost-Bayesian Network Model with Consideration of Features Interaction," *Journal of Advanced Transportation*, vol. 2022, no. 1, 2022, Art. no. 4257865, https://doi.org/10.1155/2022/4257865.

[6] I. C. Obasi and C. Benson, "Evaluating the effectiveness of machine learning techniques in forecasting the severity of traffic accidents," *Heliyon*, vol. 9, no. 8, Aug. 2023, https://doi.org/10.1016/j.heliyon.2023.e18812.

[7] S. Dong *et al.*, "Predicting and Analyzing Road Traffic Injury Severity Using Boosting-Based Ensemble Learning Models with SHAPley Additive exPlanations," *International Journal of Environmental Research and Public Health*, vol. 19, no. 5, Mar. 2022, https://doi.org/10.3390/ijerph19052925.

[8] T. Bokaba, W. Doorsamy, B. S. Paul, T. Bokaba, W. Doorsamy, and B. S. Paul, "Comparative Study of Machine Learning Classifiers for Modelling Road Traffic Accidents," *Applied Sciences*, vol. 12, no. 2, Jan. 2022, https://doi.org/10.3390/app12020828.

[9] B. K. Koo, J. W. Baek, K. Y. Chung, B. K. Koo, J. W. Baek, and K. Y. Chung, "Weight Feedback-Based Harmonic MDG-Ensemble Model for Prediction of Traffic Accident Severity," *Applied Sciences*, vol. 11, no. 11, May 2021, https://doi.org/10.3390/app11115072.

[10] T. Baykal, F. Ergezer, E. Eriskin, and S. Terzi, "Accident severity prediction in big data using auto-machine learning," *Scientia Iranica*, vol. 32, no. 7, Mar. 2025, https://doi.org/10.24200/sci.2023.60144.6626.

[11] Z. Cheng, J. Yuan, B. Yu, J. Lu, and Y. Zhao, "Crash Risks Evaluation of Urban Expressways: A Case Study in Shanghai," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 9, pp. 15329–15339, Sept. 2022, https://doi.org/10.1109/TITS.2022.3140345.

[12] S. P. Ardakani *et al.*, "Road Car Accident Prediction Using a Machine-Learning-Enabled Data Analysis," *Sustainability*, vol. 15, no. 7, Mar. 2023, https://doi.org/10.3390/su15075939.

[13] Y. Yang, K. He, Y. Wang, Z. Yuan, Y. Yin, and M. Guo, "Identification of dynamic traffic crash risk for cross-area freeways based on statistical and machine learning methods," *Physica A: Statistical Mechanics and its Applications*, vol. 595, June 2022, Art. no. 127083, https://doi.org/10.1016/j.physa.2022.127083.

[14] A. Abohassan, K. El-Basyouny, and T. J. Kwon, "Effects of Inclement Weather Events on Road Surface Conditions and Traffic Safety: An Event-Based Empirical Analysis Framework," *Transportation Research Record*, vol. 2676, no. 10, pp. 51–62, Oct. 2022, https://doi.org/10.1177/03611981221088588.

[15] C. Cao *et al.*, "Dynamic Spatiotemporal Correlation Graph Convolutional Network for Traffic Speed Prediction," *Symmetry*, vol. 16, no. 3, Mar. 2024, https://doi.org/10.3390/sym16030308.

[16] Y. Lu, Q. Lin, H. Chi, and J. Y. Chen, "Automatic incident detection using edge-cloud collaboration based deep learning scheme for intelligent transportation systems," *Applied Intelligence*, vol. 53, no. 21, pp. 24864–24875, Apr. 2023, https://doi.org/10.1007/s10489-023-04673-7.

[17] A. Grigorev, A. S. Mihăiţă, K. Saleh, and F. Chen, "Automatic Accident Detection, Segmentation and Duration Prediction Using Machine Learning," *IEEE Transactions on Intelligent Transportation Systems*, vol. 25, no. 2, pp. 1547–1568, Oct. 2024, https://doi.org/10.1109/TITS.2023.3323636.

[18] D. Santos *et al.*, "Machine Learning Approaches to Traffic Accident Analysis and Hotspot Prediction," *Computers*, vol. 10, no. 12, Nov. 2021, https://doi.org/10.3390/computers10120157.

[19] A. Grigorev, S. Shafiei, H. Grzybowska, and A. S. Mihăiţă, "Predicting the Duration of Traffic Incidents for Sydney Greater Metropolitan Area using Machine Learning Methods," *International Journal of Intelligent Transportation Systems Research*, vol. 23, no. 1, pp. 104–125, Apr. 2025, https://doi.org/10.1007/s13177-024-00437-w.

[20] Z. Yuan, K. He, and Y. Yang, "A Roadway Safety Sustainable Approach: Modeling for Real-Time Traffic Crash with Limited Data and Its Reliability Verification," *Journal of Advanced Transportation*, vol. 2022, no. 1, 2022, Art. no. 1570521, https://doi.org/10.1155/2022/1570521.

[21] "Database of Fatal Traffic Accidents 2008–2023," *National Road Safety Observatory (ONSV)*. https://www.onsv.gob.pe/datosabiertos.

[22] "Inventory of the National Road Network 2022," *Ministry of Transport and Communications (MTC)*. https://portal.mtc.gob.pe/transportes/caminos/vial.html.

[23] "Historical Record of Climate Data for Lima 2000–2014," *National Meteorology and Hydrology Service of Peru (SENAMHI)*. https://www.senamhi.gob.pe/site/descarga-datos/.

[24] A. Portal, M. Sandoval, and P. Castañeda, "Machine Learning-Based System for Predictive Management of Road Accidents in Metropolitan Lima," *Nanotechnology Perceptions*, vol. 20, no. S15, pp. 3611–3619, 2024.

[25] A. K. Chhotu and S. K. Suman, "Predicting the Severity of Accidents at Highway Railway Level Crossings of the Eastern Zone of Indian Railways using Logistic Regression and Artificial Neural Network Models," *Engineering, Technology & Applied Science Research*, vol. 14, no. 3, pp. 14028–14032, June 2024, https://doi.org/10.48084/etasr.7011.