Future Technology
Open Access Journal
ISSN 2832-0379

November 2026| Volume 05 | Issue 04 | Pages 270-283

Journal homepage: https://fupubco.com/futech

https://doi.org/10.55670/fpll.futech.5.4.22

Future Publishing LLC

Article

# Spatial–temporal machine learning for traffic violation type prediction: a case study in Qatar

**Mohammed Alshriem*, Yin Yang**

Hamad bin Khalifa University, Qatar

**ARTICLE INFO**

*Article history:*
Received 26 March 2026
Received in revised form
04 September 2026
Accepted 21 September 2026

Keywords:
Traffic violation, Multiclass classification, Spatiotemporal analysis, Random Forest, Transportation systems, Traffic safety

*Corresponding author
Email address:
malshriem@hbku.edu.qa

DOI: 10.55670/fpll.futech.5.4.22

**ABSTRACT**

Although traffic violations are among the most significant factors contributing to road accidents, little research has focused on predicting specific types of traffic violations in Qatar and the Gulf region, where the number of vehicles has increased significantly in recent years. This work addresses this gap by proposing a machine-learning framework to solve fine-grained, multi-class prediction of traffic violation types on a large-scale dataset from Qatar. The dataset is extremely imbalanced and has a high-dimensional categorical space, with 97 original violation classes. We created a preprocessing pipeline to address these issues by filtering the 97 original codes to nine well-represented violation types while retaining 99.6% of the records. The final model uses 18 spatiotemporal features, including our zone violation entropy measure, which captures violation diversity in each enforcement zone. Five classifiers are evaluated with a strict temporal train–validation–test split, training-based encoding (to avoid temporal leakage), and balanced class weighting. Random forest's macro F1 score is 0.6184 and balanced accuracy is 0.7064, outperforming the linear and neural baselines. SHAP analysis shows that location-derived features account for ~74% of the model's predictive importance, while temporal features account for ~25%. The results indicate that violation type is more strongly associated with location than with time, and they provide implications for targeted enforcement by location. The suggested framework provides a reproducible way to predict violation types in a multi-class setting. It shows that enforcement information alone can support predictive analysis without using road topology or sensor data.

## 1. Introduction

Road traffic injuries account for one of the greatest sources of mortality and morbidity in the world today, being responsible for approximately 1.3 million deaths each year. Road traffic accidents are costly for economies worldwide [1]. According to the World Health Organization, road safety is a top public health priority today, with developing nations bearing the major burden of these injuries. As urbanization and car ownership increase, the need for traffic management strategies has also grown [2]. Traffic violations are a major contributor to road accidents, with multiple studies showing a strong association between violations and accidents [3,4]. However, each type of traffic violation – e.g. speeding, red-light running, and improper lane changing – carries its own risk level. Each should therefore be addressed with specific enforcement approaches. Conventional enforcement methods remain predominantly reactive; cameras and patrols are positioned based on past violation locations or general temporal patterns, without predicting the types of violations that may occur. Nevertheless, the emergence of artificial intelligence and machine-learning technology allows traffic enforcement to shift from reactive to proactive. Recently, several research projects have used these technologies to address various traffic safety problems, with results falling into three categories. First, driver-centered risk assessment models establish relationships between driver features, behavior, and accident likelihood [5,6]. However, while such research provides some indication of human factors, it focuses on drivers rather than city enforcement. Second, spatio-temporal analysis of violations and accidents uses techniques ranging from kernel density estimation to time–series clustering [7-9]. These approaches describe historical patterns but offer little predictive power. Finally, some techniques implemented deep-learning algorithms for traffic and accident prediction using convolutional neural networks (CNNs) and long short-term memory networks (LSTMs), as well as combinations of the two [10-13]. These approaches generated good results, but they predicted risk indicators and traffic flow rather than traffic violations. While these three categories have improved substantially, one important issue remains underexplored: fine-grained, multi-class prediction of specific traffic violation types.

**Abbreviations**

| | |
|---|---|
| SHAP | SHapley Additive exPlanations |
| XGBOOST | eXtreme Gradient Boosting |
| CART | Classification and regression tree |
| MLP | Multi-layer perceptron |
| RMSE | Root mean squared error |
| Conv-LSTM | Convolutional long short-term memory |
| GNN | Graph neural network |

Most of the existing literature frames the problem as binary predictions (violation/no violation) or a simple sum of future offenses. However, neither forecast type provides sufficient detail for real-time enforcement. An algorithm whose output includes only whether a violation occurred, not the violation type, is generally insufficient to guide enforcement action. Moreover, few frameworks jointly analyze detailed spatial features (e.g., the specific street where the violation occurs), vehicle attributes, and temporal features while also addressing the severe class imbalance in enforcement data, where some violation types occur more frequently than others.

This is particularly prevalent in the Gulf region due to rapid urbanization and motorization, along with a heterogeneous driver population, creating a unique traffic safety problem [14-16]. This study is based on Qatar, which is representative of these conditions. The country has a population of about 2.8 million and more than 1.7 million registered vehicles, resulting in one of the highest motorization rates in the world. Although their roads are modern and well maintained, road traffic accidents remain a major problem, with speeding and other violations blamed for some accidents [17]. Recent studies in neighboring Abu Dhabi have also shown that culture and license origin may be key factors in behavior [18], so the models need to be locally calibrated for effective enforcement. To the best of the authors' knowledge, however, no previous work has proposed a predictive model for multi-class violation-type prediction based on large-scale enforcement data from Qatar or other Gulf countries, with realistic temporal evaluation protocols.

For this reason, this study introduces a machine-learning framework that can classify specific types of traffic violations. The task is conditional classification rather than occurrence forecasting: given a recorded violation at a specific street, time, and LP type, the model predicts its violation type. The proposed work has three primary contributions:

- A systematic feature engineering and preprocessing pipeline. We develop a frequency-based class-filtering strategy that reduces the original 97 violation classes to nine sufficiently represented types while retaining 99.6% of the dataset records at the class-filtering stage. We process only five raw features and build 18 spatial–temporal features, such as a zone violation entropy measure, which represents the diversity of violations per enforcement zone. Dominant-class frequency encoding is applied to location features so that high-cardinality categorical features are converted into continuous predictors, without causing dimensionality explosion. The final model utilizes 15 engineered and three retained raw features; street number and zone number are not used directly, but are converted to encoded spatial features.
- A comparison of five classifiers under a realistic temporal setup. Five classifiers – logistic regression, random forest, XGBoost, LightGBM, and MLP – are evaluated using a strict temporal train–validation–test data split. Data leakage is avoided by splitting the data chronologically. Balanced class weighting is used to counteract class imbalance, while maintaining the integrity of the temporal split and not resampling. To provide a more balanced assessment of performance across the nine violation classes, we use five complementary evaluation metrics (macro F1, balanced accuracy, weighted F1, ROC-AUC, and Cohen's kappa).
- Analysis of results with interpretation. To explain the results, we use SHAP, which quantifies each feature's contribution to model predictions and provides feature-level explanations. This analysis can identify the streets and zones most predictive of violation types, the effect of zone-level violation entropy on predictions, and the relative importance of spatial and temporal features.

We test our framework on large-scale, one-year enforcement information from Qatar, consisting of around 912,000 records after post-processing. The results show that random forest performs best, with a macro F1 of 0.6184 and balanced accuracy of 0.7064, compared with a naive majority-class baseline of 0.101 and 0.111, respectively. Most importantly, SHAP analysis shows a strong, consistent influence of spatial features: location-derived features account for ~74% of the importance of the model's predictions, and temporal features account for ~25%. Thus, we conclude that violation type is more strongly associated with location than with time. This finding has implications for enforcement strategy: interventions that target specific geographic areas with the right resources on the streets and within zones with a high probability of traffic violation types are likely to be more effective than broad, time-based resource reallocation.

On a practical level, the framework also shows how traffic enforcement details can predict useful violation types without needing road network topology, sensor-based information, weather information, or driver information. For enforcement agencies in resource-limited environments, this access can be highly beneficial, as they can use the records for predictive analytics. The remainder of this paper is structured as follows. Section 2 presents the relevant work of other authors. Section 3 presents the dataset and methods, including preprocessing, feature engineering, and handling class imbalance. Section 4 presents the study results. Section 5 concludes.

## 2. Related work

Road safety and violation prediction studies have developed from understanding individual driver behavior, to understanding complex interactions across the system, and most recently to interpretable, operationally actionable frameworks. This section critically reviews the literature across thematic areas; gaps persist, motivating this study.

### 2.1 Machine learning and spatio-temporal analysis for traffic violations

Machine learning in traffic violations has largely focused on binary classification problems with relatively small datasets. Alomari et al. [19] compared the performance of random forest, CART, and MLP for predicting speeding violations in the U.S. and found that random forest performed better than the other models in capturing non-linear interactions among driver, road, and environmental factors. However, this work addresses only one violation type within a binary-outcome framework; it does not address multi-class prediction of various violation types. Wan et al. [20] used random forest with SHAP analysis to explain and detect the violations committed by taxi drivers in China. They identified key spatial and operational factors that determine violation

severity, an important step toward interpretability, but they still treat violation detection as a binary classification or severity estimation problem rather than multi-class type prediction.

In parallel, spatio-temporal patterns of violations and accidents have been characterized. Li et al. [7] applied automated enforcement data to the estimation of violations at the intersection level; they found clear temporal patterns but did not apply them to prediction. Cheng et al. [8] proposed a methodology to assess spatio-temporal features at intersections and showed that certain spatial patterns generally correlate with certain types of violations. Wang and Liang [9] mined Beijing enforcement data (147,876 violations, 3,264 accidents), establishing 18 strong association rules across five accident types and four violation types. There were obvious time periods in the morning and evening with a high correlation between violations and accidents. This work validates the use of the spatiotemporal domain, but it does not deal with predictive modeling. Consistent with the fact that place is a structural determinant of traffic safety outcomes, Hazaymeh et al. [21] offered GIS-based hotspot analysis for traffic accidents in Jordan. Together, these studies make a strong case for spatial dominance in violation patterns but do not yield deployable multi-class classifiers.

### 2.2 Graph-based deep learning and advanced architectures

One major leap forward has come in the form of graph neural network architectures which make full use of the topology of road networks. Zhou et al. [22] introduced GATR, a graph attention network for building a topological graph with each node representing a road unit and each edge representing the connectivity between them. Using historical violation features, environmental features, and urban functional features (land-use type), GATR obtained an RMSE value of 1.7078 on Chinese city data (1,827 days), which was better than Conv-LSTM. The authors further showed through GNN explainer analysis that the main factors influencing violation distribution are spatial proximity and urban functional properties. Building on this, Wang et al. [23] proposed LTVPGA, which distills GATR's knowledge into an MLP. It uses only 15% of GATR's memory and 0.6% of its training time, while achieving almost the same accuracy. Although graph-based methods have worked well, they require detailed road-network topology, which operational enforcement agencies often do not provide, and they produce only regression-type output (number of violations per road segment). This does not meet the operational requirement we are addressing: classifying the type of violation occurring on a specific road segment.

### 2.3 Traffic safety studies in Qatar and the Gulf region

The Gulf region has specific features: high motorization, high rates of urbanization, diverse driving population, and advanced road infrastructure, which call for locally adapted models. Al-Thani [24] introduced Traffic Transformer, a deep learning model that forecasted traffic accidents in 98 zones in Qatar from 2017 to 2023. Another recent study [25] used interpretable machine learning to predict accident severity in Qatar and concluded that timing issues play the most significant role in accident severity outcomes. These are the closest geographical contexts to this work, but they focus on accident outcomes, not record-level prediction of multi-class violation type. According to Tarlochan et al. [14], the main risk factors for young drivers in Qatar are age, driving experience, and road characteristics. Similarly, Akin et al. [16] used supervised machine learning on Saudi Arabia's Highway 15 and found that road segment spatial characteristics did interact with driver age to increase the probability of crashes. This aligns with the SHAP results, which demonstrated spatial dominance at street level. Maghelal et al. [18] examined traffic violation determinants in Abu Dhabi and concluded that license origin and cultural background play a significant role in traffic violations; however, the small sample size limits generalisability. According to Almurayh et al. [17], drivers' perceptions of enforcement systems showed that visibility influenced violations, implying that predictive models can support selective deployment of enforcement systems rather than blanket area deployments.

### 2.4 Explainable AI and synthesis of research gaps

In safety-critical transportation use cases, explainable AI techniques are used to enhance model transparency. SHAP-based feature attribution, pioneered by Lundberg and Lee [26] and later adapted to tree-based models by Lundberg et al. [27], is widely used to identify influential road segments and factors that contribute to traffic prediction. Majidi et al. [28] used SHAP with random forest, AdaBoost, and XGBoost on Kentucky crash data to identify modifiable crash factors. The authors found that post-hoc explainability significantly improves actionability. To better understand traffic violations, Yang et al. [29] used deep learning with geographic context factors, which again highlights the importance of location-based features. Despite these advances, interpretable methods are rarely used for the specific task of multi-class violation-type prediction: most applications target binary outcomes or continuous risk scores.

A few key gaps were identified in the literature. First, most studies turn the prediction problem into binary classification or focus on a single violation type or regression; multi-class, fine-grained prediction of diverse violation types remains underexplored. Second, graph-based methods that model the spatial topology of the road network are frequently effective, but typically require road network topology data, which many operational agencies lack. Tabular methods, on the other hand, which rely solely on simple enforcement records, are still not widely used for this purpose. Third, although traffic safety research in the Gulf region is increasing, no previous study has developed a predictive framework specifically for multi-class violation-type prediction using large-scale enforcement data collected in Qatar under realistic temporal evaluation protocols. Fourth, explainable AI has been identified as a key factor in enforcement adoption, but it has not been explored to draw violation-type-specific actionable insights. This study aims to fill these gaps by proposing a complete framework (using only basic enforcement data, a new zone violation entropy feature, a strict temporal split for benchmarking, and actionable interpretability via SHAP analysis) to predict the type of traffic violation across nine predefined violation categories.

## 3. Materials and methods

### 3.1 Dataset description

The data employed in this study consists of traffic violations in Qatar during 2019. The Qatar Ministry of Interior provided the data. Aggregated traffic statistics are also published on the Qatar Open Data Portal, which offers an official channel for requesting datasets from government entities. The raw dataset had about 1.2 million records, while the pre-processed dataset contained 912,359 records. There are five raw attributes in each record: Day (1 to 365 for the day of the year), Time (a normalized value in [0,1] for the time of day as a fraction of 24 hours), LP Type (personal or non-

personal (government/company)), Street No. (identifier for street segment where the violation was recorded), and Zone No. (identifier for the enforcement zone where the violation was recorded). Violation is the target variable that contains the type of traffic violation. The dataset had two main problems. First, the original target variable has 97 violation type codes, most of which occur very rarely; thus, they are not suitable for training a reliable classifier. Second, the distribution of classes is highly skewed (a single class type – code 45- comprises around 83% of the total number of records in the filtered dataset), which means there is a strong majority class bias, and that the standard accuracy measures are not appropriate.

### 3.2 Preprocessing pipeline

**Data cleaning:** Some records were missing the zone or street number, so these records were removed to ensure effective training of machine learning models.

**Class filtering using frequency threshold:** A frequency-based filtering strategy was employed: any violation type that appeared less than five times per day (on average, across the country) was filtered out. This reduced the number of classes from 97 to 9, retaining 99.6% of the records available at the class-filtering stage.

**Temporal train–validation–test split**: A strict chronological split is applied to preserve the temporal integrity of the data and prevent data leakage. The data is split into: Days 1–270 as the training set (approx. 74% of days); Days 271–310 as the validation set (approx. 11% of days); and Days 311–365 as the held-out test set (approx. 15% of days). This time separation ensures that all transformations are trained on the training set and then tested on the validation and test sets, mimicking the deployment scenario in which the model is trained on historical data and validated on future data. Class filtering (97 violation types reduced to nine) was performed at the dataset preparation stage to define the modeling label space.

**Categorical encoding:** The training set contained 544 unique streets and 84 unique zones. The validation set contained 339 streets and 79 zones, and the test set contained 376 streets and 82 zones, all of which were present in the training period. We applied two encoding strategies to spatial categorical variables. For Zone No. and Street No., dominant-class frequency encoding was computed on the training set only by mapping each category to the relative frequency of its most common violation class within that zone or street. No additive smoothing or grouping of rare streets was applied. This produced a continuous feature that captures each location's violation profile without introducing high-cardinality, one-hot vectors. The global dominant frequency of the training set (0.8228) was used as the fallback value for unseen categories in the validation and test sets. No unseen streets or zones occurred in either set. Therefore, the fallback rate was 0% for both street and zone encodings in the validation and test sets. For the remaining categorical variables, standard label encoding was applied using a training-fitted LabelEncoder, with unseen categories mapped to an arbitrary but consistent training category in order to prevent encoding failures.

### 3.3 Feature engineering

The raw dataset contains only five attributes (day, time, LP type, street no., zone no.); therefore, feature engineering is essential in exposing the temporal and spatial structure latent in these variables. The final model uses 18 features (15 engineered and three raw) organized into three groups.

**Temporal features:** Since 2019 is not a leap year, Day ranges from 1 to 365, with Day 1 corresponding to 1 January 2019. From the Day variable, we derive: *month*, *day_of_week*, *is_weekend* (a binary indicator for Friday and Saturday), *week_of_year* (computed as ((Day-1)/7+1)), and season (winter, spring, summer, or fall, mapped from month). From the Time variable, we derive: *time_bin* (a four-category discretization into Night, Morning, Afternoon, and Evening using fixed boundaries at 0.25, 0.50, and 0.75) and hour (the integer hour of the day, computed as *Time*×24).

**Spatial features:** Five spatial encoding features are constructed: *zone_target_enc* and *street_target_enc* represent the dominant violation frequency of each zone and street respectively, computed via dominant-class frequency encoding as described in Section 3.2. We introduced the *zone_violation_entropy* feature to quantify the diversity of violation types within each zone. It was computed using the training set only and then applied unchanged to the validation and test sets (Equation 1):

$$H(zone) = -\sum_c \left(p_c \log_2(p_c + \varepsilon)\right) \quad (1)$$

where $p_c$ is the empirical probability of violation class $c$ within the zone and $\varepsilon = 10^{-10}$ is a small constant added for stability.

The entropy values were derived from Days 1–270, with no validation or test labels used in the computation. For a zone not observed during training, the feature was assigned a value of 0. However, this fallback was not used because all zones in the validation and test sets were already present in the training period. A high entropy value indicates a zone where differing violation types occur with comparable frequency, while a low entropy value indicates a zone dominated by a single violation type. This feature encodes a structural property of enforcement zones that is not captured by any individual record attribute. Additionally, label-encoded versions of zone no. and street no. (*zone_label_enc and street_label_enc*) are retained as complementary spatial identifiers, allowing tree-based models to distinguish between individual zones and streets directly through split conditions, independently of their violation frequency profiles captured by dominant-class frequency encoding.

**Interaction features:** Three interaction terms are constructed to capture cross-dimensional dependencies: *lp_x_zone* (*LP Type×zone_target_enc*), encoding the interaction between vehicle category and zone violation profile; *time_x_zone* (*time_bin_enc×zone_target_enc*), capturing the interaction of the temporal pattern of violations with the spatial zone profile; and *weekend_x_zone* (*is_weekend×zone_label_enc*), encoding the differential violation behavior on weekends across zones.

### 3.4 Class Imbalance handling

We had an extreme class imbalance (83% majority class). We addressed this challenge through balanced class weighting. Each model is trained with class weights inversely proportional to class frequency in the training set (Equation 2):

$$w_c = N/(K \cdot N_c) \quad (2)$$

where $N$ is the total number of training samples, $K$ is the number of classes (nine), and $N_c$ is the number of samples belonging to the class *c*. This way, a model's misclassification of minority classes is penalised during training, but the data distribution remains unchanged, which also means that statistical properties of the temporal split are not changed;

therefore, an optimistic bias caused by resampling across the train-validation boundary is avoided. Figure 1 shows the distribution of violation types after class filtering, and despite the filtering, the imbalance remains. The most common traffic offenses are provided in Table 1. The full description is provided in Appendix (Table A1).

### 3.5 Machine learning models

Five models are evaluated, spanning classical linear, ensemble tree, and neural network paradigms. Hyperparameter settings were fixed a priori, and no systematic hyperparameter search was performed. All models are trained on the 18-feature training set described above and evaluated on the held-out test set using identical metrics. Hyperparameter configurations are summarised in Table 2.

### 3.6 Evaluation metrics

Five complementary metrics are reported, selected to globally describe performance in the case of severe class imbalance and multi-class conditions:

**Macro F1 (primary metric):** an unweighted averaging of the F1 scores of each class; thus, all nine classes are treated equally regardless of their support.

**Balanced Accuracy:** arithmetic mean of per-class recall values, which is the same as normalized accuracy (each class is equally weighted).

**Weighted F1:** the mean of the per-class F1 scores weighted by the number of records in each class; indicates the overall predictive usefulness for all records.

**ROC-AUC (macro, one-vs-rest):** Average of ROC-AUC of all classes with each class compared against other classes.

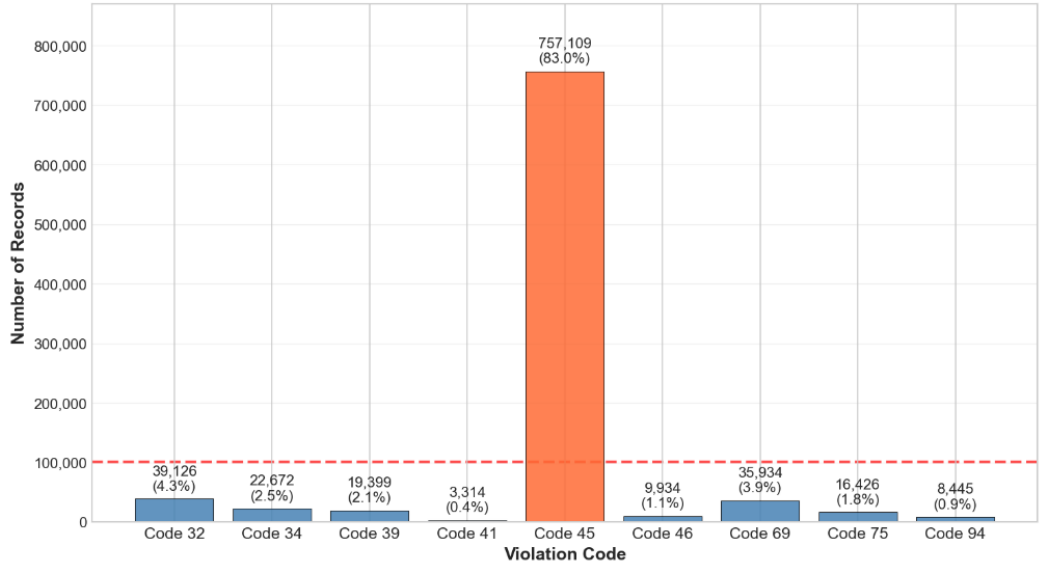


**Figure 1.** Distribution of traffic violation types after class filtering

**Table 1.** Violation-code lookup and class distribution across data splits

| Violation Short Name | Code | Train | Validation | Test | Total | % |
|---|---|---|---|---|---|---|
| Speeding | 45 | 453,868 | 127,022 | 176,219 | 757,109 | 82.98 |
| Crosswalk/sidewalk parking | 32 | 24,322 | 6,815 | 7,989 | 39,126 | 4.29 |
| Road sign violation | 69 | 22,639 | 5,402 | 7,893 | 35,934 | 3.94 |
| Prohibited parking | 34 | 13,943 | 3,694 | 5,035 | 22,672 | 2.48 |
| Undesignated parking | 39 | 12,759 | 3,285 | 3,355 | 19,399 | 2.13 |
| Traffic signal violation | 75 | 9,914 | 2,488 | 4,024 | 16,426 | 1.80 |
| Intersection obstruction | 46 | 6,700 | 1,666 | 1,568 | 9,934 | 1.09 |
| Unsafe following/overtaking | 94 | 5,416 | 1,451 | 1,578 | 8,445 | 0.93 |
| Vehicle left unsafely | 41 | 2,040 | 557 | 717 | 3,314 | 0.36 |

**Table 2.** Model configurations and hyperparameters

| Model | Key Hyperparameters | Imbalance Strategy |
|---|---|---|
| Logistic Regression | C=1.0; solver=lbfgs; max_iter=1000; multinomial | class_weight='balanced' |
| Random Forest | n_estimators=300; max_depth=20; min_samples_leaf=5 | class_weight='balanced_subsample' |
| XGBoost | n_estimators=500; lr=0.05; max_depth=8; subsample=0.8; colsample_bytree=0.8; early_stopping=20 | sample_weight per instance |
| LightGBM | n_estimators=500; lr=0.05; num_leaves=63; min_child_samples=20; feature_fraction=0.8 | class_weight_dict (encoded labels 0-8) |
| MLP | layers=[256, 256]; ReLU; Adam_lr=0.001; early_stopping_patience=10 | sample_weight per instance |

**Cohen's Kappa:** measures agreement beyond chance; very useful when there are high class imbalances, and even poor classifiers may be able to achieve high raw accuracy.

### 3.7 SHAP interpretability analysis

We use SHAP with TreeExplainer [26, 27] to quantify the contribution of each engineered feature in the prediction of the models. SHAP gives a measure of feature importance with a solid theoretical foundation. A stratified random sample of 2,000 test records was used to provide proportional representation of the nine classes. Due to the size of the dataset, applying TreeExplainer on a random forest consisting of 300 trees was computationally expensive, so LightGBM was chosen as a computationally tractable tree-based reference model for SHAP analysis. LightGBM also exhibited comparable spatial feature dominance. The mean absolute SHAP values were then averaged over classes and samples to obtain a global feature importance ranking.

### 3.8 Framework overview

As shown in Figure 2, the end-to-end pipeline consists of class filtering, basic feature derivation, train/validation/test chronological splitting, training-based categorical encoding, and spatial feature engineering of the raw enforcement records. The classifiers are then trained and evaluated using multi-class imbalance-aware metrics, and SHAP analysis is performed using LightGBM as the reference tree-based model.

## 4. Results

### 4.1 Overall model performance

The performance of the five models on the held-out test set (Days 311–365, 208378 records) is shown in Table 3. The outcome reveals that the tree-based ensemble models achieve higher performance results in comparison to linear baseline and NN models for all five metrics. Uncertainty was assessed using paired bootstrap analysis. For random forest versus XGBoost, 1,000 record-level resamples gave a macro F1 difference of +0.0203 (95% CI: [+0.0165, +0.0241]), while 1,000 day-blocked resamples over the 55 test days gave +0.0204 (95% CI: [+0.0129, +0.0275]), with random forest outperforming XGBoost in all resamples. The day-blocked 95% CIs for random forest were [0.6070, 0.6293] for macro F1, and [0.6937, 0.7201] for balanced accuracy. Full results are provided in Supplementary Appendix (Table A2 and Table A3). Sensitivity to the test window was limited. With training fixed at Days 1–270, random forest achieved macro F1 values of 0.6184 on Days 311–365 (208,378 records), 0.6149 on Days 329–365 (137,047 records), and 0.6276 on Days 342–365 (90,018 records) – a total range of 0.013. Sensitivity to the class-frequency threshold was also examined. Increasing the cutoff to an average of 10 records per day removed only code 41, leaving eight classes and 99.64% of the nine-class dataset. Macro F1 increased from 0.6184 to 0.6476 and balanced accuracy from 0.7064 to 0.7395, while Cohen's κ changed only from 0.5981 to 0.6035. This increase reflects the easier task after removing the rarest class rather than an improvement in the modeling procedure.

This is because the majority class (code 45 violation) is about 83% of the data, so a naive baseline of predicting class 45 for every record would get a weighted F1 of ~0.753. But its macro F1 would only be approximately 0.101, and its balanced accuracy 0.111. Random forest outperforms logistic regression with a macro F1 score of 0.6184 vs. 0.1536 and balanced accuracy of 0.7064 vs. 0.2818. It also shows clear improvements over the trivial majority-class baseline, suggesting it is not just a majority-class classifier. Figure 3 compares the models on the two main metrics and shows that the tree-based ensembles (random forest, XGBoost, and LightGBM) outperform the linear and neural baselines.

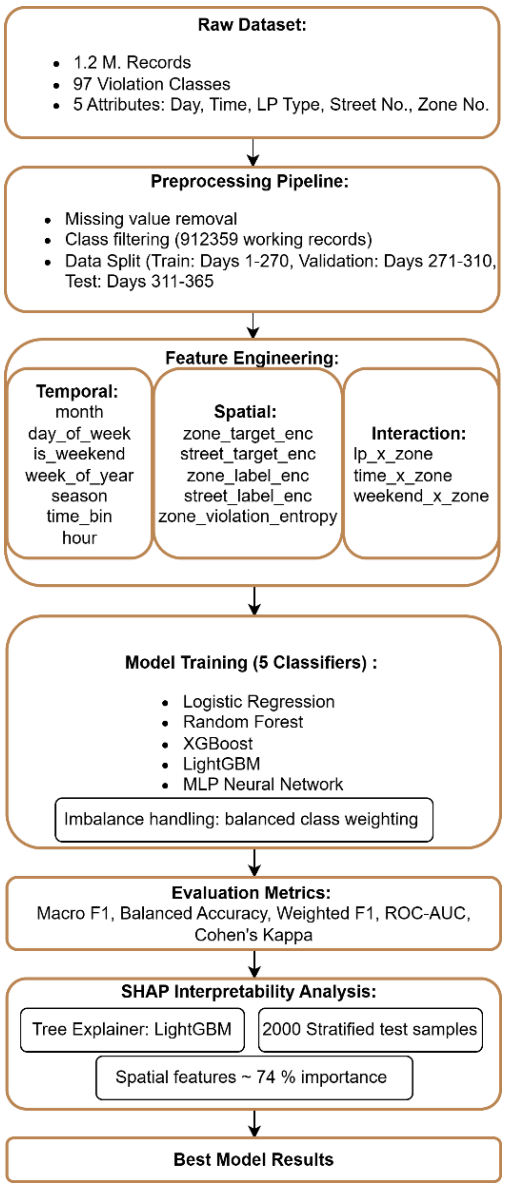


**Figure 2.** Framework architecture

**Table 3.** Model performance on the held-out test set

| Model | Macro F1 | Balanced Acc | Weighted F1 | ROC-AUC | Cohen's Kappa |
|---|---|---|---|---|---|
| Logistic Regression | 0.1536 | 0.2818 | 0.7462 | 0.8481 | 0.2660 |
| MLP Neural Network | 0.2171 | 0.3537 | 0.6395 | 0.8400 | 0.1678 |
| LightGBM | 0.5899 | 0.6907 | 0.8563 | 0.9717 | 0.5318 |
| XGBoost | 0.5981 | 0.6896 | 0.8587 | 0.9714 | 0.5374 |
| Random Forest | 0.6184 | 0.7064 | 0.8813 | 0.9747 | 0.5981 |

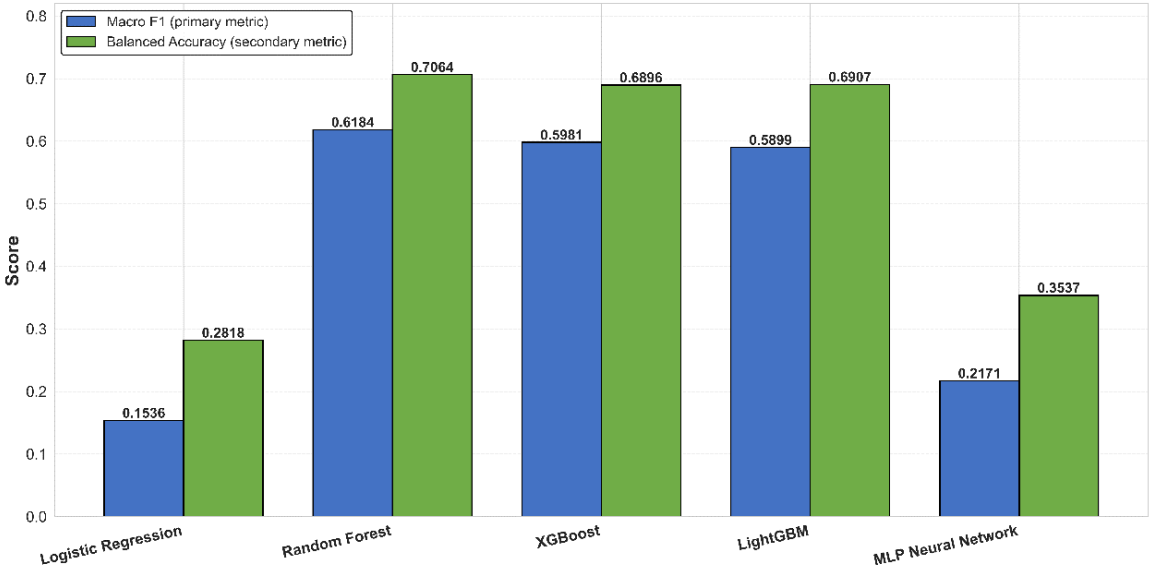


**Figure 3.** Model comparison (macro f1 and balanced accuracy)

To assess computational efficiency, Table 4 compares the training time, inference time, and serialized model size of the five evaluated models. Random forest achieved the best predictive performance but had the largest model size; XGBoost retained 96.7% of its macro F1 at only 5.3% of the serialized model size, making it a more compact option for resource-constrained deployment.

**Table 4.** Computational cost and serialized model size of the evaluated models

| Model | Training Time (s) | Inference Time (s) | Size (MB) |
|---|---|---|---|
| Logistic regression | 15.4 | 0.02 | <0.1 |
| Random forest | 86.6 | 1.74 | 838.8 |
| XGBoost | 191.9 | 4.84 | 44.3 |
| LightGBM | 82.6 | 8.97 | 20.7 |
| MLP neural network | 301.4 | 0.66 | 1.7 |

### 4.2 Ensemble models vs linear and neural baselines

The performance gap between ensemble methods and other models is consistent and substantial in all the measures. The three tree-based ensemble classifiers, namely random forest, XGBoost, and LightGBM, consistently outperform the other two classifiers, logistic regression and MLP, across all the reported evaluation metrics, achieving higher macro F1 scores. This aligns with results in the traffic safety literature, where random forest has consistently outperformed linear classifiers and single decision trees on multi-factor violation and crash datasets [19, 30]. As expected, spatial features have a high cardinality; they are non-linear and are not handled well by logistic regression. Despite the fact that using dominant-class frequency encoding flattens the zone/street identifiers into continuous values, the combination of location, time, and violation type is inherently non-linear. So without explicitly expanding the polynomial features, L2-regularised logistic regression cannot capture these interactions. Although the MLP neural network can theoretically model non-linear relationships, it did not perform as well as the ensemble models. This may be because of the data's tabular nature.

Previous research shows that gradient-boosted tree (GBT) models tend to outperform neural networks with tabular data when the number of features that provide information is relatively low compared to the number of records, and when there is a clear pattern of feature interactions [31]. In this paper, the key spatial features indicate a distinct cluster structure that can be efficiently learned by the tree models, but which might be more challenging for the shallow neural network. The MLP training loss and validation accuracy curves indicate that the MLP has converged but generalizes worse than the ensemble models.

### 4.3 Random forest as the best-performing model

Random Forest achieved the best overall performance, including a balanced accuracy of 0.7064 (versus 0.111 for the majority-class baseline) and a ROC-AUC of 0.9747. The model maintained its macro F1 advantage over XGBoost under both record-level and day-blocked bootstrap resampling. It also retained its superior ranking in the unified-weighting sensitivity analysis, indicating that differences in weighting implementation did not materially affect the comparison (Appendix (Table A4)).

### 4.4 Per-class performance analysis

Figure 4 presents the per-class F1-scores for all five models across the nine violation types. The large variations in per-class F1 scores stemmed from heterogeneous class sizes, and classifying similar spatial and temporal contexts was challenging. Table 5 presents the per-class precision, recall, F1-score, and support for the best-performing random forest model. Classes 75 (0.258) and 69 (0.346) have very low precision, whereas their recall values were relatively high (0.609 and 0.614), contributing to their lower macro F1. Violation Code 45, the majority class, achieves the highest F1-score, as expected given its large representation in the training data. For the minority classes, violations with stronger relationships to specific streets or zones generally have higher F1 scores, showing that spatial features helped to distinguish classes.

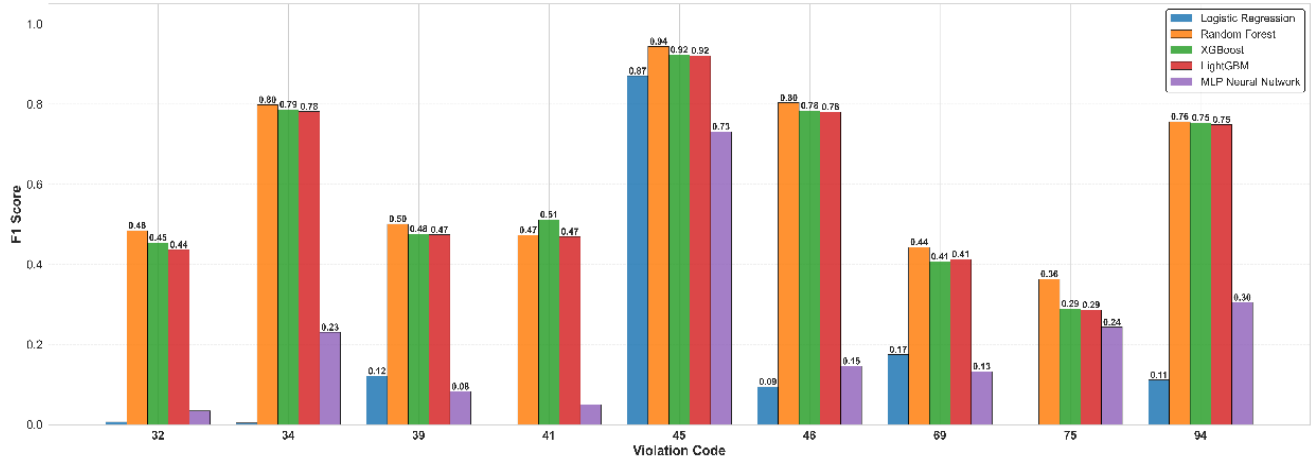


**Figure 4.** Per-class F1 scores for all models

**Table 5.** Per-class performance of Random Forest on the held-out test set

| Code | Precision | Recall | F1 | Support |
|---|---|---|---|---|
| 45 | 0.9847 | 0.9063 | 0.9439 | 176,219 |
| 32 | 0.5140 | 0.4580 | 0.4844 | 7,989 |
| 69 | 0.3461 | 0.6142 | 0.4427 | 7,893 |
| 34 | 0.7229 | 0.8916 | 0.7984 | 5,035 |
| 39 | 0.4346 | 0.5914 | 0.5010 | 3,355 |
| 75 | 0.2583 | 0.6086 | 0.3627 | 4,024 |
| 46 | 0.7320 | 0.8903 | 0.8035 | 1,568 |
| 94 | 0.6919 | 0.8340 | 0.7563 | 1,578 |
| 41 | 0.4068 | 0.5635 | 0.4725 | 717 |

In contrast, violation codes that are not strongly associated with specific locations along the road network have low F1 scores because there are fewer patterns to leverage from location-specific features. This per-class pattern also reinforced the importance of spatial features in classification, as described in Section 4. Figure 5 shows the normalized confusion matrix of the optimal model (random forest). In the case of the mean of the confusion direction pairs, the maximal confusion was found for directions 69/75 (0.150), 39/41 (0.134), 32/39 (0.113), 34/39 (0.102), and 45/69 (0.095). Zone profile similarity was also correlated with the level of confusion (Spearman $\rho$ = 0.771, p < 0.001), and street profile similarity was correlated (ρ = 0.688, p < 0.001) on the basis of 10,000 label permutations. Thus, for example, the category of violations pair 32/39 demonstrated high zone and street similarity (0.837 and 0.834) and a high level of confusion (0.113). Other pairs, such as 34/39, also showed high confusion despite low spatial similarity, indicating that spatial similarity does not explain all errors. The matrix also shows that most misclassifications occur between violation types that share similar spatial contexts, further supporting the spatial dominance finding.

### 4.5 SHAP feature importance analysis

Figure 6 shows the global feature importance of the LightGBM benchmark model, ordered by importance, based on a stratified sample of 2,000 test records. To assess whether the 2,000-record SHAP sample was sufficient for stable global importance estimates, we repeated the analysis using 10 independent samples of 2,000 test records and compared the results with a 50,000-record reference.

The rankings were highly stable, with a mean pairwise Spearman correlation of 0.998 (minimum 0.995) across the 10 samples and a mean Spearman correlation of 0.999 (minimum 0.997) with the 50,000-record reference. The highest-ranked feature (street_target_enc) and the top-five feature set were identical across all replicates. The aggregate location-derived importance share was also stable at 74.21% ± 0.09, ranging from 74.10% to 74.35%. These results support the use of the 2,000-record sample for the reported SHAP analysis.

Figure 7 shows the Gini impurity-based feature importance from the best-performing random forest model. Importance shares were normalized across all 18 features $s_i = (x_i / \sum_j x_j) * 100$. The SHAP and Gini rankings showed strong agreement (Spearman $\rho$= 0.928, p = 3.0 × $10^{-8}$; Kendall $\tau$= 0.791). Four of the five highest-ranked features were common to both measures: street_target_enc, street_label_enc, zone_label_enc, and zone_violation_entropy. The largest rank changes were observed for Day (SHAP rank 4, Gini rank 8) and weekend_x_zone (SHAP rank 12, Gini rank 16). However, Gini importance is calculated on the training data and tends to give higher weights to high-cardinality features because they provide more split points, while SHAP values are calculated on hold-out test records.

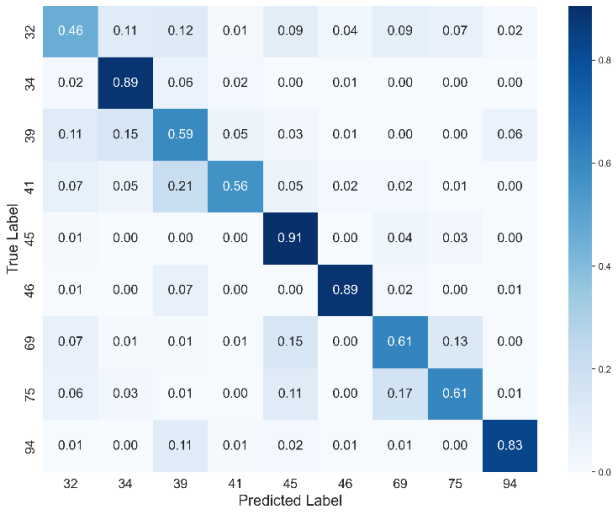


**Figure 5.** Normalized confusion matrix for the best model (RF)

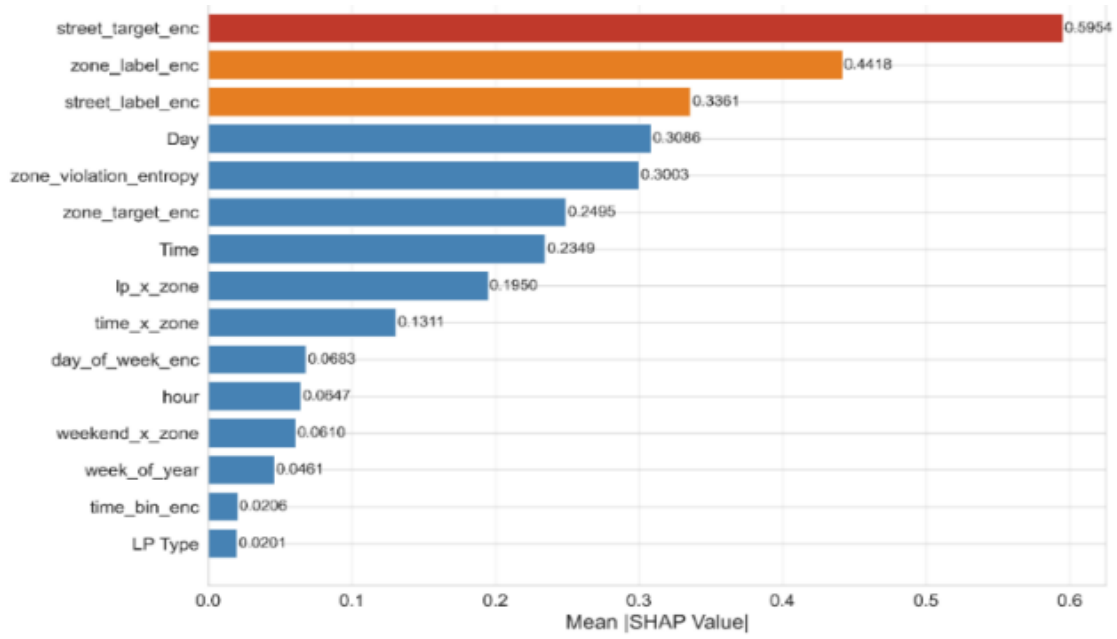


**Figure 6.** SHAP feature importance (LightGBM)

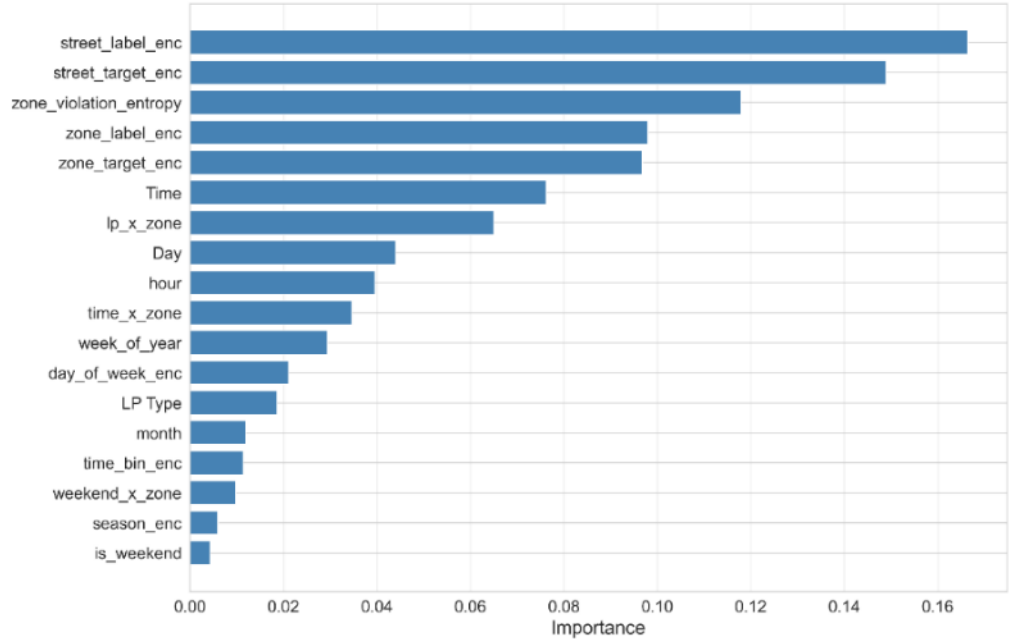


**Figure 7.** Feature importance from a random forest based on Gini impurity

In addition, the two measures come from different models (LightGBM for SHAP, Random Forest for Gini); thus, any similarity between them would indicate convergence, not confirmation. The Appendix (Table A5 and Table A6) provides the side-by-side ranking. These results show that spatial features are more important than temporal features; this has implications for enforcement strategies. Street-level encoding (street_target_enc) is the most important feature, with a mean absolute SHAP value of 0.5954 and the largest share of prediction importance among all eighteen features.

Zone-level features (zone_target_enc, zone_label_enc, zone_violation_entropy) collectively contribute the second-largest block of importance. Spatial encodings account for 61.8% of SHAP importance, while including location-derived interaction terms increases the total location-derived share to 74.2% (Figure 6). Temporal features account for 25.2% of the total importance. The corresponding Gini shares are similar (73.8% location-derived and 24.4% temporal). Variables such as *hour*, *time_bin*, *month*, *week_of_year*, *season*, and *is_weekend* show small SHAP values across samples, suggesting that they play a limited role in distinguishing violation types. LP type provides information, but it is less significant than the foremost spatial characteristics. The deletion of month,

*week_of_year,* and *season* had almost no effect on the performance, with macro F1 changing from 0.6184 to 0.6161. When all temporal features were excluded, this score decreased more significantly to 0.5990. This indicates that temporal information still provides some predictive value, but the model relies primarily on spatial features. The results suggest an important finding: the location of traffic-rule infractions, rather than the time of the violations or the type of vehicle, is the most important factor in the type of violation. The results of RF feature ablation are provided in Table 6. Complementary feature ablations show that street encodings make the largest contribution to performance. Removing street_label_enc and street_target_enc reduced macro F1 from 0.6184 to 0.4444.

In contrast, removing zone_violation_entropy or the three interaction terms had little effect, with macro F1 values of 0.6211 and 0.6190, respectively. Using label-encoded location identifiers without the dominant-class frequency encodings reduced macro F1 to 0.5850, indicating that the frequency-based location representation adds useful information but is not the sole source of predictive performance. These results point to a key finding: the type of traffic violation depends more on where it occurs than on when it occurs or the type of vehicle that is involved. This result is similar to the association rule mining analysis of the enforcement data from Beijing [9], which revealed that there are strong linkages between certain types of violations, specific locations and spatial configurations, and the GATR study [22], which demonstrated that urban land-use features (as a proxy for spatial context) are a major factor in graph-based violation forecasting. A SHAP summary plot is shown in Figure 8 to visualize the trend of SHAP values over the samples for the most important features.

### 4.6 Discussion

Three key findings of this study are relevant to traffic enforcement and urban planning in Qatar.

First, good performance of features encoded with spatial data indicates a geographical structure of traffic violations in Qatar. Some of the streets and zones exhibit consistent relationships with the types of violations, which are influenced by the roadway characteristics, the location of the enforcement cameras, traffic volumes, and traffic behavior. However, the target represents the recorded violation type, and the dataset does not include camera type, site function, or road class. The spatial pattern is clear, but it may reflect both the location of violations and the way they are detected or recorded. Furthermore, the SHAP analysis indicates that street-level spatial features contribute more to model predictions than zone-level features, suggesting that enforcement planning may benefit from targeting street segments rather than relying on broader zone-level allocation. These findings suggest that enforcement agencies may be able to predict the most likely violation type at a given location with useful accuracy using historical enforcement records, including important street-level location information, without requiring real-time sensor data or detailed road topology information. The model can thus inform resource allocation by indicating the kinds of resources that may be most needed in a particular area (e.g. enforcement personnel and equipment), rather than when or where an individual violation will occur. Second, the modest contribution of temporal features also suggests that existing enforcement practices (e.g. increased patrolling during peak hours) may have a small effect on the type of violations committed, even if they affect the overall number of violations.

**Table 6.** Random forest feature ablation results on the held-out test set

| Variant | Features | Macro F1 | Balanced Acc. | Cohen's κ |
|---|---|---|---|---|
| Full model | 18 | 0.6184 | 0.7064 | 0.5981 |
| Without street encodings | 16 | 0.4444 | 0.5503 | 0.4650 |
| Without zone violation entropy | 17 | 0.6211 | 0.7092 | 0.5992 |
| Without interaction terms | 15 | 0.6190 | 0.7028 | 0.5995 |
| Label encoding only (entropy retained) | 14 | 0.5850 | 0.6571 | 0.5782 |

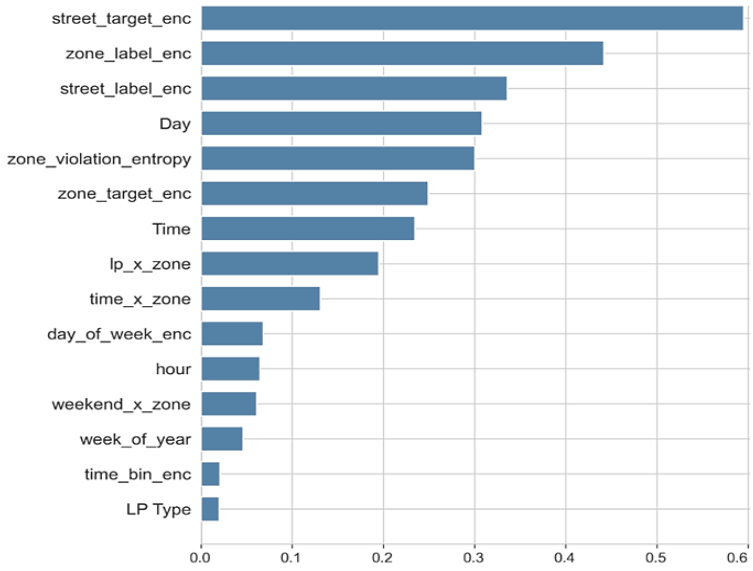


**Figure 8.** SHAP summary

Consequently, spatially targeted enforcement – namely placing type-specific resources on high-probability streets and risk zones – is likely to be more effective than a mere time-based redeployment. Interestingly, this contradicts another study [25], which deals with traffic accident severity in Qatar, and in which the time of day is the determining variable. Third, the *zone_violation_entropy* feature introduced in this study as a spatial measure ranks among the most important features in the SHAP analysis. However, removing this feature did not reduce predictive performance, suggesting that much of its information overlaps with other spatial encodings. High-entropy zones may nevertheless represent a different enforcement context from low-entropy zones dominated by a single violation type. This distinction may support general enforcement in high-entropy zones and type-targeted enforcement in low-entropy zones.

### 4.7 Comparison with related work

Comparing results with previous studies is difficult because of differences in datasets, violations addressed, and methods for evaluating violations. However, the GATR model [22] achieves an RMSE of 1.7078 when predicting the number of violations on road segments, which is a regression task, not a nine-class classification task as in our work. Likewise, Alomari et al. [19] focus on speeding violations on another national dataset, a binary classification task. Further, the transformer-based (Traffic Transformer) proposal by Al-Thani [24] aims to predict traffic accident occurrence at the zone level in Qatar, not to predict specific violation types from individual enforcement records. To date, no study (as far as we know) has analyzed multi-class prediction of specific traffic violation types using large-scale enforcement data from Qatar or the Gulf region with a chronological train–test split. Our study is based on approximately 912,000 violations during one year, providing a wide empirical foundation for analysis. The results provide a good starting point for further research into this problem.

## 5. Conclusions

This study suggested a machine learning model to classify traffic violation cases using a large-scale one-year traffic enforcement dataset from Qatar. There are about 912,000 records in the dataset in nine categories of violation. To address the high class imbalance (83% majority class) and the fine-grained multi-class prediction problem, the framework was proposed to systematically process the original 97 violation classes into nine types, keeping 99.6% of the records. The final model used 18 features, comprising 15 engineered features and three retained raw features, including a zone violation entropy measure. Five classifiers were compared under a strict temporal train–validation–test split with balanced class weighting. The best performance was obtained by random forest (macro F1 = 0.6184, balanced accuracy = 0.7064). SHAP-based interpretability showed that location-derived features were more important for predictions (~74% of prediction importance), and temporal features were ~25% important, highlighting the key takeaway that violation type is more associated with location than time. Based on this finding, spatially targeted enforcement might be beneficial, and operational enforcement data without road network topology or real-time sensor feeds could be sufficient for multi-class violation prediction. There are a number of limitations in this study. First, the data is limited to a single year (2019), which means it is unable to show trends in data across years, or impacts of policy adjustments over the long term. The temporal sensitivity analysis used nested test windows with a fixed training period rather than a full rolling-origin evaluation with multiple training cut-offs. Second, the five raw attributes do not include useful features (e.g. driver behaviour and demographics, vehicle characteristics, road geometry, weather conditions, and camera locations) that may affect the model's performance, particularly for minority classes. In addition, the authors did not perform a systematic hyperparameter search, which may have limited some models' performance. Third, the authors ran a SHAP analysis on LightGBM rather than random forest because of the computational costs. This makes the interpretation somewhat approximate. Fourth, although the model includes temporal features, it is static and therefore does not adapt to changes in violation patterns over time, which may reduce its usefulness in changing real-world settings. Fifth, because the model framework was developed and validated solely on Qatar's enforcement system, its applicability elsewhere in the Gulf or across the broader region is speculative. This model could be broadened in several ways. For instance, analyzing multi-year data can help us understand the impact of certain policies, as well as any other long-term trends. Performance may be boosted when more variables – including driver demographics, road geometry, and weather conditions – are considered; in particular, prediction for the minority class may benefit from it. Moreover, the zone_violation_entropy feature may be generalized to street-level and time-based environments. Graph-based approaches like GATR and LTVP-GA can also be used if road topology information is available. Additionally, adaptive models can monitor temporal trends in violations. Finally, it may be possible to generate heat maps that more clearly illustrate the violation hotspots and help plan location-based enforcement efforts.

### Ethical issue

The authors are aware of and comply with best practices in publication ethics, specifically regarding authorship (avoidance of guest authorship), dual submission, manipulation of figures, competing interests, and compliance with research ethics policies. The authors adhere to publication requirements that the submitted work is original and has not been published elsewhere. All survey participants provided informed consent before participating, and the anonymity and confidentiality of all respondents were strictly maintained throughout the research process.

### Data availability statement

The dataset was obtained from the Ministry of Interior of the State of Qatar following a request submitted through the Qatar Open Data Portal (https://www.data.gov.qa/pages/contact/) and was provided by email. It is not publicly available, and access is subject to the authority's approval. Reproducibility materials can be provided by the corresponding author upon reasonable request.

### Conflict of interest

The authors declare no potential conflict of interest.

**References**


[1] Gopalakrishnan, S. (2012). A public health perspective of road traffic accidents. Journal of Family Medicine and Primary Care, 1(2), 144–150. https://doi.org/10.4103/2249-4863.104987

[2] Zhang, G., Yau, K. K., Zhang, X., & Li, Y. (2016). Traffic accidents involving fatigue driving and their extent of casualties. Accident Analysis & Prevention, 87, 34–42. https://doi.org/10.1016/j.aap.2015.10.033

[3] Wu, Y. W., & Hsu, T. P. (2021). Mid-term prediction of at-fault crash driver frequency using fusion deep learning with city-level traffic violation data. Accident Analysis & Prevention, 150, 105910. https://doi.org/10.1016/j.aap.2020.105910

[4] Zahid, M., Chen, Y., Khan, S., Jamal, A., Ijaz, M., & Ahmed, T. (2020). Predicting risky and aggressive driving behavior among taxi drivers: Do spatio-temporal attributes matter? International Journal of Environmental Research and Public Health, 17(11), 3937. https://doi.org/10.3390/ijerph17113937

[5] Ding, C., Wang, D., Liu, C., Zhang, Y., & Yang, J. (2017). Exploring the influence of built environment on travel mode choice considering the mediating effects of car ownership and travel distance. Transportation Research Part A: Policy and Practice, 100, 65–80. https://doi.org/10.1016/j.tra.2017.04.008

[6] Boateng, C., Yang, K., Ghoreishi, S. G. A., Jang, J., Jan, M. T., Conniff, J., Furht, B., Moshfeghi, S., Newman, D., Tappen, R., et al. (2023). Abnormal driving detection using GPS data. In Proceedings of the 2023 IEEE 20th International Conference on Smart Communities: Improving Quality of Life Using AI, Robotics and IoT (HONET) (pp. 210–215). IEEE. https://doi.org/10.1109/honet59747.2023.10374718

[7] Li, Y., Abdel-Aty, M., Yuan, J., Cheng, Z., & Lu, J. (2020). Analyzing traffic violation behavior at urban intersections: A spatio-temporal kernel density estimation approach using automated enforcement system data. Accident Analysis & Prevention, 141, 105509. https://doi.org/10.1016/j.aap.2020.105509

[8] Cheng, Z., Zhang, L., Zhang, Y., Wang, S., & Huang, W. (2023). A systematic approach for evaluating spatiotemporal characteristics of traffic violations and crashes at road intersections: An empirical study. Transportmetrica A: Transport Science, 19(3), 2060368. https://doi.org/10.1080/23249935.2022.2060368

[9] Wang, H., & Liang, G. (2025). Association rules between urban road traffic accidents and violations considering temporal and spatial constraints: A case study of Beijing. Sustainability, 17(4), 1680. https://doi.org/10.3390/su17041680

[10] Abdullah, S. M., Periyasamy, M., Kamaludeen, N. A., Towfek, S., Marappan, R., Kidambi Raju, S., Alharbi, A. H., & Khafaga, D. S. (2023). Optimizing traffic flow in smart cities: Soft GRU-based recurrent neural networks for enhanced congestion prediction using deep learning. Sustainability, 15(7), 5949. https://doi.org/10.3390/su15075949

[11] Yin, X., Wu, G., Wei, J., Shen, Y., Qi, H., & Yin, B. (2022). Deep learning on traffic prediction: Methods, analysis, and future directions. IEEE Transactions on Intelligent Transportation Systems, 23(6), 4927–4943. https://doi.org/10.1109/TITS.2021.3054840

[12] Zheng, H., Lin, F., Feng, X., & Chen, Y. (2021). A hybrid deep learning model with attention-based conv-LSTM networks for short-term traffic flow prediction. IEEE Transactions on Intelligent Transportation Systems, 22(11), 6910–6920. https://doi.org/10.1109/TITS.2020.2997352

[13] Sayed, S. A., Abdel-Hamid, Y., & Hefny, H. A. (2025). Intelligent traffic flow prediction using deep learning techniques: A comparative study. SN Computer Science, 6(1), 60. https://doi.org/10.1007/s42979-024-03552-3

[14] Tarlochan, F., Ibrahim, M. I. M., & Gaben, B. (2022). Understanding traffic accidents among young drivers in Qatar. International Journal of Environmental Research and Public Health, 19(1), 514. https://doi.org/10.3390/ijerph19010514

[15] Shaaban, K. (2012). Comparative study of road traffic rules in Qatar compared to western countries. Procedia - Social and Behavioral Sciences, 48, 992–999. https://doi.org/10.1016/j.sbspro.2012.06.1076

[16] Akin, D., Sisiopiku, V. P., Alateah, A. H., Almonbhi, A. O., Al-Tholaia, M. M., & Al-Sodani, K. A. A. (2022). Identifying causes of traffic crashes associated with driver behavior using supervised machine learning methods: Case of Highway 15 in Saudi Arabia. Sustainability, 14(24), 16654. https://doi.org/10.3390/su142416654

[17] Almurayh, A., Bedaiwy, A., & Elsharkasy, A. (2024). Predicting the orientation of vehicle drivers towards the traffic and speed enforcement surveillance system. The Open Transportation Journal, 18. https://doi.org/10.2174/0126671212299393240603080010

[18] Maghelal, P., Li, Z., Alfarra, A., & Zhu, P. (2025). Analyzing determinants of traffic violations in a multi-cultural setting: Case of Abu Dhabi. Journal of Urban Management, 14(3), 675–689. https://doi.org/10.1016/j.jum.2025.01.006

[19] Alomari, A. H., Al-Mistarehi, B., Alnaasan, T. K., & Obeidat, M. S. (2023). Utilizing different machine learning techniques to examine speeding violations. Applied Sciences, 13(8), 5113. https://doi.org/10.3390/app13085113

[20] Wan, M., Wu, Q., Yan, L., Guo, J., Li, W., Lin, W., & Lu, S. (2023). Taxi drivers' traffic violations detection using random forest algorithm: A case study in China. Traffic Injury Prevention, 24(4), 362–370. https://doi.org/10.1080/15389588.2023.2191286

[21] Hazaymeh, K., Almagbile, A., & Alomari, A. H. (2022). Spatiotemporal analysis of traffic accidents hotspots based on geospatial techniques. ISPRS International Journal of Geo-Information, 11(4), 260. https://doi.org/10.3390/ijgi11040260

[22] Zhou, Y., Wang, Y., Zhang, F., Zhou, H., Sun, K., & Yu, Y. (2023). GATR: A road network traffic violation prediction method based on graph attention network.

International Journal of Environmental Research and Public Health, 20(4), 3432. https://doi.org/10.3390/ijerph20043432

[23] Wang, Y., Zhou, Y., & Zhang, F. (2025). LTVPGA: Distilled graph attention for lightweight traffic violation prediction. ISPRS International Journal of Geo-Information, 14(9), 332. https://doi.org/10.3390/ijgi14090332

[24] Al-Thani, M. G. (2025). Traffic accident predictive model for efficient resource allocation in Qatar: A novel transformer-based approach. Advances in Artificial Intelligence and Machine Learning, 5(2), 3975–3987. https://doi.org/10.54364/AAIML.2025.52224

[25] Alshriem, M., & Yang, Y. (2026). Prediction of Large-Scale Traffic Accident Severity in Qatar: A Binary Reformulation Approach for Extreme Class Imbalance with Interpretable AI. Future Transportation, 6(2), 88. https://doi.org/10.3390/futuretransp6020088

[26] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30, 4765–4774. https://doi.org/10.48550/arXiv.1705.07874

[27] Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S. I. (2020). From local explanations to global understanding with explainable AI for trees. Nature Machine Intelligence, 2(1), 56–67. https://doi.org/10.1038/s42256-019-0138-9

[28] Majidi, M. Z., Wang, T., & Souleyrette, R. (2025). From prediction to prevention: Identifying actionable crash factors through ML and narrative-based sensitivity testing. Future Transportation, 5(4), 190. https://doi.org/10.3390/futuretransp5040190

[29] Yang, H., Yao, X. A., Roozkhosh, F., Liu, R., & Mai, G. (2025). From theory to deep learning: Understanding the impact of geographic context factors on traffic violations. Computers, Environment and Urban Systems, 119, 102268. https://doi.org/10.1016/j.compenvurbsys.2025.102268

[30] Sneha, R., & Ebenezer, P. (2022). Prediction of traffic violation using machine learning. International Journal of Creative Research Thoughts, 10(12), b706–b711. Available at: http://www.ijcrt.org/papers/IJCRT2212193.pdf

[31] Grinsztajn, L., Oyallon, E., & Varoquaux, G. (2022). Why do tree-based models still outperform deep learning on typical tabular data? Advances in Neural Information Processing Systems, 35, 507–520. https://doi.org/10.52202/068431-0037



## Appendix

**Table A1.** Traffic violation codes and descriptions

| Violation Code | Description |
|---|---|
| 45 | Exceeding the maximum speed limit on the road. |
| 32 | Parking or waiting in places designated for pedestrian crossing or on sidewalks. |
| 69 | Failure to follow the instructions indicated by traffic signs or markings placed on the road by the licensing authority. |
| 34 | Parking or stopping in areas where parking and stopping are not permitted. |
| 39 | Parking or stopping outside designated parking areas. |
| 75 | Failure to comply with traffic signals. |
| 46 | The vehicle driver causes confusion or disrupts traffic at intersections with traffic signals. |
| 94 | Failure to maintain a safe distance from the vehicle ahead, comply with its signals, and overtake only from the left. |
| 41 | Leaving a vehicle on the road without taking the necessary precautions to prevent accidents, traffic obstruction, or unauthorized use. |

**Table A2.** Day-blocked bootstrap results for the three best-performing models

| Model | Macro F1 | 95% CI | Balanced Accuracy | 95% CI |
|---|---|---|---|---|
| Random Forest | 0.6184 | [0.6070, 0.6293] | 0.7064 | [0.6937, 0.7201] |
| XGBoost | 0.5981 | [0.5866, 0.6094] | 0.6896 | [0.6761, 0.7027] |
| LightGBM | 0.5899 | [0.5778, 0.6010] | 0.6907 | [0.6769, 0.7046] |

**Table A3.** Paired day-blocked bootstrap comparison of the three best-performing models

| Comparison | Bootstrap mean Δ Macro F1 | 95% CI | Bootstrap mean Δ Balanced Accuracy | 95% CI |
|---|---|---|---|---|
| RF – XGBoost | 0.0204 | [+0.0129, +0.0275] | 0.017 | [+0.0090, +0.0249] |
| RF – LightGBM | 0.0284 | [+0.0223, +0.0345] | 0.0157 | [+0.0096, +0.0223] |
| XGBoost – LightGBM | 0.008 | [+0.0030, +0.0131] | −0.0012 | [−0.0064, +0.0040] |

**Table A4.** Sensitivity analysis using a unified class-weighting scheme

| Model | Original scheme | Unified scheme | Δ Macro F1 |
|---|---|---|---|
| Logistic Regression | 0.1536 | 0.1536 | 0 |
| Random Forest | 0.6184 | 0.6208 | 0.0024 |
| XGBoost | 0.5981 | 0.5981 | 0 |
| LightGBM | 0.5899 | 0.5899 | 0 |
| MLP | 0.2171 | 0.2171 | 0 |

**Table A5.** Comparison of feature-group importance using LightGBM SHAP and Random Forest Gini

| Group | Features | LightGBM SHAP | RF Gini |
|---|---|---|---|
| Spatial encodings | 5 | 61.77% | 62.81% |
| Location-derived interactions | 3 | 12.43% | 10.94% |
| Location-derived total | 8 | 74.20% | 73.75% |
| Temporal | 9 | 25.15% | 24.39% |
| Vehicle (LP Type) | 1 | 0.65% | 1.86% |
| Total | 18 | 100.00% | 100.00% |

**Table A6.** Comparison of feature importance from LightGBM SHAP and Random Forest Gini (top 15)

| SHAP rank | Feature | Group | Mean \|SHAP\| | SHAP share % | Gini | Gini share % | Gini rank |
|---|---|---|---|---|---|---|---|
| 1 | street_target_enc | Spatial | 0.5954 | 19.12 | 0.1489 | 14.89 | 2 |
| 2 | zone_label_enc | Spatial | 0.4418 | 14.19 | 0.098 | 9.8 | 4 |
| 3 | street_label_enc | Spatial | 0.3361 | 10.79 | 0.1664 | 16.64 | 1 |
| 4 | Day | Temporal | 0.3086 | 9.91 | 0.044 | 4.4 | 8 |
| 5 | zone_violation_entropy | Spatial | 0.3003 | 9.65 | 0.1179 | 11.79 | 3 |
| 6 | zone_target_enc | Spatial | 0.2495 | 8.01 | 0.0968 | 9.68 | 5 |
| 7 | Time | Temporal | 0.2349 | 7.54 | 0.0762 | 7.62 | 6 |
| 8 | lp_x_zone | Spatial interaction | 0.195 | 6.26 | 0.065 | 6.5 | 7 |
| 9 | time_x_zone | Spatial interaction | 0.1311 | 4.21 | 0.0346 | 3.46 | 10 |
| 10 | day_of_week_enc | Temporal | 0.0683 | 2.2 | 0.0211 | 2.11 | 12 |
| 11 | hour | Temporal | 0.0647 | 2.08 | 0.0395 | 3.95 | 9 |
| 12 | weekend_x_zone | Spatial interaction | 0.061 | 1.96 | 0.0098 | 0.98 | 16 |
| 13 | week_of_year | Temporal | 0.0461 | 1.48 | 0.0294 | 2.94 | 11 |
| 14 | time_bin_enc | Temporal | 0.0206 | 0.66 | 0.0114 | 1.14 | 15 |
| 15 | LP Type | Vehicle | 0.0201 | 0.65 | 0.0186 | 1.86 | 13 |