

Application of Interpretable Machine Learning for Accident Risk Analysis on Jeddah Roads

An Integrated Explainable Artificial Intelligence Framework

Chiheb Eddine Ben Ncir

Department of Management Information Systems
College of Business, University of Jeddah
Jeddah, Saudi Arabia

ABSTRACT

Road traffic accident prediction can support proactive safety management only when predictive performance is evaluated together with model transparency. This study develops and evaluates an interpretable machine-learning and explainable artificial intelligence framework for daily accident-risk analysis in Jeddah, Saudi Arabia. The analysis uses 366 daily observations from 2024 containing traffic volume, average speed, accident counts, weather severity, road-work level, and a congestion index. Daily accident burden is defined as the sum of accidents recorded inside and outside the city, and a high-risk day is operationalized as seven or more accidents, corresponding to the sample 75th percentile and yielding 127 high-risk days (34.7%). Class-weighted logistic regression, random forest, XGBoost, and LightGBM are compared using repeated stratified five-fold cross-validation and a chronological 20% holdout. Mean ROC-AUC ranges from 0.493 to 0.530 and mean F1-score from 0.256 to 0.406, indicating limited discrimination with the available aggregate predictors. LightGBM provides the highest mean balanced accuracy (0.526) and ROC-AUC (0.530), while logistic regression provides the highest mean F1-score (0.406). Global explainability is examined through SHAP, permutation feature importance, rank-agreement analysis, and domain-level attribution. Congestion, temporal structure, average speed, and traffic volume are prominent in the fitted LightGBM explanations, although SHAP and permutation rankings show only moderate agreement. The findings demonstrate the value of explainable modelling as both an interpretation and diagnostic layer: it identifies the structures used by the model while also exposing the limitations of the available data. The resulting framework provides a reproducible baseline for future Jeddah road-safety studies using richer crash-level, spatial, roadway, and exposure information.

Keywords: Road safety; Accident Risk; Interpretable Machine Learning; Explainable Artificial Intelligence; SHAP; LIME; Jeddah Roads

I. INTRODUCTION

Road traffic crashes still be a major public-health and transportation-safety problem that has huge human, social and economic implications all around the world [1]. As cities grow and mobility patterns become more complex, road-safety agencies are turning increasingly to data-driven methods to identify dangerous drivers and prevent accidents and to plan for road safety and infrastructure. Machine learning methods are suited to this problem as accident levels can be linked to nonlinear correlations between traffic exposure, operating speed, congestion, environmental conditions, road conditions and time patterns. Ensemble algorithms like random forests and gradient boosting algorithms can be used to capture interactions too complex to be captured by standard parametric models, but they also present an interpretability issue when predicting and communicating the results which requires engineers and decision makers to make sense of [2].

Explainable artificial intelligence (XAI) attempts to address these challenges by adding predictive models to the model to describe how variables in the model provide an input to the prediction. SHapley Additive exPlanations (SHAP) provide additive feature attribution in global and local interpretation [3] and permutation feature importance (PFI) can assess models on the basis of the model-agnostic impact on prediction performance when the individual variables are disrupted. Local surrogate approaches like LIME can provide an additional perspective on individual predictions [4]. Explainability is important in road-safety research because an error-prone model may still be useful but not for practical application as long as it is not well understood, not well represented, or is interpreted as causal because all the information can be used in the model as the cause of the error. Recent traffic safety research has also

combined machine learning with SHAP and similar methods to make the crash risk analysis and injury severity analysis more transparent [2], [5], [6].

Jeddah is a relevant location for such analysis because traffic conditions vary across time and operational levels, and we need locally based evidence to make city-specific safety decisions. In the present paper our focus is on daily accident burden and not individual crash severity, since the 2024 dataset is organized as a daily traffic and accident series. So we want to know if aggregate daily traffic variables can distinguish high accident days, what variables the fitted models rely on most and whether the different global explanation methods are consistent. We consider four classifiers (class-weighted logistic regression, random forest, XGBoost and LightGBM) and validate them in a common way. To do this we combine SHAP, PFI, rank agreement and domain aggregation to not use a single method to determine the importance of each feature.

The contribution of this work is twofold. First, it provides a leakage-aware and reproducible baseline for daily accident-risk classification in Jeddah using repeated stratified cross-validation and chronological testing rather than a single split of train-test. Second, it treats explainability as part of model validation: agreement and disagreement among explanation methods are used to assess the stability of the model's global interpretation. This is important when the predictive discrimination is modest as explanation should reveal model behavior without overstating causal or operational conclusions. The rest of the paper is organized as follows. In Section II, we review some related machine-learning and XAI research in road safety. In Section III, we describe the dataset, target definition, feature engineering, predictive models, validation protocol, and explainability methods. We report descriptive results, predictive performance, temporal robustness and global explainability findings in Section IV. Section V explores the implications for transportation safety analysis and decision support. Section VI provides limitations and the areas of future research, and Section VII concludes the study.

II. RELATED WORK AND EXPLAINABLE ARTIFICIAL INTELLIGENCE BACKGROUND

Boosting and ensemble methods are frequently used in road safety modelling because they can capture nonlinear effects and interaction among heterogeneous predictors. Dong et al. combined boosting-based ensembles with SHAP to study highway injury-severity analysis and demonstrated that predictive evaluation can be combined with global and local attribution [2]. In Saudi Arabia, Aldhari et al. applied machine learning methods to highway crash-severity prediction and presented evidence for data-driven safety modelling for the sector in different regions. Aboulola combined transfer learning with SHAP for traffic accident severity prediction, further illustrating the use of post-hoc explanation with complex predictive models [6]. These studies differ in their data structure, outcome definition, and validation design, and as a result, they are not set to be a benchmark for daily Jeddah classification task performance. Rather, they provide a framework to evaluate predictive performance and interpretability together.

Interpretability may be intrinsic, as in a parsimonious logistic model, or post-hoc as when an explanation method is applied to a fitted ensemble. It can be global or local. SHAP allows for additive feature attributions based on Shapley-value notions and is efficient for tree ensembles [3]; LIME is capable of a local approximation of a complex predictor using a simpler surrogate [4]. PFI offers a model-agnostic way to assess the model performance after a feature is permuted. Since different explanation methods have different assumptions and failure modes, agreement between different methods can help with interpretation while disagreement can reveal instability. In this study, SHAP shows the magnitude of attribution, PFI provides an independent predictive-importance check, Spearman rank correlation indicates the agreement between the two rankings, and domain aggregation shows attribution across traffic-operation, temporal, road-work, and weather. These methods explain or probe the fitted model, and the causal effects are not established [7].

TABLE I. RELATED STUDIES AND THEIR ROLE IN THE PRESENT FRAMEWORK

Study	Predictive approach	Explainability	Role in this study
Dong et al. [2]	Boosting ensembles	SHAP	Global/local attribution with class-sensitive evaluation
Aldhari et al. [5]	Saudi highway ML	Feature importance	Regional motivation for localized modelling
Aboulola [6]	Transfer learning	SHAP	Complex-model explanation in accident severity

Study	Predictive approach	Explainability	Role in this study
Lundberg and Lee [3]	Model agnostic / tree efficient	SHAP	Additive local-to-global attribution
Ribeiro et al. [4]	Model agnostic	LIME	Local sensitivity option for future larger samples

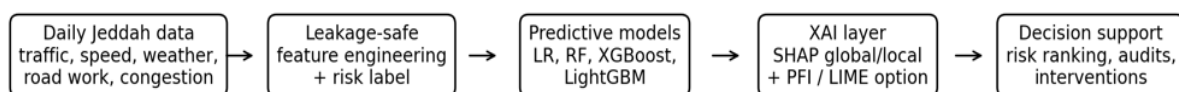
III. MATERIALS AND METHODS

The data set had 366 daily incidents from January 1 to December 31, 2024. The data set contains traffic volume, average speed, accidents inside the city, accidents outside the city, weather, road-work level and congestion index. The two accident fields were treated as outcome factors only and were excluded from the predictor set. The total daily accidents were calculated as the sum of the inside-city and outside-city accidents. The distribution had a median of 6 accidents per day and a 75th percentile of 7. A high-risk day is defined as 7 or more accidents. This resulted in 127 high-risk days (34.7% of the sample) and 239 lower-risk days. The threshold is sample-based and is not an official Saudi crash-severity category. Date was converted to sine and cosine for day of week, month and year and its cyclical time proximity was represented without any artificial boundary discontinuity.

TABLE II. DATASET VARIABLES AND OBSERVED RANGES

Variable	Role	Observed range or coding
Date	Temporal index	1 Jan–31 Dec 2024
TrafficVolume	Predictor	3,008–13,566 vehicles/day
AverageSpeed	Predictor	40.5–81.3 km/h
AccidentsInsideCity	Outcome component	0–10/day
AccidentsOutsideCity	Outcome component	0–8/day
WeatherSeverity	Predictor	Ordinal 0–4
RoadWorkLevel	Predictor	Ordinal 0–3
CongestionIndex	Predictor	21.9–87.3

Four models were evaluated: class-weighted logistic regression, random forest, XGBoost, and LightGBM. Overall, we used repeated stratified five-fold cross-validation with five repeats each resulting in 25 test folds, and a chronological robustness test, trained on the first 80% of the year and evaluated on the last 20%. We summarized the performance with accuracy, balanced accuracy, precision, recall, F1-score, receiver operating characteristic area under the curve (ROC-AUC), and Brier score for the temporal holdout. Since the high-risk class is less common, the F1 and balanced accuracy are emphasized over raw accuracy. The random seed is a fixed random number of 42 where supported. The explainability layer is applied to LightGBM and averaged mean absolute SHAP values, PFI with uncertainty, SHAP and PFI rank correlation, and domain aggregation. The resulting workflow is shown in Fig. 1.



Validation: repeated stratified CV + temporal holdout | Metrics: F1, balanced accuracy, ROC-AUC, calibration

Fig. 1. Integrated interpretable machine-learning and explainable artificial intelligence workflow for Jeddah accident-risk analysis.

II. IV. RESULTS AND GLOBAL EXPLAINABILITY

Across the 366 mean total accident burden was 5.78 accidents per day (standard deviation 2.49, range 1-14). Average traffic volume was 8,170 vehicles per day, mean average speed was 63.25 km/h, and mean congestion index was 51.94. Monthly accident rates were different across 2024 (Fig. 2) that can be described in terms of fluctuations since a single annual series cannot be used to establish a consistent seasonality. The operational high-risk threshold of seven or more accidents identified 127 days (34.7%) as high risk. These characteristics describe an exploratory daily-risk classification problem and provide the context for interpreting the validation results in Table IV.

TABLE III. DESCRIPTIVE STATISTICS FOR THE SUPPLIED JEDDAH DATA

Variable	Mean	SD	Minimum	Median	Maximum
TrafficVolume	8170.26	2038.96	3008.0	8237.0	13566.0
AverageSpeed	63.25	7.08	40.5	62.9	81.3
AccidentsInsideCity	3.53	1.92	0.0	3.0	10.0
AccidentsOutsideCity	2.24	1.48	0.0	2.0	8.0
TotalAccidents	5.78	2.49	1.0	6.0	14.0
WeatherSeverity	0.94	0.93	0.0	1.0	4.0
RoadWorkLevel	0.92	0.76	0.0	1.0	3.0
CongestionIndex	51.94	10.94	21.9	51.8	87.3

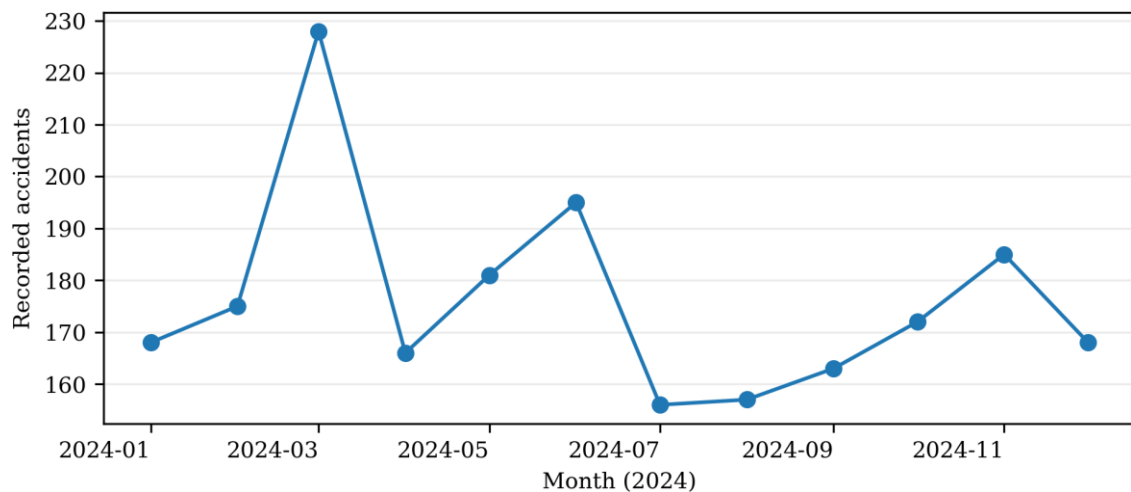


Fig. 2. Monthly total accident burden in 2024 in Jeddah roads.

TABLE IV. REPEATED FIVE-FOLD CROSS-VALIDATION PERFORMANCE

Model	Accuracy	Balanced acc.	Precision	Recall	F1	ROC-AUC
Logistic regression	0.511 ± 0.052	0.506 ± 0.060	0.350	0.488	0.406 ± 0.077	0.511 ± 0.062
LightGBM	0.562 ± 0.045	0.526 ± 0.048	0.377	0.408	0.390 ± 0.070	0.530 ± 0.066
Random forest	0.587 ± 0.054	0.512 ± 0.053	0.375	0.268	0.307 ± 0.088	0.493 ± 0.062
XGBoost	0.624 ± 0.039	0.523 ± 0.044	0.407	0.191	0.256 ± 0.093	0.522 ± 0.058

Cross-validated discrimination was very modest. Logistic regression produced the highest mean F1-score of 0.406, while LightGBM had the highest mean balanced accuracy of 0.526 and ROC-AUC of

0.530. XGBoost had the highest raw accuracy of 0.624, but recall for the high-risk class was only 0.191, demonstrating why accuracy alone is misleading in an imbalanced safety task. There was also significant repeated-fold variability with all models, so even small differences between models should not be taken too negatively. Fig. 3 compares the principal class-sensitive metrics.

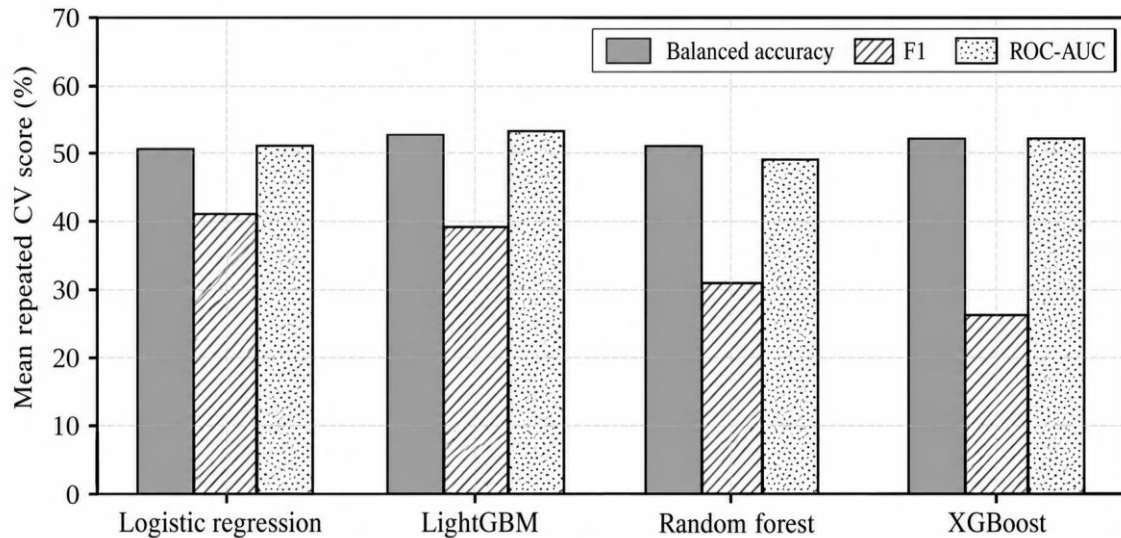


Fig. 3. Mean repeated five-fold cross-validation performance for the four evaluated models.

TABLE V. CHRONOLOGICAL HOLDOUT PERFORMANCE

Model	Accuracy	Balanced acc.	Precision	Recall	F1	ROC-AUC	Brier
Logistic regression	0.676	0.559	0.429	0.273	0.333	0.511	0.235
LightGBM	0.514	0.483	0.281	0.409	0.333	0.483	0.258
XGBoost	0.689	0.530	0.429	0.136	0.207	0.511	0.221
Random forest	0.703	0.526	0.500	0.091	0.154	0.517	0.226

The chronological holdout also yielded the same result. Logistic regression and LightGBM both achieved $F1 = 0.333$ while all ROC-AUC values remained about 0.48–0.52. The holdout has only 74 days and is therefore a stress test, and not an external validation, but it tells us that the limited discrimination we see with repeated cross-validations is not solved by a simple temporal split.

TABLE VI. GLOBAL LIGHTGBM SHAP FEATURE ATTRIBUTION

Rank	Feature	Mean absolute SHAP
1	CongestionIndex	0.341
2	dayofyear_sin	0.265
3	dayofyear_cos	0.232
4	AverageSpeed	0.231
5	TrafficVolume	0.223
6	dow_sin	0.187
7	dow_cos	0.169
8	month_cos	0.063
9	RoadWorkLevel	0.060

Rank	Feature	Mean absolute SHAP
10	month_sin	0.044
11	WeatherSeverity	0.011

The global SHAP attribution shows that the fitted LightGBM model relies most heavily on congestion index, followed by day-of-year components, average speed, traffic volume, and day-of-week components (Fig. 4). This ranking describes the internal allocation of model contribution; it does not mean that congestion or a temporal component cause accidents. Because model discrimination is weak, the appropriate interpretation is diagnostic: these variables are the principal structures the model attempted to use, but their combined information was not enough to make it reliable.

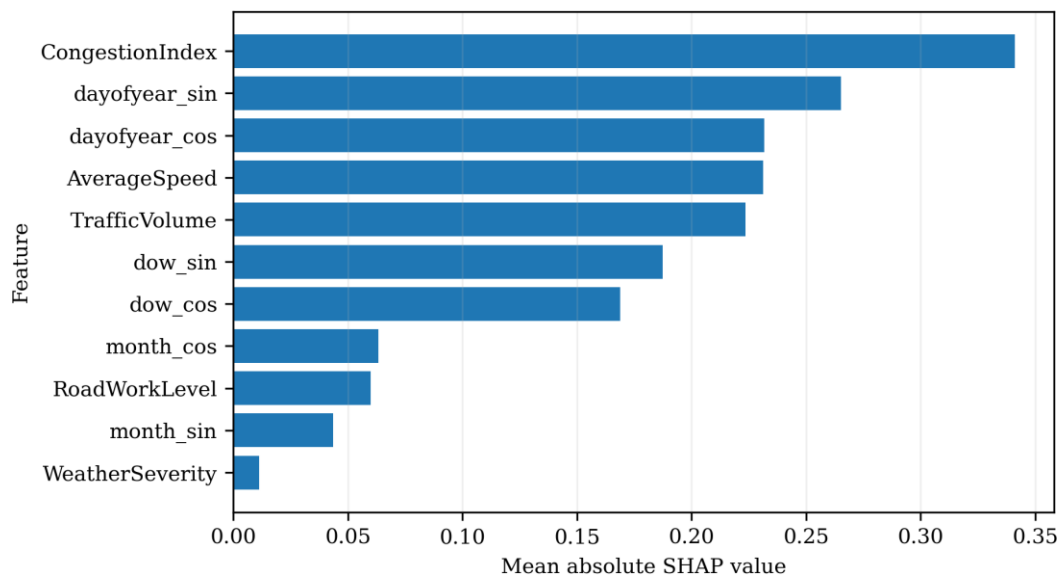


Fig. 4. Global LightGBM SHAP attribution. Larger values indicate stronger average contribution to the fitted model output, not causality.

TABLE VII. PERMUTATION FEATURE IMPORTANCE FOR LIGHTGBM

Rank	Feature	PFI mean	PFI SD
1	dow_sin	0.046	0.022
2	dayofyear_sin	0.040	0.023
3	CongestionIndex	0.033	0.018
4	dayofyear_cos	0.024	0.013
5	AverageSpeed	0.022	0.016
6	WeatherSeverity	0.020	0.014
7	month_cos	0.014	0.013
8	dow_cos	0.011	0.022
9	TrafficVolume	0.006	0.012
10	RoadWorkLevel	0.002	0.002
11	month_sin	0.000	0.000

PFI is useful as an independent check, since it asks how much predictive score is lost when each feature is disrupted. The day-of-week sine, day-of-year sine and congestion index had the highest importance of the permutation in each of these two methods, though uncertainty was large in a number of variables and some intervals approached or crossed zero (Fig. 5). SHAP and PFI ranked highly with a Spearman correlation of 0.65 ($p = 0.029$) and the two methods were moderately similar but not completely identical. Congestion and day-of-year appeared to rank very high in both methods, while weather severity and traffic volume have larger rank differences. This disagreement is interesting: it warns against taking only one global-importance graphic as a definitive one and suggests other methods to explain it.

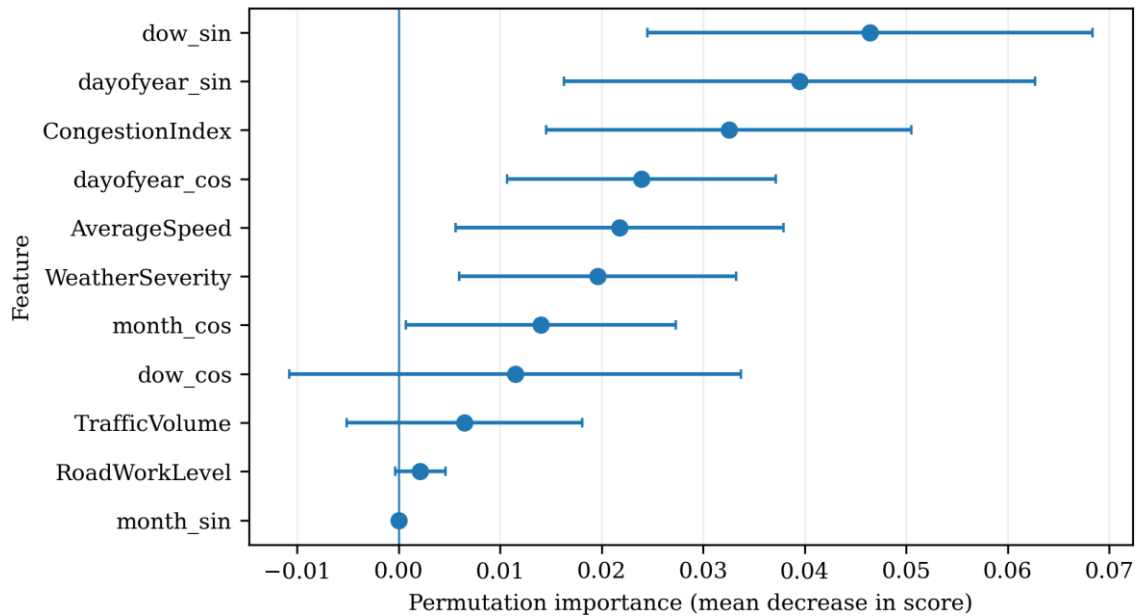


Fig. 5. Permutation feature importance with one-standard-deviation uncertainty bars. Values near zero indicate limited measurable contribution under permutation.

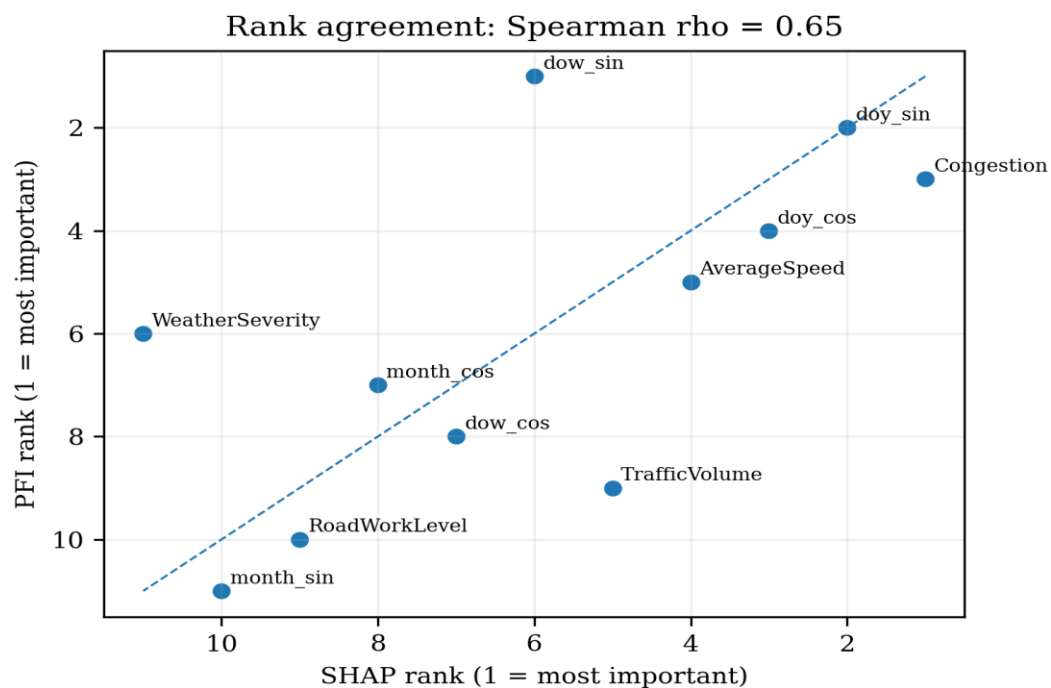


Fig. 6. Agreement between SHAP and permutation-importance rankings. Points on the dashed diagonal have identical ranks.

For a more granular systems view, the features were grouped into traffic operation (congestion, speed and traffic volume), temporal structure (day-of-year, day-of-week and month components), road work

and weather. The total absolute SHAP attribution was concentrated in the temporal and traffic operation domains (Fig. 7). This sort of grouping of features can be useful to communicate with practitioners, but it has a serious drawback: domains with more encoded variables can have more total attribution. The domain chart should therefore be viewed as an aggregate of the individual features and not a ranking of causal contributions..

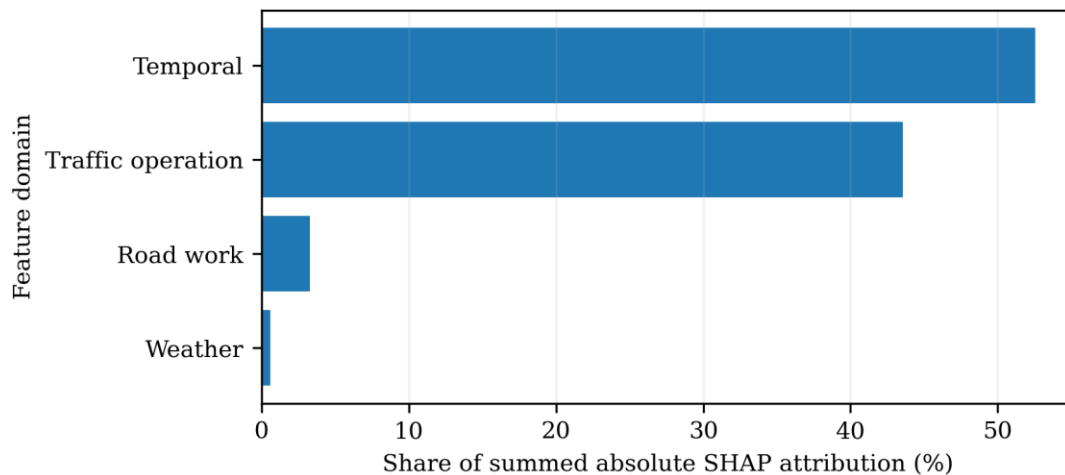


Fig. 7. Domain-level share of summed absolute SHAP attribution. Temporal domains contain multiple cyclical features and therefore have more opportunities to accumulate attribution.

V. DISCUSSION AND TRANSPORTATION IMPLICATIONS

The results indicate that the daily aggregate variables are of a clear-cut structure but are not very helpful to discern the high accident days. The cross-validation and chronological holdout produced values of ROC-AUC well close to 0.50 and the different class-sensitive metrics have significant trade-offs between the models. Logistic regression has the highest mean F1-score, while LightGBM has the highest mean balanced accuracy and ROC-AUC. This pattern indicates that the choice of models does not solve the fundamental information gap in the dataset. The daily traffic volume, mean speed, congestion, weather severity, road-work level and the number of accidents can be aggregated in a general way but the number of accidents might also depend on the intersection geometry, lane configurations, speed variation, road-user exposure, enforcement, event traffic, collision mechanism, vehicle composition, lighting, visibility and location of the accident to be attended. The global XAI analysis supports this idea. SHAP and PFI strongly emphasize temporal and traffic-operation structure but their moderate rank agreement (Spearman rho = 0.65, $p = 0.029$) shows that feature importance is method-dependent. Thus, the explanation results are most useful as a transparent account of model behavior and data sufficiency and not as evidence of causal safety mechanisms.

From a transportation management perspective, the current framework is best suited for exploratory analysis, model auditing and data prioritization rather than automated enforcement, patrol dispatch or corridor-specific engineering prescriptions. Congestion, temporal variables, speed and traffic volume should be investigated as potential signals to further explore but practical interventions have to be validated with spatially resolved crash and exposure data. A better Jeddah safety platform should link crash coordinates to roadway geometry, access points, intersection control, speed distribution, traffic exposure, lighting, weather and visibility, vehicle composition and road-user characteristics. With more detailed multi-year data the same modelling pipeline can be expanded to include spatially grouped validation, temporal testing, calibration, imbalance handling within training folds, subgroup analysis, SHAP interaction effects and local explanations. Such an extension would enable explainability for engineering interpretation while maintaining the distinction between predictive association and causal evidence [7], [8].

VI. LIMITATIONS AND FUTURE EXTENSION

Some limitations limit the interpretability. The dataset is only one year and 366 observations, which is small for machine learning and thus difficult to assess multi-year seasonality. Accidents are collected by day. As a result, the spatial and crash-level heterogeneity is lost. The high-risk threshold is sample-based and should be tested in comparison to other thresholds and count-regression scenarios. We also need to consider that some predictors could be daily summaries rather than information to be known before the day ends, which would be difficult to use in the future. In other words, SHAP and PFI are associative model-explanation tools that do not provide causal estimators. Finally, the current global explanations

are based on a LightGBM model that is not a good out-of-sample discriminator and should therefore be treated as diagnostics of model behavior. A future Jeddah safety platform should be based on multi-year crash-level and corridor-level data which are geographically separated and temporally validated, imbalance handling inside training folds, probability calibration, fairness checks for subgroups, and explanation stability analysis across SHAP, PFI, and local methods.

VII. CONCLUSION

This study developed a transparent machine-learning and explainable artificial intelligence framework for daily accident-risk analysis in Jeddah. Four classifiers were evaluated with repeated stratified five-fold cross-validation and a chronological holdout, and global explainability was assessed through SHAP, permutation feature importance, rank-agreement analysis, and domain aggregation. Predictive discrimination was limited: the highest mean F1-score was 0.406 and the highest mean ROC-AUC was 0.530. Nevertheless, the explainability analysis provided a useful diagnostic view of the fitted LightGBM model, highlighting congestion, temporal structure, average speed, and traffic volume while also showing that feature rankings vary across explanation methods. The study therefore establishes a reproducible baseline rather than a deployment-ready forecasting system. Its main value is methodological: predictive validation and explanation are considered jointly, allowing the analysis to identify both the signals present in the data and the information that is still missing. Future work using multi-year, crash-level, spatial, roadway, and exposure data can build on this framework to develop more reliable and operationally relevant road-safety decision support for Jeddah.

REFERENCES

- [1] R. Dwivedi et al., “Explainable AI (XAI): Core ideas, techniques, and solutions,” *ACM Computing Surveys*, vol. 55, no. 9, 2023.
- [2] S. Dong, A. Khattak, I. Ullah, J. Zhou, and A. Hussain, “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, p. 2925, 2022, doi: 10.3390/ijerph19052925.
- [3] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” in *Advances in Neural Information Processing Systems*, vol. 30, 2017, pp. 4765–4774.
- [4] M. T. Ribeiro, S. Singh, and C. Guestrin, “Why should I trust you? Explaining the predictions of any classifier,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 1135–1144, doi: 10.1145/2939672.2939778.
- [5] I. Aldhari et al., “Severity prediction of highway crashes in Saudi Arabia using machine learning techniques,” *Applied Sciences*, vol. 13, no. 1, p. 233, 2023, doi: 10.3390/app13010233.
- [6] O. I. Aboulola, “Improving traffic accident severity prediction using MobileNet transfer learning model and SHAP XAI technique,” *PLOS ONE*, vol. 19, no. 4, e0300640, 2024, doi: 10.1371/journal.pone.0300640.
- [7] C. Molnar, *Interpretable Machine Learning*, 2nd ed. 2022. [Online]. Available: <https://christophm.github.io/interpretable-ml-book/>
- [8] World Health Organization, *Global Status Report on Road Safety 2023*. Geneva, Switzerland: World Health Organization, 2023.
- [9] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794, doi: 10.1145/2939672.2939785.
- [10] G. Ke et al., “LightGBM: A highly efficient gradient boosting decision tree,” in *Advances in Neural Information Processing Systems*, vol. 30, 2017.