North Asian International Research Journal of Multidisciplinary

IRJIF IMPACT FACTOR: 6.012

ISSN: 2454-2326 Vol.12, Issue-5 May-2026

Index Copernicus Value: 58.12 Indian Citation Index

NAIRJC UGC-Aligned | HEI-Approved Peer Reviewed Refereed Journal (Under New UGC Guidelines)

DOI: doiglobal.org/10.2026/NAIRJCMD.010

EXPLAINABLE AI-BASED ROAD ACCIDENT SEVERITY PREDICTION FOR INDIAN HIGHWAYS USING MACHINE LEARNING

DR. JYOTI. Y. KULKARNI

Associate Professor Govt. First Grade College, Kalagi.

ABSTRACT

Road traffic crashes remain a major transportation and public-safety challenge in India. Predicting crash severity can support proactive engineering, enforcement and emergency-response decisions by identifying combinations of road, vehicle, environmental and temporal factors associated with serious outcomes. This paper proposes an explainable machine-learning framework for predicting accident severity on Indian highways. The framework integrates structured crash records, data cleaning, categorical encoding, imbalance treatment, feature engineering, Random Forest, Gradient Boosting and XGBoost-style tree ensembles, followed by SHAP-based interpretation. Rather than presenting fabricated experimental performance, the paper uses published Indian highway studies as comparative evidence and defines a reproducible protocol for future model validation. A peer-reviewed Indian-highway Random Forest study using 3,257 observations and 32 variables reported 67% training accuracy but only 41.47% test accuracy, demonstrating the danger of imbalance and weak generalization. A 2025 MethodsX framework subsequently emphasized Random Forest, Gradient Boosting and SHAP interpretation for Indian highway severity prediction. Another Indian expressway study analyzed 2,747 crashes using multinomial logit, decision tree and Random Forest models. The proposed research therefore prioritizes balanced evaluation, macro-F1, class-wise recall, calibration and explainability rather than accuracy alone. The paper argues that interpretable AI can help identify high-risk conditions and support road-safety planning, but predictive models should supplement not replace engineering judgment, police investigation and validated crash databases.

***KEYWORDS**: road accident severity; Indian highways; machine learning; Random Forest; Gradient Boosting; XGBoost; SHAP; explainable artificial intelligence; road safety*

1. INTRODUCTION

Road crashes are complex events produced by interactions among road geometry, traffic conditions, vehicle characteristics, human behavior, weather, visibility and emergency response. Their consequences range from property damage to fatal injury. For road-safety agencies, understanding not only where crashes occur but why some crashes become more severe is an important analytical problem.

Traditional statistical models such as logistic and multinomial regression remain valuable because their parameters are interpretable. However, crash data often contain nonlinear relationships, interactions and heterogeneous categorical variables. Machine-learning models can capture these patterns without imposing the same functional assumptions.

The challenge is that a high-performing black-box model may be difficult to use for policy. A road engineer needs to know whether a prediction is driven by road condition, vehicle type, collision type, weather or another factor. Explainable AI methods such as SHAP can therefore connect predictive modeling with actionable road-safety interpretation.

This paper proposes an explainable AI framework designed specifically for Indian highway crash-severity analysis. It emphasizes reproducibility, class imbalance, external validation and transparent reporting.

2. PROBLEM STATEMENT

Accident-severity datasets are commonly imbalanced because fatal and serious crashes occur less frequently than lower-severity outcomes. A classifier may achieve apparently acceptable overall accuracy while failing to identify the minority classes that matter most for safety intervention.

A second problem is transferability. A model trained on one highway corridor may learn local patterns that do not generalize to another State, road type or traffic environment. A third problem is interpretability: road agencies require explanations, not merely class predictions.

The research problem is therefore to construct a machine-learning framework that predicts severity while remaining robust to imbalance, transparent to decision-makers and testable across different Indian highway contexts.

3. OBJECTIVES

- To develop an explainable machine-learning framework for Indian road-accident severity prediction.
- To identify relevant road, environmental, temporal, crash and vehicle features from structured highway records.
- To compare Random Forest, Gradient Boosting and XGBoost-style ensemble models with simpler baselines.
- To address class imbalance through weighting, resampling or other training-only techniques.

- To evaluate models using accuracy, precision, recall, macro-F1, weighted-F1, confusion matrices and ROC/PR measures where appropriate.
- To use SHAP or equivalent explainability methods to identify global and case-specific risk factors.
- To propose a deployment approach for road-safety planning and high-risk-condition analysis.

4. RESEARCH QUESTIONS

- Which machine-learning model offers the best balance of predictive performance and interpretability for Indian highway crash severity?
- Which road, weather, vehicle and collision variables contribute most strongly to severe outcomes?
- How much does class imbalance affect fatal and serious-injury recall?
- Can SHAP explanations provide stable and policy-relevant interpretations?
- How well does a model trained on one highway corridor generalize to a different corridor or time period?

5. SCOPE AND DEFINITION OF SEVERITY

Severity may be represented as fatal, grievous/serious injury, minor injury and no-injury/property-damage-only categories, depending on the source dataset. The exact labels must follow the official recording scheme used in the selected crash database.

The paper focuses on severity conditional on a recorded crash. It does not claim to predict the exact time and location of a future crash. Risk-zone prediction and crash-frequency modeling are related but distinct research tasks.

6. INDIAN HIGHWAY DATA CONTEXT

Published Indian highway research demonstrates the type of structured data available for severity modeling. Khanum, Garg and Faheem analyzed two highway stretches: the Pune-Solapur section and the Barwa-Adda-Panagarh section. Their combined dataset contained 3,257 observations and 32 variables.

Variables included date and time, location, nature of accident, severity, cause, road feature, road condition, intersection type, weather, vehicle types and casualty counts. This illustrates the multidimensional nature of severity modeling.

For a new study, raw data should preferably be obtained from authoritative highway, police or road-safety sources with clear definitions and data-quality documentation.

7. LITERATURE REVIEW

Machine learning has increasingly been applied to crash-severity analysis because tree ensembles can represent nonlinear interactions among categorical and numerical variables. Random Forest is particularly common due to robustness and built-in feature-importance measures.

Khanum et al. (2023) applied Random Forest to 3,257 Indian highway observations. Their model reported 67% accuracy on training data with weighted F1 of 0.64 and macro-F1 of 0.53. After grid-search optimization, test accuracy was 41.47%. The authors explicitly noted possible bias and imbalance, making this study valuable as evidence that training performance can substantially overstate real-world generalization.

A 2025 MethodsX paper by Khanum and colleagues extended the methodological direction by combining Random Forest and Gradient Boosting with SHAP-based interpretation. The work emphasizes model explainability and evaluates accuracy, precision, recall, F1-score and confusion matrices.

Garg, Chatterjee and Maitra (2024) examined crash severity on Indian highways using machine learning, including artificial neural network approaches and feature-importance analysis. A separate Indian expressway study published online in 2025 analyzed 2,747 crashes from three expressways using multinomial logit, decision tree and Random Forest models.

The literature therefore shows increasing movement from single-model accuracy toward comparative, interpretable and context-aware severity analysis.

8. RESEARCH GAP

The first gap is generalization. Many studies are based on a small number of corridors. Random train-test splitting can place crashes from the same location and time period in both sets, producing optimistic estimates.

The second gap is imbalance-sensitive evaluation. Overall accuracy may conceal poor recognition of fatal crashes. Macro-F1, class-specific recall and precision-recall analysis should receive greater emphasis.

The third gap is explainability. Standard impurity-based feature importance can be biased toward variables with many possible splits. SHAP offers both global and local explanations, although its results must still be interpreted cautiously.

The fourth gap is deployment. A research classifier should be connected to a clear decision-support use case, such as identifying conditions associated with severe outcomes or prioritizing safety audits.

9. PROPOSED CONCEPTUAL FRAMEWORK

The proposed framework has eight stages: data acquisition, data-quality audit, preprocessing, feature engineering, imbalance handling, model training, evaluation and explainability. A final decision-support layer translates model findings into road-safety insights.

The framework deliberately separates predictive performance from causal interpretation. A feature associated with severe crashes is not automatically a proven cause. SHAP explains the model's prediction, not the physical causality of the crash.

10. DATA COLLECTION

A suitable dataset should include multiple highway sections and several years of records. Each observation represents a crash event. Essential metadata include location or chainage, date, time and severity label.

Predictor groups may include road geometry, surface condition, intersection type, lighting, weather, vehicle types, collision configuration, traffic context and recorded contributing factors. Variables that directly encode the outcome after the crash should be excluded if they would not be available at prediction time.

This leakage rule is critical. For example, casualty counts may be strongly related to severity because they are part of the consequence being predicted. Including such post-outcome variables can produce an unrealistic model.

11. DATA CLEANING

Crash databases may contain missing values, inconsistent category labels and duplicated records. Cleaning should standardize spelling, date-time formats and categorical coding. Missingness should be reported rather than silently discarded.

For numerical variables, robust imputation may use median values or model-based methods. Categorical missing values can be represented as an explicit unknown category when that has operational meaning.

Duplicate crashes should be identified using combinations of date, time, location and record identifiers.

12. FEATURE ENGINEERING

Time can be transformed into hour-of-day bands, weekday/weekend indicators and daylight proxies where appropriate. Road chainage can support segment-level analysis. Collision types may be grouped into interpretable categories when sparse subclasses exist.

Weather and surface variables should remain distinct unless the data show severe sparsity. Interaction features may be unnecessary for tree ensembles because these models can learn interactions automatically.

Feature engineering must be performed using information available before or at the time of the prediction target to prevent leakage.

13. ENCODING CATEGORICAL VARIABLES

Tree-based implementations differ in their support for categorical features. One-hot encoding is transparent but can produce high-dimensional matrices. Ordinal integer encoding should not imply a false numerical order unless the algorithm treats categories appropriately.

The complete preprocessing pipeline should be fitted only on training data and then applied unchanged to validation and test data.

14. CLASS IMBALANCE

Fatal and grievous classes may be underrepresented. A naïve classifier can maximize accuracy by favoring the majority class. Strategies include class weighting, random under-sampling, carefully applied over-sampling and SMOTE-like synthetic sampling.

Resampling must occur only inside the training folds. Performing it before train-test separation leaks information and invalidates evaluation.

Because synthetic examples can distort mixed categorical data, class weighting should be evaluated as a strong baseline.

15. RANDOM FOREST MODEL

Random Forest builds an ensemble of decision trees from bootstrapped samples and randomized feature subsets. For classification, the final class is determined by aggregated tree votes or probabilities.

If T_b(x) is the prediction of tree b for input x, the forest combines predictions over B trees. Important hyperparameters include number of estimators, maximum depth, minimum samples per leaf and number of candidate features per split.

Random Forest is attractive for crash data because it captures nonlinear relationships and interactions while requiring relatively limited feature scaling.

16. GRADIENT BOOSTING AND XGBOOST-STYLE MODELS

Gradient boosting builds trees sequentially, with each new learner focusing on residual errors from the existing ensemble. Modern boosted-tree systems incorporate regularization, shrinkage and efficient split optimization.

These models often perform strongly on tabular data, but extensive tuning can overfit small datasets. Hyperparameter selection should therefore use cross-validation within the training set.

17. BASELINE MODELS

Logistic or multinomial logistic regression should be included as an interpretable statistical baseline. A decision tree provides a simple nonlinear baseline. If sufficient data exist, a multilayer neural network can be evaluated, but tabular deep learning is not automatically superior to tree ensembles.

A strong research design compares models under identical data splits and metrics rather than selecting different subsets for each algorithm.

18. MATHEMATICAL EVALUATION

For multiclass prediction, Accuracy = correctly classified observations / total observations. For each class, Precision = TP/(TP+FP), Recall = TP/(TP+FN), and F1 = 2PR/(P+R).

Macro-F1 calculates the unweighted mean of class F1 values and therefore gives fatal and minority classes equal importance. Weighted-F1 weights classes by support. Reporting both reveals whether good aggregate performance depends on the majority class.

A confusion matrix should be treated as a core result because it shows whether severe crashes are systematically underestimated.

19. CROSS-VALIDATION AND TEST DESIGN

The preferred design uses a held-out test set that is geographically or temporally separated where possible. For example, one highway section can serve as an external test after training on another, or the most recent year can be held out.

Within the training set, stratified k-fold cross-validation can guide hyperparameter selection. If observations are clustered by road segment, grouped cross-validation should be considered to reduce spatial leakage.

The final test set must not influence feature selection or hyperparameter tuning.

20. EXPLAINABLE AI WITH SHAP

SHAP values are derived from cooperative game theory and attribute a model prediction to individual features relative to a baseline expectation. A global summary plot can show which variables most strongly influence predictions across the dataset.

Local explanations can examine a specific predicted fatal or serious crash and show which features increased or decreased the model's severity probability.

SHAP should not be interpreted as proof that changing a feature will causally change crash severity. It explains the fitted model's behavior.

21. CALIBRATION AND PROBABILITY QUALITY

For decision support, probability quality matters. A model that assigns 0.80 fatal-risk probability should ideally be correct at approximately that frequency among comparable predictions. Calibration curves and Brier scores can assess this property.

Poorly calibrated models may still rank risk well but should not be used to communicate absolute probabilities without correction.

22. PUBLISHED BENCHMARK RESULTS

Study	Indian data	Model	Sample size	Training result	Test / reported result	Key lesson
Khanum et al. (2023)	Two Indian highway stretches	Random Forest	3,257; 32 variables	67% accuracy; weighted F1 0.64; macro F1 0.53	41.47% test accuracy	Imbalance/generalization problem
Khanum et al. (2025)	Indian highways	Random Forest + Gradient Boosting + SHAP	Published framework	Metrics include accuracy, precision, recall, F1	SHAP-based interpretation emphasized	Explainability is central
Indian expressway study (2025)	Three expressways	MNL, Decision Tree, Random Forest	2,747 crashes	Comparative modeling	Study evaluates four severity categories	Multi-model comparison
Garg et al. (2024)	Indian highways	ANN / ML severity analysis	Indian crash records	Published comparative analysis	Feature-importance analysis	Nonlinear models useful

Table 1 contains only literature-based information. The 67%, 0.64, 0.53 and 41.47% figures are explicitly reported by Khanum et al. (2023); they are not results of the proposed model.

23. PROPOSED EXPERIMENTAL RESULTS TABLE

Model	Accuracy	Macro Precision	Macro Recall	Macro F1	Weighted F1	Fatal Recall
Multinomial Logistic Regression	To be measured	To be measured	To be measured	To be measured	To be measured	To be measured
Decision Tree	To be measured	To be measured	To be measured	To be measured	To be measured	To be measured
Random Forest	To be measured	To be measured	To be measured	To be measured	To be measured	To be measured
Gradient Boosting	To be measured	To be measured	To be measured	To be measured	To be measured	To be measured
XGBoost-style Ensemble	To be measured	To be measured	To be measured	To be measured	To be measured	To be measured

These values should be populated only after training on the selected dataset. Published numbers should not be copied into this table as if generated by the present experiment.

24. INTERPRETATION OF THE PUBLISHED RANDOM FOREST RESULT

The Khanum et al. result is methodologically instructive because the drop from 67% training accuracy to 41.47% test accuracy demonstrates why a research paper should not highlight training performance alone. The macro-F1 of 0.53 also indicates uneven class performance.

The authors reported that the model frequently underestimated severity and suggested dataset imbalance as a possible explanation. This directly motivates class-sensitive evaluation and imbalance handling in the proposed framework.

A lower but honestly measured external-test score is scientifically more valuable than an artificially high score obtained through leakage or duplicated observations.

25. FEATURE IMPORTANCE AND POLICY INTERPRETATION

Feature-importance analysis can identify variables that the model repeatedly uses, such as collision type, road condition, vehicle involvement, weather or geometric characteristics. These results can help formulate hypotheses for engineering investigation.

However, road-safety policy should not be based solely on algorithmic importance rankings. High-ranking features may reflect exposure patterns, reporting practices or correlations with omitted variables.

The recommended workflow is therefore model insight followed by domain review and, where possible, targeted field investigation.

26. HIGH-RISK CONDITION PROFILING

Instead of predicting a specific future crash, the model can characterize combinations of conditions associated with higher predicted severity. For example, certain collision configurations combined with road geometry and vehicle type may repeatedly receive elevated risk scores.

Such profiles can support road-safety audits, enforcement planning and emergency preparedness without making unrealistic claims about exact future events.

27. GEOGRAPHIC GENERALIZATION

India's highways cross diverse climatic, geometric and traffic environments. A model trained in Maharashtra may not automatically transfer to Jharkhand, West Bengal or another region.

External-corridor validation should therefore be considered a key benchmark. Performance degradation across corridors provides direct evidence about transferability and can guide regional recalibration.

28. TEMPORAL GENERALIZATION

Road conditions, vehicle fleets, enforcement practices and infrastructure change over time. A model trained on older crash records can become stale. Temporal holdout testing and periodic retraining should therefore be part of the deployment lifecycle.

Model monitoring should detect shifts in class distribution and predictor patterns.

29. PROPOSED DECISION-SUPPORT ARCHITECTURE

A practical system can ingest periodically updated crash records into a secure analytics database. A preprocessing pipeline generates validated features, and the trained model produces severity probabilities and SHAP explanations.

A dashboard can aggregate predictions by road segment, collision type or environmental condition. It should show uncertainty and historical evidence rather than presenting a single opaque 'risk score.'

The system is intended for planners and safety analysts, not for automated enforcement against individual drivers.

30. ETHICAL AND GOVERNANCE CONSIDERATIONS

Crash data may contain sensitive information about individuals and locations. Research datasets should remove unnecessary personal identifiers and follow applicable data-governance requirements.

Algorithmic bias can arise when reporting quality differs by region or crash type. Missing records are not random in many administrative datasets. Model outputs should therefore be audited before resource-allocation decisions.

Explainability improves transparency but does not eliminate bias. Documentation of data provenance, model version, validation date and limitations is essential.

31. LIMITATIONS

The proposed manuscript does not claim newly generated experimental accuracy because no raw crash dataset or model-training logs were provided. Published benchmark values are clearly separated from the proposed experimental table.

Crash-severity records can contain reporting error, missing variables and inconsistent definitions. Furthermore, observational prediction cannot establish causality.

A model may perform well on known corridors while failing under new road or traffic conditions. This limitation must be addressed through external validation.

32. FUTURE SCOPE

Future research can integrate road geometry from GIS, weather observations, traffic volume, speed data and emergency-response time with conventional crash records. Spatiotemporal models can examine how risk evolves across road segments and time periods.

Graph neural networks may model connected road networks, while survival and time-to-event methods can address different road-safety questions. Deep tabular models can also be compared with boosted trees on sufficiently large datasets.

Causal inference methods should be explored separately when the objective is to estimate the effect of a road-safety intervention rather than merely predict severity.

33. MAJOR FINDINGS

- Machine-learning models are appropriate for nonlinear crash-severity relationships but require careful validation.
- Published Indian highway evidence demonstrates substantial train-test performance gaps, reinforcing the importance of generalization testing.
- Accuracy alone is inadequate for imbalanced severity classes; macro-F1 and fatal/serious recall should be central metrics.
- Post-outcome variables must be excluded to prevent target leakage.
- SHAP can make tree-ensemble predictions more interpretable, but SHAP values are explanatory rather than causal.
- Geographic and temporal holdout tests provide stronger evidence than random splitting alone.
- Predictive models are most useful as decision-support tools for safety audits, resource prioritization and risk-factor investigation.

34. RECOMMENDATIONS

- Develop standardized, machine-readable crash databases with consistent severity and road-condition definitions.
- Record reliable road geometry, weather, lighting, vehicle and collision information at crash locations.
- Use class-sensitive metrics and explicitly report minority-class performance.
- Evaluate class weighting before aggressive synthetic oversampling of mixed categorical data.
- Maintain a completely locked external or temporal test set.
- Report model calibration when probabilities are used for prioritization.

• Use SHAP or comparable tools alongside conventional feature-importance measures.

• Validate models across multiple Indian States and highway types before operational use.

• Separate prediction from causality when communicating findings to policymakers.

• Document data provenance, preprocessing, model version and limitations for every deployed model.

35. CONCLUSION

Explainable machine learning offers a promising approach to road-accident severity analysis on Indian highways. Tree ensembles can model complex interactions among road, vehicle, environmental and crash characteristics, while SHAP can make their predictions more transparent to engineers and policymakers.

The published Indian literature also provides an important warning. A Random Forest model can appear reasonably successful on training data yet perform substantially worse on held-out observations. The reported decline from 67% training accuracy to 41.47% test accuracy in a peer-reviewed Indian highway study illustrates the need for strict evaluation, imbalance handling and external validation.

The proposed framework therefore prioritizes scientific reliability over inflated performance claims. It uses leakage-aware preprocessing, class-sensitive metrics, cross-validation, held-out testing, probability calibration and explainability. These practices are necessary if machine learning is to contribute meaningfully to road-safety decisions.

The ultimate objective is not to replace road engineers, police records or safety audits with an algorithm. It is to transform complex crash databases into interpretable evidence that can help identify high-risk conditions, guide targeted investigation and support more effective road-safety interventions across India's diverse highway network.

REFERENCES

1. Khanum, H., Garg, A., & Faheem, M. I. (2023). Accident severity prediction modeling for road safety using random forest algorithm: An analysis of Indian highways (Version 2). F1000Research, 12, 494. https://doi.org/10.12688/f1000research.133594.2
2. Khanum, H., Garg, A., Faheem, M. I., & Kulkarni, R. (2025). A methodological framework for road accident severity prediction for Indian highways using machine learning models. MethodsX, 15, 103728. https://doi.org/10.1016/j.mex.2025.103728
3. Garg, T., Chatterjee, S., & Maitra, B. (2024). Analysis of road traffic crash severity through machine learning on Indian highways. Journal of the Eastern Asia Society for Transportation Studies, 15, 3228–3244. https://doi.org/10.11175/easts.15.3228
4. Analysing crash severity on expressways in India: Statistical and machine learning models. (2025). Proceedings of the Institution of Civil Engineers – Transport. https://doi.org/10.1680/jtran.24.00071
5. Breiman, L. (2001). Random forests. Machine Learning, 45, 5–32.

6. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30.
7. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794.
8. Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. Annals of Statistics, 29(5), 1189–1232.
9. Ministry of Road Transport and Highways, Government of India. Road Accidents in India reports.
10. National Highways Authority of India. Highway safety and accident-data resources.