International Journal of Engineering Science and Advanced Technology (IJESAT)
Vol 26 Issue 02(February),2026
Received: 18-12-2025 Accepted: 27-01-2026 Published: 03-02-2026

EXPLAINABLE GRADIENT BOOSTING PIPELINE FOR MULTI-CLASS TRAFFIC ACCIDENT SEVERITY PREDICTION

Zoya Fatima[1], Dr. Waseema Masood[2]
[1]*Student, Department of Computer Science & Engineering, Deccan College of Engineering and Technology, Hyderabad, Telangana, India*
[2]*Supervisor, Department of Computer Science & Engineering, Deccan College of Engineering and Technology, Hyderabad, Telangana, India*
Email: [1]f.zoya2209@gmail.com, [2]waseema@deccancollege.ac.in,

Abstract— Traffic accident severity prediction supports emergency response and road safety policy, but real-world adoption requires reliable performance and transparent explanations. This paper presents an end-to-end, reproducible machine learning pipeline for multi-class accident severity prediction using a large-scale crash dataset. The workflow includes schema validation, missing-value handling, stable categorical encoding, temporal feature extraction, class-imbalance-aware learning, cross-validated model selection, and explainability using SHAP. We benchmark three gradient-boosted tree ensembles (XGBoost, LightGBM, CatBoost) using balanced accuracy, macro precision/recall, and macro F1-score. LightGBM achieves the best overall results (balanced accuracy 0.7028, macro F1 0.4054). System testing is performed across pipeline integrity, preprocessing stability, performance reliability, and explainability consistency to validate correctness and reproducibility. The resulting framework provides both strong predictive baselines and actionable interpretations for severity-specific safety interventions.

Keywords— Traffic safety analytics; accident severity prediction; gradient boosting; LightGBM; class imbalance; explainable AI; SHAP; system testing.

I. INTRODUCTION

Road traffic crashes remain a major public safety challenge. Predicting the severity of an accident can support faster triage, better allocation of emergency resources, and evidence-based prevention strategies. However, severity modelling is difficult because severe outcomes are relatively rare and depend on non-linear interactions among temporal, spatial, infrastructure, and environmental conditions. Traditional statistical approaches offer interpretability but can be restrictive on high-dimensional real-world data. Modern machine learning (ML), particularly gradient-boosted decision trees, often improves predictive power on tabular data, yet the resulting models must be made explainable to build trust and enable actionable safety insights. This paper targets both goals: (i) benchmark multi-class severity prediction using state-of-the-art boosting models; and (ii) ensure end-to-end reliability and interpretability through system testing and SHAP-based explanations.

II. RELATED WORK

Recent research reports strong performance of ensemble learning for severity prediction on structured crash datasets, while emphasizing the importance of explainability and robustness in safety-critical deployments. Explainable ML using SHAP is increasingly used to quantify feature contributions at global and local levels, bridging predictive accuracy with transparency. Our study follows this direction by benchmarking XGBoost, LightGBM, and CatBoost, and by validating the pipeline through system-level tests that check both data and explanation stability.

III. DATASET AND PROBLEM FORMULATION

A. Dataset

International Journal of Engineering Science and Advanced Technology (IJESAT)
Vol 26 Issue 02(February),2026
Received: 18-12-2025 Accepted: 27-01-2026 Published: 03-02-2026

Experiments use a large-scale US traffic accident dataset with millions of records. Each record includes time and location information and contextual attributes such as road elements and weather. Severity is labelled as an ordinal impact level from 1 (least impact) to 4 (most severe impact).

B. Learning Task

Given a feature vector x describing accident context, the objective is to predict the multi-class severity label y ∈ {1,2,3,4. Because class frequencies are imbalanced, model evaluation prioritizes balanced accuracy and macro-averaged metrics to reflect minority-class performance.

IV. METHODOLOGY

A. Preprocessing and Feature Engineering

The pipeline applies schema validation, consistent datatype enforcement, and missing-value treatment. Temporal fields are expanded into categorical features (start_month, start_year, start_hour, start_day) to capture seasonality and exposure. Categorical variables are encoded with stable mappings learned from the training set and re-used unchanged at inference to prevent leakage and drift.

B. Models and Training

We train three gradient-boosted tree models: XGBoost, LightGBM, and CatBoost. Randomized search with cross-validation is used to reduce sensitivity to a single split and to obtain stable estimates. Class imbalance is handled using class/sample weighting during training and cross-validation.

C. Explainability (SHAP)

To interpret predictions, SHAP explanations are computed for the selected best model and validated for stability. Global SHAP summaries rank feature contributions across the dataset, while local explanations highlight factors driving a specific high-severity prediction.

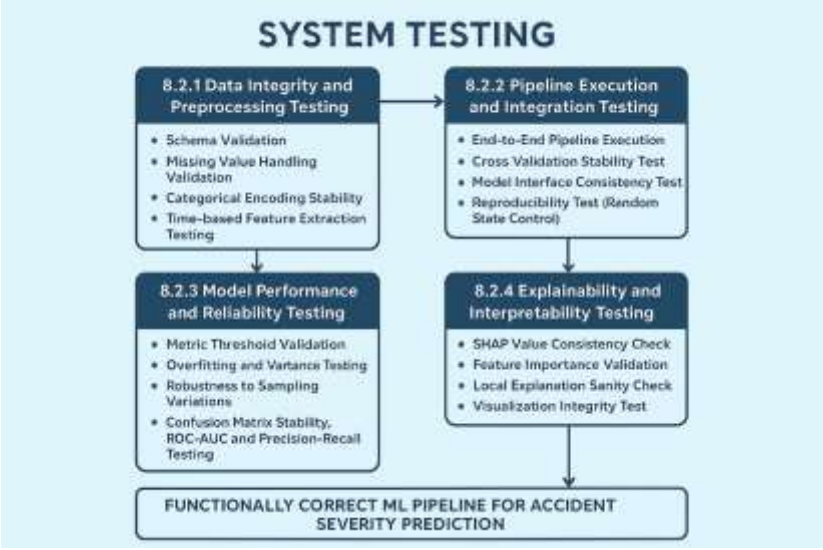

Fig. 1 End-to-End ML Pipeline and System Testing Design

V. EXPERIMENTS AND RESULTS

A. Benchmark Metrics

Models are evaluated on a held-out test set using balanced accuracy, macro precision, macro recall, and macro F1-score. In addition, confusion matrices, ROC curves, and precision–recall curves are used to diagnose class-wise behaviour for high-severity classes.

International Journal of Engineering Science and Advanced Technology (IJESAT)
Vol 26 Issue 02(February),2026
Received: 18-12-2025 Accepted: 27-01-2026 Published: 03-02-2026

TABLE I
OVERALL BENCHMARK SUMMARY (TEST SET)

Model	Balanced Accuracy	Macro Precision	Macro Recall	Macro F1-Score
XGBoost	0.6894	0.3852	0.6894	0.3919
LightGBM	0.7028	0.3930	0.7028	0.4054
CatBoost	0.6337	0.3587	0.6337	0.3454

TABLE II
CLASS-WISE PERFORMANCE FOR LIGHTGBM (TEST SET)

Class	TP	FP	FN	TN	Precision	Recall	F1-Score
1	12,654	93,755	786	1,433,957	0.1189	0.9415	0.2108
2	717,965	42,964	509,782	270,441	0.9435	0.5849	0.7210
3	179,536	263,657	79,701	1,018,258	0.4051	0.6925	0.5132
4	24,122	206,499	16,606	1,293,925	0.1046	0.5923	0.1765
Macro Avg	-	-	-	-	0.3930	0.7028	0.4054

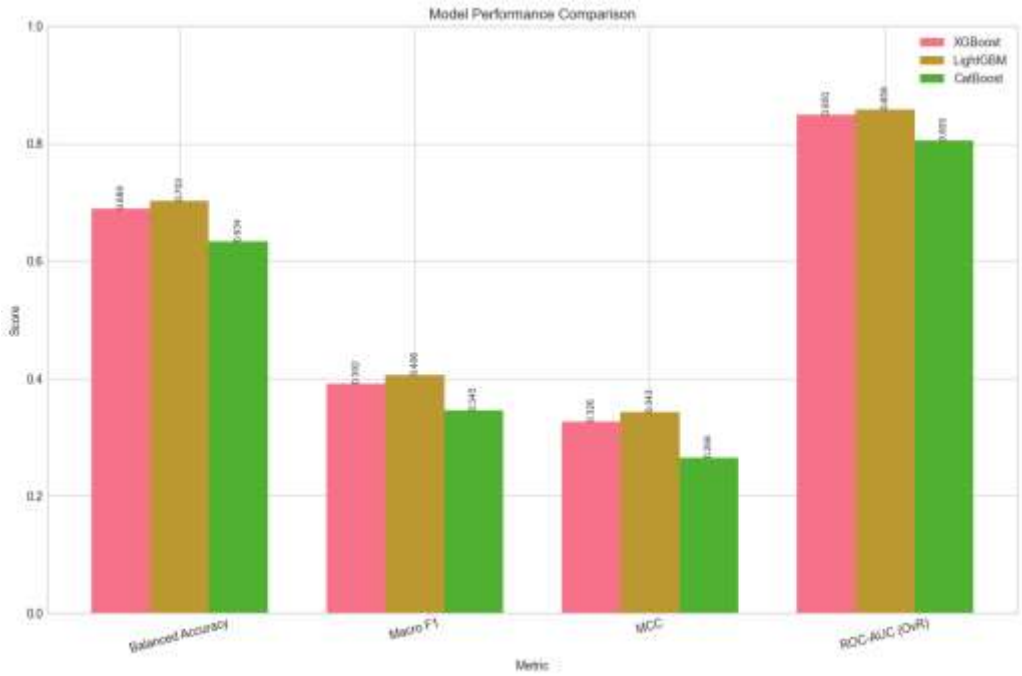

Fig. 2 Model Performance Comparison (Balanced Accuracy, Macro F1, MCC, ROC-AUC)

2250-3676 ijesatj@gmail.com https://www.ijesat.com/

International Journal of Engineering Science and Advanced Technology (IJESAT)
Vol 26 Issue 02(February),2026
Received: 18-12-2025 Accepted: 27-01-2026 Published: 03-02-2026

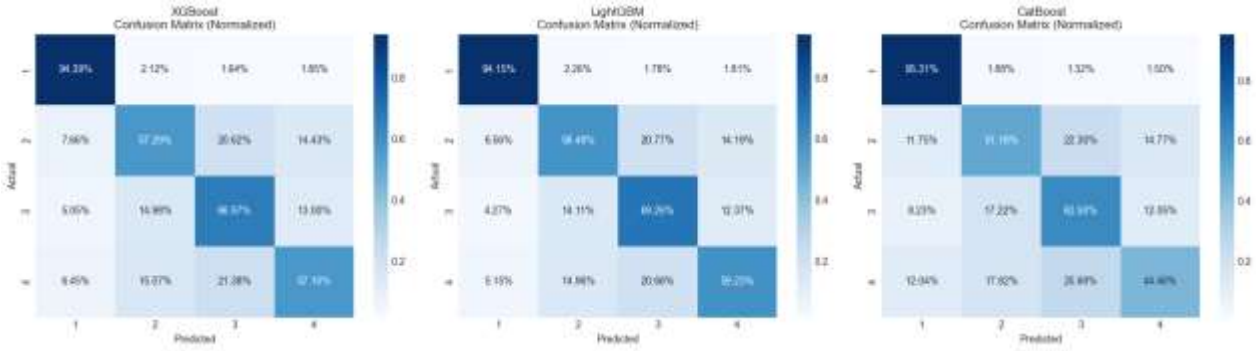

Fig. 3 Normalized Confusion Matrices (XGBoost, LightGBM, CatBoost)

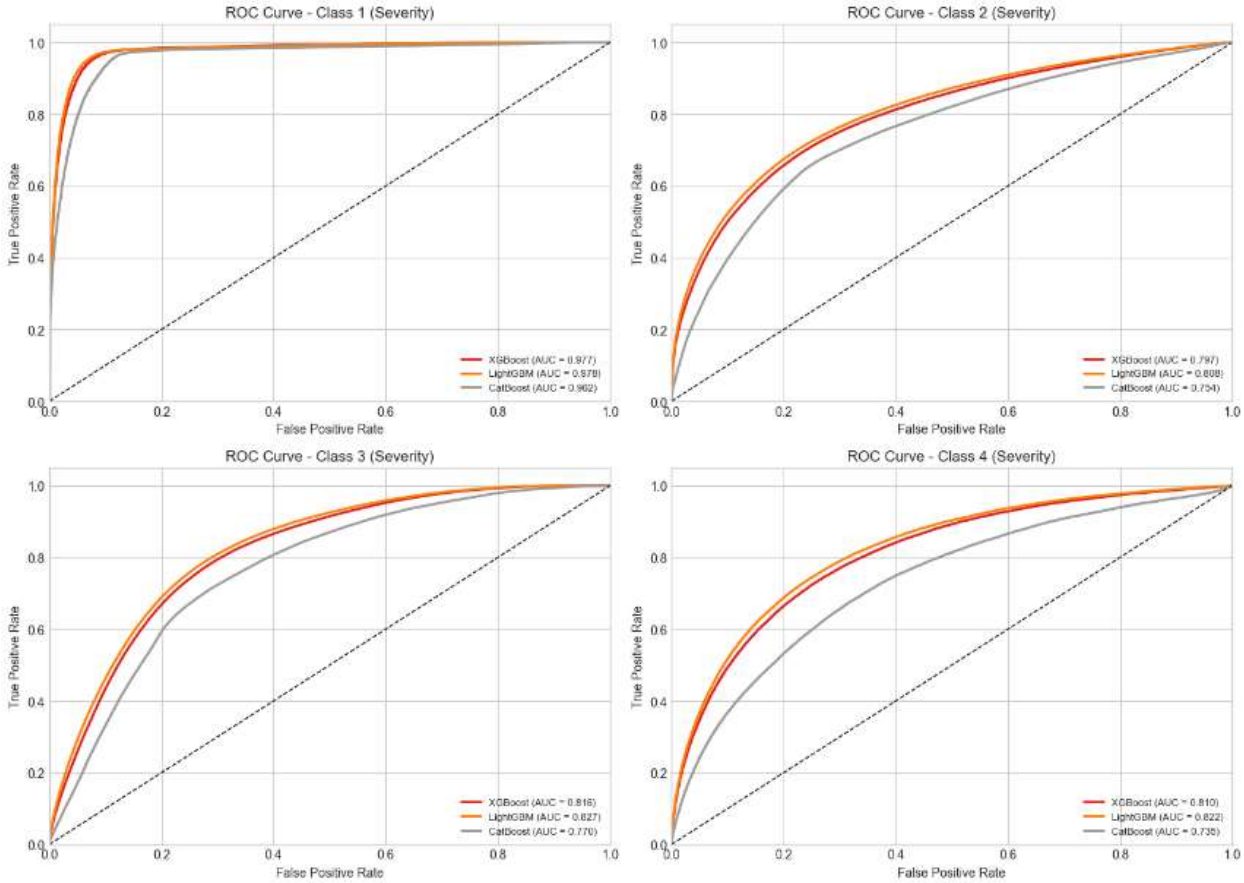

Fig. 4 ROC Curves (One-vs-Rest) for Severity Classes

International Journal of Engineering Science and Advanced Technology (IJESAT)
Vol 26 Issue 02(February),2026
Received: 18-12-2025 Accepted: 27-01-2026 Published: 03-02-2026

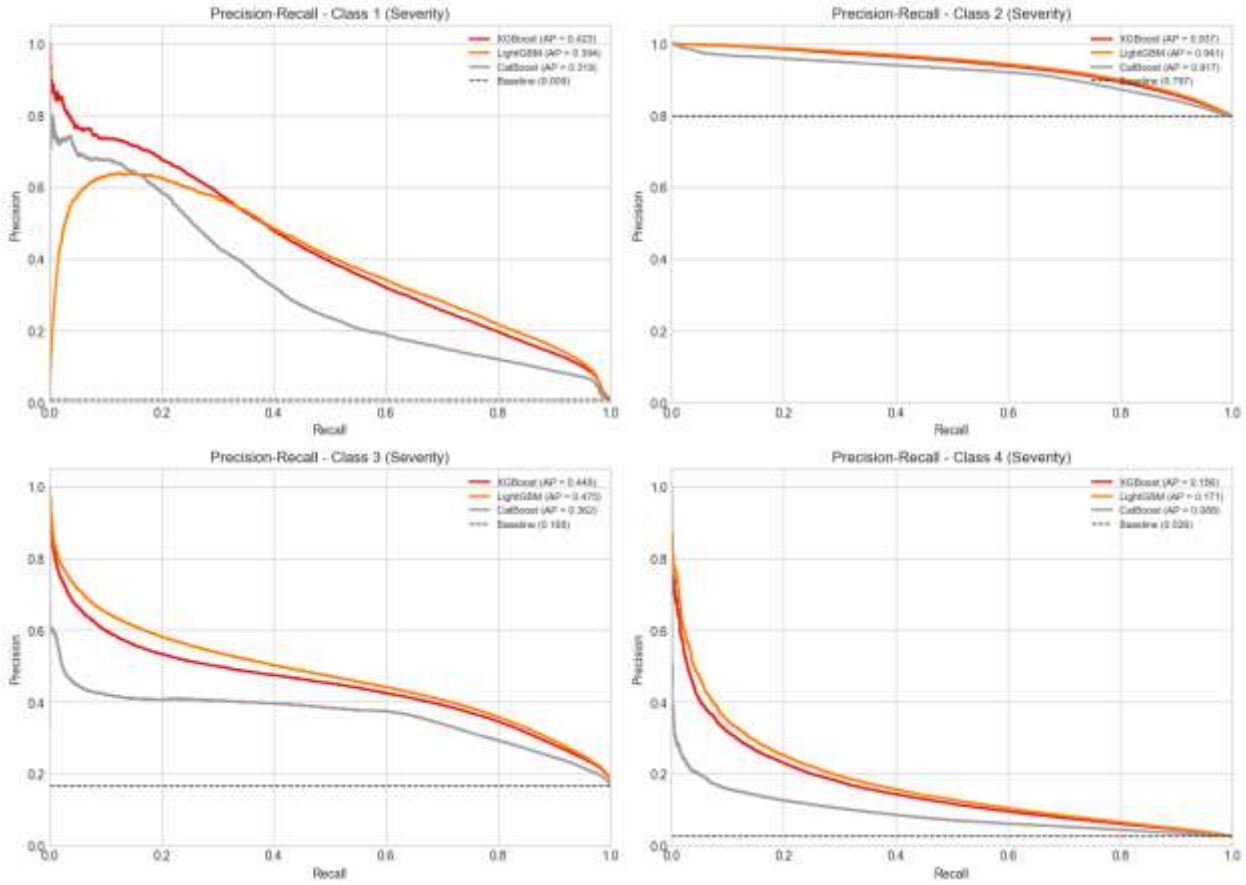

Fig. 5 Precision–Recall Curves for Severity Classes

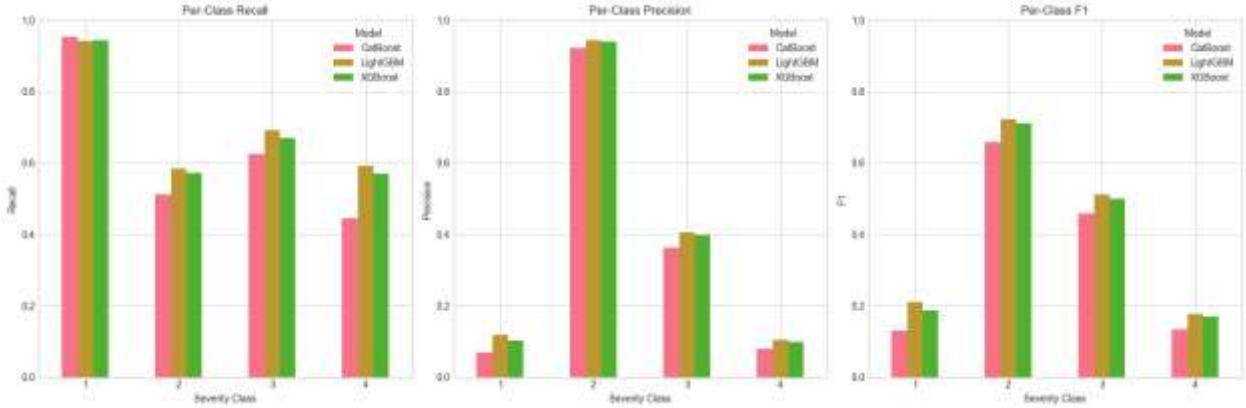

Fig. 6 Per-Class Recall and Precision across Models

B. Benchmark Findings

LightGBM achieved the best macro-averaged performance (balanced accuracy 0.7028, macro F1 0.4054), followed by XGBoost and CatBoost. Across models, class 2 and class 3 show substantially higher F1-scores than classes 1 and 4, reflecting the difficulty of the most severe and least severe classes under class imbalance. The confusion matrices and PR curves confirm that minority high-severity cases remain challenging, motivating both targeted imbalance-aware training and careful operational thresholds.

VI. EXPLAINABILITY AND SYSTEM RELIABILITY

A. Global and Local Explainability

2250-3676 ijesatj@gmail.com https://www.ijesat.com/

International Journal of Engineering Science and Advanced Technology (IJESAT)
Vol 26 Issue 02(February),2026
Received: 18-12-2025 Accepted: 27-01-2026 Published: 03-02-2026

SHAP explanations are used to translate model behavior into feature-level contributions. Global SHAP summaries for the best model highlight the most influential drivers, including temporal exposure variables (start_year, start_month, start_hour), geographic proxies (state), and infrastructure indicators (traffic_signal, crossing, stop). Local explanations for high-severity predictions demonstrate how these factors combine to increase predicted risk for individual instances.

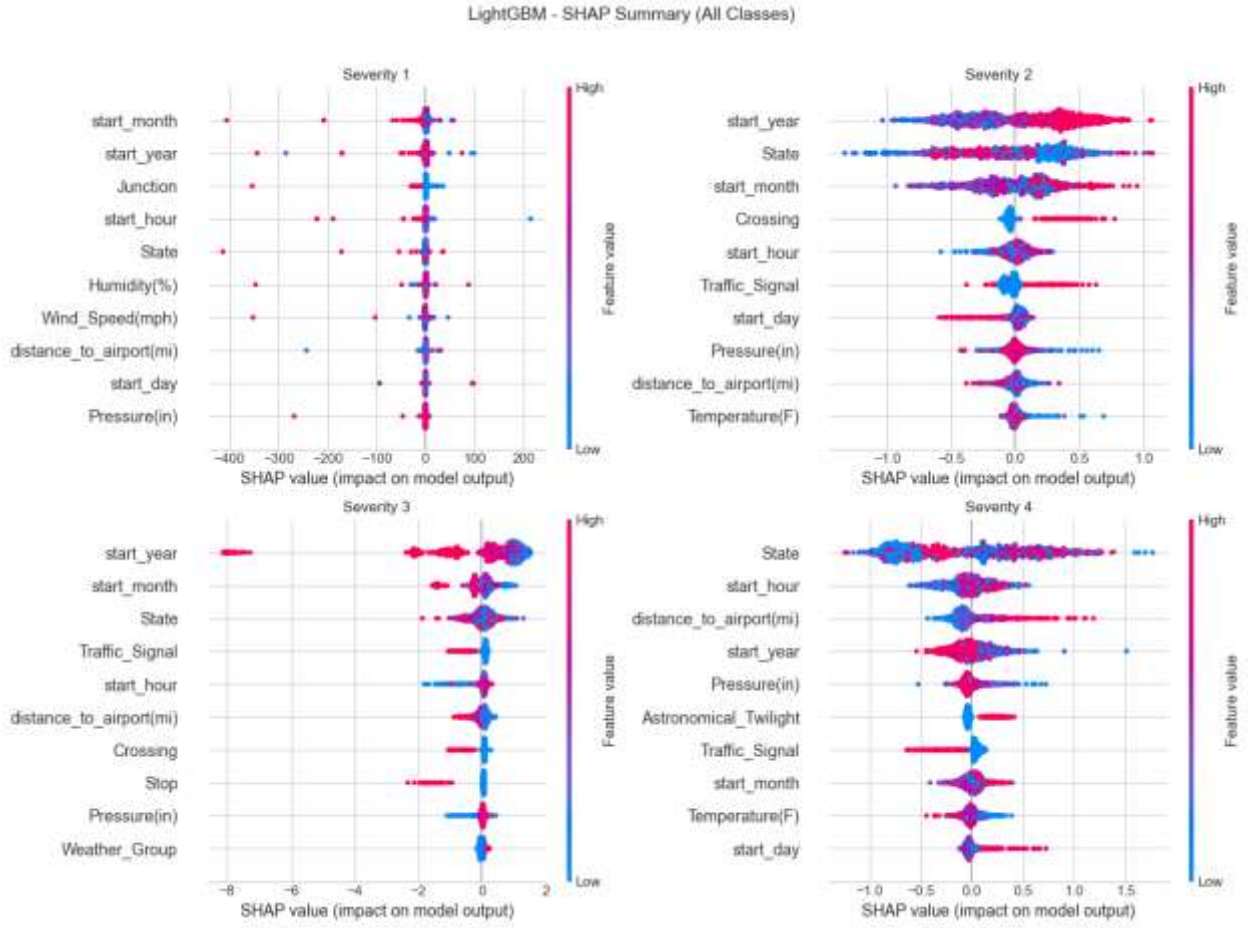

Fig. 7 LightGBM SHAP Summary Plot (All Classes)

B. System Testing and Reproducibility

System testing validates end-to-end correctness beyond traditional software tests. The pipeline is tested in four dimensions: (1) pipeline integrity (successful raw-data-to-SHAP execution), (2) data quality and preprocessing stability (schema, missing values, categorical mappings, time-feature extraction), (3) model performance and reliability (metric thresholds, variance/overfit checks, stability under sampling), and (4) explainability verification (SHAP consistency, alignment with model importances, visualization integrity). These checks confirm that results are reproducible under fixed random seeds and robust to minor data variations.

VII. CONCLUSION

This paper presented a tested, explainable ML pipeline for multi-class traffic accident severity prediction. Benchmarking shows that LightGBM provides the strongest overall macro-averaged performance on the evaluated dataset, while the full suite of plots and per-class metrics reveal remaining challenges for the most severe class. By integrating SHAP-based interpretability and system-level tests for

2250-3676 ijesatj@gmail.com https://www.ijesat.com/

reproducibility and stability, the work delivers both reliable baselines and actionable insights for severity-specific road safety interventions.

ACKNOWLEDGMENT

The authors thank supervisors, reviewers, and the open-source community for tools and libraries that supported this work.

REFERENCES

[1] S. Moosavi, M. H. Samavatian, S. Parthasarathy, and R. Ramnath, "A Countrywide Traffic Accident Dataset," arXiv:1906.05409, 2019.

[2] S. Moosavi, M. H. Samavatian, S. Parthasarathy, R. Teodorescu, and R. Ramnath, "Accident Risk Prediction based on Heterogeneous Sparse Data: New Dataset and Insights," in Proc. 27th ACM SIGSPATIAL Int. Conf. on Advances in Geographic Information Systems, 2019.

[3] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in Proc. 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, 2016, pp. 785–794, doi: 10.1145/2939672.2939785.

[4] Alang, K., Vikram, S., Peddi, S., Gangavarapu, R., & Kandula, N. P. (2025). Cloudless AI: Redesigning AI Infrastructure for Decentralized, Edge-First Architectures. 2025 5th Asian Conference on Innovation in Technology (ASIANCON), 1–8. https://doi.org/10.1109/asiancon66527.2025.11280905

[5] G. Ke et al., "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," in Advances in Neural Information Processing Systems 30 (NeurIPS 2017), 2017.

[6] Todupunuri, A. (2025). Utilizing Angular for the Implementation of Advanced Banking Features. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.5283395.

[7] Rongali, L. P. (2025). DevSecOps for Critical Energy Infrastructure: A Secure and Sustainable Paradigm. https://doi.org/10.36227/techrxiv.175433224.49519285/v1

[8] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in Advances in Neural Information Processing Systems 30 (NeurIPS 2017), 2017.

[9] S. Ahmed, M. A. Hossain, S. K. Ray, M. M. I. Bhuiyan, and S. R. Sabuj, "A Study on Road Accident Prediction and Contributing Factors Using Explainable Machine Learning Models: Analysis and Performance," Transportation Research Interdisciplinary Perspectives, vol. 19, p. 100814, 2023, doi: 10.1016/j.trip.2023.100814.

[10] P. Lagias, G. D. Magoulas, Y. Prifti, and A. Provetti, "Predicting Seriousness of Injury in a Traffic Accident: A New Imbalanced Dataset and Benchmark," arXiv:2205.10441, 2022, doi: 10.48550/arXiv.2205.10441.