Machine Learning for Car Accident Severity Prediction in Thailand: A Comprehensive Data-Driven Study with Critical Analysis and Integration Pathways

Nariman Tursaliev
Asian Institute of Technology
st125983@ait.ac.th

Gaurav Karki
Asian Institute of Technology
st126522@ait.ac.th

Abstract

Thailand experiences some of the highest road fatality rates globally, making early detection and prioritization of high-severity crashes crucial. This study analyzes 81,735 accident records from 2019–2022, applies advanced temporal–spatial–environmental feature engineering, and trains several machine learning models, focusing on XGBoost and LightGBM. Our Exploratory and Comparative Analysis (ECA) reveals important behavioral and environmental risk patterns, while boosted models achieve high recall for identifying severe accidents under significant class imbalance. Beyond prediction performance, we critically examine the real-world usefulness of such models—their limitations, their dependence on data quality, and how they become impactful only when integrated with external technologies such as IoT, CCTV, telematics, navigation systems, and emergency dispatch pipelines. We conclude with deployment pathways and the ethical considerations required for responsible use.

1 Introduction

Road injuries remain a persistent public health crisis in Thailand, with fatality rates exceeding 25 deaths per 100,000 population annually [WHO, 2021]. Many cases involve young motorcyclists, severe speeding, impairment, and nighttime driving. Traditional accident-analysis tools lack the ability to model complex interactions between time, location, weather, and behavior. Machine learning, when applied responsibly, can provide a complementary risk-sensing layer for emergency services, traffic planners, and policy agencies.

However, the usefulness of such models depends on more than accuracy—it depends on data quality, interpretability, operational embedding, and integration with other digital systems. This report therefore provides not only a technical evaluation but also a critical reflection on when and how such systems matter in the real world.

2 Related Works

2.1 Thai ML Studies

EECSS [2024] applied XGBoost, Random Forest, Bagging, Decision Tree, and MLP to 81,735 Thai accident records. XGBoost delivered the highest performance (F1 = 66%, AUC 0.90), especially when using class weighting to handle imbalance.

Thai Journal [2025] analyzed 31,817 accidents using KNN, RF, and XGBoost under undersampling, oversampling, and combined sampling. They demonstrated that recall and precision shift dramatically with resampling, implying operational consequences for emergency planning.

2.2 Global Reviews

A review of 56 studies from 2001–2021 [Review, 2023] found Random Forest and GBMs consistently outperform statistical baselines. The main challenges across countries include:

- chronic class imbalance,
- temporal drift (changing mobility patterns),
- missing behavioral and contextual attributes,
- reporting inconsistencies.

2.3 Our Contribution

Our study expands on previous works by:

- creating cyclic temporal, grid-based spatial, and environmental features,
- training both recall-optimized and F1-optimized boosting models,
- providing a detailed Exploratory and Comparative Analysis (ECA),
- adding a critical perspective on usefulness, risks, and integration with real systems.

3 Dataset Description

We analyze 81,735 accident records from 2019–2022, containing temporal, spatial, environmental, road, and behavioral variables. The target is:

$$\text{severity} = \{0 = \text{low},\ 1 = \text{high}\}.$$

Table 1: Feature Groups in the Dataset

Feature Group	Examples
Temporal	hour, month, weekday, weekend, holiday
Spatial	latitude, longitude, province, grid density
Environmental	weather, lighting (day/night/dawn)
Road type	slope, road category, route group
Vehicle	vehicle type, vehicle count
Behavioral	cause description, hit object
Target	low vs. high severity

4 Exploratory and Comparative Analysis (ECA)

5 Principal Component Analysis (PCA) for Structural Diagnosis

While EDA reveals surface correlations, Principal Component Analysis (PCA) is used here as a diagnostic tool to inspect the intrinsic geometry of the feature space and to verify whether severity labels form natural groupings under linear projection.

All numeric attributes were standardized before PCA. The first two principal components were visualized under different target definitions to examine class compactness and overlap.

5.1 PCA with 3 Severity Levels

When accident severity is encoded into three classes (e.g., low, medium, high), PCA reveals strong geometric overlap among all levels. Despite engineered temporal, spatial, and environmental features, the projected samples collapse into a single dense manifold without reliable separation boundaries.

This indicates that the "medium" class is not statistically separable, but rather a noisy interpolation region. As a consequence:

- class boundaries become unstable,
- decision surfaces fluctuate across folds,
- confusion between adjacent classes rises,
- both precision and recall degrade.

5.2 PCA with 2 Severity Levels

After reformulating the target into two classes (non-severe vs. severe), PCA shows clear structural improvement. Although the data is not linearly separable, extreme cases begin to concentrate along dominant variance directions instead of diffusing into the center region.

This results in:

- reduced class overlap,
- tighter grouping,
- improved downstream modeling stability,
- better bias for tree-based ensemble learners.

5.3 Geometric Interpretation

PCA does not prove class separability. It diagnoses information geometry.

The failure of the 3-class projection implies that the middle category does not represent a natural regime but rather an artificial segmentation imposed on a continuous severity spectrum.

Binary encoding consequently yields a cleaner decision problem that aligns with the true statistical structure of the dataset and improves robustness under imbalance and noise.

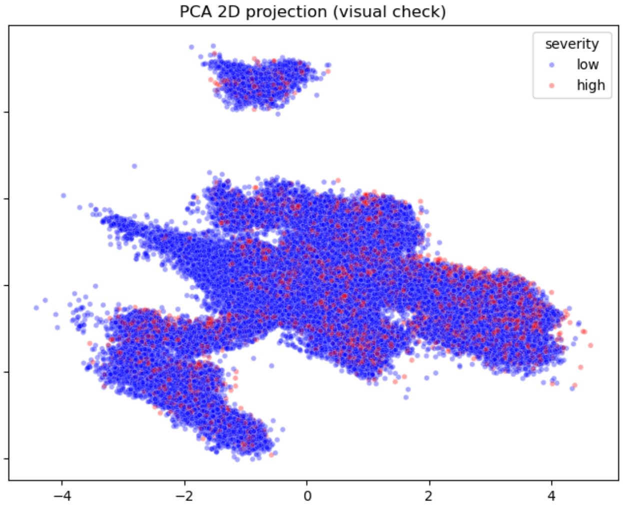

Figure 1: PCA projection (first two components) under binary severity encoding. Compared to the three-class formulation The dataset for accidents according to related works are generally non linear as shown in PCA. Hence, tree based models are preferred.

Understanding risk patterns is crucial for meaningful feature engineering.

5.4 Temporal Patterns

Late-night and early-morning hours show the highest severity proportions, likely due to fatigue, reduced visibility, and alcohol involvement.

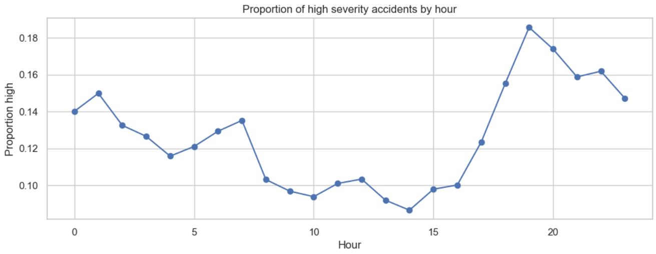

Figure 2: Proportion of high-severity accidents by hour.

5.5 Weather and Environmental Effects

Adverse weather significantly increases severity risk.

5.6 Cause-Based Analysis

Speeding is the most common cause, but alcohol-related crashes have disproportionately higher fatality ratios.

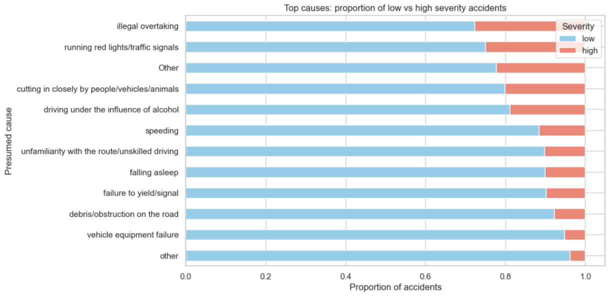

Figure 3: Major causes and their severity proportions.

5.7 Class Imbalance

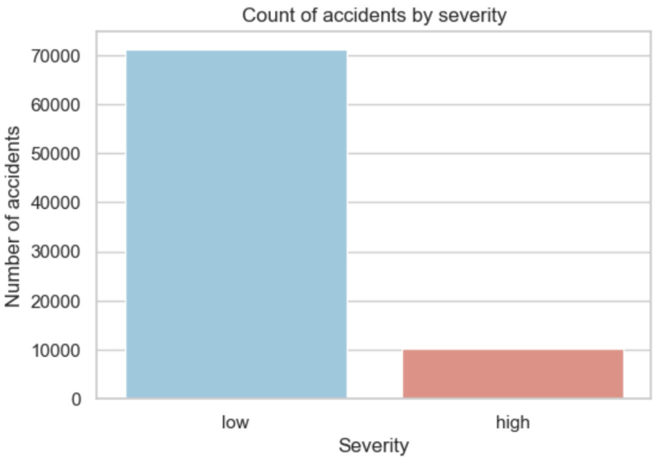

Figure 4: [Upload Image] Severe vs. non-severe class distribution.

Only 13% of accidents are severe, making naive accuracy misleading.

6 Methodology

6.1 Preprocessing Workflow

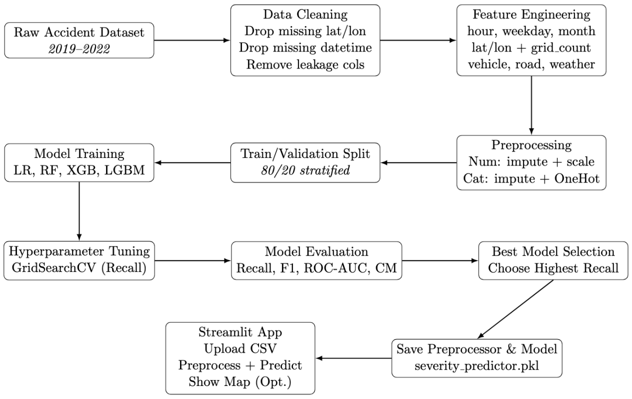

Figure 5: Preprocessing and modeling pipeline.

Key steps:

- cleaning invalid timestamps,
- OneHot + frequency encoding of categorical variables,
- scaling where necessary,
- stratified train/test split.

6.2 Feature Engineering

Table 2: Engineered Features Summary

Feature Type	**Examples**
Cyclic time	hour_sin/cos, month_sin/cos
Temporal context	weekend, holiday, peak hour
Spatial grids	lat/lon rounding, accident density
Environmental	weather, lighting
Behavioral	cause, hit object, vehicles involved

7 Modeling and Training Setup

Models evaluated:

- Logistic Regression,
- Random Forest,
- XGBoost,
- LightGBM.

Each boosting model is trained under:

- recall-optimized objective,
- F1-optimized objective.

Hyperparameters tuned via grid search.

8 Evaluation Results

8.1 Performance Comparison

Table 3: Model Performance Across Metrics

Model	Recall	Prec.	F1	AUC
LR	0.52	0.19	0.28	0.70
RF	0.61	0.24	0.34	0.79
XGB (Recall)	0.76	0.30	0.43	0.82
XGB (F1)	0.71	0.36	0.48	0.84
LGBM (Recall)	0.74	0.32	0.45	0.83
LGBM (F1)	0.49	0.45	0.47	0.83

8.2 Confusion Matrices

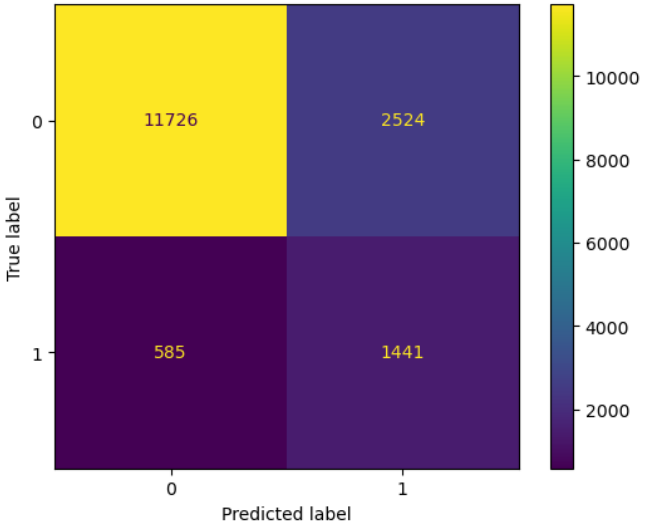

Figure 6: XGBoost Recall-Optimized Confusion Matrix.

9 Feature Importance

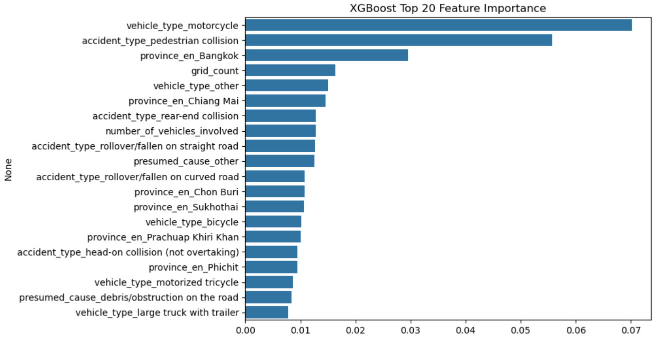

Figure 7: Feature importance from XGBoost.

10 Critical Reflection: Usefulness and Integration

A severity prediction model **does not prevent accidents** by itself. Its usefulness depends on where it is embedded.

10.1 Limitations and Risks

The model is constrained by:

- reporting bias,
- missing causal variables,
- temporal drift,
- rural under-reporting,
- limited behavioral detail.

False precision can mislead agencies if predictions are not contextualized.

10.2 Where It Becomes Useful

The model becomes meaningful only when integrated into broader systems:

1. Emergency Dispatch Systems. Severity predictions at call time can help allocate ambulances, advanced medical teams, or route patients to specialized hospitals.

2. IoT and Vehicle Telematics. Integrating real-time speed, braking, weather, and GPS signals enables personalized risk alerts to drivers.

3. Smart Traffic Infrastructure. Severity-risk maps combined with CCTV/ANPR allow dynamic control of signal timing, speed limits, and hazard warnings.

4. Navigation Apps (Google Maps, HERE, Waze). The model can trigger high-risk warnings ("dangerous curve ahead during rain") based on patterns in the data.

5. Policy and Urban Planning. GIS-based dashboards using model outputs help planners target high-risk zones for lighting improvements, guardrails, or enforcement.

10.3 Conditions for Real Impact

The model becomes truly useful when:

- it is retrained at a certain time interval and monitored,
- predictions are combined with human decision-making,
- ethical constraints (privacy, surveillance concerns) are respected,
- it is deployed within existing workflows rather than as an isolated dashboard.

11 Conclusion

Our model provides a strong severity classification pipeline under severe imbalance. However, its real-world value emerges only when embedded into larger technological ecosystems such as smart infrastructure, emergency dispatch, or navigation systems. Stand-alone predictions have limited impact; integrated predictions can meaningfully support life-saving decisions.

A Appendix: Figures (Placeholders)

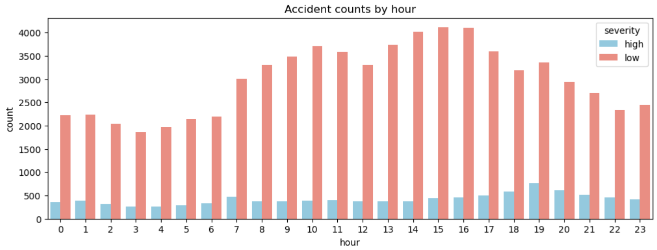

Figure 8: Hourly accident distribution.

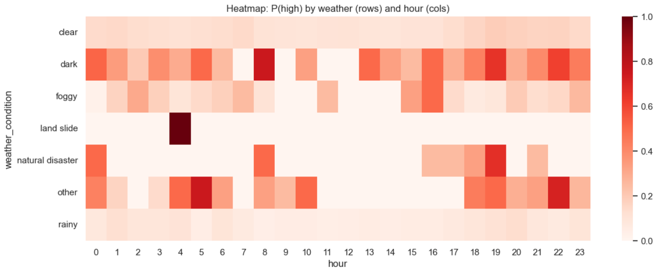

Figure 9: Severity by weather category.

References

World Health Organization (2021). Global Status Report on Road Safety.

EECSS (2024). Prediction of Severity Level of Road Traffic Accidents in Thailand Using Machine Learning.

Thai Journal Online (2025). Effect of Resampling Techniques on ML Models for Accident Severity in Thailand.

Author(s) (2023). A Review of ML Algorithms for Crash Injury Severity (2001–2021).