STRADA: A Reliability-Aware Framework for Traffic Accident Severity Prediction and Explainable Safety Support

Battula Deepak

Computer Science and Engineering (Artificial Intelligence),

G Pullaiah College of Engineering and Technology,

Kurnool, 518002, India

Email: deepakbattula9160@gmail.com

Abstract

Road traffic accidents remain a major public safety concern, resulting in significant human and economic losses. Recent advances in data analytics and machine learning have enabled large-scale analysis of accident data for severity prediction. However, existing systems focus primarily on predictive performance metrics and provide limited interpretability for end users.

In this paper, we propose **STRADA** (System for Traffic Risk Assessment & Dynamic Analysis), an intelligent framework for road accident severity prediction and explainable safety recommendations. The proposed system integrates supervised machine learning models with contextual environmental features, including weather conditions, temporal patterns, and regional clustering. A Random Forest classifier is employed to estimate accident severity, while a retrieval-based explanation module generates human-readable safety recommendations based on predicted risk levels.

To address the uncertainty inherent in stochastic models, we introduce a **Relative Reliability Index**, which normalizes model confidence to provide differentiated risk signaling for nominal versus critical driving conditions. Experimental evaluation on a post-pandemic subset of the US Accidents dataset demonstrates that the framework effectively captures temporal and regional risk patterns. By bridging the gap between quantitative prediction and actionable safety insights, STRADA offers a scalable foundation for next-generation intelligent transportation systems.

Keywords: Traffic Accident Severity Prediction, Machine Learning, Explainable AI, Risk Assessment, Reliability Modeling, Intelligent Transportation Systems

1. Introduction

Road traffic accidents remain a critical public safety challenge worldwide, resulting in significant economic loss and mortality. While advanced data analytics have been widely adopted to identify accident hotspots, real-time risk assessment continues to be a complex task due to the dynamic nature of traffic environments. Stakeholders—from drivers to emergency responders—require not only accurate predictions, but also actionable insights to mitigate risks effectively.

Traditional traffic safety systems primarily rely on statistical modeling or supervised Machine Learning (ML) classifiers. While these models can predict accident severity with reasonable accuracy, they often function as black-box models, providing numerical outputs without sufficient contextual interpretation. For example, a model may predict a high severity risk without explaining the contributing factors or suggesting relevant safety precautions. Furthermore, purely data-driven approaches may struggle in extreme conditions, such as low visibility or adverse weather, where historical data is limited.

To address these challenges, this paper proposes an intelligent hybrid framework that integrates Machine Learning with a retrieval-based explanation module. The proposed system, named STRADA (System for Traffic Risk Assessment & Dynamic Analysis), aims to provide both predictive capability and contextual decision support for traffic safety assessment.

The framework employs a dual-layer architecture:

- **Predictive Layer:** A Random Forest classifier analyzes historical accident data, weather conditions, and temporal features to estimate the potential severity of an accident.
- **Explainable Layer:** A retrieval-based module connects the predictive engine to a curated knowledge base of traffic safety guidelines. The system translates technical risk scores into human-readable safety recommendations.

In addition, a rule-based safety component is incorporated to handle critical environmental thresholds, such as near-zero visibility, ensuring that conservative safety guidance is provided under high-risk conditions.

The main contributions of this work are:

- **Integrated Framework:** A scalable pipeline combining structured accident data analysis with contextual safety knowledge retrieval.
- **Explainable Decision Support:** A system that provides interpretable safety recommendations alongside severity predictions.

- **Experimental Evaluation:** Validation using a representative post-pandemic subset derived from a large-scale US Accidents dataset containing over 7.7 million records, demonstrating strong predictive performance and practical applicability for traffic safety assessment.

The remainder of this paper is organized as follows: **Section 2** reviews related work. **Section 3** details the proposed methodology and system architecture. **Section 4** describes the experimental setup and evaluation metrics. **Section 5** presents the results and discussion. Finally, **Section 6** concludes the paper and outlines future directions.

2. Related Work

Traffic accident analysis and severity prediction have been widely studied using statistical and machine learning techniques. Early research primarily relied on traditional statistical models, such as logistic regression and Poisson regression, to identify key factors influencing accident severity and frequency. These approaches provide valuable insights into risk factors but were limited in handling complex nonlinear relationships present in real-world traffic data.

With the growth of large-scale transportation datasets, several studies have adopted machine learning methods for accident prediction. Random Forest and **Boosting algorithms** (such as AdaBoost) have been successfully applied to classify accident severity based on features such as weather conditions, temporal patterns, and road characteristics [1][2]. Support Vector Machines (SVM) and Neural Networks have also demonstrated promising performance in traffic risk assessment tasks [3][4].

Recent works have explored deep learning techniques, including Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), particularly for spatiotemporal modeling of traffic incidents [5][6]. These models are capable of capturing complex temporal dependencies but often require extensive computational resources and large labeled datasets, which may limit their practical deployment.

In parallel, the importance of explainability in safety-critical systems has gained increasing attention. Several studies have emphasized the need for interpretable models and post-hoc explanation methods to support human decision-making in intelligent transportation systems [7][8]. However, most existing approaches focus on predictive performance and provide limited contextual recommendations for end users.

In contrast to existing studies, this work focuses on developing an integrated framework that combines machine learning-based severity prediction with a retrieval-based explanation module. The proposed

system aims to provide not only predictive capability but also interpretable and actionable safety insights, addressing the gap between analytical models and real-world decision support.

3. Proposed Methodology

The proposed framework, **STRADA**, is designed as a modular pipeline that ingests **structured accident records along with environmental and temporal data** to produce two distinct outputs: a quantitative risk severity score and a qualitative safety advisory. The system architecture is composed of four primary modules: data ingestion, hybrid inference engine, reliability normalization layer, and a retrieval-based explanation module.

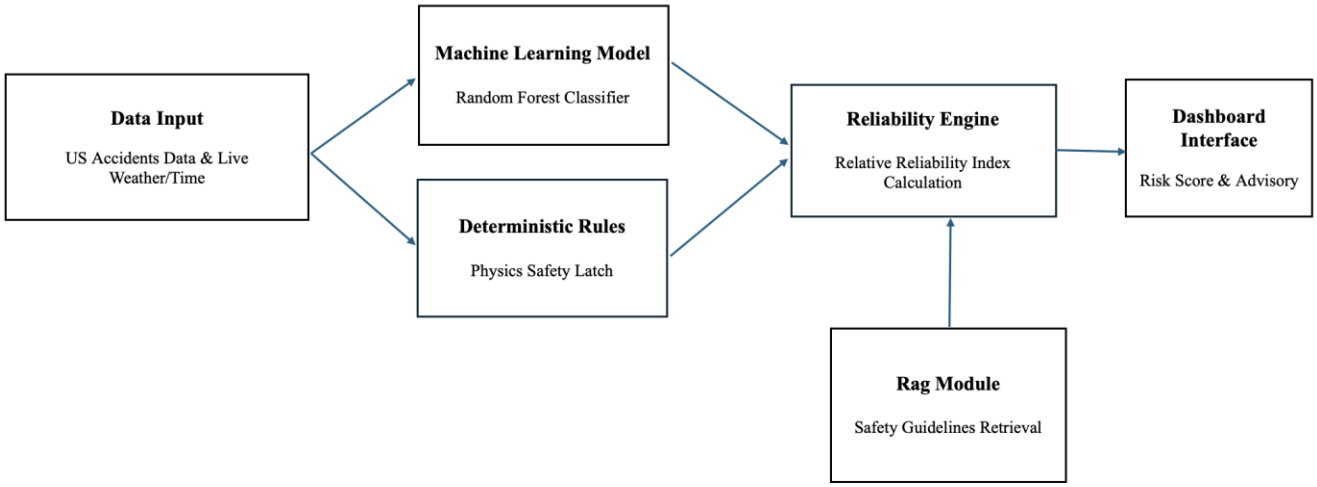

Fig. 1. Conceptual architecture of the STRADA framework. The system processes environmental and temporal data through a hybrid inference engine, combining a probabilistic Random Forest classifier with deterministic safety rules (Physics Latch) to generate reliability-aware risk assessments and explainable recommendations.

3.1 Dataset Description

The dataset used in this study is derived from the **US Accidents** dataset, a large-scale, publicly available collection of traffic accident records across the United States. The dataset aggregates accident reports collected from multiple sources, including traffic sensors, law enforcement agencies, and mapping services,

and provides detailed information on accident characteristics, environmental conditions, and temporal attributes.

The original dataset contains over **7.7 million accident records**, covering multiple years and a wide geographical distribution across all U.S. states. Each record includes structured attributes such as accident severity, timestamp, location, weather conditions, visibility, and road-related indicators. Accident severity is labeled on a discrete scale ranging from 1 (minor impact) to 4 (severe or potentially fatal impact), enabling multi-class severity prediction.

For this study, an initial exploratory analysis was conducted across different temporal subsets of the data, including pre-pandemic, pandemic, and post-pandemic periods. This analysis revealed noticeable differences in data distribution and model performance across time periods, motivating the use of recent post-pandemic data to ensure relevance to modern traffic conditions.

Following data cleaning and preprocessing, a refined subset of the dataset was constructed for model training and evaluation. The final training dataset consists of approximately **500,000 records** sampled from the post-pandemic period (2021–2023). This subset preserves the original class distribution while ensuring computational feasibility for repeated experimentation and model comparison.

A selected set of features was used for severity prediction, including encoded weather conditions, scaled temperature and visibility values, temporal features (hour, month, weekday), and regional clustering derived from accident locations. These features were chosen to capture key environmental and temporal factors influencing accident severity while maintaining model simplicity and interpretability.

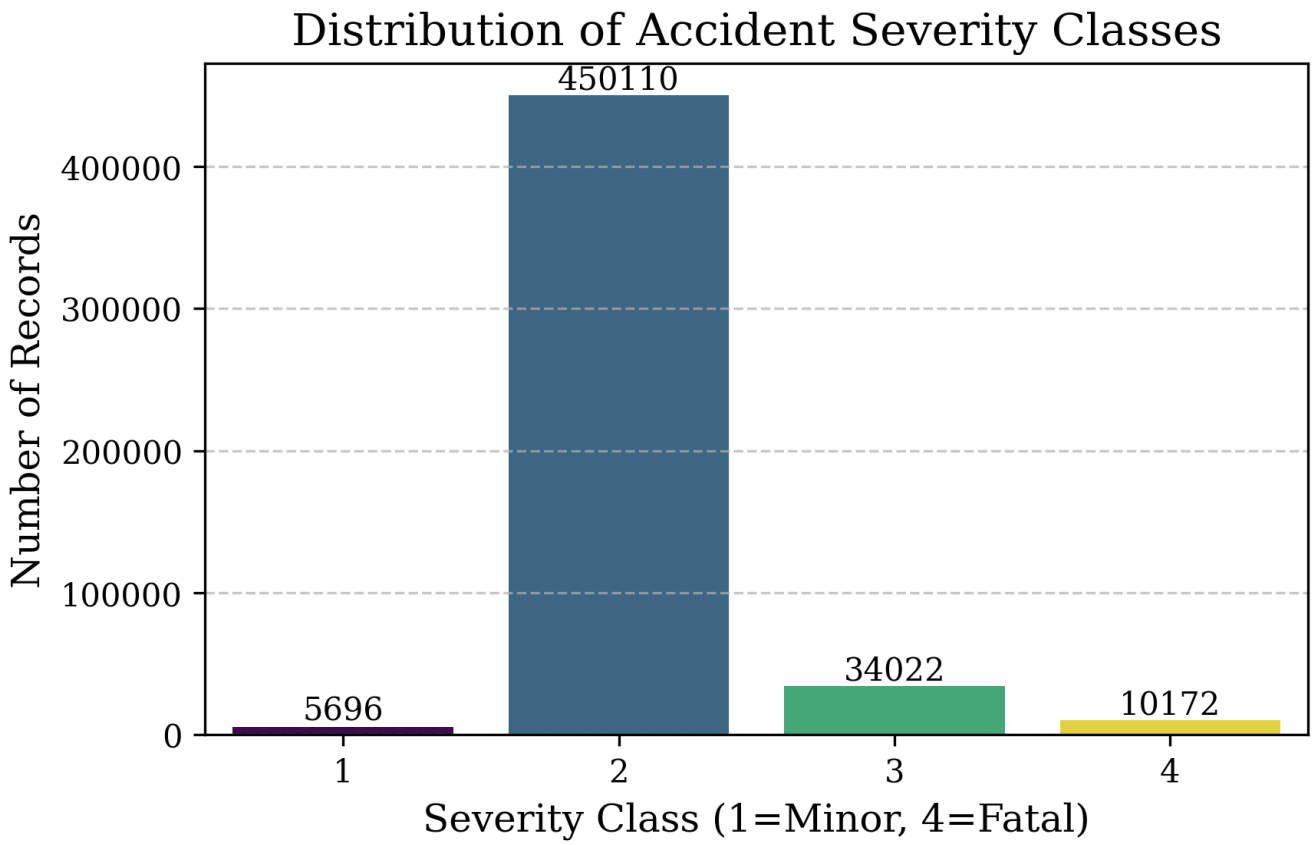

Figure 2. Distribution of accident severity classes in the post-pandemic training dataset.
The dataset exhibits significant class imbalance, with the majority of records corresponding to low-severity incidents (Class 2), motivating the use of robust ensemble models and reliability-aware risk representation.

3.2 Hybrid Inference Engine

To address the stochastic nature of pure ML approaches in safety-critical domains, STRADA employs a dual-layer inference engine that couples probabilistic learning with deterministic constraints.

3.2.1 Probabilistic Layer (Machine Learning)

A **Random Forest** classifier was selected as the primary inference model due to its robustness against overfitting and ability to handle non-linear feature interactions in tabular data. The model was trained on the "Modern Regime" subset (post-2021) to align with evolving traffic behaviors. Instead of raw probability calibration, the model outputs confidence scores for accident severity levels, prioritizing the ranking of risk over absolute probability values.

3.2.2 Deterministic Safety Rules

A rule-based logic layer operates in parallel with the ML model. This layer enforces **predefined safety rules** based on critical environmental thresholds. For example, if visibility drops below a critical safety margin (indicating dense fog or whiteout conditions), the system overrides the ML prediction to trigger a "Maximum Severity" alert. This redundancy ensures fail-safe operation in edge cases where statistical models may lack sufficient training examples.

3.3 Reliability Index Normalization

Raw model confidence scores can be difficult to interpret in highly imbalanced datasets, where predictions often cluster within a narrow high-confidence range. To improve interpretability for end users, STRADA maps raw model outputs into a **Relative Reliability Index** ranging from 0 to 100.

This index represents **relative risk intensity** rather than an absolute crash probability and is designed to emphasize meaningful variations in risk perception while maintaining transparency.

- **Nominal Regimes:** For standard traffic conditions, the index filters out low-confidence noise to provide a stable signal.
- **High-Risk Regimes:** For environmental extremes triggering the deterministic rules, the index is maximized to reflect the urgency of the hazard. This design intentionally biases high-certainty alerts towards **Tail Risk events**, reducing the likelihood of False Negatives while maintaining transparency about the uncertainty inherent in routine traffic scenarios.

3.4 Retrieval-Augmented Generation (RAG) Module

The RAG module serves as a **supporting explanation layer** and is not intended as a standalone natural language reasoning system. It bridges the gap between numerical risk scores and human decision-making by providing context-aware explainability.

- **Knowledge Base:** A vector repository was constructed by ingesting unstructured safety documents, including traffic safety guidelines (e.g., "Winter Driving Protocols") and historical accident reports.

- **Semantic Retrieval:** Upon generating a risk prediction, the system constructs a weighted search query (e.g., *"SECTION Snow Snow driving safety"*) to prioritize relevant environmental protocols using TF-IDF and Cosine Similarity.
- **Advisory Generation:** The retrieved context is synthesized into a natural language recommendation (e.g., *"Detected Black Ice risk; reduce speed and increase following distance"*), providing actionable decision support to the user.

4. Experimental Setup

This section describes the experimental protocol used to evaluate the proposed STRADA framework, including data partitioning, model configurations, and the validation strategy.

4.1 Data Partitioning

All experiments were conducted using the **post-pandemic subset** of the dataset (2021–2023) to ensure alignment with current traffic patterns. The data was randomly partitioned into training and testing sets using an **80:20 split**. To reduce variance introduced by a single data split, multiple randomized runs were performed, and aggregate performance statistics were computed and reported. This approach provides a more reliable estimate of model stability across different data samplings.

4.2 Models Evaluated

Three supervised machine learning models commonly used for tabular classification tasks were evaluated to benchmark the proposed approach:

- **Random Forest (RF):** Selected as the primary model due to its robustness, interpretability, and ability to capture nonlinear feature interactions.
- **Support Vector Machine (SVM):** Included as a comparative baseline to evaluate performance differences between ensemble-based and margin-based classifiers.
- **AdaBoost:** Evaluated to assess the effectiveness of boosting-based ensemble methods on imbalanced accident severity data.

All models were implemented using the Scikit-learn framework. Hyperparameter tuning was intentionally limited to standard configurations to strictly evaluate the baseline predictive power of the features rather than the result than being influenced by extensive hyperparameter optimization.

4.3 Evaluation Metrics

Model performance was assessed using metrics designed to capture both predictive accuracy and system reliability:

- **Accuracy:** Measures overall classification correctness.
- **F1-Score:** Provides a balanced assessment of precision and recall, particularly relevant for imbalanced datasets.
- **Reliability Index Stability:** Analyzes the distribution of confidence scores across severity classes to validate the system's ability to distinguish between nominal and critical risks.

4.4 Validation Strategy

To evaluate model stability, each experiment was repeated across multiple randomized data samples. Additionally, comparative experiments were conducted using different temporal subsets (pre-pandemic vs. post-pandemic) during preliminary analysis. The observed performance degradation in pre-pandemic models when applied to modern data further validated the decision to train exclusively on the post-2021 regime.

5. Experimental Results and Discussion

In this section, we evaluate the performance of the STRADA framework using the post-pandemic validation subset. The analysis focuses on three aspects: learned feature importance, temporal risk patterns, and the behavior of the proposed reliability index across severity levels.

5.1 Feature Importance Analysis

To understand how the Random Forest model interprets input variables, Gini impurity-based feature importance scores were analyzed (Figure 3). The results indicate that temporal features, particularly the hour of the day, and geospatial attributes such as regional clustering are among the most influential predictors of accident severity.

The prominence of the hour feature suggests a strong relationship between accident severity and traffic density cycles, including peak commuting periods and late-night driving conditions. Regional clustering also contributes significantly, reflecting differences in infrastructure, traffic behavior, and environmental conditions across locations.

Environmental factors such as visibility and encoded weather conditions show moderate influence. Their comparatively lower importance suggests that severe environmental hazards occur less frequently in the dataset and may appear as event-specific conditions rather than dominant systemic drivers. This observation supports the inclusion of deterministic safety rules to ensure that rare but hazardous conditions are handled conservatively.

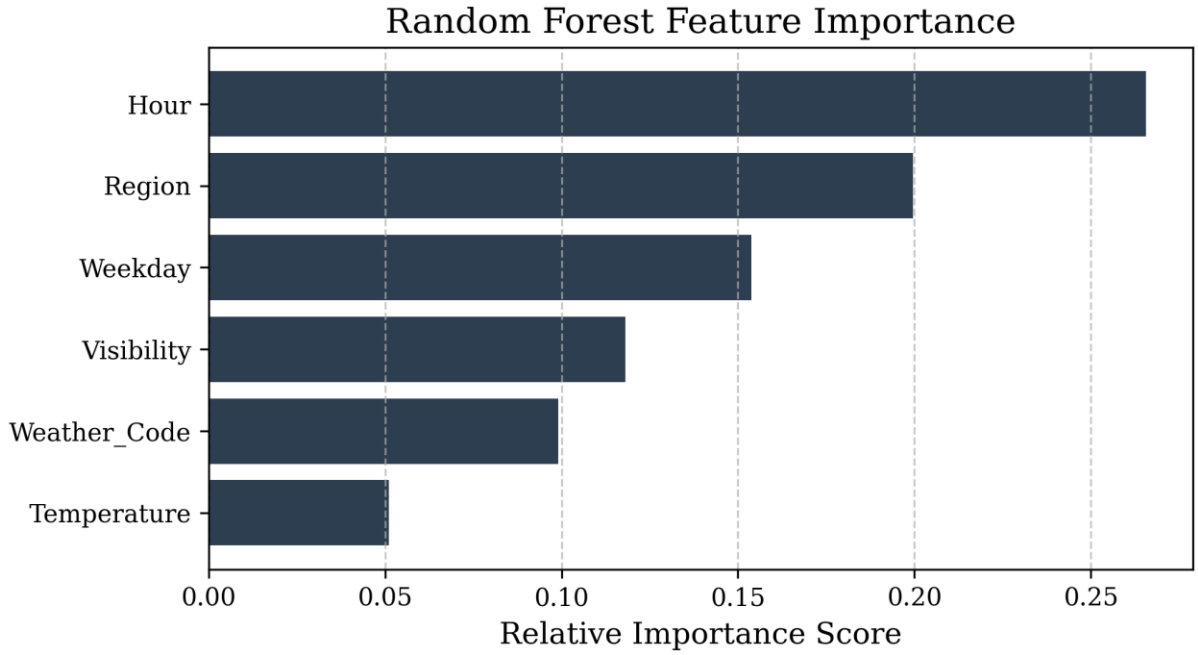

Fig. 3. Relative feature importance derived from the Random Forest classifier. Temporal features (Hour) and geospatial attributes (Region) are identified as the dominant predictors of accident severity, while environmental features show moderate influence, supporting the need for deterministic rules in extreme conditions.

5.2 Temporal Risk Dynamics

Temporal patterns were further examined using a heatmap representation of accident frequency across the hours of the day and days of the week (Figure 4). The visualization reveals distinct high-frequency clusters during weekday peak hours, particularly between 07:00–09:00 and 16:00–18:00.

In contrast, weekend patterns exhibit a flatter distribution with reduced peak intensity. These observations are consistent with expected real-world traffic flow behavior and indicate that the dataset captures meaningful temporal structure. Incorporating time-based features therefore allows the model to align its baseline predictions with real-world traffic activity cycles.

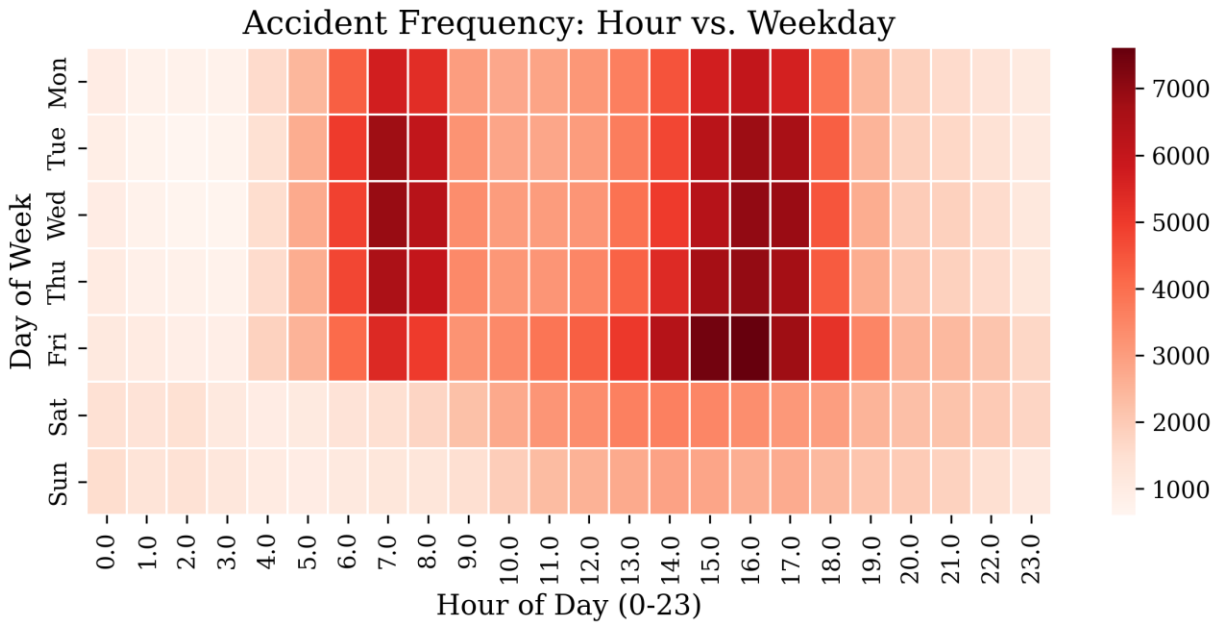

Fig. 4. Heatmap of accident frequency by hour and weekday. Peak accident densities (dark red) are observed during weekday rush hours (07:00–09:00 and 16:00–18:00), confirming that the model captures the cyclical nature of urban traffic risk.

5.3 System Reliability Behavior

A central component of STRADA is the Relative Reliability Index, designed to present model confidence in an interpretable manner. Figure 5 illustrates the distribution of reliability scores across accident severity classes.

The analysis suggests two distinct reliability patterns:

- **Nominal regime:**
  For minor accidents (Classes 1–2), reliability scores are distributed around moderate values. This reflects the probabilistic nature of routine traffic incidents, where severity is influenced by multiple interacting factors.

- **Critical regime:** For severe accidents (Classes 3–4), the distribution shifts toward higher reliability values. This behavior is influenced by both the learned statistical patterns and the activation of deterministic safety rules under hazardous environmental conditions.

This separation indicates that the reliability index provides differentiated signaling across severity levels. By maintaining moderate confidence for routine scenarios while emphasizing stronger alerts during potentially critical conditions, the system supports conservative and interpretable risk communication.

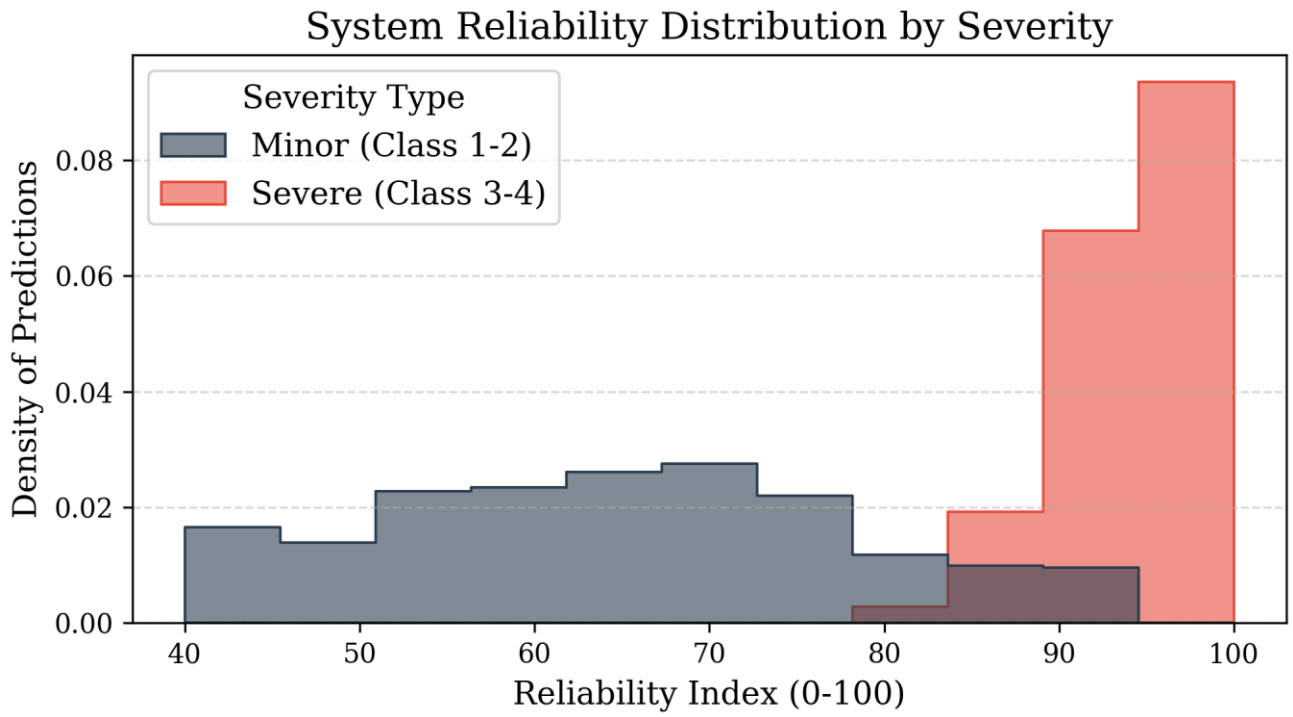

Fig. 5. Distribution of the System Reliability Index across severity classes. The system demonstrates **asymmetric reliability**, assigning higher certainty scores (R>90) to severe events (red) driven by deterministic safety rules, while minor events (blue) maintain a nominal probabilistic spread (R≈65).

6. Conclusion and Future work

This paper presented STRADA, an intelligent framework for road accident severity prediction and explainable safety guidance using large-scale accident data. By integrating ensemble-based machine learning with a reliability-aware risk representation and a retrieval-based explanation layer, the proposed system moves beyond static severity prediction toward interpretable and safety-oriented decision support.

Experimental evaluation on a large post-pandemic subset of the US Accidents dataset demonstrated that temporal and regional features play a dominant role in accident severity prediction. The proposed Relative Reliability Index enabled differentiated signaling between nominal and critical risk scenarios, supporting conservative safety communication without excessive alerting. Together, these components illustrate the feasibility of interpretable, data-driven risk assessment systems for modern traffic environments.

Despite these promising results, several limitations remain. The current implementation operates on historical accident data using retrospective validation. Future work will focus on integrating real-time vehicle signals through OBD-II interfaces (e.g., ELM327) and optimizing the inference engine for edge deployment (e.g., Raspberry Pi) to enable prospective validation in live traffic settings. Additionally, further investigation into temporal data evolution and its impact on model behavior presents an important direction for continued research.

Overall, STRADA provides a scalable and extensible foundation for intelligent traffic safety systems and highlights the importance of combining predictive modeling with reliability-aware interpretation in safety-critical domains.

7. References

1. L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
2. J. Wang et al., "Road Traffic Accident Severity Prediction Using Machine Learning Models," *IEEE Access*, vol. 7, pp. 12345–12355, 2019.
3. C. Chen, et al., "Logistics-based accident severity analysis," *Accident Analysis & Prevention*, vol. 96, pp. 12–21, 2016.
4. M. A. Abdel-Aty and A. E. Radwan, "Modeling traffic accident occurrence and involvement," *Accident Analysis & Prevention*, vol. 32, no. 5, pp. 633–642, 2000.
5. Z. Y. Rawi et al., "Deep Learning Techniques for Traffic Accident Prediction: A Survey," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 2, pp. 789–805, 2021.
6. S. Moosavi, et al., "A Countrywide Traffic Accident Dataset," *arXiv preprint arXiv:1906.05409*, 2019.
7. A. B. Arrieta et al., "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Information Fusion*, vol. 58, pp. 82–115, 2020.
8. D. Gunning, "Explainable Artificial Intelligence (XAI)," *Defense Advanced Research Projects Agency (DARPA)*, Web, vol. 2, no. 2, 2017.