The American Journal of Engineering and Technology
ISSN 2689-0984 | Open Access

TYPE Original Research
PAGE NO. 194-202

Check for updates

Integrated Framework for Reliable Work Zone Crash Classification: Combining Data Validation, Machine Learning Ensembles, and Natural Language Methods

Dr. Mateo Alvarez
Global Institute for Transport Safety, University of Lisbon

OPEN ACCESS

SUBMITTED 11 October 2025
ACCEPTED 18 October 2025
PUBLISHED 31 October 2025
VOLUME Vol.07 Issue 10 2025

CITATION

Dr. Mateo Alvarez. (2025). Integrated Framework for Reliable Work Zone Crash Classification: Combining Data Validation, Machine Learning Ensembles, and Natural Language Methods. *The American Journal of Engineering and Technology*, 7(10), 194–202.

Abstract: This paper presents a comprehensive, publication-ready investigation into the problem of reliable work zone crash classification and risk prediction using an integrated pipeline that emphasizes rigorous data validation, modern machine learning ensembles, and natural language processing of crash narratives. Work zones are high-risk environments on road networks and accurate identification and classification of work zone crashes is essential to enable targeted safety interventions, resource allocation, and reliable research (Yang, 2015; Blackman et al., 2020). Yet, existing operational crash datasets suffer from misclassification, incomplete fields, and inconsistent semantics arising from heterogeneous reporting practices (Swansen et al., 2013; Carrick et al., 2009). We argue that improving data quality through systematic validation and hybrid AI-augmented checks is a prerequisite for robust predictive modeling (Van Der Loo & De Jonge, 2020; Redman, 1998). Building on advances in ensemble learning and hyperparameter optimization (Almahdi et al., 2023; Asadi & Wang, 2023), together with text-mining approaches for narrative analysis (Sayed et al., 2021), we design and describe an end-to-end methodology: (1) a layered data validation and correction module that uses deterministic rules and large language model-assisted anomaly detection; (2) a multimodal feature engineering strategy that integrates structured traffic and environmental data with unstructured narrative-derived features; (3) an ensemble classifier framework that uses stacked learners with hyperparameter tuning to achieve robust classification across varying traffic conditions; and (4) a human-in-the-loop verification stage to capture residual errors and provide continuous feedback for model retraining (Malviya & Parate, 2025; OpenAI, 2023). We

present a descriptive analysis of modeled experimental outcomes and sensitivity studies, discuss theoretical implications, confront limitations, and outline future research directions. The findings demonstrate that combining principled data validation with ensemble learning and narrative text mining materially reduces misclassification rates, produces better calibrated crash-risk scores, and yields interpretability benefits valuable for practitioners and policymakers (Pande et al., 2011; Sayed et al., 2021). This article contributes a detailed procedural blueprint and theoretical rationale for transportation researchers seeking reliable, defensible analytics for work zone safety.

Keywords: Work zone safety; crash classification; data validation; ensemble learning; text mining; machine learning; large language models

Introduction

Background and importance of work zone crash classification

Work zones—temporary modifications to roadway geometry and traffic operations due to construction, maintenance, or incident response—consistently rank among the most hazardous environments for road users and roadway workers (Yang et al., 2015). The presence of narrowed lanes, altered signage, changed speed regimes, and unfamiliar road geometry concentrates exposure and elevates conflict risk. Accurate identification of crashes that occur in work zones is fundamental for multiple downstream tasks: calculating work zone crash incidence and trends, evaluating the safety effects of different work zone configurations, prioritizing investments in countermeasures, and building predictive systems that support real-time risk mitigation (Blackman et al., 2020; Yang et al., 2015). Operationally, transport agencies rely on crash databases aggregated from police reports, employer reports, and administrative logs. However, these sources were not designed with automated classification or machine learning in mind; they are heterogeneous, contain inconsistent semantics, and often omit or mislabel key indicators about work zone involvement (Swansen et al., 2013; Carrick et al., 2009).

Persistent challenges: data quality and misclassification

Several studies have documented systematic misclassification of work zone crashes in official databases (Carrick et al., 2009; Swansen et al., 2013). Misclassification arises for many reasons: transcription errors, ambiguous report fields, the difficulty of geospatial matching when work zones move or are temporary, and human interpretation variance when the decision to label a crash as "work zone-related" is subjective (Blackman et al., 2020). Moreover, narrative fields—free-text sections of crash reports that contain rich contextual information—are underused due to their unstructured nature; yet, they often contain critical cues that can disambiguate borderline cases (Sayed et al., 2021). Beyond misclassification, fundamental data quality problems such as missing values, inconsistent field formats, and temporal mismatches undermine model development and operational adoption (Redman, 1998; Pipino et al., 2002).

Opportunities: machine learning, ensembles, and text mining

Recent advances in machine learning provide powerful tools to address classification and prediction tasks in transportation. Ensemble learning techniques, which combine multiple base learners into a single predictive model, have demonstrated improved robustness and generalization for crash classification across varying traffic conditions (Almahdi et al., 2023; Asadi & Wang, 2023). Hyperparameter optimization and systematic model stacking further enhance predictive performance and stability (Almahdi et al., 2023). Complementary to structured-data modeling, natural language processing (NLP) techniques applied to crash narratives enable automated extraction of salient features such as mentions of work activity, presence of temporary signs, presence of construction vehicles, or explicit statements about closure or flagging operations (Sayed et al., 2021; Swansen et al., 2013). Moreover, modern large language models (LLMs) and fine-tuning approaches can augment text-mining pipelines, improving entity recognition and semantic normalization of incident descriptions (OpenAI, 2023; Achiam et al., 2023).

Need for rigorous data validation as a foundation

While sophisticated models can extract signal from complex datasets, their value is fundamentally constrained by the quality of the inputs. Data validation is not a perfunctory preprocessing step—it is an ongoing, systemic process that must combine automated checks, domain-informed rules, and human review (Van Der Loo & De Jonge, 2020; Pipino et al., 2002). The literature on data quality emphasizes multi-dimensional assessment—completeness, accuracy, consistency, timeliness, uniqueness, and validity—that

must precede trustworthy analytics (Batini et al., 2009; Redman, 1998). Recent proposals to augment data validation with AI-assisted frameworks and hybrid rule-based agents propose to combine the scalability of machine learning with the interpretability and domain constraints of deterministic checks (Malviya & Parate, 2025). These approaches are particularly salient for work zone crash classification because mislabeling not only corrupts model training but also directly influences policy decisions and resource allocation.

Problem statement and research gap

Despite the recognition of these separate components—data validation, ensemble modeling, and narrative mining—there is a paucity of research that synthesizes them into a single, operationalizable pipeline that addresses the end-to-end challenges of work zone crash classification. Existing studies either focus narrowly on improving classifiers with particular algorithms (Almahdi et al., 2023; Asadi & Wang, 2023) or on narrative mining techniques separate from structured-data quality issues (Sayed et al., 2021). The literature lacks a detailed, theoretically grounded, and practically oriented framework that articulates how integrated validation procedures, narrative-derived features, and tuned ensemble methods interact to minimize misclassification and produce calibrated, interpretable predictions for deployment in agency workflows. Additionally, the potential of LLMs to assist in anomaly detection and semantically rich narrative normalization remains underexplored in the work zone safety domain despite promising developments in AI fine-tuning methods (OpenAI, 2023; Achiam et al., 2023; Touvron et al., 2023).

Contributions of this paper

This paper addresses the identified gap by developing and describing an integrated framework that treats data validation as the first-class requirement, leverages narrative text mining, and employs ensemble learning with hyperparameter optimization to improve work zone crash classification. The contributions are fourfold: (1) a principled data validation architecture design that blends deterministic rules with AI-augmented anomaly detection; (2) a multimodal feature engineering approach combining structured fields and narrative-derived semantic features; (3) an ensemble modeling methodology using stacked learners and hyperparameter tuning adapted to heterogeneous traffic conditions; and (4) a descriptive evaluation and sensitivity analysis that characterizes the interplay between validation rigor and classification performance. The methodological exposition is detailed and reproducible, emphasizing transparent decisions and the theoretical rationale underpinning design choices.

Methodology

Conceptual overview

The proposed methodology organizes the problem into sequential modules executed as an integrated pipeline: data ingestion, validation and correction, narrative processing, feature engineering, model training (including ensemble construction and tuning), evaluation and calibration, and deployment-ready verification. The pipeline is intentionally modular: validation outputs feed feature engineering; narrative-derived variables complement structured predictors; ensemble learners are trained on validated data; and a human-in-the-loop verification mechanism continuously monitors deployed model outputs for drift and residual misclassification. The overall architecture is inspired by established data-quality frameworks (Batini et al., 2009; Pipino et al., 2002) and recent hybrid AI approaches for automated validation (Malviya & Parate, 2025; Van Der Loo & De Jonge, 2020).

Data sources and types

The framework presumes access to routinely collected crash data from police and administrative reports, agency work zone logs, and optionally, roadway sensor streams. Typical structured fields include crash date/time, geographic coordinates, road segment identifiers, weather conditions, road class, number of vehicles involved, injury severity indicators, contributing factors (if encoded), and categorical flags purportedly indicating work zone involvement. Additionally, free-text incident narratives, officer remarks, and witness statements provide unstructured context. Prior work has shown the value of narratives for correcting misclassifications and for enriching feature sets (Swansen et al., 2013; Sayed et al., 2021).

Layered data validation module

At the heart of the pipeline is a layered data validation module designed to detect and remedy common defects. The module is organized into the following tiers:

1. Schema and format validation: deterministic checks verify that required fields are present, data types match expectations, and categorical fields conform to valid enumerations. This stage enforces baseline

syntactic integrity (Van Der Loo & De Jonge, 2020).

2.Consistency and cross-field logic checks: this tier applies domain-informed rules such as verifying that a recorded work zone flag is consistent with the presence of an active work-permit record for the location and time, or that injury severity codes align with reported injuries in narrative text. Such rules capture logical consistency and can detect transposition or reporting errors (Pipino et al., 2002).

3.Geospatial-temporal matching: rules reconcile coordinates and roadway identifiers against agency-maintained work zone deployment logs. Moving or mobile work zones are particularly challenging; where special logs exist (e.g., scheduled lane closures with time windows), automated spatiotemporal overlap queries flag potential matches or mismatches (Carrick et al., 2009).

4.Statistical anomaly detection: unsupervised methods scan for outliers across key dimensions—e.g., crashes labeled as work zone-related at atypical speeds or at off-hours incompatible with reported work times. Unsupervised clustering or density estimation detects records with low conformity to population patterns and flags them for review (Van Der Loo & De Jonge, 2020).

5.AI-augmented narrative validation: this innovative tier leverages modern LLMs (fine-tuned on labeled narrative corpora where available) to extract structured assertions from text, such as whether construction activity was present, whether temporary traffic control devices were noted, and whether the crash happened within the active zone. The LLM output is compared against the structured work zone flag; disagreement triggers a confidence-weighted correction suggestion or human review (OpenAI, 2023; Achiam et al., 2023).

6.Human-in-the-loop reconciliation: high-confidence automatic corrections are applied where rules and AI agree; low-confidence or high-impact corrections are queued for human adjudication. This stage preserves accountability and provides labeled examples for subsequent model retraining (Malviya & Parate, 2025).

This layered approach builds on established data-quality principles, emphasizing that rule-based checks reduce trivial errors while AI methods can address semantic inconsistencies that deterministic rules cannot easily capture (Batini et al., 2009; Van Der Loo & De Jonge, 2020).

Narrative text-mining pipeline

The unstructured narrative processing pipeline is tasked with extracting features that capture the semantics of the crash context. The pipeline is designed with the following components:

1.Preprocessing: tokenization, de-identification (where necessary for privacy), normalization of abbreviations and domain-specific shorthand, and sentence segmentation. This stage addresses the idiosyncratic styles of officer-written narratives (Sayed et al., 2021).

2.Entity and event extraction: rule-enhanced named-entity recognition detects mentions of construction equipment, temporary signage, presence of flaggers, closure types (e.g., lane closure vs. shoulder work), and actors (worker, contractor vehicle). Combining rules and statistical models balances precision with recall (Sayed et al., 2021).

3.Relation and temporality resolution: extracting temporal relationships (e.g., "before work started," "during active paving") is crucial to disambiguate whether reported work was active at crash time. Temporal normalization maps relative phrases to absolute timestamps where possible, aligned with structured date/time fields (Swansen et al., 2013).

4.Sentiment and causality cues: although not traditional sentiment tasks, identifying causal language (e.g., "hit a cone," "struck by maintenance truck") helps infer contributory mechanisms. Causal cue detection supplements feature space with mechanism-level descriptors (Sayed et al., 2021).

5.Embedding and semantic feature generation: text embeddings produce dense vector representations capturing narrative semantics. Semantic clusters (e.g., "vehicle struck work vehicle", "struck temporary barrier", "worker struck") become categorical or continuous predictors for downstream classifiers.

6.LLM-assisted normalization: LLMs fine-tuned on domain-specific narrative corpora standardize variant expressions into canonical labels (e.g., mapping "cone" and "traffic delineator" to the same device class). This reduces vocabulary fragmentation and enhances downstream model stability (OpenAI, 2023; Achiam et al., 2023).

Sayed et al. (2021) demonstrated that text-mining techniques can systematically identify misclassified work-zone crashes; our pipeline builds upon that work, extending it with LLM normalization and temporal grounding.

Feature engineering and multimodal fusion

Feature engineering integrates validated structured fields and narrative-derived features. The design emphasizes interpretability and resilience to missingness. Key categories include:

- Static scene attributes: road class, number of lanes, posted speed limit, shoulder condition.
- Dynamic context: time-of-day, day-of-week, weather conditions, traffic volume estimates (if available), and moving average speed measures.
- Work zone descriptors: canonicalized work zone presence (from validation module), closure type (full, partial, shoulder), presence of temporary devices, presence of work vehicles, and flagger activity (derived from narratives and logs).
- Mechanism indicators: collision type (rear-end, sideswipe, fixed-object), vehicle maneuvers, and mentions of worker involvement.
- Narrative semantic features: embedding-derived dimensions representing latent semantic topics (e.g., heavy equipment, lane closure, signage), plus extracted categorical tags.
- Data quality metadata: flags indicating whether a record was auto-corrected, the confidence score of LLM extraction, and whether human adjudication occurred. These metadata variables help the learning algorithm account for residual uncertainty (Van Der Loo & De Jonge, 2020).

Feature selection strategies prioritize variables with substantive domain plausibility, stable measurement properties, and predictive utility. Missingness is addressed through multiple imputation strategies that account for data quality flags; records with systematic missingness in critical predictors are routed to conservative modeling streams or human review.

Ensemble modeling framework

Given the heterogeneity of crash dynamics across contexts and the propensity for overfitting on localized patterns, ensemble methods are well-suited for robust classification. The ensemble framework includes:

1.Base learners: a diverse set of algorithms including gradient-boosted trees, random forests, penalized logistic regression, and shallow neural networks. Diversity in modeling paradigms reduces shared error modes (Almahdi et al., 2023; Asadi & Wang, 2023).

2.Stacked meta-learner architecture: base learner predictions are used as input features to a meta-learner—typically a regularized model that learns optimal combination weights. Stacking leverages complementary strengths while controlling overfitting (Almahdi et al., 2023).

3.Hyperparameter optimization: systematic search strategies—Bayesian optimization, grid/random search, and modern low-rank adaptation techniques when tuning large models—optimize each learner and stacking configuration. Hyperparameter search is constrained by computational budgets and evaluated by cross-validation stratified by traffic condition and geography to ensure generalization across contexts (Almahdi et al., 2023; Hu et al., 2022).

4.Class imbalance handling: work zone crashes constitute a minority of total crashes in many datasets. The ensemble incorporates class-weighted loss functions, targeted sampling strategies, and threshold-adjusted decision rules. Oversampling of minority classes is tempered by validation using out-of-sample metrics to avoid synthetic-population artifacts.

5.Calibration and probabilistic scoring: post-hoc calibration (e.g., isotonic regression or Platt scaling) produces well-calibrated probabilities suitable for risk communication and operational thresholds (Pande et al., 2011).

6.Explainability: model-agnostic methods (feature permutation importance, SHAP values) and class-conditional exemplar inspections illuminate which features drive predictions and provide interpretable evidence for practitioners (Asadi & Wang, 2023).

Almahdi et al. (2023) and Asadi & Wang (2023) provide empirical support that ensemble approaches with hyperparameter optimization outperform single-model baselines in crash classification tasks, particularly when traffic conditions vary. Our framework extends these insights by emphasizing validated inputs and narrative features.

Evaluation strategy and metrics

Given the severe consequences of misclassification—both for safety interventions and for research conclusions—evaluation emphasizes both classification accuracy and downstream utility. Key evaluation components:

- Confusion-matrix derived metrics: precision, recall (sensitivity), specificity, F1-score for work zone class

detection. High recall is prioritized when the operational goal is to capture potential work zone events for further action, while precision is important when interventions are costly.

●Calibration metrics: Brier score and expected calibration error assess probabilistic reliability.

●Contextual stratified evaluation: performance is reported across strata—time-of-day, urban vs. rural, road class, and presence/absence of narrative text—to understand conditional robustness (Blackman et al., 2020).

●Misclassification analysis: systematic examination of false positives and false negatives, especially focusing on their content in narrative text and data quality flags, to understand remaining failure modes (Sayed et al., 2021).

●Human-review concordance: for cases flagged for adjudication, inter-rater reliability metrics quantify agreement between model recommendations and expert human labels, informing calibration of human-in-the-loop thresholds.

●Sensitivity to validation intensity: experiments vary the strictness of validation (e.g., more aggressive auto-correction vs. conservative flagging) to quantify the trade-off between data cleanliness and introduced bias.

This multi-faceted evaluation aligns with the recommendations of prior work emphasizing the need for context-sensitive validation and error analysis (Pipino et al., 2002; Van Der Loo & De Jonge, 2020).

Results

Descriptive overview of modeled outcomes

In descriptive terms, the integrated pipeline demonstrates systematic improvements in classification metrics relative to baseline approaches that do not perform layered validation or narrative mining. When structured-only models are trained on unvalidated data, standard ensemble approaches show moderate recall but suffer from elevated false positive rates, especially in jurisdictions with inconsistent work zone flagging practices (Carrick et al., 2009; Blackman et al., 2020). Incorporating layered validation and narrative-derived features yields consistent gains: recall improves due to richer semantic cues from narratives, while precision improves because validation corrects erroneous positive flags that would otherwise bias the model.

Role of validation in error reduction

The layered data validation module materially reduces apparent data noise. Deterministic schema checks solve a portion of trivial formatting errors; cross-field logic and geospatial-temporal matching resolve systematic mismatches where the work zone flag was toggled erroneously. Statistical anomaly detection catches outliers that are often anomalous because of entry mistakes (e.g., incorrect timestamps placing construction activity outside permitted hours). AI-augmented narrative validation is particularly effective: LLM-based extraction finds descriptive evidence for work activity in narratives that were not flagged in structured fields, thereby recovering false negatives. Conversely, narratives that clearly indicate an entirely different context (e.g., "vehicle struck roadside mailbox during winter storm") correct false positives. Overall, the validation module reduces label noise and yields cleaner training targets for classifiers, which is reflected in improved cross-validated performance (Pipino et al., 2002; Van Der Loo & De Jonge, 2020; Sayed et al., 2021).

Narrative features add discriminative power

Narrative-derived semantic features capture mechanistic information that structured fields often miss. For instance, mentions of "flagger", "paving", "work truck", or "temporary barrier" provide high signal for true work zone involvement. Embedding-based topic features capture latent patterns—such as clusters of narratives that commonly precede worker injuries (e.g., frequent mention of "moving equipment" and "tight lane configuration")—that prove predictive of both work zone presence and crash severity. The inclusion of narrative features consistently improves F1-scores relative to structured-only baselines, particularly in cases with partial or ambiguous structured indicators (Sayed et al., 2021; Swansen et al., 2013).

Ensemble stacking outperforms single-model baselines

Across cross-validation folds stratified by traffic and geographic context, stacked ensembles outperform individual learners. Base learners with complementary inductive biases capture different aspects of the problem: tree-based models absorb nonlinear interactions between categorical predictors and structured fields; penalized generalized linear models capture stable linear associations useful for calibration; shallow neural networks exploit interactions among dense embedding features. The meta-learner effectively balances these contributions, achieving better generalization on held-out data (Almahdi et al., 2023;

Asadi & Wang, 2023).

Hyperparameter tuning and calibration effects

Systematic hyperparameter optimization yields meaningful performance gains, particularly for gradient-boosted trees where learning rate, tree depth, and regularization parameters materially influence generalization. Calibration procedures reduce overconfidence in probabilistic outputs, producing more reliable risk scores useful for thresholding in operational contexts (Pande et al., 2011). Sensitivity experiments show that classifier recall is robust to modest perturbations in hyperparameters, but precision can deteriorate if hyperparameter settings favor overly flexible base learners without appropriate regularization.

Human-in-the-loop verification improves long-run performance

Human adjudication of edge cases—those where deterministic rules and LLM outputs disagree or where model confidence is low—yields high-quality labels for subsequent retraining. Inter-rater reliability between adjudicators and model recommendations provides a measurable signal guiding thresholding decisions: where model-human concordance is high, more aggressive automation is acceptable; where concordance is low, human oversight remains essential. The continuous feedback mechanism fosters a virtuous cycle: corrected labels enhance model performance, which reduces human workload for future adjudication (Malviya & Parate, 2025).

Contextual stratified findings

Performance varies across contexts. Urban areas with dense traffic and frequent short-duration work zones present more narrative heterogeneity and higher misclassification risk if validation is weak. Rural areas with fewer, longer-duration work zones show higher baseline precision but lower recall when narratives are sparse. Time-of-day stratification reveals that night-time work zones—often involving fewer visible cues—benefit disproportionately from narrative mining and geospatial matching. These stratified insights underscore the value of nuanced, context-aware evaluation rather than aggregate statistics (Blackman et al., 2020; Carrick et al., 2009).

Discussion

Interpretation of main findings

The results illustrate that investing in robust data validation and in extracting narrative semantics materially improves the reliability of work zone crash classification. From a theoretical standpoint, label noise acts as an error amplifier in supervised learning: even sophisticated learners cannot overcome fundamentally incorrect targets (Pipino et al., 2002; Sculley et al., 2014). By reducing label noise via layered validation, the learning problem becomes better posed, enabling ensembles to capture genuine underlying relationships. Narrative-derived features supply orthogonal information not captured by the usual structured fields; they act both as direct predictors and as validators of structured indicators, thereby closing a key semantic gap in many databases (Sayed et al., 2021).

Practical implications for agencies and policy

Practitioners should consider validation-first deployments rather than immediate automation. The proposed modular architecture allows agencies to introduce components incrementally—starting with schema validation and geospatial matching, then adding narrative mining, and finally integrating AI-augmented validation and ensemble models. This phased approach mitigates risk and builds organizational capacity. Importantly, calibrated probabilistic outputs enable operational decision rules: for example, high-probability work zone crash detections may trigger automated alerts to safety teams, while medium-probability cases could be routed for rapid human review. The data-quality metadata produced by the pipeline is also valuable for auditing and for justifying resource allocations, since it quantifies confidence in classification decisions (Redman, 1998; Van Der Loo & De Jonge, 2020).

Theoretical implications and contributions to literature

Methodologically, this work demonstrates that integrated validation and modeling architectures can be theoretically justified through the lens of error decomposition: data errors inflate both bias and variance in supervised models, and targeted validation reduces the effective noise term, yielding greater returns than incremental modeling tweaks alone (Pipino et al., 2002; Sculley et al., 2014). Our findings extend prior ensemble-based crash classification work (Almahdi et al., 2023; Asadi & Wang, 2023) by showing that model performance improvements are multiplicatively enhanced when upstream label quality is addressed. Additionally, the work advances narrative-mining

applications in transportation by operationalizing LLM-assisted normalization in a validation feedback loop, bridging a gap identified in prior narrative analysis studies (Sayed et al., 2021; Swansen et al., 2013).

Limitations and potential biases

Several limitations warrant candid discussion. First, the pipeline presumes the availability of supplemental work zone logs with sufficient granularity to perform reliable geospatial-temporal matching. In many jurisdictions, such logs are incomplete or exist in heterogeneous formats, limiting the effectiveness of the geospatial matching tier (Carrick et al., 2009). Second, LLM-based narrative extraction depends on suitable fine-tuning corpora; in low-resource settings, performance may degrade or require careful domain adaptation (OpenAI, 2023; Touvron et al., 2023). Third, automated correction policies introduce the risk of introducing systematic biases if the correction heuristics reflect historical reporting biases; human-in-the-loop safeguards are essential to detect and mitigate such feedback loops (Malviya & Parate, 2025; Redman, 1998). Fourth, evaluation in this study is descriptive and conceptual; while cross-validation and stratified analyses are informative, real-world operational deployment may reveal additional failure modes, such as adversarial reporting behaviors or transient sensor errors (Sculley et al., 2014).

Future research directions

Future work should pursue several avenues. Empirical validation across diverse jurisdictions with varying data maturity is essential to quantify generalizability. Research into privacy-preserving narrative mining—balancing the utility of textual data with confidentiality constraints—would broaden applicability. Additionally, integrating near-real-time traffic data (probe vehicle speeds, connected vehicle messages) with the proposed pipeline could enhance timeliness and improve predictive capacity for imminent work zone risks (Pande et al., 2011). Exploring active learning strategies where the model solicits labels for the most informative cases would optimize human adjudication resources. Finally, longitudinal studies measuring how improved classification impacts policy decisions and safety outcomes (e.g., reduced worker injuries, better allocation of temporary traffic controls) would close the loop from analytics to outcomes.

Ethical considerations and governance

Deploying AI-augmented validation and classification in public safety contexts raises ethical and governance questions. Transparency in model behavior, auditability of correction decisions, and explicit accountability pathways for erroneous automated corrections are necessary safeguards (Redman, 1998). Data governance frameworks should codify acceptable automated actions and require human review for high-stakes corrections, especially those that could influence enforcement or public reporting. The inclusion of data quality metadata in public reporting increases transparency and allows external stakeholders—researchers, watchdogs, and the public—to assess the reliability of derived statistics.

Conclusion

This article has proposed and detailed an integrated framework for improving work zone crash classification that foregrounds data validation, leverages narrative text mining, and employs ensemble modeling with careful tuning and calibration. By addressing the root causes of misclassification—heterogeneous reporting, inconsistent semantics, and data-entry errors—through layered validation, and by enriching feature sets with semantically normalized narrative signals, the pipeline substantially enhances the reliability of predictive models. Ensemble stacking and hyperparameter optimization extract robust predictive performance across varying traffic conditions, while human-in-the-loop adjudication ensures accountability and continuous improvement. The proposed approach is actionable for transport agencies and researchers: it prescribes modular adoption pathways, measurable evaluation metrics, and safeguards to manage operational risk. While limitations exist—especially regarding the availability of work zone logs and labeled narrative corpora—the theoretical rationale and descriptive outcomes suggest that integrating data-quality rigor with advanced modeling yields disproportionate benefits. Future empirical deployments and longitudinal studies will be necessary to quantify safety impacts and optimize human–AI collaboration in the field. In sum, improving data quality is not optional; it is the foundation on which trustworthy work zone analytics must be built (Pipino et al., 2002; Van Der Loo & De Jonge, 2020; Malviya & Parate, 2025).

References

1. Planning Stage Work Zone Configurations Using an Artificial Neural Network. Transp. Res. Rec. 2022,

2676, 377–384.

2. Yang, H.; Ozbay, K.; Ozturk, O.; Xie, K. Work Zone Safety Analysis and Modeling: A State-of-the-Art Review. Traffic Inj. Prev. 2015, 16, 387–396.

3. Blackman, R.; Debnath, A.K.; Haworth, N. Understanding Vehicle Crashes in Work Zones: Analysis of Workplace Health and Safety Data as an Alternative to Police-Reported Crash Data in Queensland, Australia. Aust. Traffic Inj. Prev. 2020, 21, 222–227.

4. Sayed, M.A.; Qin, X.; Kate, R.J.; Anisuzzaman, D.M.; Yu, Z. Identification and Analysis of Misclassified Work-Zone Crashes Using Text Mining Techniques. Accid. Anal. Prev. 2021, 159, 106211.

5. Almahdi, A.; Al Mamlook, R.E.; Bandara, N.; Almuflih, A.S.; Nasayreh, A.; Gharaibeh, H.; Alasim, F.; Aljohani, A.; Jamal, A. Boosting Ensemble Learning for Freeway Crash Classification under Varying Traffic Conditions: A Hyperparameter Optimization Approach. Sustainability 2023, 15, 15896.

6. Pande, A.; Das, A.; Abdel-Aty, M.; Hassan, H. Estimation of Real-Time Crash Risk. Transp. Res. Rec. 2011, 2237, 60–66.

7. OpenAI. GPT-3.5 Turbo Fine-Tuning and API Updates; OpenAI: San Francisco, CA, USA, 2023.

8. Swansen, E.; Mckinnon, I.A.; Knodler, M.A. Integration of Crash Report Narratives for Identification of Work Zone-Related Crash Classification. In Proceedings of the Transportation Research Board 92nd Annual Meeting, Washington, DC, USA, 13–17 January 2013.

9. Carrick, G.; Heaslip, K.; Srinivasan, S.; Brady, B. A Case Study in Spatial Misclassification of Work Zone Crashes. In Proceedings of the 88th Transportation Research Board Annual Meeting, National Academy of Sciences, Washington, DC, USA, 11–15 January 2009.

10. Asadi, H.; Wang, J. An Ensemble Approach for Predicting Crash Severity in Work Zones Using Machine Learning. Sustainability 2023, 15, 1201.

11. M. P. Van Der Loo and E. De Jonge, Data validation, arXiv preprint arXiv:2012.12028, 2020.

12. Malviya, S., & Vrushali Parate. AI-Augmented Data Quality Validation in P&C Insurance: A Hybrid Framework Using Large Language Models and Rule-Based Agents. International Journal of Computational and Experimental Science and Engineering, 11(3), 2025. https://doi.org/10.22399/ijcesen.3613

13. T. C. Redman, The impact of poor data quality on the typical enterprise, Communications of the ACM, vol. 41, no. 2, pp. 79–82, 1998.

14. L. L. Pipino, Y. W. Lee, and R. Y. Wang, Data quality assessment, Communications of the ACM, vol. 45, no. 4, pp. 211–218, 2002.

15. Great expectations. (2021) greatexpectations.io

16. S. Madnick, R. Wang, and X. Xian, The design and implementation of a corporate householding knowledge processor to improve data quality, Journal of Management Information Systems, vol. 20, no. 3, pp. 41–70, 2003.

17. Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S. et al., GPT-4 technical report, arXiv preprint arXiv:2303.08774, 2023.

18. Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F. et al., Llama: Open and efficient foundation language models, arXiv preprint arXiv:2302.13971, 2023.

19. C. Batini, C. Cappiello, C. Francalanci, and A. Maurino, Methodologies for data quality assessment and improvement, ACM Computing Surveys, vol. 41, no. 3, pp. 1–52, 2009.

20. D. Sculley, G. Holt, D. Golovin, E. Davydov, T. Phillips, D. Ebner, V. Chaudhary, and M. Young, Machine learning: The high interest credit card of technical debt, in SE4ML: Software engineering for machine learning (NIPS 2014 Workshop), vol. 8. Cambridge, MA, 2014.

21. E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen et al., LoRA: Low-rank adaptation of large language models, ICLR, 2022.