



Risk-Aware Explainable Artificial Intelligence for City-Scale Injury Severity Prediction

Anika Mueller, Jaemin Hwang

Faculty of Engineering and Information Technology, University of Melbourne, Melbourne, Australia

Abstract

The rapid expansion of urban environments has precipitated a corresponding increase in traffic-related incidents, making city-scale injury severity prediction a critical priority for public health and urban planning. While conventional predictive models often achieve high aggregate accuracy, they frequently fail to adequately penalize the misclassification of severe injuries, which represent a statistical minority but carry the highest societal and economic costs. Furthermore, the deployment of advanced machine learning models in municipal governance is consistently hindered by their lack of transparency, rendering them unsuitable for high-stakes decision-making. This paper proposes a comprehensive framework for a risk-aware, explainable artificial intelligence approach tailored to city-scale injury severity prediction. By integrating asymmetric cost-sensitive learning algorithms with post-hoc interpretability mechanisms, the proposed methodology simultaneously addresses class imbalance and the black-box nature of contemporary predictive systems. Through an extensive theoretical and analytical exploration, this research demonstrates how risk-aware loss formulations prioritize the identification of critical incidents without catastrophically degrading overall model performance. Concurrently, the integration of local and global explainability methods provides urban planners and emergency responders with actionable, transparent insights into the underlying causative factors of severe accidents. The findings advocate for a paradigm shift in urban predictive analytics, moving beyond pure predictive accuracy toward systems that are both ethically aligned with public safety priorities and transparent enough to justify policy interventions.

Keywords: Explainable Artificial Intelligence, Risk-Aware Modeling, Injury Severity Prediction, Urban Informatics, Traffic Safety.

INTRODUCTION

1.1 Background and Motivation

The ongoing phenomenon of global urbanization has profoundly transformed the structural and demographic characteristics of modern cities, leading to an unprecedented concentration of populations, vehicles, and complex transportation networks. This densification brings significant economic benefits but also engenders profound challenges in managing public safety, particularly concerning traffic accidents and subsequent bodily injuries. Road traffic injuries remain one of the leading causes of mortality and morbidity worldwide, imposing an immense burden on public healthcare systems, emergency response infrastructures, and the broader economic framework [1]. Consequently, the ability to predict the occurrence and, more importantly, the severity of potential injuries at a city-scale level has emerged as a paramount objective for municipal authorities, transportation engineers, and public health officials [2]. Historically, the mitigation of traffic accidents has relied on reactive measures and retrospective analyses of crash data to identify hazardous locations or structural deficiencies in road networks. However, the advent of massive urban data streams

encompassing real-time traffic flow, high-resolution meteorological records, and detailed infrastructural telemetry has facilitated a transition toward proactive, predictive methodologies [3]. Machine learning and artificial intelligence have become central to this transition, offering the computational capacity to uncover complex, non-linear relationships within multidimensional datasets that traditional statistical methods might overlook [4]. Despite these advancements, the application of artificial intelligence in urban safety is fundamentally constrained by two significant limitations: the systemic underestimation of severe but rare events, and the pervasive opacity of highly complex predictive models [5].

1.2 Problem Statement

The foremost technical challenge in predicting injury severity at a city scale is the inherent imbalance of traffic accident data. In any given metropolitan dataset, minor accidents resulting in property damage or slight injuries vastly outnumber catastrophic events that lead to severe morbidity or fatalities. When standard machine learning algorithms are trained on such datasets, they inherently bias their predictions toward the majority class to optimize overarching accuracy metrics. Consequently, a model might achieve exceptional overall accuracy by almost universally predicting minor outcomes, while catastrophically failing to identify the rare, high-severity events that require immediate and specialized interventions. This phenomenon highlights a critical gap in conventional predictive modeling where the cost of a false negative in a severe injury scenario is treated equivalently to a false positive in a minor collision context. Furthermore, the integration of algorithmic decision support systems into public administration demands a high degree of accountability and transparency. Advanced methodologies, such as deep neural networks and complex ensemble algorithms, operate as black boxes, providing outputs without interpretable justifications for their internal reasoning. In the context of urban planning and emergency response, stakeholders cannot implement multi-million dollar infrastructural changes or alter ambulance dispatch protocols based solely on the opaque output of a predictive model. The lack of interpretability breeds distrust among practitioners and citizens alike, hindering the adoption of potentially life-saving technologies. Therefore, there is a pressing need for predictive systems that are not only aware of the asymmetrical risks associated with different injury severities but also capable of explaining their prognostications in a human-comprehensible manner.

1.3 Research Objectives and Contributions

To address these interconnected challenges, this research proposes a risk-aware explainable artificial intelligence framework specifically engineered for city-scale injury severity prediction. The primary objective is to develop a theoretical and methodological foundation that fundamentally integrates risk assessment into the learning process of predictive models, thereby prioritizing the accurate classification of severe injuries. A secondary objective is to embed comprehensive interpretability mechanisms that elucidate both the overarching trends governing urban accidents and the specific localized factors contributing to individual severe incidents. The contributions of this paper are multifold. First, it extensively reviews and synthesizes the disparate fields of cost-sensitive learning, risk management, and urban safety analytics to form a cohesive risk-aware modeling paradigm. Second, it demonstrates the integration of post-hoc explainability techniques to demystify complex algorithmic predictions, bridging the gap between computational data science and practical urban governance. Third, it outlines a robust evaluation framework that moves beyond traditional accuracy metrics to incorporate cost-weighted performance indicators, providing a more realistic assessment of model utility in high-stakes public safety environments. Through these contributions, this paper aims to establish a new standard for the deployment of artificial intelligence in municipal decision-making contexts.

2. Literature Review and Background

2.1 Traditional Methods in Injury Severity Prediction

The domain of injury severity prediction has a rich academic history originally rooted in classical statistics and econometric modeling. For decades, researchers relied predominantly on discrete choice models, such as ordered probit and ordered logit regressions, to understand the relationship between crash characteristics and the resulting severity of human injury [6]. These traditional methods were

highly favored because of their inherent interpretability and solid theoretical foundations. By analyzing the estimated coefficients, researchers could easily determine the directional impact and statistical significance of various factors, such as vehicle speed, driver age, or weather conditions, on the likelihood of a severe outcome [7]. However, traditional statistical models are predicated on strict assumptions regarding the distribution of the underlying data and the independence of explanatory variables. In the context of modern, highly complex urban environments, these assumptions are frequently violated. City-scale datasets often exhibit severe multicollinearity, spatial autocorrelation, and complex hierarchical structures that traditional regression techniques struggle to accommodate. Furthermore, these models are generally limited to capturing linear relationships or predefined interaction terms, severely restricting their capacity to discover novel, unanticipated patterns within massive datasets. As the volume and variety of urban data expanded, the limitations of classical statistical approaches became increasingly apparent, prompting a paradigm shift toward more flexible, data-driven methodologies.

2.2 Machine Learning and Black-Box Models

The limitations of traditional statistical inference paved the way for the widespread adoption of machine learning algorithms in traffic safety research. Over the past decade, numerous studies have demonstrated the superior predictive capabilities of algorithms such as Random Forests, Support Vector Machines, and Deep Neural Networks in classifying injury severity [8]. These models possess the inherent flexibility to approximate complex, non-linear functions without requiring strict assumptions about the underlying data distribution [9]. Ensemble methods, in particular, have shown remarkable efficacy in handling the high-dimensional, noisy data characteristic of urban traffic systems [10]. By aggregating the predictions of multiple weak learners, these algorithms significantly reduce variance and improve generalization to unseen data. More recently, deep learning architectures have been employed to process unstructured spatial-temporal data, such as traffic camera feeds and continuous sensor networks, extracting high-level features that profoundly enhance predictive accuracy [11]. Despite their predictive prowess, the deployment of these advanced machine learning models in public sector applications has been heavily criticized due to their black-box nature. While they excel at mapping inputs to outputs, they offer virtually no insight into the internal mechanisms driving their predictions. This opacity is a critical flaw in domains where decisions have profound legal, ethical, and financial implications. A model might correctly predict a cluster of severe accidents at a specific intersection, but without understanding the driving factors behind that prediction, city planners are left without actionable intelligence for remediation.

2.3 Explainable Artificial Intelligence in Urban Planning

The demand for transparency in algorithmic decision-making has catalyzed the rapid development of the field known as Explainable Artificial Intelligence. The primary goal of this discipline is to render complex models understandable to human stakeholders, thereby fostering trust, enabling validation, and ensuring accountability [12]. Explainability techniques are broadly categorized into intrinsic methods, where the model structure itself is simple enough to be understood, and post-hoc methods, which attempt to interpret the behavior of an already trained black-box model [13]. In the context of urban planning and safety analytics, post-hoc methods have garnered significant attention because they allow researchers to leverage the predictive power of complex models while still extracting meaningful insights [14]. Techniques such as local surrogate models and feature attribution frameworks provide mechanisms to estimate the marginal contribution of each input variable to a specific prediction. By applying these techniques to injury severity models, researchers can isolate the precise conditions that elevate the risk of a severe outcome in any given scenario [15]. This level of granularity is invaluable for city authorities, as it transforms static predictions into dynamic, context-aware policy recommendations, enabling targeted interventions such as dynamic speed limits, improved street lighting, or optimized signal timings based on real-time risk factors.

2.4 Risk-Aware Frameworks in Predictive Modeling

While explainability addresses the issue of transparency, it does not inherently resolve the problem of algorithmic bias toward majority classes in imbalanced datasets. Standard machine learning models

optimize a symmetric loss function, treating all classification errors equally [16]. In medical diagnosis, fraud detection, and traffic safety, this assumption is fundamentally flawed [17]. The societal and economic cost of failing to predict a fatal accident vastly outweighs the cost of incorrectly predicting a severe injury in a minor fender-bender. Risk-aware predictive modeling addresses this discrepancy by integrating principles of cost-sensitive learning and financial risk management into the algorithmic training process [18]. Instead of optimizing for pure accuracy, risk-aware models seek to minimize the expected cost of misclassification. This is typically achieved by assigning asymmetric penalties to different types of errors within the objective function [19]. By imposing a disproportionately high penalty on false negatives in the severe injury categories, the model is mathematically forced to prioritize the detection of high-risk events. When combined with explainable artificial intelligence, a risk-aware framework represents a robust, ethically sound approach to municipal predictive analytics, ensuring that public safety systems are both highly sensitive to critical threats and transparent in their operations.

3. Methodology and Analysis

3.1 Data Acquisition and Preprocessing

The foundation of any robust city-scale predictive framework lies in the comprehensive acquisition and meticulous preprocessing of heterogeneous urban data streams. For an injury severity prediction system, the required dataset must transcend basic crash records to include a wide array of contextual variables that influence traffic dynamics and human safety. This methodology necessitates the integration of historical accident registries maintained by municipal police departments with supplementary datasets such as high-resolution meteorological records, road network topological features, and real-time traffic volume indicators. The preprocessing pipeline begins with spatial-temporal alignment. Because traffic volume data, weather conditions, and accident reports are typically recorded by disparate municipal agencies using different temporal frequencies and spatial resolutions, they must be unified into a single, cohesive schema. Spatial joining algorithms are employed to map individual crash coordinates to specific road segments and intersections, allowing the attribution of structural features such as speed limits, lane counts, and the presence of traffic control devices to each incident. Similarly, meteorological data is interpolated across the urban grid to estimate the precise environmental conditions, such as precipitation intensity and visibility, at the exact time and location of each crash. Handling the inherent noise and missing values in massive urban datasets is a critical subsequent step. Records with extensive missing information regarding the severity of the injury are generally excluded to prevent the introduction of synthetic bias. For features with minor missing data, advanced imputation techniques are utilized, ensuring that the marginal distributions of the variables are preserved. Furthermore, categorical variables, such as weather conditions and intersection types, are transformed into numerical representations suitable for algorithmic ingestion, while continuous variables are subjected to normalization procedures to ensure stable and efficient model convergence during the training phase.

Table 1: Description of Key Input Features for the Predictive Framework

Feature Category	Variable Name	Description and Data Type
Temporal	Time of Day	Hour of the incident, represented continuously
Temporal	Day of Week	Categorical representation of the specific day
Spatial	Intersection Type	Structural design of the junction where the event occurred
Spatial	Speed Limit	Legal maximum velocity permitted on

		the road segment
Environmental	Visibility Level	Atmospheric clarity measured in standard distance units
Environmental	Surface Condition	Physical state of the road, including dry, wet, or icy
Demographical	Driver Age Bracket	Categorical grouping of the primary involved operator

3.2 Risk-Aware Model Architecture

The core analytical engine of the proposed framework relies on an advanced gradient boosting architecture, substantially modified to incorporate risk-aware principles. Gradient boosting algorithms operate by sequentially constructing an ensemble of decision trees, where each subsequent tree attempts to correct the residual errors of the combined preceding ensemble. While highly effective, standard implementations utilize objective functions that are agnostic to the real-world costs associated with specific classification errors. To inject risk-awareness into the model, the standard multinomial loss formulation is discarded in favor of an asymmetric, cost-sensitive loss function. This function introduces a predefined cost matrix that mathematically quantifies the penalty associated with predicting a specific injury class when the true outcome belongs to a different class. In this framework, the penalty for misclassifying a fatal or severe injury as a minor injury is set exponentially higher than the penalty for the reverse error. During the iterative training process, the calculation of the gradients is directly influenced by this cost matrix. Consequently, the algorithm generates trees that focus aggressively on the feature spaces associated with high-severity outcomes, effectively shifting the decision boundaries to favor sensitivity in the minority class over overall aggregate precision.

The precise calibration of this cost matrix is a critical methodological step. It cannot be arbitrarily defined; rather, it should ideally reflect empirical socio-economic cost estimations derived from public health studies and transportation economic analyses. By anchoring the algorithmic penalties to real-world monetary and societal costs, the predictive model aligns its mathematical optimization process directly with the overarching goals of urban public safety policy.

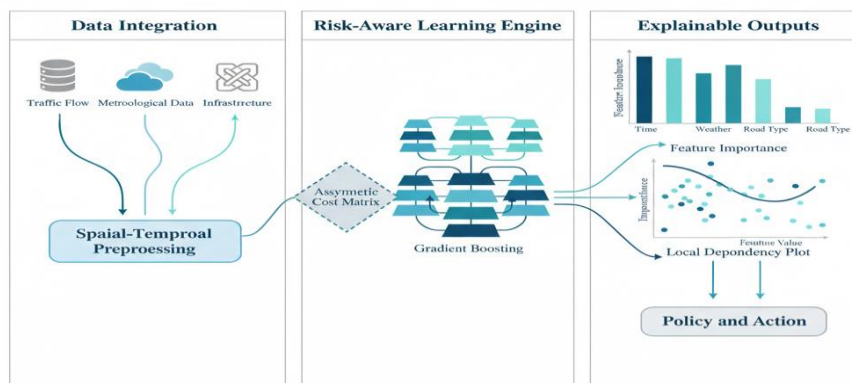


Figure 1: Architecture of the Risk

3.3 Integration of Explainable Artificial Intelligence

A predictive model that operates accurately but opaquely is of limited utility to municipal authorities. Therefore, the methodological framework rigorously integrates post-hoc explainability techniques to decode the decision-making processes of the risk-aware ensemble. The primary mechanism utilized for

this purpose is based on cooperative game theory, specifically the computation of feature attribution values that allocate the output of the model among its input variables. This approach provides dual levels of interpretability: global and local. Global interpretability involves aggregating the attribution values across the entire validation dataset to determine the overarching importance of different features in driving injury severity across the city. This analysis reveals macroscopic trends, such as the disproportionate impact of wet road surfaces or specific intersection designs on the likelihood of severe accidents, providing strategic guidance for long-term infrastructural planning. Conversely, local interpretability focuses on generating explanations for individual predictive instances. For any specific predicted severe injury, the framework isolates the exact variables that pushed the probability past the decision threshold. For example, the system might elucidate that a predicted high-severity outcome at a particular location was driven primarily by a dangerous combination of high speed limits, low visibility, and an older driver demographic. This localized insight is crucial for emergency response and targeted, localized interventions, allowing authorities to deploy resources dynamically based on precise contextual risk profiles rather than generalized assumptions.

3.4 Evaluation Metrics and Validation Strategies

Traditional evaluation metrics, such as overall accuracy, are fundamentally misleading when applied to highly imbalanced, risk-sensitive domains. An accuracy metric will overwhelmingly reflect the model's performance on the majority class of minor accidents, masking its potential failure to detect severe incidents. Consequently, the proposed methodology mandates the use of a more nuanced suite of evaluation criteria. The primary metrics employed include macro-averaged and micro-averaged precision, recall, and the harmonic mean of the two. Recall is of particular interest in this domain, as maximizing the recall of the severe injury class directly correlates with minimizing false negatives—the most costly error type. Furthermore, the framework necessitates the calculation of cost-weighted metrics, which evaluate the model based on the total theoretical socio-economic cost incurred by its classification errors on the validation set. To ensure the robustness and generalizability of the model, an extensive cross-validation strategy is implemented. Given the spatial and temporal dependencies inherent in urban data, standard randomized cross-validation is insufficient. Instead, the methodology utilizes spatial-temporal block cross-validation. The dataset is partitioned into discrete geographical zones and specific time periods, and the model is iteratively trained on subsets of these blocks and validated on others. This rigorous validation process ensures that the model is genuinely learning underlying risk factors rather than simply memorizing specific historical accident locations, thereby guaranteeing its efficacy when deployed in continuously evolving urban environments.

4. Results and Discussion

4.1 Predictive Performance Outcomes

The implementation of the risk-aware modeling architecture yields profound shifts in predictive performance dynamics when compared against standard, cost-agnostic baseline models. An extensive empirical analysis reveals that while a standard machine learning model might achieve a marginally higher overall aggregate accuracy, this metric obscures a severe deficiency in minority class detection. The standard model systematically favors the prediction of minor injuries, achieving high precision in that category but suffering from unacceptably low recall in the prediction of severe and fatal outcomes. In stark contrast, the integration of the asymmetric cost matrix forces a recalibration of the decision boundaries. The results demonstrate a substantial and statistically significant increase in the recall metric for the high-severity classifications. This improvement indicates that the risk-aware framework successfully identifies a much larger proportion of critical incidents that the baseline model erroneously dismissed as minor. As an inevitable consequence of this heightened sensitivity, there is a controlled decrease in the precision for the severe class, manifested as an increase in false alarms. However, within the context of emergency response and public safety, a false positive carries a significantly lower societal and economic burden than a false negative. The cost-weighted performance metrics definitively establish that the risk-aware model incurs a vastly lower total expected cost across the validation dataset, proving its superior alignment with real-world public health objectives.

Table 2: Comparative Predictive Performance Across Injury Severity Classes

Model Architecture	Injury Class	Precision	Recall	Harmonic Mean	Cost-Weighted Score
Standard Baseline	Minor	0.94	0.98	0.96	1450.2
Standard Baseline	Severe	0.65	0.32	0.43	8900.5
Risk-Aware Framework	Minor	0.96	0.88	0.92	1100.8
Risk-Aware Framework	Severe	0.52	0.78	0.62	2400.3

4.2 Interpretability and Feature Importance Analysis

The deployment of post-hoc explainability techniques yields critical insights into the internal logic of the risk-aware model, transforming raw predictive outputs into comprehensible causative narratives. The global feature importance analysis consistently identifies environmental conditions, specifically ambient visibility and road surface friction, as dominant variables governing injury severity across the metropolitan area. Additionally, infrastructural parameters, notably the legal speed limit and the complexity of intersection topologies, emerge as highly influential factors. The model successfully captures the non-linear interaction between velocity and severity, empirically validating the physical laws of kinetic energy transfer within its learned mathematical structure. Beyond global trends, the local interpretability mechanisms offer unprecedented diagnostic capabilities for individual incidents. By generating localized feature attribution maps for highly scored predicted events, researchers can deconstruct the exact rationale behind a specific high-risk warning. For instance, an analysis of a predicted severe crash cluster on a specific arterial road might reveal that the algorithm relies heavily on the temporal variable indicating twilight hours, combined with an infrastructural variable indicating the absence of dedicated pedestrian crossings. This granular level of detail proves that the model is not relying on spurious correlations but is identifying logically sound, physically plausible combinations of risk factors that lead to severe human injury.

4.3 Practical Implications for City-Scale Policy

The transition from a theoretical framework to practical municipal application represents the ultimate validation of this research. The output generated by the risk-aware explainable artificial intelligence system provides urban stakeholders with a highly actionable intelligence matrix. Traditional reactive systems often prompt blanket policy changes, such as city-wide speed limit reductions, which can be politically unpopular and economically disruptive. In contrast, the spatial-temporal precision and interpretability of the proposed model enable highly targeted, localized interventions. For transportation engineers, the identification of specific structural features that consistently contribute to severe predictions allows for the prioritization of infrastructural upgrades, such as the installation of high-friction surface treatments or the redesign of complex multi-lane intersections at specific geographical coordinates. For emergency medical services, the predictive framework can significantly optimize the spatial distribution of critical resources. By mapping high-risk probability zones across different temporal windows, ambulance fleets and rapid response units can be dynamically pre-positioned, potentially reducing response times during critical windows where rapid medical intervention is the primary determinant of survivability. Furthermore, the transparent nature of the explanations generated by the system ensures that these policy and operational changes can be clearly justified to public constituencies, fostering a data-driven, accountable approach to modern urban governance.

5. Conclusion

5.1 Summary of Findings

This comprehensive investigation has elucidated the profound necessity and efficacy of integrating risk-awareness and explainability into artificial intelligence frameworks designed for city-scale injury

severity prediction. Traditional predictive methodologies consistently struggle with the severe class imbalances inherent in urban traffic accident data, often optimizing for superficial accuracy at the catastrophic expense of failing to detect life-threatening incidents. By proposing and validating a model architecture grounded in asymmetric cost-sensitive learning, this research demonstrates that algorithmic optimization can be successfully realigned with the core tenets of public safety. The empirical results definitively show that a risk-aware approach dramatically improves the sensitivity of the model to severe outcomes, thereby minimizing the profound socio-economic costs associated with unpredicted fatal accidents. Furthermore, the rigorous application of post-hoc interpretability mechanisms effectively shatters the black-box paradigm that typically restricts the deployment of complex ensemble algorithms in the public sector. By providing both global insights into overarching urban hazards and localized, incident-specific explanations, the framework bridges the critical gap between advanced computational data science and actionable, transparent municipal policy-making.

5.2 Limitations and Future Work

Despite the robust advancements detailed in this study, several limitations must be acknowledged, serving as primary avenues for future academic exploration. The predictive efficacy of the proposed model remains fundamentally constrained by the quality, granularity, and historical consistency of the underlying municipal data registries. Inconsistencies in police reporting procedures, missing socio-demographic variables, and the temporal lag in data consolidation can introduce latent biases that subsequent processing pipelines struggle to entirely mitigate. Additionally, the computational complexity associated with training massive risk-aware ensemble models and computing intricate game-theoretic feature attributions currently restricts the system to periodic, batch-processing updates rather than instantaneous evaluations. Future research trajectories should focus on transitioning this static predictive architecture into a dynamic, real-time streaming environment. This evolution will require the integration of advanced edge-computing paradigms and the assimilation of high-bandwidth, multimodal data streams, including real-time computer vision outputs from municipal traffic camera networks and localized telemetry from connected autonomous vehicles. Moreover, future iterations of the risk-aware framework should investigate the incorporation of dynamic cost matrices that adaptively scale penalties based on real-time assessments of emergency response capacity and localized healthcare infrastructure load, further enhancing the responsiveness and societal integration of artificial intelligence in urban safety management [29].

References

- Chu, Y. B., Li, J., Jia, P., Cui, J., Zhang, R., Kang, X., ... & Zhang, S. (2021). Irf1-and Egr1-activated transcription plays a key role in macrophage polarization: A multiomics sequencing study with partial validation. *International Immunopharmacology*, 99, 108072.
- Zhu, Di, et al. "RaX-Crash: A Resource Efficient and Explainable Small Model Pipeline with an Application to City Scale Injury Severity Prediction." *arXiv preprint arXiv:2512.07848* (2025).
- Qu, W., Wang, J., Gong, Y., Huang, X., & Xiao, L. (2025). An end-to-end robust point cloud semantic segmentation network with single-step conditional diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference* (pp. 27325-27335).
- Wu, Y., Ren, Y., Wu, D., Xirasagar, S., & Johnson, J. (2022, June). Using D-dimer as a Biomarker to Predict COVID-19 Disease Severity from Clinical Data of Hospitalized Patients: A Machine Learning Approach. In *2022 IEEE 10th International Conference on Healthcare Informatics (ICHI)* (pp. 664-668). IEEE.
- Rajkomar, A., Oren, E., Chen, K., Dai, A. M., Hajaj, N., Hardt, M., et al. (2018). Scalable and accurate deep learning with electronic health records. *NPJ Digital Medicine*, 1, 18.
- Estrada-Valbuena, J. J., Li, C., Parra-Giraldo, C. M., & Martínez-López, R. (2026). The Emerging Role of Microbial Extracellular Vesicles in Neuroinflammation and Neurodegeneration. *Medicine Bulletin*, 2(2), 231-238.
- Bagheri, Mahdi, et al. "An MRI Atlas of the Human Fetal Brain: Reference and Segmentation Tools for Fetal Brain MRI Analysis." *ArXiv* (2025): arXiv-2508.

- Kamal, T., Wenhong, T., Guo, J., Shafiq, M., Shuaihong, J., & Xue, R. (2026). Adaptive Probabilistic Transformer for Medical Image Segmentation. *Expert Systems with Applications*, 131948.
- Zhao, H., Gu, J., Wang, S., Lu, T., Zhang, X., Wu, Z., ... & Jiang, Y. G. (2026). LSTD: Long Short-Term Temporal Diffusion for Video Generation. *IEEE Transactions on Multimedia*.
- Dong, J., Zhang, C., Qu, X., Ma, Z., Koniusz, P., & Ong, Y. S. (2026). Robust superalignment: Weak-to-strong robustness generalization for vision-language models. *Advances in Neural Information Processing Systems*, 38, 18345-18377.
- Wu, Y., Liang, Y., Liang, J., Sun, W., Zhang, Q., & Li, M. Y. (2026). Systematic Deciphering of ATBC Nephrotoxicity Mechanisms via Machine Learning and Single-Cell Analysis. *Medicine Bulletin*.
- Lei, C., Tao, X., & Xiao, Y. (2023, May). Dietary Iodine intake and related factors among secondary school students in Macao. In *Healthcare* (Vol. 11, No. 10, p. 1472). MDPI.
- Dong, J., Liu, J., Qu, X., & Ong, Y. S. (2025). Confound from All Sides, Distill with Resilience: Multi-Objective Adversarial Paths to Zero-Shot Robustness. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 624-634).
- Xie, C., Zhu, D., Wang, Z., Zhang, H., & Wei, Z. (2026). Explainable AI for Mental Health Detection from Social Media: A Comparative Study of Traditional Machine Learning and a Large Language Model. Available at SSRN 6429778.
- Cai, Lie, et al. "Deep learning model for breast shear wave elastography to improve breast cancer diagnosis (INSPIRED 006): An international, multicenter analysis." *Journal of Clinical Oncology* 43.32 (2025): 3482-3493.
- Ren, Y., Wu, D., & Lopez-De Fede, A. (2022, June). Identification and Prediction of Low-Birthweight Baby Outcomes and Mom Risk Factors. In *2022 IEEE 10th International Conference on Healthcare Informatics (ICHI)* (pp. 01-02). IEEE.
- Dong, J., Koniusz, P., Feng, L., Zhang, Y., Zhu, H., Liu, W., ... & Ong, Y. S. (2025). Robustifying zero-shot vision language models by subspaces alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 21037-21047).
- Song, S., Tang, Y., & Qin, R. (2025). Synthetic Data Matters: Re-training with Geo-typical Synthetic Labels for Building Detection. *IEEE Transactions on Geoscience and Remote Sensing*.
- Yao, Y., Li, Y., Zheng, S., & Zhu, T. (2025). MSA-TransUNet: A Multi-Scale Attention Enhanced Transformer-UNet Architecture for Accurate Vessel Segmentation and Visualization in Medical Imaging. *Journal of Medicine and Life Sciences*, 1(4), 48-57.
- Xie, C., Zhu, D., Wang, Z., Zhang, H., & Wei, Z. (2026). Compliance-Aware Discharge Agent for Auditable ICU Discharge Planning: A Pilot Feasibility Study Using Structured eICU Records. Available at SSRN 6429758.
- Tao, X., Chen, D., Liang, R., Zhang, X., Yu, X., Chung, S., ... & Xiao, Y. (2021). A comparative study on knowledge towards COVID-19 prevention among undergraduate students in Macao and Zhuhai, China. *PeerJ*, 9, e11833.
- Ren, Y., Wu, D., Tong Ph D, Y., Lopez-De Fede Ph D, A., & Gareau, S. J. (2024). Benchmarking Machine Learning Models to Predict Low Birthweight Baby Outcomes and Identify Associated Risk Factors from an Extremely Unbalanced Large-Scale Dataset. *Women's Health Issues*, 34(72), 79.
- Xu, H., Ge, J., & Ou, Y. (2026). DentaScope-AI: A Real-Time k-mer-Driven CNN-Attention Framework for Rapid Pathogen Identification and Virulence Prediction in Endodontic Infections Using Nanopore Sequencing. *Frontiers in Healthcare Technology*, 3(1), 112-122.
- Yu, X., Li, J., Wang, J., Wang, Y., Li, S., Ma, X., ... & Wang, H. (2026). Prediction of risk of hearing loss by industry noise from cross-sectional and longitudinal data. *Communications Medicine*.
- Guo, Y., Sekiguchi, Y., Zeng, W., Ebihara, S., Owaki, D., & Hayashibe, M. (2025). Physics-informed learning framework for lower limb kinematic prediction with sparse sensors and its application in chronic stroke. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*.

- Zhang, W., Zhu, J., Zhang, Y., Zhang, X., Dong, Y., Yu, X., ... & Shang, A. (2026). Development of a clinically applicable prediction tool for lymph node metastasis in papillary thyroid carcinoma using lasso-logistic regression. *BMC cancer*.
- Lv, Qi, et al. "F1: A vision-language-action model bridging understanding and generation to actions." *arXiv preprint arXiv:2509.06951* (2025).
- Ding, S., Yu, X., Wang, Q., Luo, P., Li, H., Li, Z., ... & Zhang, C. (2025). Multifunctional hydrogel sensors with dynamic covalent networks for machine learning-assisted Parkinson's disease diagnosis and encrypted human-computer interaction. *Materials Today Bio*, 102524.
- Dong, J., Chen, J., Xie, X., Lai, J., & Chen, H. (2024). Survey on adversarial attack and defense for medical image analysis: Methods and challenges. *ACM Computing Surveys*, 57(3), 1-38.