Beta Distribution Learning for Reliable Roadway Crash Risk Assessment

Ahmad Elallaf[1], **Nathan Jacobs**[2], **Xinyue Ye**[3], **Mei Chen**[4], **Gongbo Liang**[1]

[1]Texas A&M University-San Antonio, San Antonio, TX, USA
[2]Washington University in St. Louis, St. Louis, MO, USA
[3]University of Alabama, Tuscaloosa, AL, USA
[4]University of Kentucky, Lexington, KY, USA

aelal01@jaguar.tamu.edu, jacobsn@wustl.edu, xye10@ua.edu, mei.chen@uky.edu, gliang@tamusa.edu

Abstract

Roadway traffic accidents represent a global health crisis, responsible for over a million deaths annually and costing many countries up to 3% of their GDP. Traditional traffic safety studies often examine risk factors in isolation, overlooking the spatial complexity and contextual interactions inherent in the built environment. Furthermore, conventional Neural Network-based risk estimators typically generate point estimates without conveying model uncertainty, limiting their utility in critical decision-making. To address these shortcomings, we introduce a novel geospatial deep learning framework that leverages satellite imagery as a comprehensive spatial input. This approach enables the model to capture the nuanced spatial patterns and embedded environmental risk factors that contribute to fatal crash risks. Rather than producing a single deterministic output, our model estimates a full Beta probability distribution over fatal crash risk, yielding accurate and uncertainty-aware predictions—a critical feature for trustworthy AI in safety-critical applications. Our model outperforms baselines by achieving a 17-23% improvement in recall, a key metric for flagging potential dangers, while delivering superior calibration. By providing reliable and interpretable risk assessments from satellite imagery alone, our method enables safer autonomous navigation and offers a highly scalable tool for urban planners and policymakers to enhance roadway safety equitably and cost-effectively.

Project Page — https://www.gb-liang.com/projects/betarisk

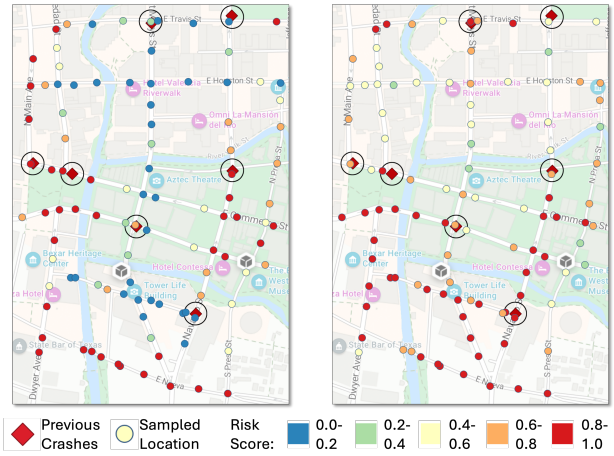


Figure 1: A case study of crash risk assessment for the San Antonio River Walk. Historical fatal crashes (red diamonds) serve as ground truth. (Left) The baseline `MSCM-MS` model exhibits low recall and spatially inconsistent predictions, with abrupt risk changes between adjacent points. (Right) Our `Prob-MS` model demonstrates superior recall by correctly identifying more crash sites and generates a more nuanced and spatially coherent risk field, providing a more realistic safety assessment. See the Result section for more.

Introduction

Roadway traffic accidents claim over 1.3 million lives annually (WHO 2023) and impose economic burdens of 3% of the GDP in many countries (WHO 2018). Transportation safety has garnered significant research (Caliendo et al. 2007; Tamerius et al. 2016; Zhu et al. 2024), yet accurately estimating crash risk remains a challenge due to its inherent uncertainties and the sparse nature of crash events.

Conventional safety research often analyzes individual factors separately, such as driver behavior (Simons-Morton et al. 2014), road infrastructure (Pembuain et al. 2019), traffic patterns (Huang et al. 2020), and weather (Jaroszweski and McNamara 2014), overlooking the complex interplay between these elements (Gu et al. 2022). Since crash occurrences frequently result from intricate multi-factor interactions, methods that analyze these factors in isolation struggle to predict risk holistically (Carrodano 2024). Furthermore, data limitations have constrained the scope of most studies to highways (Ahmed 2013; Song et al. 2018; Cheng et al. 2019; Ma et al. 2020; Joo et al. 2023), leaving comprehensive crash risk analysis for local roads, where data is often less available, relatively unexplored.

To overcome these limitations, we introduce a novel deep learning framework that learns a full Beta probability distribution, moving beyond simple point-estimates of fatal crash risk. Our primary contributions are threefold:

- A **probabilistic formulation** that yields well-calibrated, uncertainty-aware predictions, a critical feature for trustworthy AI in high-stakes, safety-critical domains.

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

- A **highly scalable and equitable methodology** that uses near-globally available satellite imagery, enabling risk assessment for both highways and previously under-assessed local roads.

The proposed probabilistic model is evaluated through extensive experiments conducted over four major metropolitan areas, which have a population of ≈ 20 million. Our model achieves a 17-23% improvement in recall over baselines, a crucial metric for any safety-critical task, while also delivering superior model calibration and F1 scores. By producing reliable and interpretable risk assessments from satellite imagery alone, this work provides a foundational tool for enhancing traffic safety, from enabling safer route selection for drivers and autonomous vehicles to empowering urban planners and policymakers to mitigate high-risk areas.

Background

Estimate Roadway Crash Risk A primary challenge in data-driven roadway safety is formulating the risk estimation task. Existing methods often frame it as classifications, such as predicting a crash occurrence within a short time frame (Huang et al. 2020). While valuable, these approaches do not estimate the inherent, continuous crash risk of a given road segment. A more nuanced approach is to directly estimate a crash probability, such as using Monte Carlo simulations (de Almeida Guimarães et al. 2018; Al-Sharif et al. 2012; Jeon and Hong 2016). However, this is fundamentally challenged by the extreme sparsity of crash data, rendering traditional estimation techniques unreliable, as they can obscure high-risk areas while falsely flagging safe ones (He et al. 2021), leading to false negatives that are dangerous in any safety-critical application.

Deep Neural Networks (DNNs) offer a powerful alternative, as they can learn complex, task-specific features directly from data and provide near-instantaneous inference. However, supervised DNNs typically rely on large, manually labeled datasets, such as manually assigned risk levels (e.g., low, neutral, high) (Najjar et al. 2017). Creating these datasets is prohibitively expensive, and the manual labels can suffer from human bias, potentially misrepresenting the true risk (Li et al. 2024; Chen and Sundar 2023). These challenges motivate the need for a new approach that can learn a continuous risk score from objective crash data while effectively handling the probabilistic nature of the task.

Deep Neural Network Miscalibration Over the recent years, DNNs have shown promising performance on various domains, such as medical imaging (Xing et al. 2023; Liu et al. 2022), cybersecurity (Zulu et al. 2024), transportation (Jonnala et al. 2025), and astrophysics (Lin et al. 2022). However, for a predictive model to be trustworthy in high-stakes applications, its predicted confidence must accurately reflect its probability of being correct. However, modern DNNs are often miscalibrated, tending to produce overconfident predictions (Pereyra et al. 2017; Guo et al. 2017).

Mathematically, a model is perfectly calibrated if, for any given confidence level p, the long-run accuracy of predictions with that confidence is indeed p. For DNNs, the calibration error, the difference between a model's predicted confidence and its actual accuracy, is often significantly greater than zero (Hinton et al. 2015). This miscalibration is a critical failure point in high-stakes applications where decisions depend on the model's self-assessed certainty.

While various techniques can mitigate this issue, they often have limitations. Post-processing methods like temperature scaling (Guo et al. 2017) adjust model outputs without altering the learned features, while in-training regularization (Kumar et al. 2018; Liang et al. 2020) requires careful tuning for the weight scaler. Given that model complexity is a key contributor to miscalibration (Chidambaram and Ge 2023), we argue that an effective solution must be deeply integrated into the learning process. Our work achieves this by reformulating the risk estimation task as learning a full probability distribution, a method that inherently encourages better-calibrated and more reliable predictions.

Method

Probabilistic Modeling Framework Our method recasts roadway crash risk estimation from a standard classification task into a probabilistic learning problem, motivated by the limitations of conventional models that provide a single point-estimation. Consider a fatal crash, a stochastic occurrence, at a specific point in spacetime, $C = (x, y, t, d)$, where (x, y) is the geolocation and (t, d) is the time and date. While any single crash is a random event, its location provides the strongest available evidence for a local maximum in the underlying, continuous risk field, $R(\cdot)$. Therefore, it is intuitive that the inferred risk should be higher at or near the crash site and should decay smoothly as one moves away in space or time. For nearby points, such as a spatially displaced point $C' = (x-\delta, y, t, d)$, the risk should be lower, i.e., $R(C') < R(C)$. Standard point-estimate classifiers fail to capture this continuous field, as they are trained to predict a binary outcome for each location independently.

While a complete model would account for both spatial and temporal decay, this work focuses on the challenging and foundational task of estimating the **static, inherent risk** of a location based on its geographic and structural features. Our goal is to model the spatial component of this uncertainty by learning a distribution over possible risk values, capturing the intuition that for a nearby point C', the risk is attenuated but non-zero: $0 < R(C') < R(C)$.

We specifically employ the Beta distribution for this task due to its natural support on the $[0, 1]$ interval and its flexibility in representing diverse risk profiles. Instead of a single value, our model $h(x)$ maps an input image x to the two positive scalar parameters, (α, β), which define a Beta distribution, $P_p \sim Beta(\alpha, \beta)$. This formulation allows the model to express its uncertainty through the shape of the distribution: a sharp peak indicates high confidence, while a wide distribution signifies high uncertainty. The final risk score R is the mean of this predicted distribution:

$$R = \mathbb{E}[P_p] = \frac{\alpha}{\alpha + \beta}. \quad (1)$$

To achieve this, our framework integrates three key technical contributions: 1) a novel procedural labeling technique

Algorithm 1: Target Beta Distribution Generation

Require: Original image x, binary label $l \in \{0, 1\}$, base concentration K_{base}, minimum positive risk mean μ_{min}, minimum positive concentration k_{min}, distance weight w_{dist}, size weight w_{size}, and $\epsilon = 1e^{-5}$

if $l = 0$ **then** ▷ For negative samples, create a low-risk, ▷ high-certainty distribution

$\alpha_t \leftarrow \epsilon$

$\beta_b \leftarrow K_{base}$

else ▷ For positive samples (l=1), generate labels ▷ based on crop geometry

$x' \leftarrow$ random crop of x

$d_{norm} \leftarrow$ normalized distance of x' from center of x

$s_{norm} \leftarrow \frac{size(x')}{size(x)}$

$influence \leftarrow w_{dist} \cdot (1 - d_{norm}) + w_{size} \cdot s_{norm}$

$\mu_t \leftarrow \mu_{min} + (1 - \mu_{min}) \cdot influence$

$k_t \leftarrow k_{min} + (K_{base} - k_{min}) \cdot influence$

$\alpha_t \leftarrow \mu_t \cdot k_t$

$\beta_t \leftarrow \epsilon$

end if

return (α_t, β_t) ▷ Return the target Beta distribution

that generates the targeting Beta distributions from data augmentation, 2) a multi-scale deep neural network architecture, and 3) a compound loss function for joint optimization.

Target Beta Distributions Generation A key innovation of our framework is the procedural generation of supervisory signals in the form of target Beta distributions. Instead of using static labels, we dynamically create a target Beta distribution, $P_t \sim Beta(\alpha_t, \beta_t)$, for each training sample based on the properties of the random crop augmentation. Specifically, given an input image, we first apply a random crop. The target Beta distribution is, then, generated using Algorithm 1. This process acts as a sophisticated form of structured label smoothing, transforming data augmentation from a simple regularizer into a rich source of continuous supervision for risk and uncertainty.

For **negative samples** (no crash), the objective is to predict low risk with high confidence. The target distribution is therefore constant: α_t is set to a small positive value ϵ and β_t is set to a large value representing high certainty K_{base}, creating a distribution sharply peaked at zero.

For **positive samples** (crash), the target distribution reflects the quality of the visual evidence in the random crop. This is quantified by an `influence` score, which modulates the target distribution's mean and concentration to generate a supervision signal that is proportional to the information content of the augmented image. The score is a weighted combination of two geometric properties of the crop: its centrality relative to the crash location and its size.

We set the weights to 0.7 for centrality (w_{dist}) and 0.3 for relative size (w_{size}). This weighting scheme is based on the strong intuition that the visual features most critical to understanding risk are spatially concentrated around the event's location. A crop that is well-centered on the crash point provides the clearest and most relevant evidence, thus

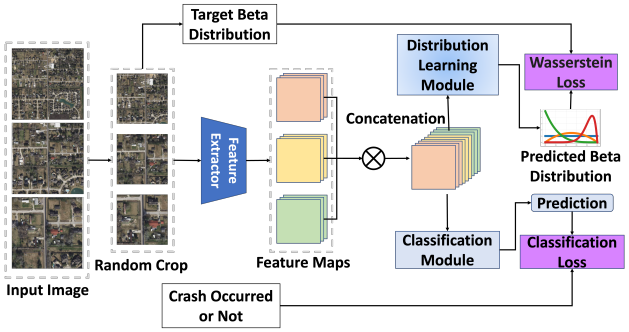


Figure 2: Training Architecture with Joint Optimization

deserving a higher `influence` score and a more confident target distribution. The relative size of the crop provides useful, but secondary, broader context about the surrounding environment. This principled approach transforms data augmentation into a rich source of supervision, teaching the model to dynamically associate higher risk and confidence with visual samples that contain the most informative evidence.

This `influence` score then modulates the target mean μ_{min} and the target concentration k_t, which in turn define the final Beta parameters. For positive samples, the β_t is set to the small constant ϵ, ensuring the distribution is always skewed towards high risk, with the `influence` score controlling the precise shape and confidence.

Model Architecture The architecture of our model, illustrated in Figure 2, is designed to process multi-scale satellite imagery. During training, a random crop is sampled from the input, which consists of image slices of the same location at different resolutions.

The cropped images are, then, passed through a shared feature extractor backbone to produce multiple corresponding feature maps. These maps are concatenated along the channel dimension to form a unified feature representation, serving as the input for two parallel prediction heads:

- A Distribution Learning Head, which outputs the two Beta parameters (α, β).
- An auxiliary Classification Head, which outputs a single logit for the binary crash/no-crash task.

Training and Optimization The model is trained end-to-end by jointly optimizing the two parallel heads with a compound loss function. The primary distribution learning head is supervised by the a mean-variance loss that inspired by the squared Wasserstein-2 (W_2^2) distance (Vaserstein 1969), which measures the dissimilarity between the predicted (P_p) and the target (P_t) Beta distributions:

$$\mathcal{L}_{W_2^2}(P_p, P_t) = (\mu_p - \mu_t)^2 + (\sigma_p - \sigma_t)^2, \qquad (2)$$

where the μ and σ are the mean and standard deviation.

We empirically selected this W_2^2 surrogate over true W_2^2 distance and other distribution divergence metrics, including KL-Divergence (Csiszár 1975) and the Cramér-von Mises criterion (Cramér 1928). As a true metric, our W_2^2 surrogate

loss provides a more stable gradient than KL-Divergence, especially when the predicted and target distributions have little overlap. Most importantly, for one-dimensional distributions like the Beta, the W_2^2 surrogate loss directly optimize of the risk score (the mean) and confidence level (the standard deviation) simultaneously. Our experimental analysis also shows this surrogate is a close approximation of the true W_2^2, deviating only in extreme cases.

The auxiliary classification head is supervised by a Binary Cross-Entropy loss, which encourages the shared backbone to learn discriminative features relevant to the safety task:

$$\mathcal{L}_{BCE} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log(p_i) + (1-y_i)\log(1-p_i)\right], \tag{3}$$

where y_i and p_i are the label and predicted probability.

The overall objective function is a weighted combination of the two losses, balanced by hyperparameters, λ_1 and λ_2:

$$\mathcal{L} = \lambda_1 \cdot \mathcal{L}_{BCE} + \lambda_2 \cdot \mathcal{L}_{W_2^2}. \tag{4}$$

Inference Process The inference process is direct and computationally efficient. The random crop augmentation and the auxiliary classification head are removed. The full, uncropped multi-scale image is passed through the feature extractor backbone and the distribution head. The risk score R is calculated as the mean of the distribution, per Equation 1. This feed-forward process allows for rapid and scalable risk assessment of any location.

Experiment Setup

This study utilizes the MSCM dataset (Liang et al. 2024), a large-scale collection of multi-scale satellite images from Texas, USA, with 16,451 locations labeled with historical fatal crashes. All models use a ResNet-50 (He et al. 2016) backbone, $\lambda_1 = 5, \lambda_2 = 1$, and were trained on NVIDIA A100 GPUs. The best checkpoint for each model was selected based on the model accuracy on the validation set. See the **our project page** for more information about the dataset, implementation details, and hyperparameters.

Evaluation Methodology

Quantitative Metrics We first evaluate our model's practical effectiveness by framing the risk estimation as a binary classification task to identify historical crash locations. The model's predicted risk score R, derived from Equation 1 is thresholded at 0.5 to yield a binary prediction. We then assess the model's predictive performance using standard metrics: F1-Score, Precision, Recall, AUC (Area Under the Receiver Operating Characteristic curve), and PRC (area under the precision-recall curve); and assess model's calibration using Expected Calibration Error (ECE) and Brier score. Due to safety-oriented, **we consider Recall to be the most critical metric** that answers the question: "*Of all crash locations, what fraction did our model successfully identify?*"

We also evaluate our method against a Deep Ensemble (DE) of the strongest baseline, constructed from three independent training runs. The final predicted risk score of a DE model, R_{DE}, is calculated as the mean of the predictions from each individual model in the ensemble. This single score for each sample is then used to compute all the aforementioned performance and calibration metrics.

The ensemble's predictive uncertainty is quantified in two ways: the variance of the risk scores and the disagreement rate among the final binary predictions. A higher value in either metric reflects greater disagreement among the models and thus higher uncertainty in the final prediction.

Qualitative Analysis To intuitively understand the value of our probabilistic approach, we conduct a qualitative analysis of the model's outputs from two perspectives.

First, we analyze the aggregate behavior of the model's outputs by comparing the overall distribution of predicted probabilities from our model against the baselines. By plotting a histogram of all risk scores, we can visually assess model confidence. A well-calibrated, uncertainty-aware model is expected to utilize the full [0, 1] probability range, whereas overconfident models will show predictions heavily clustered at the extremes (near 0 and 1).

Second, we visualize the predicted Beta distributions for four distinct scenarios: true positives (TP), true negatives (TN), false negatives (FN), and false positives (FP). The goal of this analysis is to provide an intuitive understanding of the model's behavior by interpreting its successes and failures.

Case Study: San Antonio River Walk To demonstrate the model's utility, we conduct a case study of the San Antonio River Walk, providing practical insight into the model's performance in a challenging, safety-critical area.

Baseline Models

We evaluate our method against three baselines to isolate our framework's contributions. Our primary benchmark is the Multi-Scale Cross-Matching (MSCM) model (Liang et al. 2024), the current state-of-the-art for fatal crash risk estimation using only satellite imagery.

`ImageNet` Baseline: A standard model pre-trained on ImageNet (Krizhevsky et al. 2012) that takes single-scale satellite images as input, providing us the performance of a generic, non-domain-specific feature extractor on our task.

`MSCM-SS` (Single-Scale): The same single-scale architecture but using weights generated by the self-supervised pre-training through cross-matching, proposed in the MSCM paper, to test the value of domain-specific features.

`MSCM-MS` (Multi-Scale): The full MSCM model, which uses both its domain-specific pre-training and multi-scale imagery as input, represents the strongest available baseline, allowing us to compare against the current state-of-the-art classification approach directly.

Results

Quantitative Analysis Table 1 summarizes the quantitative results. The single-scale baselines (`ImageNet` and `MSCM-SS`) achieve a < 0.5 precision and recall scores, indicating their predictions for positive cases are close to random and exhibit little ability to identify high-risk areas. While the `MSCM-MS` model achieves high precision

Methods	Pre-Train	Probabilistic	Multi-Scale	Performance (↑)					Uncertainty (↓)	
				F1	Precision	Recall	AUC	PRC	ECE	Brier
ImageNet	ImageNet	✗	✗	0.4753	0.4968	0.4555	0.7980	0.4862	0.1281	0.1600
MSCM-SS	MSCM	✗	✗	0.4966	0.4981	0.4950	0.8165	0.5185	0.1006	0.1458
MSCM-MS	MSCM	✗	✓	0.5409	**0.6731**	0.4521	0.8572	0.6269	0.1067	0.1296
Prob-SS (Ours)	MSCM	✓	✗	0.5001	0.4252	**0.6070**	0.7749	0.4409	0.1731	0.1922
Prob-MS (Ours)	MSCM	✓	✓	**0.5762**	0.6296	0.5311	**0.8663**	**0.6489**	**0.0881**	**0.1211**

Table 1: Main Quantitative Results (**bold**: best performance; underlined: second best performance)

Methods	Performance (↑)					Uncertainty (↓)		Disagreement (↓)	
	F1	Precision	Recall	AUC	PRC	ECE	Brier	Variance	Disagr. Rate
Ensemble MSCM-MS	0.5966	**0.7062**	0.5165	**0.8839**	**0.6890**	0.0787	0.1112	0.0925	16.93%
Ensemble Prob-MS (Ours)	**0.5976**	0.6750	**0.5361**	0.8761	0.6886	**0.0605**	**0.1075**	**0.0822**	**15.14%**

Table 2: Deep Ensemble Results over Three Training Trails (**bold**: best performance)

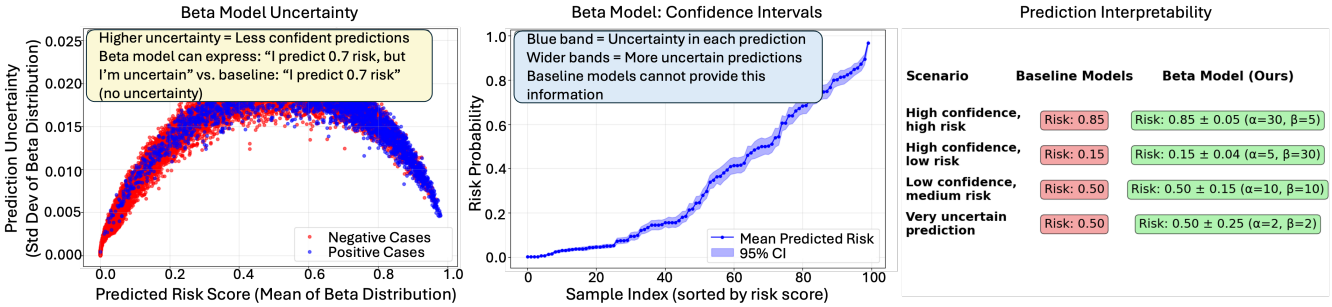


Figure 3: Uncertainty Quantification and Interpretability

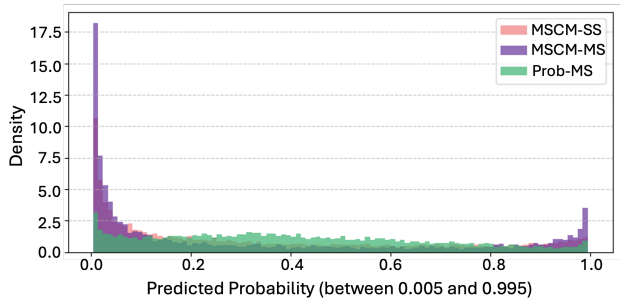


Figure 4: Analysis of Predicted Probability Distributions

(0.6731), its poor recall (0.4521) means it fails to identify over half of all crash locations, rendering it unreliable for safety-critical applications.

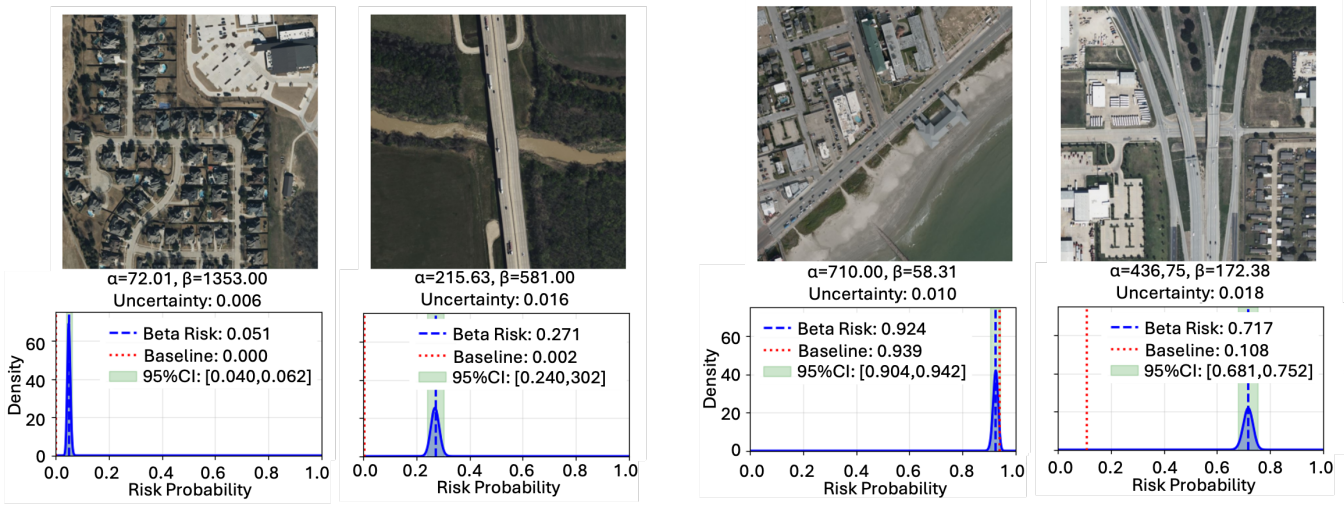
In contrast, our models demonstrate a significant improvement in identifying potential dangers. Our multi-scale model, `Prob-MS`, achieves the best overall balance of performance, attaining the highest F1-score. Its most significant contribution is boosting the recall to 0.5311, a 17% relative improvement over `MSCM-MS`, drastically reducing the number of hazardous sites that would be missed. Our single-scale model, `Prob-SS` (0.6070 recall score), significantly improves the metric by 23% over the best baseline (0.4950).

Crucially, `Prob-MS` is also the most trustworthy model, achieving the lowest (best) ECE of 0.881 and Brier of 0.1211. This confirms that our model's probabilistic outputs are more statistically sound and reliable for real-world decision-making.

We also evaluate our method against a Deep Ensemble of the strongest baseline (Table 2). When comparing our single Prob-MS model against the baseline Ensemble MSCM-MS, we find that our single model achieves competitive performance, including a 3% higher recall, better calibration, and lower uncertainty at only 1/3 the computational cost at both training and inference times. This highlights the efficiency and practical advantage of our approach.

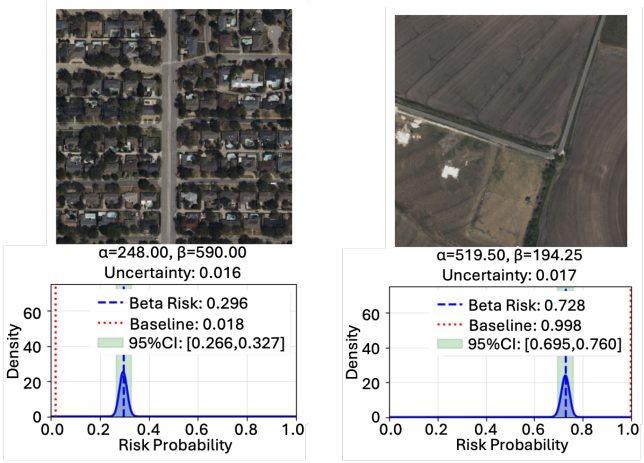
In an apples-to-apples comparison between ensembled methods, our Ensemble Prob-MS demonstrates the clear superiority of our probabilistic framework. It outperforms the baseline ensemble on the most critical metrics for this task, achieving a higher F1-score and recall. Most importantly, it is significantly better calibrated and exhibits lower uncertainty, as evidenced by its superior (lower) ECE, Brier, Variance, and Disagreement Rate scores.

Qualitative Analysis Our qualitative analysis highlights the superior interpretability and trustworthiness of our probabilistic framework. As shown in Figure 3, our model provides a comprehensive and practical understanding of risk



(a) (b)

Figure 5: Qualitative Results for Unambiguous ("Easy") Cases (a: True Negatives, b: True Positives)



(a) (b)

Figure 6: Interpreting Model Behavior on Ambiguous ("Hard") Cases (a: False Negative, b: False Positive)

that standard classifiers cannot offer. The "Beta Model Uncertainty" plot (left) confirms the model's rational behavior, showing that prediction uncertainty is lowest for highly confident predictions and highest for ambiguous ones around a 0.5 risk score. The "Confidence Intervals" plot (center) demonstrates that every prediction is accompanied by a 95% confidence interval, with the interval's width directly communicating the model's certainty on a per-prediction basis. Finally, the "Prediction Interpretability" table (right) crystallizes this key advantage, showing how our Beta model resolves the ambiguity of a baseline's "Risk: 0.50" output by distinguishing between a low-confidence prediction (e.g., with $\alpha = 10, \beta = 10$) and a very uncertain one (e.g., with $\alpha = 2, \beta = 2$). This additional context is invaluable for any safety-critical application.

This nuanced, per-prediction behavior leads to a more rational distribution of predictions in aggregate (Figure 4). While baseline models behave like overconfident black boxes with predictions heavily clustered at the extremes of 0 and 1, our model utilizes the full probability spectrum to express varying degrees of certainty. This ability to be "less confident" is not a weakness but a hallmark of a more honest and trustworthy risk assessment tool.

Visualizing Model Uncertainty To be a trustworthy tool for risk assessment, a model must not only make accurate predictions but also provide a reliable measure of its own uncertainty. We visualize this uncertainty using a Beta distribution for each prediction. As shown in Figure 5, our model demonstrates well-calibrated confidence across a spectrum of cases, a crucial feature for real-world deployment.

For visually unambiguous locations, the model produces predictions with high confidence. For example, in a simple suburban neighborhood (Figure 5a, left), it predicts a low risk (0.051) with a correspondingly low uncertainty score (0.006), reflected in a sharp Beta distribution. Likewise, for a coastal road with high traffic density and high potential of distractions (Figure 5b, left), it correctly predicts a high risk (0.924) with high confidence (uncertainty of 0.010).

The model's utility is further demonstrated in more complex scenarios where it appropriately reduces its confidence. For a visually complex but safe highway overpass (Figure 5a, right), the model still correctly predicts low risk, but the wider Beta distribution indicates higher uncertainty. This nuanced confidence is also evident when assessing a complex highway interchange (Figure 5b, right); the model correctly predicts a high risk of 0.717, but acknowledges the significant uncertainty due to the challenging visual features.

Crucially, the model's rational expression of uncertainty extends to its failures, a characteristic vital for establishing trust. For false negatives (Figure 6a), where the model

misses a crash, the low-risk predictions are consistently paired with wider, higher-uncertainty distributions. This correctly signals that the visual evidence was ambiguous, containing conflicting features, e.g., an arterial road with many intersections within an otherwise low-risk residential area.

Similarly, for false positives (Figure 6b), the model flags locations as high-risk despite no recorded crashes, but again with reduced confidence. This behavior is highly interpretable, as the model correctly identifies latent risk factors, such as sharp (L-shape) turns or high-density highways. The prediction thus reflects a successful identification of hazardous features, while the increased uncertainty correctly marks them as borderline cases. This ability to temper certainty in response to visual complexity, especially when incorrect, distinguishes our model as a more reliable and interpretable system for risk assessment.

Case Study To demonstrate the practical utility of our model, we conducted a case study of the San Antonio River Walk, a major tourist destination that presents a challenging environment with a complex mix of vehicular, pedestrian, and cyclist traffic. We generated risk predictions for over 140 locations in this area using `Prob-MS` and `MSCM-MS`.

The results, shown in Figure 1, highlight the superior performance of our approach. The baseline `MSCM-MS` model (middle panel) fails to identify close to half of the historical fatal crash locations (red diamonds), assigning them erroneously low risk scores. The baseline's predictions also lack spatial coherence, exhibiting sharp, unrealistic gradients between adjacent points and producing polarized risk scores with few intermediate values.

In contrast, our `Prob-MS` model (right panel) correctly assigns elevated risk scores (yellow and orange) to a greater number of the known crash sites. This is exemplified at the intersection near Navarro St and Villita St, a known fatal crash location at the bottom-right in the map. While the baseline model misses this site, ours correctly assigns the area a high-risk score. An analysis of the satellite and ground-level imagery reveals a confluence of latent risk factors not apparent from an overhead view alone. The location, a major entry point to the River Walk, is surrounded by numerous parking facilities. Ground-level images show that these structures, combined with dense trees and building columns, create significant visual obstructions and blind spots for both drivers and pedestrians. This environment forces complex interactions: vehicles constantly enter and exit parking garages across wide pedestrian walkways as tourists navigate narrow sidewalks. Our model likely learned to associate this specific combination of visual clutter, unpredictable vehicle maneuvers, and high pedestrian-vehicle conflict with an elevated risk of a fatal crash.

Furthermore, our model generates a more nuanced and spatially coherent risk map where predictions transition smoothly across locations. This case study demonstrates that our model's strong quantitative performance translates into more reliable, interpretable, and actionable safety assessments for complex urban environments.

Discussion and Conclusion

Our evaluation demonstrates that the proposed probabilistic framework yields a risk assessment model that is not only more effective but also more reliable and interpretable than deterministic baselines. By predicting a full Beta probability distribution instead of a single point-estimate, our model learns a more nuanced and less overconfident representation of risk. This trustworthiness is reinforced by its interpretable behavior; the model's "mistakes" are often rational, such as flagging visually complex but historically safe highway interchanges as high-risk. This capacity to reason about visual factors and express nuanced confidence is highly valuable for practical applications, from enabling more sophisticated path planning for autonomous vehicles to allowing urban planners to confidently prioritize infrastructure improvements. Furthermore, by relying solely on publicly available satellite imagery, our method circumvents the significant privacy concerns associated with other data sources

Ethical Considerations and Responsible Deployment The ethical implications of deploying an AI tool for public safety are significant. As historical crash data may contain undiscovered biases, such as under-reporting in certain socioeconomic or geographic areas, a model used without critical oversight could perpetuate inequities. We therefore emphasize that this model is designed as a decision-support tool to augment, not replace, human expertise.

A key feature for responsible, human-in-the-loop deployment is the model's ability to signal its own uncertainty, which can serve as a bias and fairness mitigation tool. High uncertainty in any prediction (whether high-risk or low-risk) can flag regions with potential data disparities or underreporting. For instance, a visually complex area with high uncertainty and a low-risk prediction may indicate a dangerous false negative due to a lack of historical crash data. These uncertain predictions should act as a flag for human experts to conduct a more detailed investigation, enabling a more equitable allocation of safety resources.

Limitations This study has several limitations that open avenues for future research. Our model estimates static risk and does not account for dynamic variables like real-time traffic or weather; future work should focus on integrating these data streams. Our study is also geographically constrained to Texas, and validation on diverse international datasets is a critical next step to ensure generalizability. Furthermore, this work can be extended by exploring a learned weighting mechanism for the centrality and size components of our procedural labeling scheme. Finally, while our model identifies strong correlations, future work could explore methods for moving toward causal inference.

In conclusion, this work demonstrates that moving from deterministic point-estimates to a full probabilistic framework is a crucial step toward creating more reliable and trustworthy AI for public safety. By learning to predict a Beta probability distribution from satellite imagery, our model not only outperforms existing baselines in identifying high-risk locations but also provides the well-calibrated uncertainty estimates that are vital for interpretable, human-in-

the-loop decision-making in applications from urban planning to autonomous navigation.

Acknowledgments

This material is partially based upon work supported by the National Science Foundation under 2401860 and 2526487. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and the funders have no role in the study design, data collection, analysis, or preparation of this article.

Portions of this research utilized high-performance computing resources provided by the High Performance Computing Research Center at Texas A&M University–San Antonio, led by Dr. Izzat Alsmadi. We gratefully acknowledge his support in facilitating access to these resources.

References

Ahmed, I. 2013. Road infrastructure and road safety. *Transport and Communications Bulletin for Asia and the Pacific*, 83: 19–25.

Al-Sharif, L.; et al. 2012. The use of Monte Carlo simulation in evaluating the elevator round trip time under up-peak traffic conditions and conventional group control. *Building Services Engineering Research and Technology*, 33(3): 319–338.

Caliendo, C.; et al. 2007. *Accident Analysis & Prevention*, 39(4): 657–670.

Carrodano, C. 2024. Data-driven risk analysis of nonlinear factor interactions in road safety using Bayesian networks. *Scientific Reports*, 14(1): 18948.

Chen, C.; and Sundar, S. S. 2023. Is this ai trained on credible data? the effects of labeling quality and performance bias on user trust. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–11.

Cheng, G.; Cheng, R.; Zhang, S.; and Sun, X. 2019. Risk evaluation method for highway roadside accidents. *Advances in Mechanical Engineering*, 11(1): 1687814018821743.

Chidambaram, M.; and Ge, R. 2023. On the Limitations of Temperature Scaling for Distributions with Overlaps. In *International Conference on Learning Representations*.

Cramér, H. 1928. On the composition of elementary errors. *Scandinavian Actuarial Journal*, 1928(1): 13–74.

Csiszár, I. 1975. I-divergence geometry of probability distributions and minimization problems. *The annals of probability*, 146–158.

de Almeida Guimarães, V.; et al. 2018. Evaluating the sustainability of urban passenger transportation by Monte Carlo simulation. *Renewable and Sustainable Energy Reviews*, 93: 732–752.

Gu, C.; Xu, J.; Gao, C.; Mu, M.; E, G.; and Ma, Y. 2022. Multivariate analysis of roadway multi-fatality crashes using association rules mining and rules graph structures: A case study in China. *Plos one*, 17(10): e0276817.

Guo, C.; Pleiss, G.; Sun, Y.; and Weinberger, K. Q. 2017. On calibration of modern neural networks. In *International conference on machine learning*, 1321–1330. PMLR.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

He, S.; Sadeghi, M. A.; Chawla, S.; Alizadeh, M.; Balakrishnan, H.; and Madden, S. 2021. Inferring high-resolution traffic accident risk maps based on satellite imagery and GPS trajectories. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11977–11985.

Hinton, G. E.; et al. 2015. Distilling the Knowledge in a Neural Network. *ArXiv*, abs/1503.02531.

Huang, T.; et al. 2020. Highway crash detection and risk estimation using deep learning. *Accident Analysis Prevention*, 135: 105392.

Jaroszweski, D.; and McNamara, T. 2014. The influence of rainfall on road accidents in urban areas: A weather radar approach. *Travel behaviour and society*, 1(1): 15–21.

Jeon, S.; and Hong, B. 2016. Monte Carlo simulation-based traffic speed forecasting using historical big data. *Future generation computer systems*, 65: 182–195.

Jonnala, R.; Liang, G.; Yang, J.; and Alsmadi, I. 2025. Exploring the potential of large language models in public transportation: San antonio case study.

Joo, Y.-J.; et al. 2023. A generalized driving risk assessment on high-speed highways using field theory. *Analytic Methods in Accident Research*, 40: 100303.

Krizhevsky, A.; et al. 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25.

Kumar, A.; et al. 2018. Trainable calibration measures for neural networks from kernel mean embeddings. In *International Conference on Machine Learning*, 2805–2814.

Li, Y.; et al. 2024. Label Bias: A Pervasive and Invisibilized Problem. *Notices of the American Mathematical Society*, 71(8): 1069–1077.

Liang, G.; Zhang, Y.; Wang, X.; and Jacobs, N. 2020. Improved Trainable Calibration Method for Neural Networks on Medical Imaging Classification. In *British Machine Vision Conference (BMVC)*.

Liang, G.; Zulu, J.; Xing, X.; and Jacobs, N. 2024. Unveiling roadway hazards: Enhancing fatal crash risk estimation through multiscale satellite imagery and self-supervised cross-matching. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 17: 535–546.

Lin, S.-C.; Su, Y.; Liang, G.; Zhang, Y.; Jacobs, N.; and Zhang, Y. 2022. Estimating cluster masses from SDSS multiband images with transfer learning. *Monthly Notices of the Royal Astronomical Society*, 512(3): 3885–3894.

Liu, L.; Wang, Y.; Chang, J.; Zhang, P.; Liang, G.; and Zhang, H. 2022. LLRHNet: multiple lesions segmentation using local-long range features. *Frontiers in Neuroinformatics*, 16: 859973.

Ma, Q.; Yang, H.; Wang, Z.; Xie, K.; and Yang, D. 2020. Modeling crash risk of horizontal curves using large-scale auto-extracted roadway geometry data. *Accident Analysis Prevention*, 144: 105669.

Najjar, A.; et al. 2017. Combining satellite imagery and open data to map road safety. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

Pembuain, A.; et al. 2019. The effect of road infrastructure on traffic accidents. In *11th Asia Pacific Transportation and the Environment Conference (APTE 2018)*, 176–182. Atlantis Press.

Pereyra, G.; Tucker, G.; Chorowski, J.; Kaiser, Ł.; and Hinton, G. 2017. Regularizing neural networks by penalizing confident output distributions. In *International Conference on Learning Representations*.

Simons-Morton, B. G.; Guo, F.; Klauer, S. G.; Ehsani, J. P.; and Pradhan, A. K. 2014. Keep your eyes on the road: Young driver crash risk increases according to duration of distraction. *Journal of Adolescent Health*, 54(5): S61–S67.

Song, W.; Workman, S.; Hadzic, A.; Zhang, X.; Green, E.; Chen, M.; Souleyrette, R.; and Jacobs, N. 2018. FARSA: Fully Automated Roadway Safety Assessment. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 521–529.

Tamerius, J.; Zhou, X.; Mantilla, R.; and Greenfield-Huitt, T. 2016. Precipitation effects on motor vehicle crashes vary by space, time, and environmental conditions. *Weather, Climate, and Society*, 8(4): 399–407.

Vaserstein, L. N. 1969. Markov processes over denumerable products of spaces, describing large systems of automata. *Problemy Peredachi Informatsii*, 5(3): 64–72.

WHO. 2018. Global status report on road safety. *World Health Organization*.

WHO. 2023. Road Traffic Injuries. *World Health Organization*. Accessed: 2025-05-22.

Xing, X.; Liang, G.; Wang, C.; Jacobs, N.; and Lin, A.-L. 2023. Self-supervised learning application on covid-19 chest x-ray image classification using masked autoencoder. *Bioengineering*, 10(8): 901.

Zhu, C.; Dadashova, B.; Lee, C.; Ye, X.; and Brown, C. T. 2024. Equity in non-motorist safety: Exploring two pathways in Houston. *Transportation research part D: transport and environment*, 132: 104239.

Zulu, J.; Han, B.; Alsmadi, I.; and Liang, G. 2024. Enhancing machine learning based sql injection detection using contextualized word embedding. In *Proceedings of the 2024 ACM Southeast Conference*, 211–216.