ALCo-FM: Adaptive Long-Context Foundation Model for Accident Prediction

Pinaki Prasad Guha Neogi [1] **Ahmad Mohammadshirazi** [1] **Rajiv Ramnath** [1]

Abstract

Traffic accidents are rare, yet high-impact events that require long-context multimodal reasoning for accurate risk forecasting. In this paper, we introduce **ALCo-FM**, a unified adaptive long-context foundation model that computes a volatility pre-score to dynamically select context windows for input data and encodes and fuses these multimodal data via shallow cross attention. Following a local GAT layer and a BigBird-style sparse global transformer over H3 hexagonal grids, coupled with Monte Carlo dropout for confidence, the model yields superior, well-calibrated predictions. Trained on data from 15 U.S. cities with a class-weighted loss to counter label imbalance, and fine-tuned with minimal data on held-out cities, ALCo-FM achieves 0.94 accuracy, 0.92 F1, and an ECE of 0.04—outperforming 20+ state-of-the-art baselines in large-scale urban risk prediction. Code and dataset are available at: https://github.com/PinakiPrasad12/ALCo-FM

1. Introduction

Road traffic accidents account for over 1.19 million deaths worldwide each year and generate substantial economic losses (World Health Organization, 2025). Appendix A shows a 25% rise in U.S. traffic fatalities over the last decade despite recent marginal declines, revealing a dynamic landscape that fixed-horizon predictors cannot fully capture.

Existing forecasting methods—from ensemble and transformer-based time-series models to Vision Transformers on map tiles and spatio-temporal GNNs—typically (1) assume a short, fixed history window, (2) operate on a single data modality, or (3) omit uncertainty calibration, and most critically, (4) fail to model long-range context jointly across modalities. These gaps limit their effectiveness in real-world, large-scale deployments.

To address these challenges, we introduce the **Adaptive Long-Context Foundation Model (ALCo-FM)**, whose key innovations are:

- **Volatility-Driven Context Selection** We compute a lightweight pre-score from each 1h ContiFormer (Chen et al., 2024b) + T2T-ViT (Yuan et al., 2021) embedding to gate between 1h, 3h, or 6h look-back windows, ensuring more history is used when and where it matters.
- **Unified Dual-Transformer Encoding Fusion** Numerical time-series and map imagery are encoded in parallel by a ContiFormerand a T2T-ViT, then seamlessly fused via shallow cross-attention to capture rich temporal–spatial interactions.
- **Scalable Hybrid Attention on H3 Grids** We first propagate local neighborhood information with a GAT layer, then apply a BigBird-style sparse global transformer for efficient, city-wide context aggregation across all H3 cells (Uber Technologies, 2023).
- **Foundation-Scale Calibration Generalization** A 2-layer MLP head is calibrated with Monte Carlo dropout for reliable uncertainty estimates, and ALCo-FM is pretrained on 15 U.S. cities with a class-weighted loss, then fine-tuned on new regions using minimal data.

These advances combine to deliver robust, long-context multimodal accident-risk forecasting at urban scale. Evaluated on 1,771 regions across 15 U.S. cities—and, after minimal fine-tuning, on 3 held-out cities—our model sets new benchmarks in accuracy, F1, and calibration, demonstrating both long-context capability and foundation-model versatility.

2. Related Work

Traffic-risk prediction has been primarily approached via graph-based and attention-driven spatio-temporal models. Diffusion and convolution-based GNNs such as DCRNN (Li et al., 2018), STGCN (Yu et al., 2018), and Graph WaveNet (Wu et al., 2019) combine graph convolutions with recurrent or dilated temporal filters to capture local propagation but struggle with long-range dependencies. Attention and ODE-based methods—including ASTGCN (Guo et al., 2019), STSGCN (Song et al., 2020), and

[1]Department of Computer Science and Engineering, Ohio State University, Ohio, US. Correspondence to: Pinaki Prasad Guha Neogi <guhaneogi.2.2@osu.edu>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

STGODE (Fang et al., 2021)—introduce adaptive reweighting and continuous-time dynamics, yet often scale poorly on large graphs and lack multi-modal integration.

On the other hand, neural architecture search approaches like AutoSTG (Pan et al., 2021) and Auto-DSTSG (Jin et al., 2022) automate spatio-temporal block selection but remain confined to predefined search spaces. Hierarchical and multi-scale networks (e.g. SST-DHL (Cui et al., 2024), VSTGCN (Gan et al., 2024), SST-GCN (Kim et al., 2024), DGCRN (Li et al., 2023), HetConvLSTM (Yuan et al., 2018), DSTGCN (Yu et al., 2021), MVMT-STN (Wang et al., 2021), UTAASTRL (Bao et al., 2020), UTARPR (Chen et al., 2024a), GLST-TARP (Alhaek et al., 2025), FC-STGNN (Wang et al., 2024)) capture multi-resolution patterns and sometimes quantify uncertainty, but typically focus on single modalities and fixed horizons.

By contrast, our proposed framework unites uncertainty-driven context window selection, cross-modal fusion of time-series and imagery, and a hybrid local–global attention mechanism, providing scalable long-range reasoning with calibrated risk estimation in a single end-to-end model.

3. Dataset Description

We construct a high-resolution, multi-source dataset tailored for adaptive long-context accident prediction.

3.1. Data Preparation

We aggregate four heterogeneous data sources on an H3 hexagonal grid at resolution $R = 7$ (Uber Technologies, 2023). First, we incorporate traffic events from the nationwide dataset of Moosavi et al. (Moosavi et al., 2019), which spans 49 U.S. states and provides precise timestamps, GPS locations, and contextual details per incident (Appendix B.1). Second, we enrich each cell with demographic attributes—150 socio-economic variables such as income, population density, and age distribution—sourced for 45,000 ZIP codes from the U.S. ZIP Code database (United States ZIP Codes) (Appendix B.2). Third, we align hourly meteorological observations (2016–2023) from the Iowa Environmental Mesonet's ASOS network (Iowa State University), including temperature, precipitation, and wind speed, to each accident record (Appendix B.3). Finally, for spatial context we download 256×256 map tiles from OpenStreetMap (OpenStreetMap contributors, 2017) centered at each hexagon's centroid (Appendices B.4, D.3). Together, these inputs form a rich, multi-modal foundation for our adaptive long-context modeling.

These data sources, combined with rigorous preprocessing and advanced feature engineering, allow us to model the intricate dependencies influencing accident risk. Further dataset details and feature descriptions are provided in Appendix B. Also, a detailed statistical and structural analysis of our dataset is provided in Appendix C, which highlights its high dimensionality, severe class imbalance, weak feature correlations, and nonlinear separability—underscoring its real-world complexity and the need for advanced modeling techniques. The final dataset for our experiment consists of 1,771 regions across 15 U.S. cities (Appendix D.4).

Long-context. Although in our experiment we consider only 1-6 hours window (4.1), each hour yields a high-dimensional, multimodal embedding: up to hundreds of numeric features and P visual patch tokens per cell. A 6h window therefore produces on the order of $6T + 6P$ tokens per node, and we must reason over all 1,771 cells simultaneously. This naturally gives rise to a "long-context" modeling problem, motivating our adaptive windowing, cross-modal fusion, and sparse global attention mechanisms (Section 4).

3.2. Data Preprocessing

All sources are first merged on the H3 cell index and timestamp and then aligned into uniform **1-hour** atomic windows. We enrich each atomic window with temporal indicators (rush hour, part of day, U.S. holidays) and simple geographic summaries (neighbor counts, road density), impute any remaining gaps via FAISS k-NN (Johnson et al., 2019)(Refer Appendix D for full preprocessing pipeline and rationale). These 1-hour windows are the input units that get dynamically grouped into adaptive long-context spans in the model (Section 4.1).

4. Methodology

In this section, we step through our unified adaptive long-context framework for accident prediction, which seamlessly blends spatial, temporal, and visual information within a single model. Detailed descriptions of each component presented in Sections 4.1–4.5.

4.1. H3 Grid Mapping & Adaptive Long-Context Dual Encoding

We tessellate the target area into H3 hexagons, yielding nodes v_i. For each v_i, we aggregate traffic, weather, and demographic records over a one-hour atomic window and encode them via paired encoders:

Numerical Encoder (ContiFormer). The 1-hour numerical time series data at node v_i is processed by a continuous-time state-space transformer (ContiFormer), producing T tokens $X_{num} \in \mathbb{R}^{T \times d}$ that capture temporal dynamics and serve both for uncertainty scoring and downstream fusion.

Visual Encoder (T2T-ViT). Concurrently, we crop a 256×256 map tile around v_i and tokenize it with a Token-to-Token Vision Transformer into P patch tokens

$X_{vis} \in \mathbb{R}^{P\times d}$, encoding spatial context.

The resulting dual embedding $\big[X_{num}; X_{vis}\big]$ is then used to compute a volatility signal u, which, in turn is used to decide how many hours of history to use. This strategy—similar to (Graves, 2017) and (Sukhbaatar et al., 2019)—ensures that more history is used when volatility is high, and compute is conserved otherwise):

$$\sigma_{\text{num}} = \tfrac{1}{d}\sum_{j=1}^{d} \text{std}_t\big(X_{\text{num}}[:, j]\big)$$

$$\sigma_{\text{vis}} = \tfrac{1}{d}\sum_{j=1}^{d} \text{std}_p\big(X_{\text{vis}}[:, j]\big)$$

$$u = \tfrac{1}{2}\big(\sigma_{\text{num}} + \sigma_{\text{vis}}\big)$$

Here std_t is computed across the T temporal tokens, std_p across the P visual patches. We then set thresholds $\tau_{\text{low}}, \tau_{\text{high}}$ to the 33rd/67th percentiles of all training-set u, refining via validation grid-search to optimize F1 and calibration. And, based on the volatility score u, we choose a look-back window w (in hours) as:

$$w \in \{1, 3, 6\}, \quad w = \begin{cases} 6 & u > \tau_{high}, \\ 3 & \tau_{low} \le u \le \tau_{high}, \\ 1 & u < \tau_{low}. \end{cases}$$

This adaptive long-context mechanism uses more history when volatility is high, and conserves compute when the signal is stable; during training, we randomly sample w to build robustness across spans.

4.2. Cross-Modal Long-Context Fusion

Here, we fetch the last w atomic windows and re-encode them to obtain $(X_{\text{num}}, X_{\text{vis}}) \in \mathbb{R}^{wT\times d} \times \mathbb{R}^{wP\times d}$. We then apply $L = 2$ shallow cross-attention layers (single head, $d_h = 128$):

$$\text{CM}(X_{num}, X_{vis}) = \mathcal{S}\big(\tfrac{X_{num}W_Q\,(X_{vis}W_K)^\top}{\sqrt{d_k}}\big)\,(X_{vis}W_V)$$

with residuals and LayerNorm, symmetrically in both directions. Here $\mathcal{S}(\cdot)$ denotes a softmax function. Mean-pooling yields fused embeddings $\widetilde{x}_{num}, \widetilde{x}_{vis} \in \mathbb{R}^d$ per node.

4.3. Spatio-Temporal Graph Construction & Local GAT

We build a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ over the N H3 cells at the current time window, where edges connect each cell to its k immediate hex-grid neighbors. Given the fused embeddings $\widetilde{x}^i_{\text{num}}, \widetilde{x}^i_{\text{vis}} \in \mathbb{R}^d$ for node i, we apply one GAT layer:

$$h'_i = \sigma\Big(\sum_{j\in\mathcal{N}(i)} \alpha_{ij}\, W_g\,[\widetilde{x}^j_{\text{num}}; \widetilde{x}^j_{\text{vis}}]\Big),$$

$$\alpha_{ij} = \frac{\exp\big(\text{LeakyReLU}(a^\top[\,W_g[\widetilde{x}^i];\, W_g[\widetilde{x}^j]\,])\big)}{\sum_{k\in\mathcal{N}(i)}\exp\big(\text{LeakyReLU}(a^\top[\ldots])\big)}.$$

Here $[\cdot;\cdot]$ denotes concatenation, W_g and a are learnable, and σ is a nonlinearity (ReLU). We then stack these outputs into a matrix

$$Z = \begin{bmatrix} (h'_1)^\top \\ (h'_2)^\top \\ \vdots \\ (h'_N)^\top \end{bmatrix} \in \mathbb{R}^{N\times d},$$

which serves as the input to the global attention module.

4.4. Spatio-Temporal Sparse Global Attention

To capture long-range, city-wide interactions efficiently, we apply a BigBird-style sparse transformer to Z. The attention is defined as

$$\text{STA}(Z) = \text{softmax}\Big(\tfrac{ZW_Q\,(ZW_K)^\top}{\sqrt{d_k}} + M\Big)\,(ZW_V),$$

where $W_Q, W_K, W_V \in \mathbb{R}^{d\times d_k}$ are projection matrices; and $M \in \mathbb{R}^{N\times N}$ is a binary mask that preserves full connectivity between each of the G learnable global tokens and all N cells in a city, and restricts each cell to attend only to its k (= 6 in H3) spatial neighbors (as in the GAT).

A single sparse-attention block with residual connections and LayerNorm produces refined representations $\{z'_i\}_{i=1}^N$, which combine both local GAT context and long-range city-wide signals.

4.5. Uncertainty-Aware Risk Calibration

We train the final 2-layer MLP with a class-weighted binary cross-entropy

$$\mathcal{L} = -w_1\, y \log \hat{y} - w_0\,(1-y)\log(1-\hat{y}),$$

where $w_1 > w_0$ are set inversely proportional to class frequencies, giving more weight to the rare "risky" events to boost recall ("risky" events are rarer as shown in Figure 5).

- **MC Dropout**: We keep dropout active at inference (dropout rate $p = 0.2$) and perform $K = 10$ stochastic forward passes. For outputs $\{y_1, \ldots, y_K\}$ we compute

  $$\hat{y} = \frac{1}{K}\sum_{i=1}^{K} y_i, \quad \sigma = \sqrt{\frac{1}{K}\sum_{i=1}^{K}(y_i - \hat{y})^2},$$

  and report a 95% confidence interval $\hat{y} \pm 1.96\,\sigma$.

This single-model approach is both efficient and effective at capturing epistemic uncertainty over our long-context risk estimates.

5. Experiments and Results

Building on our adaptive long-context framework, we evaluate three aspects: (1) module-wise gains via progressive ablation, (2) end-to-end comparison to state-of-the-art baselines, and (3) transfer to unseen cities using fine-tuning.

5.1. Experimental Setup

All variants train on 6× NVIDIA H100 GPUs (80 GB each) with a global batch size of 12,288 (nodes × time windows). We vary the history window $w \in \{1\,\mathrm{h}, 3\,\mathrm{h}, 6\,\mathrm{h}\}$ via our adaptive gating and train for 40 epochs using AdamW with linearly-scaled learning rates ($\mathrm{LR}_{\mathrm{ContiFormer}} = 7.5 \times 10^{-4}$, $\mathrm{LR}_{\mathrm{ViT}} = 1.5 \times 10^{-5}$, weight decay 1×10^{-4}). We report **Accuracy**, **F1**, **Precision**, **Recall**, and **Expected Calibration Error (ECE)** to evaluate both long-context performance and confidence calibration.

5.2. Progressive Ablation

We start from a fixed-context baseline (ContiFormer + T2T-ViT, no graph, $w = 3h$), then add components in turn progressively. Table 1 shows that the addition of each module leads to consistent gains in both F1 and confidence scores.

Table 1. Ablation results: F1 / ECE (↓)

Variant	F1	ECE
Baseline	0.82	0.12
+ Local GAT	0.84	0.11
+ Cross-Modal Fusion	0.86	0.10
+ Sparse Global Attention	0.88	0.08
+ MC-Dropout Calibration	0.88	0.05
+ Adaptive Gating (complete pipeline)	**0.92**	**0.04**

5.3. Comparison to State-of-the-Art

We compare our full adaptive long-context model against leading baselines-all using fixed 3h context and no cross-modal fusion. Table 2 presents the comparison with the state-of-the-art models, and it is evident that our approach yields a +1–2 pp F1 boost and slashes ECE from 0.15 to 0.04, highlighting superior long-range reasoning and uncertainty control. Note that, SST-GCN shows better accuracy than ours, but F1 score—being the harmonic mean of precision and recall—provides a more meaningful assessment under imbalanced datasets, as it directly measures the model's ability to correctly identify rare but critical accident events.

5.4. Generalization to Unseen Cities

We fine-tune on three held-out cities (Columbus, Portland, Oklahoma City) by unfreezing only the final GAT layer and MLP head for 5 epochs with early stopping. Table 3 shows that F1 remains high and ECE stays under 0.06, confirming robust transfer of our adaptive long-context framework.

Table 2. Performance on 15 Cities

Method	Accuracy	F1	Precision	Recall
DCRNN	0.55	0.46	0.42	0.50
STGCN	0.64	0.49	0.51	0.47
ASTGCN	0.74	0.60	0.58	0.63
GWN	0.72	0.53	0.55	0.52
STSGCN	0.85	0.72	0.78	0.67
STFGNN	0.82	0.65	0.66	0.64
STGODE	0.84	0.65	0.59	0.72
AutoSTG	0.71	0.49	0.50	0.48
Auto-DSTSG	0.73	0.59	0.63	0.55
SST-DHL	0.89	0.78	0.85	0.72
VSTGCN	0.80	0.74	0.78	0.70
SST-GCN	**0.95**	0.80	0.89	0.73
DGCRN	0.89	0.72	0.74	0.71
HetConvLSTM	0.66	0.48	0.52	0.45
DSTGCN	0.84	0.66	0.67	0.65
MVMT-STN	0.79	0.59	0.60	0.59
UTAASTRL	0.81	0.59	0.58	0.61
UTARPR	0.90	0.76	0.66	0.89
GLST-TARP	0.92	0.85	0.83	0.88
FC-STGNN	0.94	0.79	0.81	0.78
ALCo-FM	0.94	**0.92**	**0.91**	**0.93**

Table 3. Fine-Tuning on Unseen Cities

City	Accuracy	F1	Precision	Recall
Columbus	0.92	0.90	0.89	0.91
Portland	0.93	0.90	0.88	0.93
Oklahoma City	0.90	0.89	0.88	0.91

6. Conclusion and Future Work

We present ALCo-FM, an adaptive long-context foundation model that dynamically selects history lengths via uncertainty-driven H3-hexagon aggregation and fuses numerical, environmental, and spatial-temporal signals within a single backbone. Evaluated on 1,771 regions across 15 U.S. cities, ALCo-FM achieves 0.94 accuracy and 0.92 F1—surpassing over 20 strong baselines—and generalizes to three unseen cities with minimal fine-tuning. Its uncertainty-aware design delivers not only state-of-the-art performance but also well-calibrated risk estimates, making it both interpretable and reliable for real-world deployments. Future work will explore adaptive spatial indexing beyond fixed H3 grids, incorporate telematics and mobile-sensor streams, and develop cross-city domain-adaptation techniques to further bolster robustness and generalization.

References

Alhaek, F., Li, T., Rajeh, T. M., Javed, M. H., and Liang, W. Encoding global semantic and localized geographic spatial-temporal relations for traffic accident risk prediction. *Information Sciences*, 697:121767, 2025.

Bao, W., Yu, Q., and Kong, Y. Uncertainty-based traffic accident anticipation with spatio-temporal relational learning. In *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 2682–2690, 2020.

Breiman, L. Random forests. *Machine Learning*, 45 (1):5–32, October 2001. ISSN 1573-0565. doi: 10.1023/A:1010933404324. URL https://doi.org/10.1023/A:1010933404324.

Chen, M., Yuan, H., Jiang, N., Bao, Z., and Wang, S. Urban traffic accident risk prediction revisited: Regionality, proximity, similarity and sparsity. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, CIKM '24, pp. 281–290. ACM, October 2024a. doi: 10.1145/3627673.3679567. URL http://dx.doi.org/10.1145/3627673.3679567.

Chen, T. and Guestrin, C. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pp. 785–794. ACM, August 2016. doi: 10.1145/2939672.2939785. URL http://dx.doi.org/10.1145/2939672.2939785.

Chen, Y., Ren, K., Wang, Y., Fang, Y., Sun, W., and Li, D. Contiformer: Continuous-time transformer for irregular time series modeling. *Advances in Neural Information Processing Systems*, 36, 2024b.

Cui, P., Yang, X., Abdel-Aty, M., Zhang, J., and Yan, X. Advancing urban traffic accident forecasting through sparse spatio-temporal dynamic learning. *Accident Analysis & Prevention*, 200:107564, 2024.

Du, W., Côté, D., and Liu, Y. Saits: Self-attention-based imputation for time series. *Expert Systems with Applications*, 219:119619, June 2023. ISSN 0957-4174. doi: 10.1016/j.eswa.2023.119619. URL http://dx.doi.org/10.1016/j.eswa.2023.119619.

Fang, Z., Long, Q., Song, G., and Xie, K. Spatial-temporal graph ode networks for traffic flow forecasting. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery amp; Data Mining*, KDD '21, pp. 364–373. ACM, August 2021. doi: 10.1145/3447548.3467430. URL http://dx.doi.org/10.1145/3447548.3467430.

Gan, J., Yang, Q., Zhang, D., Li, L., Qu, X., and Ran, B. A novel voronoi-based spatio-temporal graph convolutional network for traffic crash prediction considering geographical spatial distributions. *IEEE Transactions on Intelligent Transportation Systems*, 2024.

Graves, A. Adaptive computation time for recurrent neural networks, 2017. URL https://arxiv.org/abs/1603.08983.

Guo, C., Yang, W., Liu, C., and Li, Z. Iterative missing value imputation based on feature importance. *Knowledge and Information Systems*, 66(10):6387–6414, July 2024. ISSN 0219-3116. doi: 10.1007/s10115-024-02159-7. URL http://dx.doi.org/10.1007/s10115-024-02159-7.

Guo, S., Lin, Y., Feng, N., Song, C., and Wan, H. Attention based spatial-temporal graph convolutional networks for traffic flow forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 922–929, 2019.

Insurance Institute for Highway Safety (IIHS). Yearly snapshot – fatality statistics, 2024. URL https://www.iihs.org/topics/fatality-statistics/detail/yearly-snapshot. Latest data released by IIHS, accessed: 2025-01-28.

Iowa State University. Iowa environmental mesonet (iem) asos-awos-metar data. https://mesonet.agron.iastate.edu/request/download.phtml. Accessed: 2025-01-26.

Jin, G., Li, F., Zhang, J., Wang, M., and Huang, J. Automated dilated spatio-temporal synchronous graph modeling for traffic prediction. *IEEE Transactions on Intelligent Transportation Systems*, 24(8):8820–8830, 2022.

Johnson, J., Douze, M., and Jégou, H. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7 (3):535–547, 2019.

Kim, T.-w., Lee, H.-j., Jung, H.-J., Yang, J.-W., and Hong, E. J. Sst-gcn: The sequential based spatio-temporal graph convolutional networks for minute-level and road-level traffic accident risk predictio. *arXiv preprint arXiv:2405.18602*, 2024.

Lam, N. S.-N. Spatial interpolation methods: a review. *The American Cartographer*, 10(2):129–150, 1983.

Li, F., Feng, J., Yan, H., Jin, G., Yang, F., Sun, F., Jin, D., and Li, Y. Dynamic graph convolutional recurrent network for traffic prediction: Benchmark and solution. *ACM Transactions on Knowledge Discovery from Data*, 17(1):1–21, 2023.

Li, Y., Yu, R., Shahabi, C., and Liu, Y. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting, 2018. URL https://arxiv.org/abs/1707.01926.

Maćkiewicz, A. and Ratajczak, W. Principal components analysis (pca). *Computers Geosciences*, 19(3):303–342, 1993. ISSN 0098-3004. doi: https://doi.org/10.1016/0098-3004(93)90090-R. URL https://www.sciencedirect.com/science/article/pii/009830049390090R.

Moosavi, S., Samavatian, M. H., Parthasarathy, S., and Ramnath, R. A countrywide traffic accident dataset. *arXiv preprint arXiv:1906.05409*, 2019.

OpenStreetMap contributors. Planet dump retrieved from https://planet.osm.org . https://www.openstreetmap.org, 2017.

Pan, Z., Ke, S., Yang, X., Liang, Y., Yu, Y., Zhang, J., and Zheng, Y. Autostg: Neural architecture search for predictions of spatio-temporal graph. In *Proceedings of the Web Conference 2021*, WWW '21, pp. 1846–1855, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383127. doi: 10.1145/3442381.3449816. URL https://doi.org/10.1145/3442381.3449816.

Song, C., Lin, Y., Guo, S., and Wan, H. Spatial-temporal synchronous graph convolutional networks: A new framework for spatial-temporal network data forecasting. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(01):914–921, Apr. 2020. doi: 10.1609/aaai.v34i01.5438. URL https://ojs.aaai.org/index.php/AAAI/article/view/5438.

Sukhbaatar, S., Grave, E., Bojanowski, P., and Joulin, A. Adaptive attention span in transformers, 2019. URL https://arxiv.org/abs/1905.07799.

Uber Technologies, I. Uber h3: Hexagonal hierarchical geospatial indexing system. https://uber.github.io/h3/, 2023. Accessed: 2025-01-13.

United States Postal Service. Zip code lookup and address information. https://www.usps.com. Accessed: 2025-01-13.

United States ZIP Codes. United states zip codes database. https://www.unitedstateszipcodes.org/. Accessed: 2025-01-13.

U.S. Census Bureau. United states census data. https://www.census.gov. Accessed: 2025-01-13.

Wang, S., Zhang, J., Li, J., Miao, H., and Cao, J. Traffic accident risk prediction via multi-view multi-task spatio-temporal networks. *IEEE Transactions on Knowledge and Data Engineering*, 35(12):12323–12336, 2021.

Wang, Y., Xu, Y., Yang, J., Wu, M., Li, X., Xie, L., and Chen, Z. Fully-connected spatial-temporal graph for multivariate time-series data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 15715–15724, 2024.

World Health Organization. Global road safety data – yearly snapshot, 2021. URL https://apps.who.int/gho/data/node.main.A997. Latest data released by WHO, accessed: 2025-01-28.

World Health Organization. Road traffic injuries, 2025. URL https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries. Accessed: 2025-01-09.

Wu, Z., Pan, S., Long, G., Jiang, J., and Zhang, C. Graph wavenet for deep spatial-temporal graph modeling, 2019. URL https://arxiv.org/abs/1906.00121.

Yu, B., Yin, H., and Zhu, Z. Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pp. 3634–3640. International Joint Conferences on Artificial Intelligence Organization, 7 2018. doi: 10.24963/ijcai.2018/505. URL https://doi.org/10.24963/ijcai.2018/505.

Yu, L., Du, B., Hu, X., Sun, L., Han, L., and Lv, W. Deep spatio-temporal graph convolutional network for traffic accident prediction. *Neurocomputing*, 423:135–147, 2021.

Yuan, L., Chen, Y., Wang, T., Yu, W., Shi, Y., Jiang, Z.-H., Tay, F. E., Feng, J., and Yan, S. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 558–567, October 2021.

Yuan, Z., Zhou, X., and Yang, T. Hetero-convlstm: A deep learning approach to traffic accident prediction on heterogeneous spatio-temporal data. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 984–992, 2018.

Appendix

A. Statistics Showing the Death Rates Due to Accidents in the US and Other Countries

According to the latest available data from the World Health Organization (WHO) (World Health Organization, 2021) and the most recent fatality statistics from the Insurance Institute for Highway Safety (IIHS) (Insurance Institute for Highway Safety (IIHS), 2024), despite concerted policy efforts and technological advancements, traffic fatality rates have not declined uniformly across all regions. Figure 1 compares annual road deaths in various developed and first-world countries from 2000 to 2022. Although many show a downward trend, attributed to improvements in vehicle safety standards, stricter enforcement, and public awareness campaigns, the United States has shown a significantly higher and more erratic pattern.

Even when we examine road fatality rates per 100,000 population for these same countries (Figure 2), the United States remains notably higher. Furthermore, the downward trajectories observed for most countries over this period do not apply to the United States, whose rates continue to hover at elevated levels. In Figure 2, we also compare the U.S. trend with populous nations such as India and China, revealing that U.S. fatality rates are not only comparable to these high-population countries but also follow a stubbornly flat or upward pattern, in contrast to the slow yet discernible downward trends in India and China. When we trried to compare rapidly transformed economies like Qatar and UAE, a steep decline in recent years is evident; although these nations once exhibited rates much higher than those of the U.S., they have now fallen below U.S. levels.

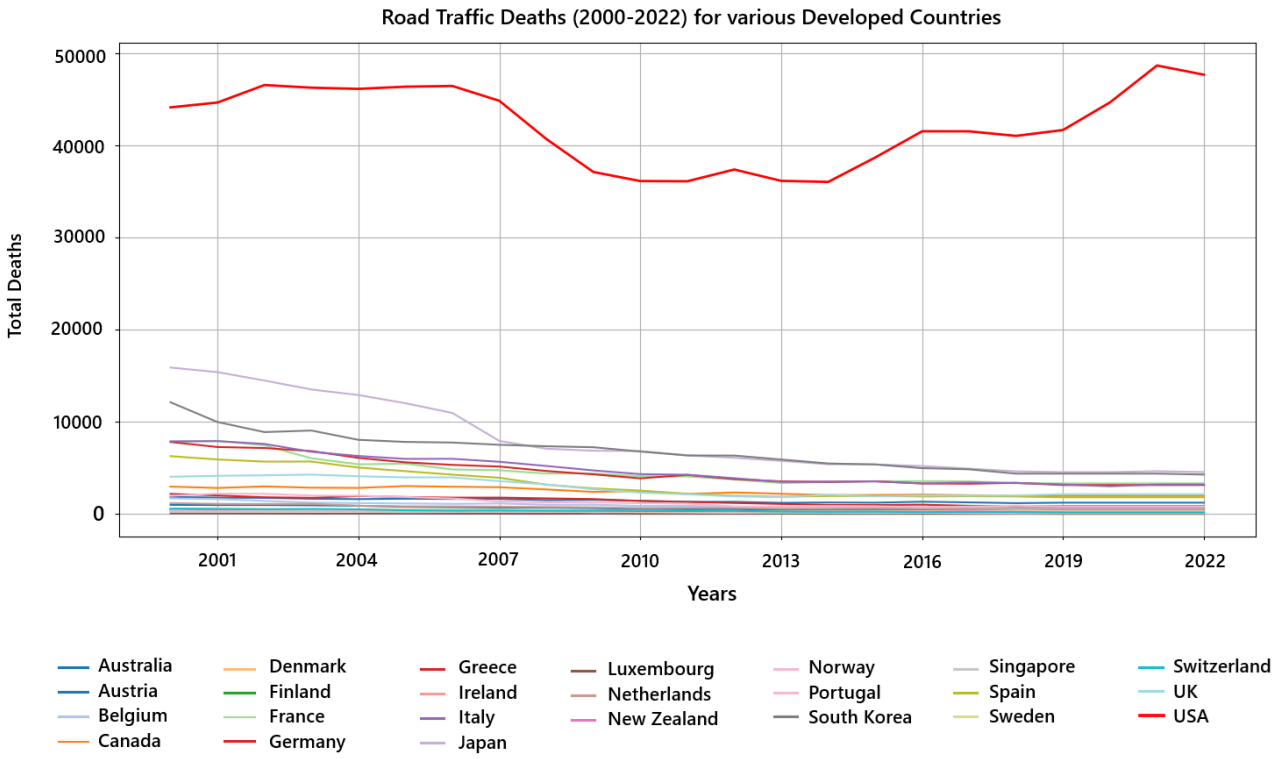

Figure 1. Deaths due to road accidents in various developed countries (2000–2022).

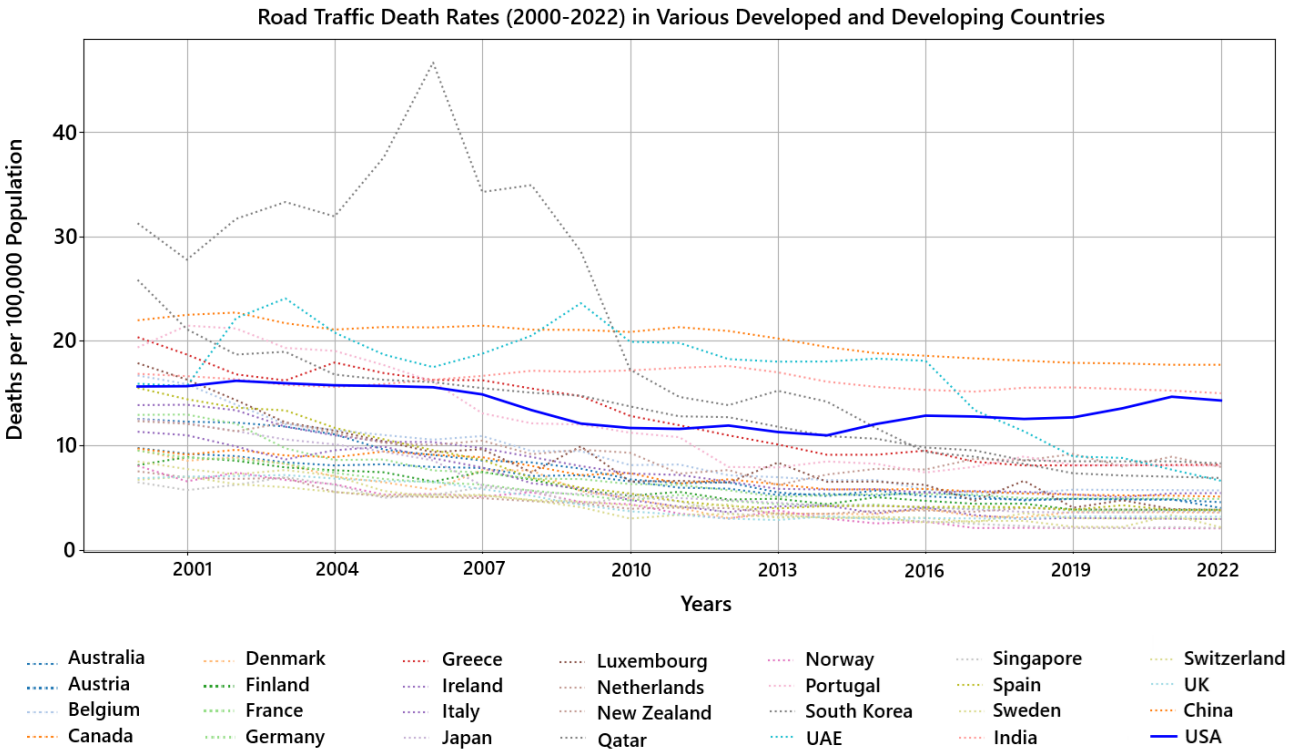

Figure 2. Road accident fatality rates (per 100,000 population) in various countries (2000–2022).

B. Details of Dataset Descriptions

B.1. Traffic Events Data

To build a robust dataset for accident risk prediction, we leverage a comprehensive car accident dataset by Moosavi et al. (Moosavi et al., 2019), covering 49 states across the United States. This dataset, collected from February 2016 to March 2023, comprises approximately 7.7 million accident records. The data is aggregated from multiple sources, including Bing, MapQuest, real-time traffic APIs that collect information from transportation departments, law enforcement agencies, traffic cameras, and in-road traffic sensors.

The dataset provides a rich set of attributes for each recorded accident, such as precise time and location of occurrence (including city, state, and ZIP code), severity levels, duration, and the length of resulting traffic impact. These features enable a granular analysis of traffic incidents, offering insights into possible patterns and contributing factors. The accident data is continuously updated in real time through multiple Traffic APIs, providing high temporal resolution and nationwide coverage of the contiguous United States. In addition to core variables like time, location, and severity, each record also includes relevant "Point of Interest (POI)" annotation tags, sourced from OpenStreetMap (OSM) (OpenStreetMap contributors, 2017). Table 4 presents an overview of the different POI categories found in the dataset, along with their corresponding descriptions.

Table 4. Definition of Point-Of-Interest (POI) annotation tags based on OpenStreetMap (OSM).

Type	**Description**
Amenity	Denotes specific locations such as restaurants, libraries, colleges, bars, etc.
Bump	Represents speed bumps or humps designed to reduce vehicle speed.
Crossing	Indicates designated pedestrian or cyclist crossings across roads.
Give-way	Road sign that dictates priority at intersections.
Junction	Represents highway ramps, exits, or entry points.
No-exit	Marks a point where travel cannot continue further along a designated path.
Railway	Identifies locations where railway tracks are present.
Roundabout	Indicates a circular road junction facilitating smooth traffic flow.
Station	Denotes public transportation hubs such as bus stops or metro stations.
Stop	Signifies stop signs at intersections or along roads.
Traffic Calming	Encompasses road features designed to slow down vehicle speed.
Traffic Signal	Represents traffic lights at intersections or pedestrian crossings.
Turning Loop	Designates widened sections of a highway featuring a non-traversable island for turning.

B.2. Demographic Attributes Data

To account for socioeconomic and population-based factors, we incorporate a comprehensive demographic dataset sourced from an online geographic data resource providing detailed postal and demographic information across the U.S. (United States ZIP Codes). It aggregates authoritative data from reputable sources such as the U.S. Census Bureau (U.S. Census Bureau) and the United States Postal Service (USPS) (United States Postal Service), ensuring the accuracy and reliability of the information presented. The platform offers extensive datasets, including ZIP code boundaries, population statistics, income levels, housing data, and geographic coordinates, making it a valuable tool for researchers, policymakers, and businesses. The dataset encompasses approximately 45,000 U.S. ZIP Codes and contains 150 demographic variables, offering a multi-faceted perspective on local socio-economic conditions. The data collection process involved API-based retrieval methods, ensuring a wide coverage of geographical regions while maintaining data consistency and accuracy.

Each ZIP code entry in the dataset includes information across four primary categories: (1) Population Characteristics, covering attributes such as age distribution, gender ratios, and racial composition; (2) Housing Information, detailing aspects such as homeownership rates, rental statistics, and average household sizes; (3) Employment and Income Statistics, providing insights into median household income, employment rates, and occupational distribution; and (4) Education Levels, offering data on educational attainment and literacy rates within each region.

By leveraging this rich demographic information, our study aims to analyze the potential influence of socio-economic factors on traffic patterns, accident frequencies, and severity levels. The inclusion of such granular data enhances our model's ability to capture regional disparities and provide more accurate predictive insights.

B.3. Weather Conditions Data

Since meteorological factors significantly influence traffic flow and accident likelihood, incorporating weather data is crucial for developing robust accident prediction models. Weather conditions such as temperature, precipitation, wind speed, and visibility can directly impact driving behavior, road surface conditions, and vehicle performance. Adverse weather, including rain, snow, fog, and extreme temperatures, can increase the risk of accidents by reducing visibility, affecting braking distances, and leading to hazardous road conditions. Moreover, sudden weather changes can disrupt normal traffic patterns, leading to congestion and increased accident probabilities. Hence, integrating meteorological data allows for more accurate and context-aware accident prediction models.

For our study, we obtained hourly weather data spanning from 2016 to 2023 for selected stations from the Iowa State University Iowa Environmental Mesonet (IEM) - ASOS Network ASOS-AWOS-METAR Data Download platform (Iowa State University). The IEM archives automated airport weather observations from various global locations, primarily collected from Automated Surface Observation System (ASOS) and Automated Weather Observation System (AWOS) stations. These observations provide detailed insights into prevailing atmospheric conditions at a high temporal resolution. The processed weather dataset was structured as a time series spanning the study period, with each row representing an hourly observation for the respective stations. This structured representation enables downstream modeling tasks to analyze temporal correlations between weather patterns and accident occurrences effectively.

By leveraging this comprehensive weather dataset, our accident prediction framework integrates meteorological variables alongside traffic and demographic data to capture the complex interplay between environmental conditions and accident risks. The inclusion of weather data enriches model accuracy by accounting for seasonal variations, adverse weather events, and localized microclimates that might otherwise be overlooked in traditional predictive models. Table 5 contains the list of weather features we collected.

Table 5. Description of Weather Features

Feature	Description
tmpf	Air Temperature in Fahrenheit, typically measured at 2 m above the ground.
dwpf	Dew Point Temperature in Fahrenheit, typically measured at 2 m above the ground.
relh	Relative Humidity expressed as a percentage.
drct	Wind Direction in degrees, measured from true north.
sknt	Wind Speed in knots.
p01i	One-hour precipitation amount in inches, recorded from the observation time to the previous hourly precipitation reset. May include melted frozen precipitation.
alti	Pressure altimeter measurement in inches.
mslp	Sea Level Pressure measured in millibars.
vsby	Visibility in miles.
skyc1	Sky Level 1 Coverage.

B.4. Map Image Representation

To capture the spatial context and road-network features associated with accident locations, we utilize a systematic approach based on hexagonal spatial segmentation and map image retrieval. Each accident location is mapped to a unique hexagonal region using Uber's Hexagonal Hierarchical Spatial Indexing (H3) (Uber Technologies, 2023) with a resolution of $R = 7$. This resolution results in hexagons with an approximate edge length of 2,604 meters and a total area of about 5.16 km^2, effectively covering the vicinity of the accident site.

Once the hexagonal zoning is established, the center coordinates of each hexagonal region are used to retrieve corresponding

map tiles from OpenStreetMap (OSM) (OpenStreetMap contributors, 2017). Specifically, we extract square map tiles at a zoom level of 14, which results in images of size 256×256 pixels. At this zoom level, each pixel represents approximately 9.547 meters, making the total coverage of each tile approximately 2.44 km per side, corresponding to an area of 5.95 km^2. The spatial coverage of these map tiles ensures that the entire hexagonal zone is sufficiently represented, allowing the model to incorporate comprehensive geographic features.

Figure 3 presents some examples of map tiles retrieved from OSM at zoom level 14, which are used to approximate H3 zones (R = 7) in regions of Columbus, Ohio. These map images encapsulate critical environmental elements, including roads, intersections, buildings, and other infrastructure, offering valuable insights into potential accident-prone areas. By leveraging these geospatial insights with rich textural data (e.g., road names), our model can identify spatial patterns, such as the density of road networks, the complexity of intersections, and the presence of high-traffic zones, which contribute to the occurrence of traffic incidents. This multi-faceted spatial encoding provides an essential context for enhancing the accuracy of our accident prediction framework.

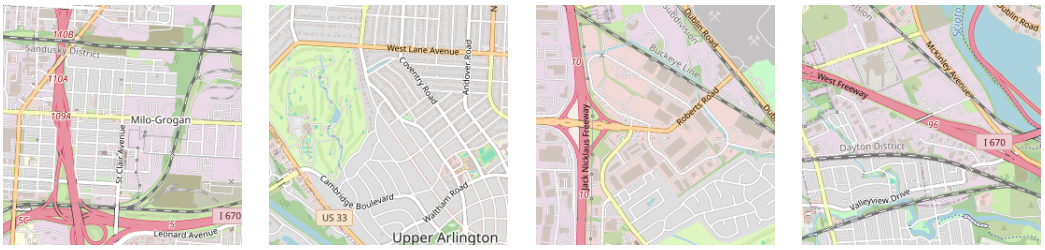

Figure 3. Examples of map tiles obtained from OSM (zoom level = 14), representing approximate H3 zones (R = 7) in Columbus, Ohio

C. Dataset Characterization and Statistical Analysis

To address any concerns regarding the perceived simplicity of the dataset, we provide a comprehensive statistical and structural analysis. The following figures highlight the underlying heterogeneity, feature complexity, and imbalanced nature of the accident prediction problem—underscoring that the dataset reflects real-world challenges rather than a synthetic or overly simplified scenario.

C.1. Feature Correlation and Redundancy

Figure 4 presents a feature correlation heatmap showing the pairwise Pearson correlation coefficients among all structured features. The presence of low to moderate correlations across most features indicates that the dataset is neither redundant nor trivially separable. Notably, some weather attributes (e.g., temperature, humidity, wind speed) show weak or no linear relationships with accident occurrence, emphasizing the need for nonlinear modeling strategies like GNNs and Transformers. Temporal and demographic features also remain weakly correlated, supporting the multimodal learning requirement.

C.2. Class Imbalance Across Cities

As illustrated in Figure 5, the dataset is highly imbalanced, with a significant dominance of "no accident" records across all cities. This mirrors real-world scenarios where accidents are rare events, despite dense traffic volumes. The imbalance varies across cities, requiring the model to generalize under skewed label distributions—a challenge not typically encountered in toy or synthetic datasets.

C.3. Feature Complexity and Dimensionality

To assess the dimensional structure of the data, we apply Principal Component Analysis (PCA). Figure 6 shows that over 20 principal components are needed to capture 95% of the variance, indicating that the dataset is high-dimensional and not compressible into a simple subspace. This further confirms the necessity of deep learning-based feature extraction as opposed to shallow models or linear classifiers.

C.4. Feature Importance Distribution

Figure 7 ranks features by their importance scores using an XGBoost classifier. A wide range of features—including weather attributes (e.g., temperature, humidity), temporal segments (e.g., date, rush hour), and demographic variables—contribute meaningfully to the accident prediction task. This confirms the dataset's multimodal nature and the nontrivial interaction between features, invalidating any assumptions of simplicity.

C.5. Nonlinear Separability in Latent Space

Figures 8 and 9 show 2D projections of the high-dimensional dataset using t-SNE and PCA respectively. Both visualizations reveal that accident vs. non-accident data points are not linearly separable and form dense, overlapping clusters. The sparse presence of accident cases amidst abundant "no accident" points highlights the difficulty of learning decision boundaries and underscores the need for advanced modeling techniques capable of capturing spatio-temporal dependencies and multimodal feature interactions.

Thus, the dataset presents real-world challenges such as high-dimensionality, severe class imbalance, weak feature correlations, and complex nonlinear boundaries. These characteristics necessitate the use of sophisticated architectures like BTS and invalidate concerns regarding dataset simplicity. This appendix offers transparent evidence that accident prediction in our context is a challenging and realistic task.

D. Details of Numerical Data Preprocessing and Feature Engineering

D.1. Merging the Data Sources

We start by loading the Traffic Event Data, which provides detailed accident records including ZIP codes and geographic coordinates (latitude, longitude). Next, we load the Demographic Attributes Data, keyed by ZIP codes.

Using the shared ZIP code field, we perform a join operation to enrich each accident record with socio-economic attributes (e.g., population density, income levels, and additional demographic features). This integration is crucial for capturing the broader context in which accidents occur, ensuring that each record in the traffic event dataset is accompanied by pertinent demographic indicators. The merged dataset thus combines spatio-temporal accident details with socio-economic factors, serving as a foundation for subsequent filtering and feature engineering steps.

D.2. City Selection Based on Missing Data Ratio

Upon investigating the dataset, we found that missing values are scattered across various records and feature columns, rather than being concentrated in specific rows or attributes. This pattern suggests that there are no dominant bulk-missing features or entirely incomplete records, making it suitable to measure missingness at the city level and apply a filtering criterion accordingly.

A common strategy in data science for handling significant missingness is to measure the fraction of all possible data entries that are missing and then filter out entities (in this case, cities) with excessive missingness. Let R_c be the total number of records for city c, and let D be the total number of columns after merging all relevant attributes. Define M_c as the sum of missing feature values across all records in city c. Formally, the missingness ratio δ_c is:

$$\delta_c = \frac{M_c}{R_c \times D}. \tag{1}$$

We then rank the cities in ascending order of δ_c and retain the top 15 cities with the lowest missing ratios. By eliminating cities that surpass a certain threshold of missingness, we reduce the need for heavy imputation, which can introduce additional noise and uncertainty. This filtering step thus preserves higher data fidelity and provides a more reliable foundation for subsequent analyses. In the sebsequent subsections we have shown the list of selected cities used for training the model.

D.3. Spatial Partitioning with H3

Although demographic attributes are keyed to ZIP codes, these codes vary greatly in both shape and size, making them suboptimal for fine-grained spatial analysis. To address this limitation, we employ Uber's H3 library at a resolution of $R = 7$ to partition the study area into a grid of hexagonal cells (referred to as Area IDs). At this resolution, each cell covers

approximately 5.16 km^2 and has edges of about 2.6 km, yielding a uniform spatial framework that applies consistently across all cities under consideration.

This hexagonal partitioning offers several distinct benefits:

- **Uniformity:** H3 cells are designed to be nearly identical in shape and area, facilitating more consistent distance-based comparisons than ZIP codes, which were originally developed for mail delivery rather than spatial analytics.
- **Neighbor Identification:** By encoding each accident's latitude and longitude into a hexagonal cell, we can readily establish adjacency relationships among cells, which is crucial for subsequent graph-based modeling.
- **Scalability:** The hierarchical nature of H3 allows for flexible adjustments in spatial granularity, supporting diverse analytical resolutions without altering the fundamental grid structure.

In this study, we favor H3 over ZIP-code boundaries to avoid the irregularities and non-uniform areas that arise from using postal regions. The standardized hexagonal cells not only simplify distance calculations but also offer a consistent basis for comparing spatial relationships across different cities and states.

D.4. Filtering by H3 Regions

Even within a single city, certain H3 cells can contain very few valid (non-missing) accident records. Excessive reliance on imputation for these sparse cells may introduce noise and bias, ultimately obscuring meaningful patterns. To mitigate this risk, we apply two additional filters:

(1) Minimum Record Threshold. We first enforce a lower bound on the number of records per H3 cell. Specifically, any cell with

$$\text{TotalRecords} < 100$$

is excluded from further analysis. This criterion ensures that each cell considered has sufficient coverage for a robust evaluation of local traffic and environmental attributes.

(2) Non-Missing Ratio Threshold. Next, we measure the proportion of records in each H3 cell that are fully complete (i.e., contain no missing values in the core fields). Formally, we define:

$$\text{NonMissingRatio}_{\text{H3}} = \frac{\text{NonMissingRecords}}{\text{TotalRecords}}. \tag{2}$$

To maintain high data fidelity, we discard any cell with

$$\text{NonMissingRatio}_{\text{H3}} < 0.95.$$

Requiring at least 95% of records to be fully observed limits the extent of imputation, thereby preserving the integrity of the underlying signals.

Outcome. By applying these two filters, we arrive at a final subset of H3 cells (drawn from the 15 selected cities) that is both dense (in terms of record count) and predominantly complete (in terms of observed values). This ensures that subsequent analyses focus on geographically localized regions where the data quality is sufficiently high to support reliable pattern discovery and prediction. Table 6 shows the final number of nodes in each selected cities.

D.5. Time-Series Construction & Label Definition

In this step, we transform the accident data into discrete time slices spanning June 1, 2016 to March 31, 2023, grouped in 3-hour intervals. The choice of a 3-hour window is guided by two principal considerations:

- **Balancing sparsity and resolution:** Using a shorter interval (e.g., 1 hour) risks producing excessively sparse data in many H3 cells, as accidents do not occur frequently enough in every region. Conversely, a coarser interval (e.g., 6 hours) can obscure important intra-day variations, such as rush-hour patterns.

Table 6. Final Number of Nodes in Each Selected City

City	State	No. of Nodes
Atlanta	GA	88
Austin	TX	139
Baton Rouge	LA	64
Charlotte	NC	183
Dallas	TX	140
Houston	TX	246
Los Angeles	CA	85
Miami	FL	152
Minneapolis	MN	84
Nashville	TN	93
Orlando	FL	150
Phoenix	AZ	106
Raleigh	NC	97
Sacramento	CA	74
San Diego	CA	70
Total	-	**1,771**

- **Worst-case severity labeling:** If multiple accidents occur within the same 3-hour window, we assign the highest observed severity to that window. This "worst-case" labeling strategy ensures that models can account for the most critical safety scenarios, thereby providing a conservative estimate for risk-focused applications.

Once grouped by these 3-hour windows, we aggregate all accidents occurring in a given H3 cell within the same time slice. The aggregated record is then enriched with additional features (e.g., weather variables), creating a comprehensive event representation for further modeling.

D.6. Temporal Representation

Temporal attributes are pivotal for modeling accident risk, weather variations, and traffic congestion. Once the accident records have been aggregated into 3-hour intervals, we further enrich each time slice with cyclical and categorical indicators that capture intra-day, weekly, and seasonal patterns. Table 7 summarizes the principal temporal features extracted from the dataset.

By incorporating these temporal markers, our predictive models gain finer-grained insights into traffic-related phenomena, enabling more accurate modeling of short-term fluctuations, diurnal patterns, and seasonal variations that influence accident probabilities.

D.7. Weather Data Integration

In order to incorporate meteorological context into our accident prediction framework, we align each 3-hour accident window with the relevant weather observations collected over the same time span. Specifically, we aggregate numerical weather variables (e.g., temperature, humidity) by computing their arithmetic mean within each 3-hour period. For the categorical variable `skyc1`, indicating cloud cover, we convert its discrete labels to an integer encoding, compute the average of these encoded values across the window, and then round back to the nearest integer to determine a single representative category. The weather dataset exhibited a minimal missingness rate of only 0.01%, making it impractical to discard records due to sparsity. Instead, we applied bidirectional interpolation (Lam, 1983) to impute missing values before aggregating weather attributes over each 3-hour window.

This procedure yields a compact set of weather attributes for each (H3 cell, 3-hour time slice) tuple, enabling downstream

Table 7. Summary of temporal features derived from timestamps spanning 2016–2023. Numeric values used for categorical encoding are indicated in parentheses.

Feature	Description
Season	Categorized based on the month: **Winter (0)** [Dec, Jan, Feb]; **Spring (1)** [Mar, Apr, May]; **Summer (2)** [Jun, Jul, Aug]; **Fall (3)** [Sep, Oct, Nov].
Month	Month number **(1–12)**.
Date	Day of the month **(1–31)**.
Day	Day of the week, encoded as **Monday (0)** to **Sunday (6)**.
Weekday	Binary label: **Weekday (1)** if Monday–Friday; **Weekend (0)** if Saturday or Sunday.
Holiday	Binary indicator: **Holiday (1)** if the date is a recognized U.S. public holiday; otherwise **Non-Holiday (0)**.
Part-of-Day	**Morning (0)** [6:00–11:59]; **Afternoon (1)** [12:00–17:59]; **Evening (2)** [18:00–23:59]; **Night (3)** [00:00–5:59].
Rush-Hour	**Morning Rush (0)** [6:00–8:59]; **Evening Rush (1)** [15:00–17:59]; **Non-Rush Hours (2)** [all other times].

models to account for variations in atmospheric conditions. By jointly modeling traffic, demographic, and weather information, our framework benefits from a more holistic view of the factors influencing accident risk.

D.8. Handling Missing Data

Despite comprehensive data collection efforts, the large-scale nature of our dataset inevitably resulted in a non-trivial percentage of missing entries.

Overview of Imputation Techniques. Rather than discarding incomplete rows, we opted to impute missing values using several state-of-the-art algorithms. This choice balances data retention against the risk of inaccurate imputation, ensuring we utilize as much of the available information as possible.

The following methods were evaluated, with hyperparameters selected via **GridSearchCV** over an appropriate range:

- **FAISS kNN (k=5) (Johnson et al., 2019)**: Scalable similarity search, well-suited for large datasets. It addresses the slow query issue of classical kNN by using efficient nearest-neighbor lookups. We chose $k = 5$ after tuning on $k \in \{3, 5, 7, 9, 15\}$, as it yielded the lowest error.
- **SAITS (Du et al., 2023)**: A self-attention-based time-series imputation approach that adaptively models temporal dependencies. We fine-tuned the transformer layers and attention heads in the range $\{1, 2, 4, 8\}$ and selected optimal values based on validation performance.
- **XGBoost (max_depth=6, learning_rate=0.05) (Chen & Guestrin, 2016)**: A tree-ensemble method that leverages non-linear relationships to estimate missing feature values. Hyperparameters such as `max_depth` and `learning_rate` were selected using GridSearch over $\{3, 6, 9\}$ and $\{0.01, 0.05, 0.1\}$, respectively.
- **Iterative Imputation (max_iter=10, estimator=XGBoost) (Guo et al., 2024)**: A model-based iterative scheme that treats each feature with missing values as a regression target, iteratively refining estimates using other features as predictors. We experimented with linear regression, random forests, and XGBoost as the estimator, finding XGBoost to perform best.

Comparison and Selection. We conducted an internal evaluation of these techniques, assessing imputation performance by comparing reconstructed values against known observations in a holdout sample. Table 8 summarizes the imputation accuracy and post-imputation data consistency for each method. Based on these findings, we selected **FAISS kNN** with **k=5** as the best-performing method.

Table 8. Missing Values Imputation Results

Models Used	MSE	MAE	RMSE
Iterative Imputation	0.46214	0.36617	0.67981
XGBoost	0.01447	0.02451	0.12029
SAITS	0.00392	0.00244	0.091684
FAISS KNN	**0.00053**	**0.00026**	**0.02310**

Most missing values originate from demographic attributes for certain ZIP codes. Since demographic data tends to be more correlated with geographic proximity rather than temporal trends, using kNN for imputation makes the most sense in this context. By leveraging nearest-neighbor similarity, FAISS kNN effectively reconstructs missing demographic features using information from surrounding regions. While transformer-based techniques like SAITS generally excel in time-series imputation, they underperform in this case because demographic features exhibit stronger spatial dependencies rather than temporal variations.

D.9. Feature Selection

Although our merged dataset is richly detailed, certain attributes may be redundant or exhibit minimal correlation with accident outcomes, thereby inflating both computational costs and the risk of overfitting. To address this, we performed a systematic feature selection process using multiple techniques:

- Random Forest Feature Importance (Breiman, 2001): We trained a Random Forest model and extracted the Gini importance scores for each feature. This provided an interpretable ranking based on how much each feature contributed to reducing impurity in decision trees.
- XGBoost Feature Importance (Chen & Guestrin, 2016): We leveraged XGBoost's built-in feature ranking, which assigns importance scores based on split frequency, gain, and coverage across decision trees.
- Principal Component Analysis (PCA) (Maćkiewicz & Ratajczak, 1993): We applied PCA to examine variance contributions across features, helping us identify and remove attributes with minimal independent information content.

Each method produced a highly similar ranking, with only minor variations (1-2 positions for some features). The next challenge was determining an optimal subset of features to retain. We employed the following experiment to select the appropriate number of features:

Step 1: Performance vs. Feature Count Analysis. To systematically determine an appropriate feature subset size, we trained a baseline accident prediction model using subsets of increasing feature counts (from the most important feature up to all 115 features). We evaluated model performance using Accuracy, F1-score, and AUC.

Step 2: Identification of the Performance Plateau. We observed that model performance increased significantly when adding the first few features, then gradually plateaued beyond a certain threshold. Specifically, we noted that beyond 21 features, the gain in predictive performance was marginal (less than a 0.5% increase in AUC) while computational complexity continued to rise.

Step 3: Feature Stability Across Methods. We further validated our choice by analyzing the stability of selected features across Random Forest, XGBoost, and PCA. The top 21 features were consistently ranked highly across all three methods, reinforcing their importance, and beyond 21 features, we observed increased variance in validation performance, suggesting potential overfitting. Hence, by limiting the feature count to 21, we achieved a balance between model generalization and efficiency. Table 9 lists all the selected 21 features.

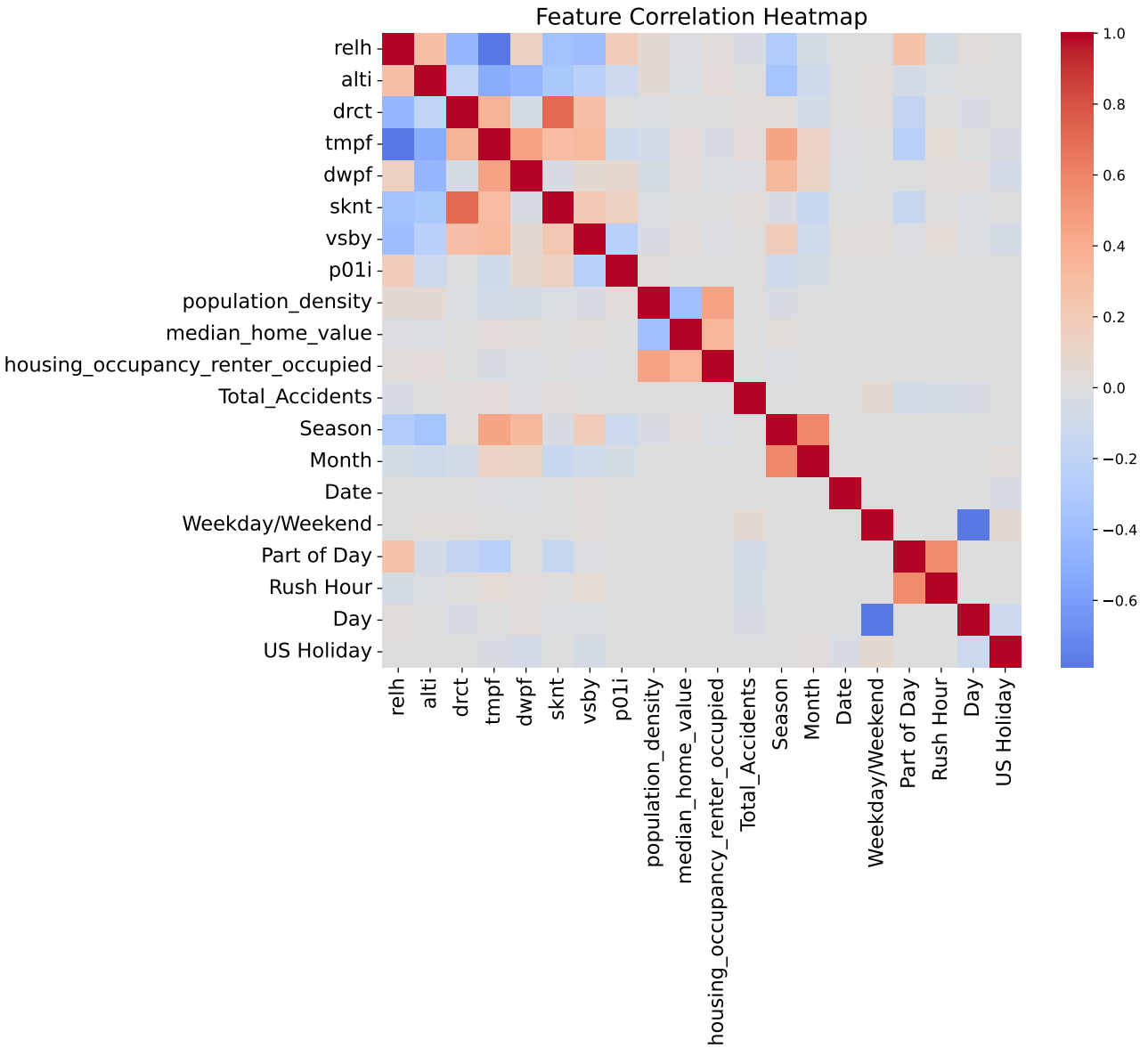

Figure 4. Pairwise correlation matrix of numerical and categorical features.

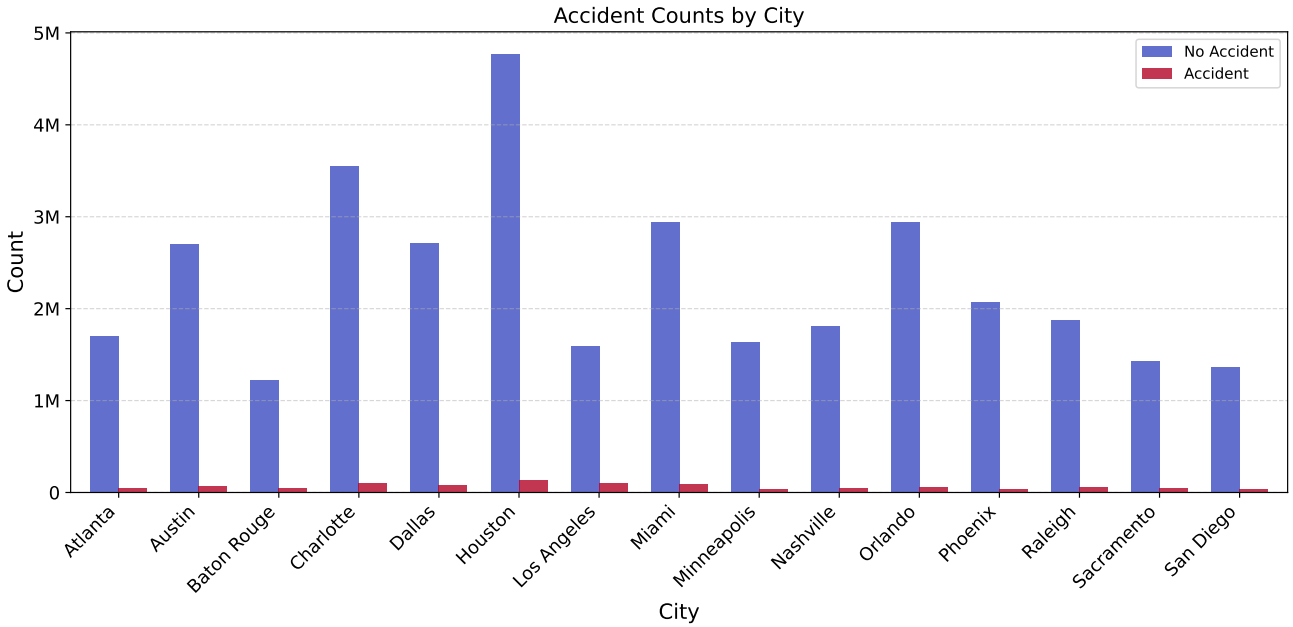

Figure 5. Distribution of accident vs. no-accident instances across different cities.

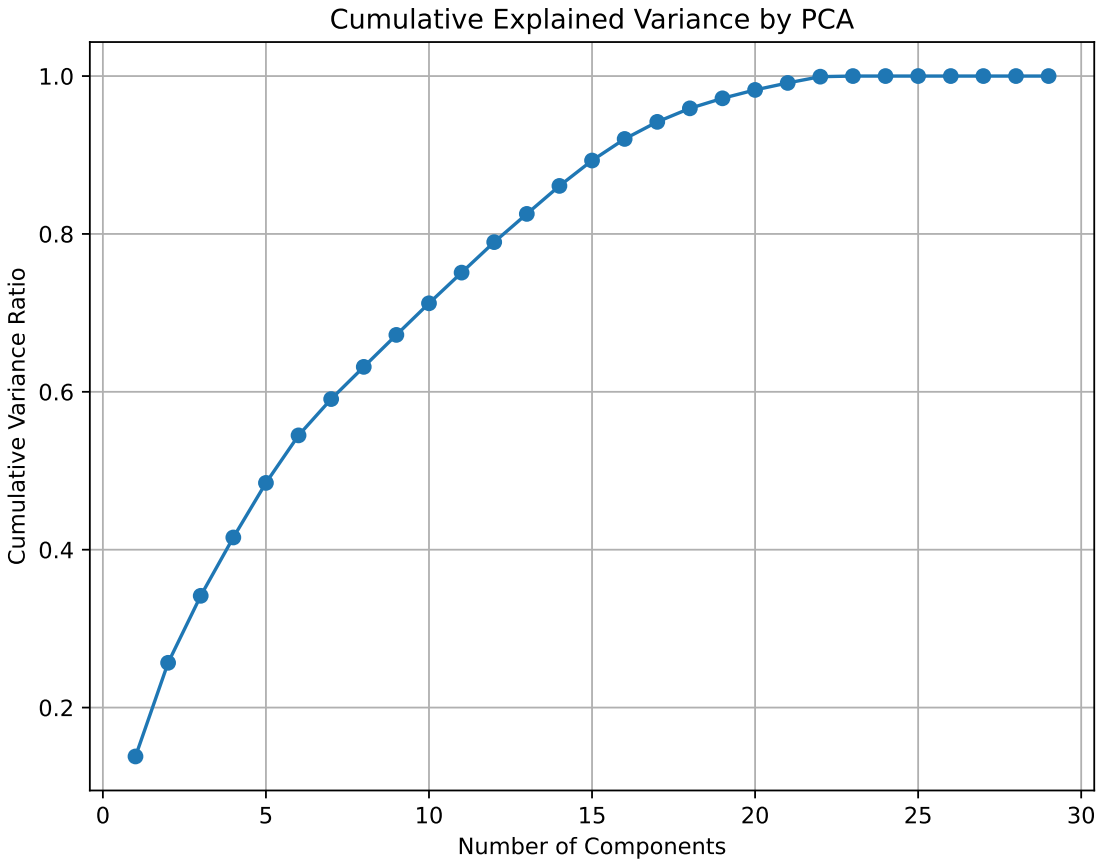

Figure 6. Cumulative explained variance from PCA across top 30 components.

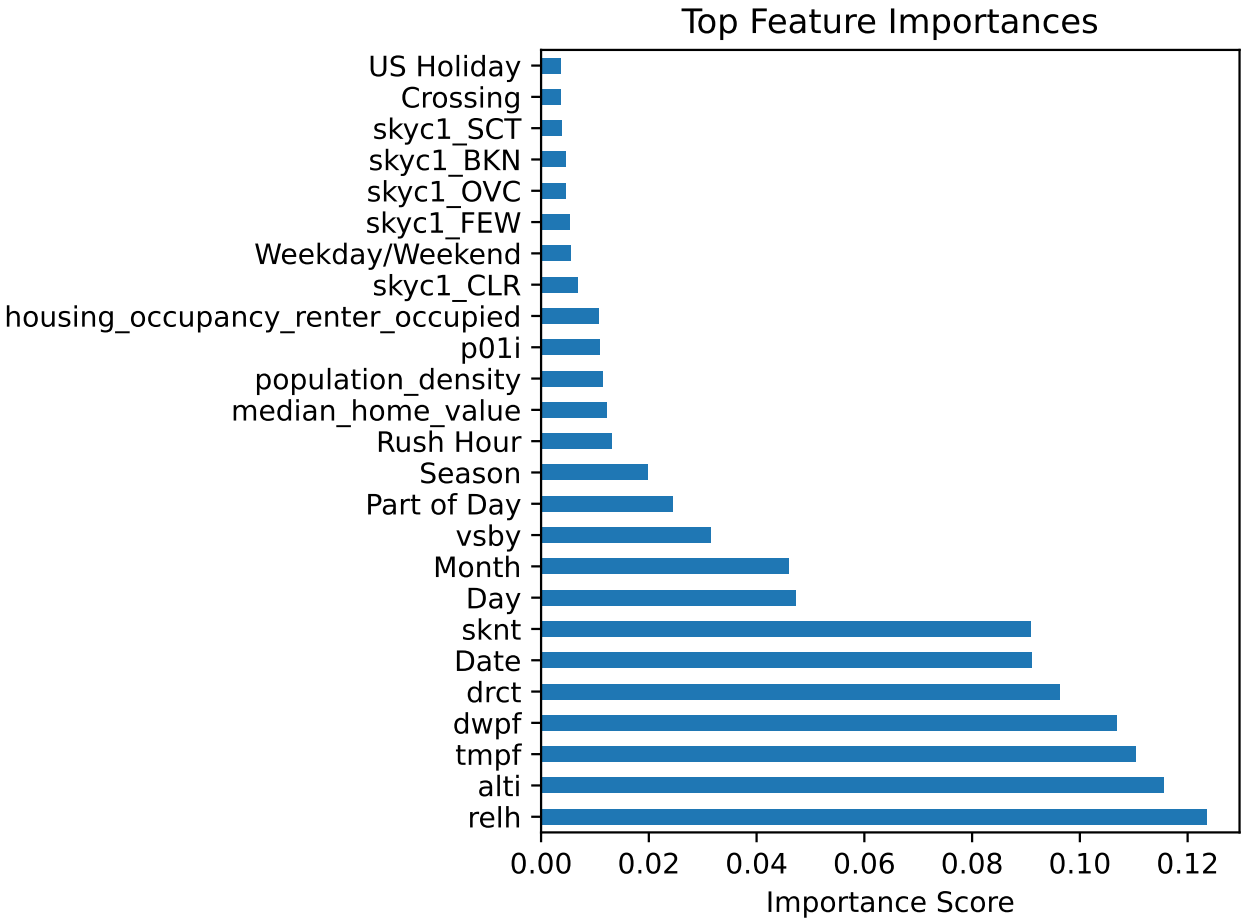

Figure 7. XGBoost-based feature importance ranking.

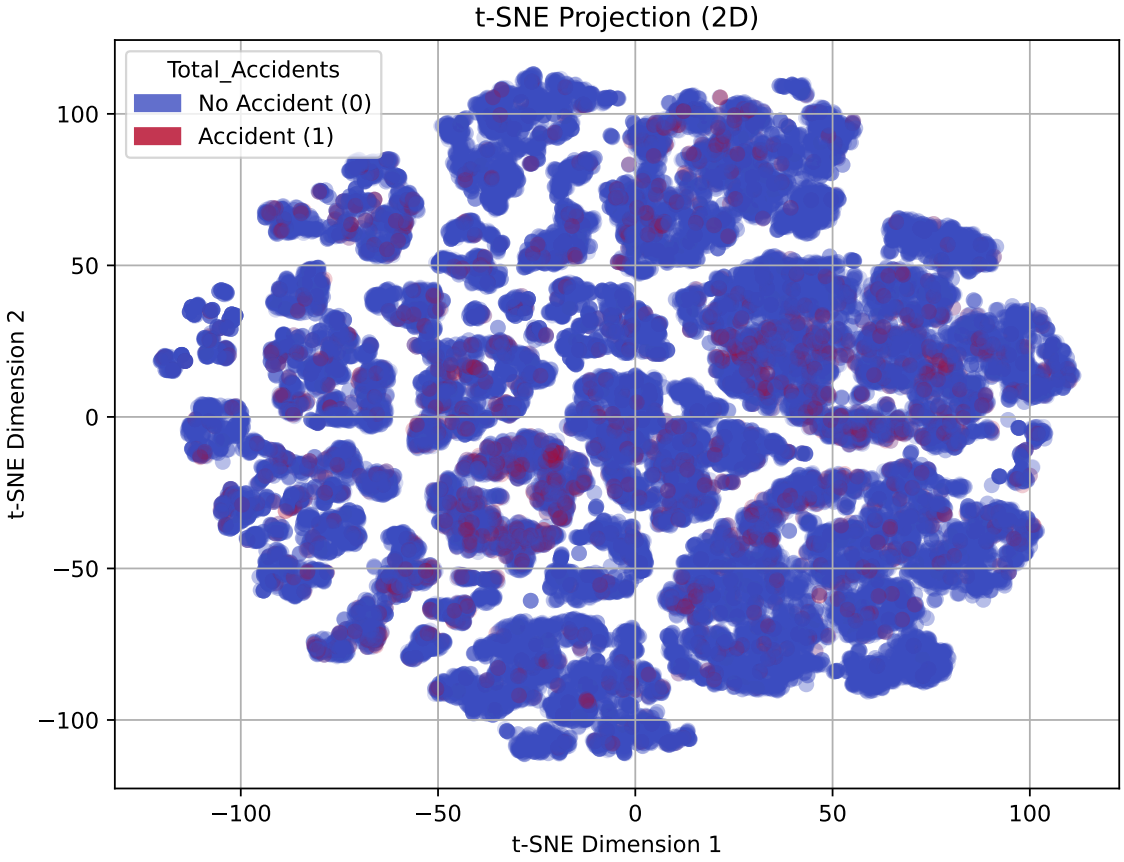

Figure 8. t-SNE projection of the dataset in 2D space, colored by accident labels.

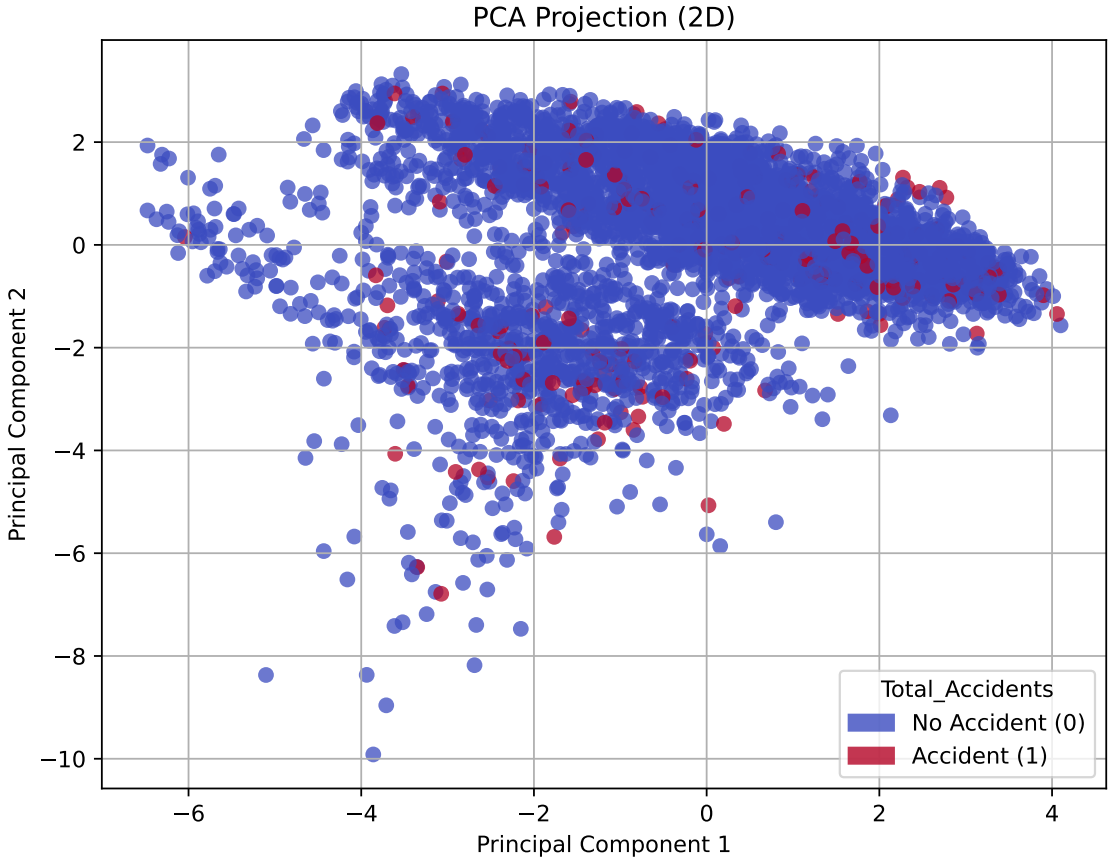

Figure 9. PCA projection of the dataset in 2D space, colored by accident labels.

Table 9. Selected Features for Traffic Accident Prediction

Feature Name	**Description**
Date	Captures overall time trends and seasonality.
Day	Identifies whether the accident occurred on a weekday or weekend.
Month	Helps model seasonal effects on traffic patterns.
relh	Relative Humidity expressed as a percentage.
alti	Pressure altimeter measurement in inches.
drct	Wind Direction in degrees, measured from true north.
tmpf	Air Temperature in Fahrenheit, typically measured at 2 meters above the ground.
dwpf	Dew Point Temperature in Fahrenheit, typically measured at 2 meters above the ground.
sknt	Wind Speed in knots.
Rush Hour	Indicates whether the accident occurred during peak traffic hours.
Season	Represents broader seasonal trends impacting road safety.
vsby	Visibility in miles.
skyc1	Sky Level 1 Coverage.
Traffic_Signal	Presence of a traffic signal, affecting vehicle stopping and interactions.
Part of Day	Distinguishes between morning, afternoon, evening, and night-time driving patterns.
p01i	One-hour precipitation amount in inches, recorded from the observation time to the previous hourly precipitation reset. This measurement may include melted frozen precipitation depending on the sensor.
Crossing	Indicates the presence of pedestrian crossings, increasing accident risks.
US Holiday	Identifies national holidays, which influence traffic patterns and congestion.
population_density	Measures urbanization levels, which impact accident risk and traffic volume.
median_home_value	Socioeconomic proxy for regional traffic infrastructure and urban development.
housing_occupancy _renter_occupied	Captures transient populations, influencing local driving behavior.