Geospatial Risk Mapping of Urban Traffic Accidents Using Earth Observation and Machine Learning: A Case Study of Kolkata

Moumita Mishra
Indian Institute of Technology Kharagpur
Kharagpur, India
moumitamishra15@gmail.com

Diptesh Das
Vellore Institute of Technology
Vellore
ddiptesh15@gmail.com

Soumya K Ghosh
Indian Institute of Technology Kharagpur
Kharagpur, India
skg@cse.iitkgp.ac.in

Bhargab Maitra
Indian Institute of Technology Kharagpur
Kharagpur, India
bhargab@civil.iitkgp.ac.in

***Abstract*—Rapid urban growth, coupled with increased motorization, has significantly elevated road traffic accident risks, posing substantial challenges to public safety and urban management. Traditional approaches—primarily focused on descriptive accident mapping and historical data analysis—fall short of delivering the predictive insights required for proactive risk mitigation. To address these limitations, this research introduces a comprehensive predictive geospatial analytics framework that integrates machine learning, and Earth Observation technologies via Google Earth Engine (GEE). The urban landscape was systematically divided into uniform 50-meter grid cells, each enriched with spatially derived attributes such as historical accident data (2017–2023) and proximity to critical urban infrastructures including schools, hospitals, markets, and residential complexes. Based on these attributes, accident frequencies were categorized into defined risk levels: low, moderate, high, and very high—forming the foundation for predictive modeling. Advanced machine learning algorithms, specifically Random Forest and XGBoost Boosting regressors, were employed to forecast the spatial distribution of accident risk for 2023. The Random Forest model demonstrated superior predictive performance, achieving an R^2 value of 0.88, underscoring the effectiveness of this approach. By providing precise, predictive maps of accident-prone areas, the proposed methodology equips urban planners and policymakers with actionable intelligence for strategic safety interventions and efficient resource allocation—contributing significantly to improved urban traffic management practices.**

***Index Terms*—Urban Traffic Safety, Predictive Analytics, Google Earth Engine, Machine Learning, Random Forest, Accident Risk Mapping, GeoSpatial Analysis, Proactive Traffic Management**

I. Introduction

In crowded urban areas, the combination of rapid growth, high vehicle density, and intricate infrastructure has significantly raised road traffic accident risks. The World Health Organization (WHO, 2018) highlights road traffic injuries as a major global health issue, disproportionately affecting low- and middle-income countries. Urban areas face increased traffic accident risks due to high vehicle density, inadequate infrastructure, and vulnerable road users. Traditional methods like black spot analysis and KDE offer limited, descriptive insights and lack predictive power. While GIS has improved hotspot detection, it often omits contextual factors. Machine learning (ML) models—such as Random Forest and XGBoost Boosting—offer better prediction but face challenges with data quality and spatial resolution. Earth Observation (EO) tools like Google Earth Engine (GEE) address these gaps by providing scalable, high-resolution geospatial data. This study integrates GEE with a grid-based ML framework using 50-meter cells and features like proximity to key infrastructures and historical accident data (2017–2023) to predict urban accident risks. The modular approach improves interpretability, generalizability, and future risk forecasting. Despite limitations in real-time data integration, the model supports proactive safety planning . The major contribution of this paper are as follows:

Grid-Based Predictive Framework: Developed a 50meter grid-based accident risk prediction model for urban areas.

Integration of GEE and ML: Combined Google Earth Engine's geospatial capabilities with machine learning (Random Forest) for enhanced risk mapping.

Rich Spatial Feature Engineering: Incorporated proximity to critical infrastructures (e.g., hospitals, schools, markets) into spatial risk modeling.

Shift from Descriptive to Predictive: Transitioned from traditional hotspot analysis to proactive accident prediction using historical data (2017–2023).

Enhanced Interpretability and Scalability: Adopted a modular, interpretable, and computationally efficient model suitable for replication in other urban settings.

SDG Alignment and Overcame Data Limitations: Contributed toward Sustainable Development Goal 11.2 by supporting data-driven, safe, and sustainable urban transport planning and Addressed issues of inconsistent and low-resolution

urban accident data using Earth Observation and proxy indicators.

II. Related Work

Several studies have used different methods to predict the accident count and accident risk .In 2024 adewopo used deep learning ensemble model to predicts accident-prone zones using sensor data but lacks geospatial analysis[1].In this paper, the author used multimodal data to predict accidents but overlooks spatial relationships between traffic density and accidents[2]. Similarly Authors used Deep learning to detect real-time traffic anomalies[2],Authors used graph convolutional recurrent networks to detect of hazardous driving behaviors [3].In 2020 zhang identifies accidents in video footage, focusing on real-time detection over spatial distribution using Mask R-CNN techniques[4]. Authors introduced a machine learning that predicts traffic flow but ignores the link between traffic flow and accident-prone areas[5]. Authors used ARM for idendentifying blackspot analysis[6]. Authors analyzed the Road Network Deformation using SAR data[7] The author used HMM model to risk assessment[8][9] but they did not consider the spatial relationship of high-density traffic zones, accident counts, and points of interest.The authors identified blackspots using various parameters and applied k-means clustering to determine the most dangerous zones[10]. Traditional methods like KDE and black-spot analysis [11] offer descriptive insights into accident-prone zones but lack predictive power and often ignore key contextual factors such as land use, population density, and infrastructure .Google Earth Engine (GEE) has advanced traffic safety analytics by enabling large-scale geospatial analysis [12] but its predictive use in road safety remains limited due to challenges in data integration and high-resolution processing .Machine learning models like Random Forest and XGBoost have improved traffic accident predictionbut challenges in data quality and real-time integration persist. This study addresses these gaps by combining ML, Earth Observation via GEE, and spatial analytics for accurate, scalable urban accident risk prediction .

Our propose work integrates these advancements—spatial analytical techniques, Earth observation capabilities via GEE, and robust ML models—to construct an innovative framework aimed at accurately mapping and predicting urban accident risks. By synthesizing these distinct but complementary methodologies, the proposed approach addresses many of the limitations identified in existing literature, contributing significantly towards enhanced urban traffic safety management.

III. Proposed Methodology

A novel geospatial analytics is proposed for detecting high-risk accident zones that integrates machine learning, and Earth Observation technologies via Google Earth Engine (GEE). Currently, the approach is applied to the Kolkata region but can be extended to other locations, provided that the necessary data is available.

A. Data Preprocessing

The dataset contains accident data (fatal and nonfatal) of Kolkata from 2017 to 2020 and 2021-2023. In addition to accident data, we have also used Google imagery, shapefiles of points of interest (POIs), and road features.The detailed overview of the crash datasets has been given in Table-I.

TABLE I
Dataset description

Dataset Name	Features	Instances
2017-2020 Accident Dataset		
Fatal Case	145	1057
Fatal Person	22	1139
Non-fatal Case	146	3558
Non-fatal Person	18	2558
2021-2023 Accident Dataset		
Fatal	32	1000
Non-fatal	26	5145
Weather Data all		
2017-2020 weather data	8	2857
2021-2023 weather data	8	3878

B. Feature Extraction

Road features, road intersection points , and points of interest can significantly impact the identification of high-density traffic accident roads and road safety.To enrich this identification additionally, those feature were extracted using OSMnx's,GeoPandas libraries. First we have extracted the Kolkata road network shape file where roads are different types like primary,motorary and trunk road.Then we have extracted high-traffic road and high-density intersections. If the road has more than 4 intersections classified as a high-density intersection. Additionally, points of interest (POIs) such as residential zones, schools, and hospitals were considered to understand areas contributing to significant traffic.

C. Visualization

The work used the open source software QGIS to visualize Kolkata's traffic-prone zones, highlighting high-traffic roads, intersections, school locations,market, apartments and hospitals. The visualization is given below:

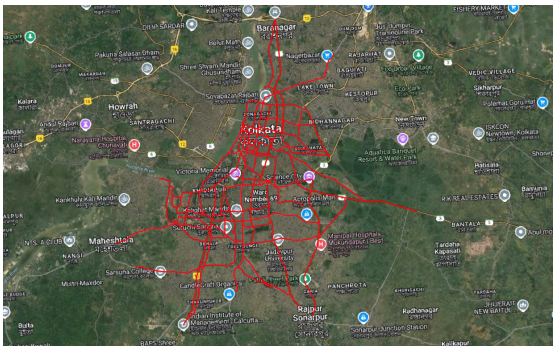

Fig. 1. High Density Roads

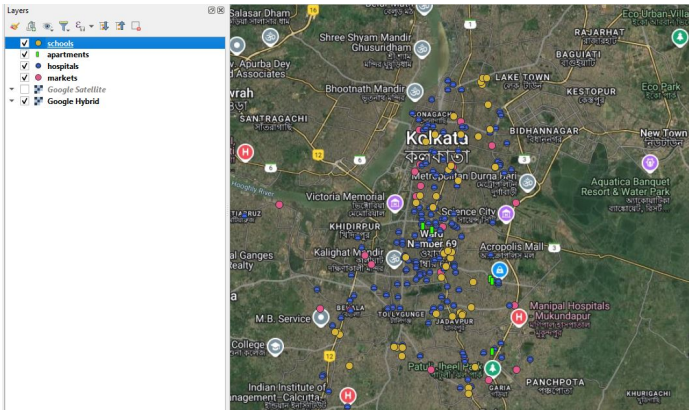

Fig. 2. POIs(School,Market,Hospitals and apartments) Locations

In Fig. 5, the yellow color represents schools, the blue color represents hospitals, the red color represents markets, and the green color represents apartments locations, respectively.

D. Grid Generation

The study region was overlaid with a uniform grid of 50m by 50m polygon cells created using GEE's ee.Image.pixelCoordinates() and ee.Image.focal-min() functions. This grid balances spatial resolution with computational efficiency, ensuring sufficient granularity to capture local variations in accident occurrence. Each cell is assigned a unique identifier (cell-id) for subsequent spatial joins and feature aggregation.

E. Accident Counting and POI Distance Calculation

A spatial join was performed between the accident FeatureCollection and the grid. The ee.Join.saveAll() and ee.Filter.bounds() methods were used to count the total number of accidents falling within each cell for each study year. For every grid cell centroid, the minimum Euclidean distance to the nearest school, hospital, market, and apartment cluster was computed using GEE's distance() reducer. These distances serve as proxy indicators for pedestrian activity and emergency response accessibility.The proposed framework integrates geospatial processing, feature engineering, risk classification, and machine learning to predict urban accident risk. A 50m grid is generated using Google Earth Engine (GEE), with each cell enriched by spatially joining accident data and distances to key points of interest (POIs) such as schools, hospitals, markets, and residential areas. These features are used to classify each cell into risk levels—low, moderate, high, and very high—based on accident counts. Thematic maps for different years are developed for visual analysis and validated against known hotspot regions. Using historical accident data, POI distances, and temporal trends,various ML models, Random Forest and XGBoost Boosting models are trained to forecast 2023 accident counts. The predicted values are then converted into risk categories and visualized, supporting proactive identification and management of emerging high-risk zones.

IV. Result and Discussion

Figures 3 and 4 show risk maps for previous year(2017-20222), In 2017-2020 high-risk zones were concentrated along central corridors and major intersections. In 2021-2022, these zones expanded northward with a 15% rise in very-high-risk cells. New hotspots also emerged near peripheral residential developments, indicating shifting risks due to urban growth and traffic redistribution. Figure 5 shows the 2023 risk map, highlighting both stable and shifting accident zones. High-risk areas persisted in central corridors, while eastern moderate-risk cells escalated. Southern cells saw reduced risk, likely from previous span traffic calming efforts. Overall, accident counts per cell rose by 8%, signaling a need for stronger safety measures in growing urban areas.

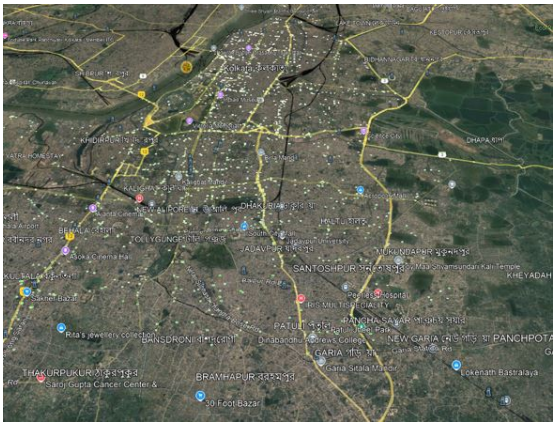

Fig. 3. Riskmap1

Fig. 4. Riskmap2

A. Predictive Model Performance and Forecasted Risk Map for 2023

The Random Forest regressor and complementary XGBoost Boosting model were evaluated on a held-out validation set drawn from the 2022 grid. Table 1 summarizes key performance metrics. XGBoost regressor achieved R^2 of 0.88 and NRMSE of 0.14. It depicts the predicted 2023 risk map generated by the XGBoost model. The spatial pattern closely aligns with the observed map correctly identifying 88% of

very-high–risk cells and high-risk cells. Predicted emergent hotspots in the northern and western fringes correspond to zones where rapid land-use change occurred, underscoring the model's ability to capture the influence of POI proximities and temporal accident trends. The CSV file containing per-cell predicted accident counts (Appendix A) reports a mean of 2.8 accidents per cell (SD = 1.9), closely aligning with the observed mean of 2.7 (SD = 1.8).

Model Name	MAE	RMSE	NRMSE	R2 Score
Linear Regressor	15.03	17.90	0.25	0.004
Lasso Regressor	14.31	18.90	0.25	0.004
Ridge Regressor	14.43	18.87	0.26	0.004
SVR	13.58	19.24	0.23	0.003
KNN Regression	14.08	18.95	3.56	0.31
Decision Tree	1.88	7.67	0.05	0.79
Random Forest Regressor	2.04	5.94	0.05	0.81
XGBoost Regressor	1.34	3.93	0.04	0.88

TABLE II
PREDICTION RESULT USING ML TECHNIQUES

Fig. 5. Actual Riskmap

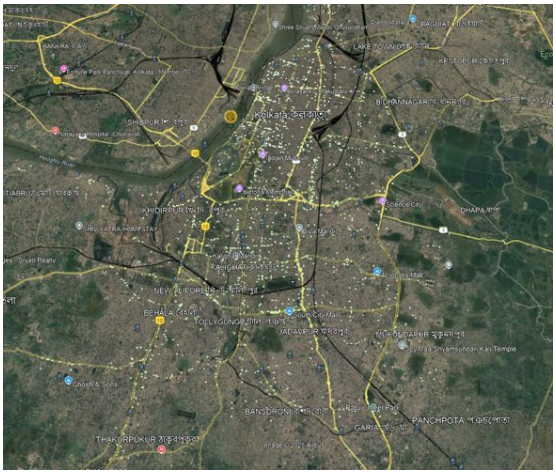

Fig. 6. Predicted Riskmap

V. CONCLUSION

The results show that integrating historical accident data with POI distances and temporal trends creates a strong predictive model, with Random Forest effectively capturing complex spatial patterns. The expansion of high-risk zones into new urban areas underscores the need for dynamic land-use planning. Limitations include reliance on static POI data and potential underreporting in accident records. Future improvements should involve real-time traffic and weather data, finer grid resolution (25m), and advanced models like spatio-temporal graph neural networks. Overall, the framework offers a scalable, interpretable tool to guide urban traffic safety planning and intervention.

REFERENCES

[1] V. A. Adewopo and N. Elsayed, "Smart city transportation: Deep learning ensemble approach for traffic accident detection," *IEEE Access*, 2024.

[2] M. A. Khasawneh, M. Daraghmeh, A. Awasthi, and A. Agarwal, "Multilevel learning for enhanced traffic congestion prediction using anomaly detection and ensemble learning." 2024.

[3] P. Khosravinia, T. Perumal, and J. Zarrin, "Enhancing road safety through accurate detection of hazardous driving behaviors with graph convolutional recurrent networks," *IEEE Access*, vol. 11, pp. 52 983–52 995, 2023.

[4] Q. Zhang, X. Chang, and S. B. Bian, "Vehicle-damage-detection segmentation algorithm based on improved mask rcnn," *IEEE Access*, vol. 8, pp. 6997–7004, 2020.

[5] J. Zhang, C. Song, Z. Mo, and S. Cao, "A transfer learning-based approach to estimating missing pairs of on/off ramp flows," *IEEE Transactions on Intelligent Transportation Systems*, 2023.

[6] A. Mbarek, M. Jiber, A. Yahyaouy, and A. Sabri, "Accident black spots identification based on association rule mining," *Bulletin of Electrical Engineering and Informatics*, vol. 13, no. 3, pp. 2075–2085, 2024.

[7] S. Yu, B. Zhang, T. Luo, S. Xiong, C. Wang, S. Wu, J. Zhu, and Q. Li, "Analysis of road network deformation and sinkhole hazards with sentinel-1 sar data: A case study of longgang district in shenzhen, china," in *IGARSS 2024-2024 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 2024, pp. 11 149–11 152.

[8] Z. Hu, J. Zhou, and E. Zhang, "Improving traffic safety through traffic accident risk assessment," *Sustainability*, vol. 15, no. 4, p. 3748, 2023.

[9] X. Zheng, D. Zhang, H. Gao, Z. Zhao, H. Huang, and J. Wang, "A novel framework for road traffic risk assessment with hmm-based prediction model," *Sensors*, vol. 18, no. 12, p. 4313, 2018.

[10] M. Mishra, B. Maitra, and S. K. Ghosh, "Estimation of road accident severity using k-means clustering of spatio-temporal data with backend spatial data infrastructure," in *IGARSS 2024-2024 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 2024, pp. 3142–3146.

[11] M. He, G. Meng, X. Wu, X. Han, and J. Fan, "Road traffic accident prediction based on multi-source data–a systematic review," *Promet-Traffic&Transportation*, vol. 37, no. 2, pp. 499–522, 2025.

[12] Y. Fu, C. Li, F. R. Yu, T. H. Luan, and Y. Zhang, "A survey of driving safety with sensing, vehicular communications, and artificial intelligence-based collision avoidance," *IEEE transactions on intelligent transportation systems*, vol. 23, no. 7, pp. 6142–6163, 2021.