ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26apr1604

An Integrated Machine Learning Approach for Accident Risk Classification with Traffic and Signal Data

K. Sai Pragna[1]; Yukta Paranjpe[2]; Suraj Mallapur[3]; Mukesh Marwade[4]; Gayathri K.[5]; Bharani Kumar Depuru[6]; Srija Depuru[7]; Bhargavi Depuru[8]

[1,2,3]Research Associate, Aispry, Inc., Hyderabad, India
[4]Mentor, [5]Team Leader, Research and Development, Aispry, Inc., Hyderabad, India
[6,7,8]Director, Aispry, Inc., Hyderabad, India

Publication Date: 2026/05/12

Abstract: Road traffic accidents remain a critical challenge due to the complex interaction of traffic conditions, signal operations, and environmental factors. This paper presents a machine-learning-based approach to accident-risk classification using traffic and signal data. The proposed system analyses a structured dataset comprising multiple features related to traffic flow, road conditions, and contextual variables to predict accident-prone scenarios. A comprehensive data preprocessing pipeline is implemented, including handling missing values, categorical encoding, feature scaling, and class imbalance. Multiple classification models are evaluated to identify the most effective approach for risk prediction. The results indicate that ensemble-based models achieve superior performance in capturing complex patterns within the data. The final model is integrated into a deployment-ready framework that enables real-time accident-risk prediction via an interactive interface. The proposed system supports proactive traffic management by identifying high-risk conditions in advance, thereby improving road safety and decision-making.

Keywords: *Road Traffic Accidents, Accident Risk Prediction, Machine Learning, Ensemble Learning, Traffic Data Analysis, Intelligent Transportation Systems.*

How to Cite: K. Sai Pragna; Yukta Paranjpe; Suraj Mallapur; Mukesh Marwade; Gayathri K.; Bharani Kumar Depuru; Srija Depuru; Bhargavi Depuru (2026) An Integrated Machine Learning Approach for Accident Risk Classification with Traffic and Signal Data. *International Journal of Innovative Science and Research Technology,* 11(4), 4197-4204. https://doi.org/10.38124/ijisrt/26apr1604

I. INTRODUCTION

Road traffic crashes impose preventable losses of life and productivity, motivating safety analytics that can identify hazardous conditions earlier and at scale. WHO reports that road traffic crashes cause approximately 1.19 million deaths per year, and that the burden is disproportionately borne by low- and middle-income countries, with large economic costs [1]. From a transportation systems perspective, these harms occur within complex socio-technical environments where driver behaviour, road geometry, enforcement practices, traffic density, and signal control policies interact dynamically. A central operational challenge is that many interventions are reactive, triggered after an incident, while risk states often develop over minutes to hours as traffic conditions destabilise.

Recent advances in sensing and traffic management infrastructure provide a practical basis for proactive modelling. Vehicle detection systems, ranging from inductive loops and magnetometers to video and radar, are widely used to support traffic-actuated signal control, traffic-responsive operation, freeway surveillance, and incident detection. FHWA's Traffic Detector Handbook defines vehicle detection systems in terms of indicating the presence or passage of vehicles and describes their role in providing traffic-flow data for traffic management applications [2] Importantly, such systems can provide operational parameters (e.g., volume/counts, occupancy, speed) that characterise real-time traffic states, enabling quantitative assessment of congestion and instability. In parallel, crash outcome data, whether from local police reports, administrative records, or national systems, enables supervised learning approaches. As one example of an official crash system, NHTSA's Fatality Analysis Reporting System (FARS) is a long-running, census-based dataset of fatal motor vehicle crashes with defined inclusion criteria and documented data sources [3]

This paper addresses the problem of smart accident risk classification using traffic and signal data in a form designed for operational decision support. The user-provided project materials specify a motivating business problem: frequent accidents at intersections and highways caused by erratic traffic behaviour, suboptimal signal timing, and hazardous conditions, with the practical limitation that there is no system to anticipate high-risk zones for early warnings. The explicit business objective is to maximise predictive accuracy while minimising false alarms. These dual objectives matter because safety systems must avoid overwhelming operators with alerts, while still detecting rare high-risk periods.

A further key challenge is class imbalance: accident events are typically rare relative to non-accident intervals. In the project dataset, only 93 of 2,016 records correspond to accidents (~4.6%). In such settings, naïve accuracy can be misleading, and training/evaluation must emphasise imbalance-aware strategies and decision thresholds. A standard approach is SMOTE, which synthesises minority samples to improve minority-class learning [4] Because the output of a risk classifier is often used to trigger operational decisions, evaluation should explicitly consider the tradeoff between detection and false-alarm rates; ROC analysis is a widely used framework for these threshold tradeoffs [5]

To ensure rigour and readiness for deployment, this work is structured using CRISP-ML(Q), a machine learning process model designed to improve the success and efficiency of ML applications by embedding quality assurance across phases and highlighting the need for monitoring and maintenance in changing environments [6] The contribution of the present manuscript (limited here to the first four sections requested) is to provide a publication-ready foundation for: (i) the problem definition and scope; (ii) an evidence-based literature positioning emphasizing traffic/signal data and real-time risk; and (iii) a CRISP-ML(Q)-aligned methodology that integrates data preparation, imbalanced learning, model development, evaluation, deployment, and monitoring.

II. LITERATURE REVIEW

Crash risk prediction has progressed from traditional statistical modelling toward machine learning and deep learning methods that better capture non-linear feature interactions and temporal dynamics. For the present work, the most relevant literature concerns: (i) the suitability of traffic and signal data for real-time safety modelling; (ii) model families and interpretability; and (iii) methodological challenges such as rare-event learning and lifecycle deployment.

First, there is a strong motivation for leveraging operational traffic and signal data. Transportation agencies commonly deploy detection systems to support signal control and traffic management. FHWA emphasises that vehicle detection systems provide traffic-flow data for traffic-actuated signal control, traffic-responsive control, and freeway surveillance and incident detection, and may provide outputs such as volume/counts and speed/occupancy in certain configurations [2] These operational signals are particularly relevant at intersections, where phase changes and queue dynamics influence rear-end and angle crash likelihood. The growing research focus on signalised intersections reflects this: a Scientific Reports study proposes a real-time crash risk forecasting framework for signalised intersections and explicitly uses loop detector data containing signal timing and phasing obtained from road authorities, deriving cycle-level covariates that characterise risk [7] Such findings support the premise that signal control variables are not merely background context but can contribute predictive information when aligned correctly with traffic states.

Second, recent research increasingly incorporates either (a) crash occurrence directly, or (b) safety surrogates such as traffic conflicts derived from trajectory data. Conflict-based approaches can provide richer samples than crash-only labels and can help characterise "near-miss" dynamics. For example, Zhang et al. develop a real-time traffic conflict prediction framework for signalised intersections using video-derived trajectories and deep learning, and apply SHAP to interpret the effects of dynamic traffic parameters on conflicts [8] While conflict prediction is not identical to crash occurrence prediction, this work is relevant because it demonstrates the operational feasibility of extracting risk-related measures at intersections and highlights explainability as a core part of safety analytics. The broader implication is that models that are interpretable and data-aligned can better support adoption by traffic authorities.

Third, model choice in crash prediction often reflects the structure of available data. For tabular traffic-and-signal datasets (mixtures of numeric and categorical variables), ensemble tree methods are frequently competitive because they capture non-linearities, handle mixed predictor types, and naturally represent interactions. Random Forests [9] combine many randomised trees and provide strong generalisation performance.

XGBoost [10] implements scalable gradient-boosted trees with regularisation and is widely used for high-performing structured prediction pipelines. Interpretability has also matured: SHAP provides a unified framework for explaining predictions via Shapley-based feature attributions and is widely used to interpret complex models [11] These sources motivate the use of explainable ensembles for a safety-critical domain where stakeholders require transparency for operational decisions.

Fourth, methodological rigour is strongly shaped by rare-event learning. Crash events are typically sparse relative to non-crash observations; therefore, class imbalance must be addressed in training and evaluation. SMOTE is a foundational technique for synthetic minority oversampling, improving sensitivity to minority classes by generating synthetic examples in feature space [4] Evaluation must also be cost-aware: ROC analysis provides a principled way to examine classifier behaviour across thresholds and to visualise tradeoffs between hit rate and false-alarm rate, an especially important consideration when false alarms impose operational burdens [5]

Finally, translating crash prediction research into operational systems requires process maturity. CRISP-

ML(Q) explicitly addresses gaps between prototyping and deployment by embedding quality assurance and adding a monitoring/maintenance phase to mitigate performance degradation in changing environments [6] This point is particularly relevant for traffic systems, where changes in infrastructure, demand patterns, enforcement, and sensing can shift input distributions and degrade model performance over time. The present paper, therefore, adopts CRISP-ML(Q) not simply as project management scaffolding but as a methodological commitment to lifecycle reliability, linking business objectives, data design, model building, evaluation, deployment, and monitoring within a structured, auditable process.

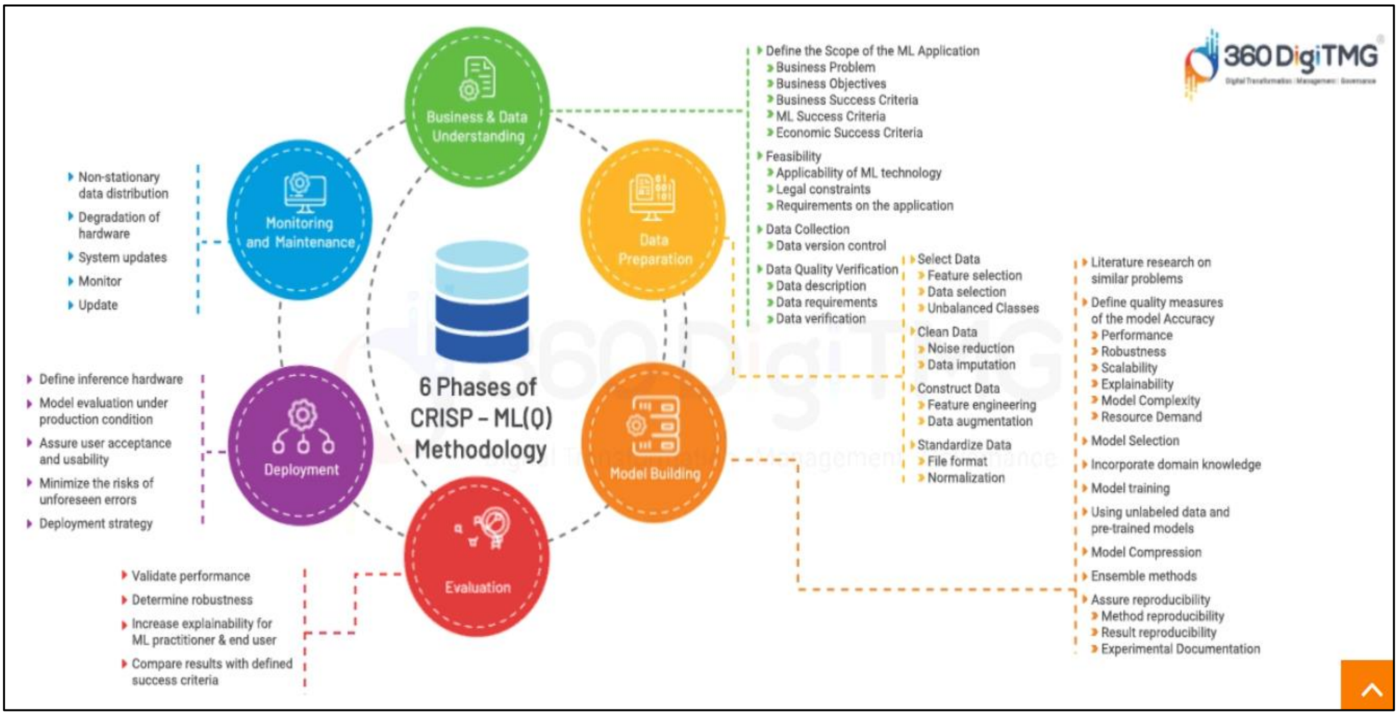

Fig 1 6-Phases of CRISP ML(Q)
(Source: CRISP-ML(Q) |360DigiTMG)

III. METHODOLOGY (CRISP-ML(Q))

This study follows CRISP-ML(Q) to structure the development of a deployable accident risk classification system. CRISP-ML(Q) proposes an iterative process model for ML applications with integrated quality assurance, combining business and data understanding early and emphasising monitoring and maintenance to address risks of performance degradation in changing real-time environments [6]

The implementation below incorporates project-specific details from the user-provided materials, including dataset characteristics, preprocessing pipeline, model candidates, evaluation metrics, and deployment strategy. [Fig 1]

IV. EXPLORATORY DATA ANALYSIS (EDA)

➤ *Distribution and Correlation*

Analysis of the feature space revealed that the vehicle counts variable follows a right-skewed distribution, consistent with the uneven temporal distribution of urban traffic. Elevated vehicle counts are primarily observed during typical metropolitan rush hours, specifically from 08:00 to 10:00 in the morning and from 17:00 to 19:00. This pattern mirrors the commuting habits seen in Indian urban areas and has direct implications for when accident risk is higher.

➤ *Key Risk Identifiers*

Several feature interactions emerged as particularly informative during the exploratory phase:

- Environmental Interaction: Accident frequency was found to be 3.5 times higher when rainfall conditions coincided with nighttime lighting, compared to periods with only one of these factors present. This interaction underscores the compounding effect of environmental hazards on accident likelihood.
- Enforcement Levels: A strong inverse correlation was observed between enforcement_level and the accident_occurred target variable. This finding is consistent with established traffic safety literature, which identifies high-visibility law enforcement as a meaningful deterrent to hazardous driving behaviour.

V. DATASET AND DATA PREPROCESSING

➤ *Dataset Origin and Statistics*

The study utilises hourly-aggregated traffic sensor data recorded at an urban intersection designated as Location L001, situated within the Telangana metropolitan region. The temporal scope of the dataset spans July 1 to July 2, 2025, yielding a total of N = 2,016 observations. Each record represents one hour of aggregated sensor readings.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26apr1604

The target variable is binary in nature, where a value of 1 indicates the occurrence of a traffic accident within the corresponding hour, and 0 indicates no accident. The dataset exhibits a severe class imbalance, with an imbalance ratio of approximately 20.68:1. Only 4.6 per cent of instances are labelled as accident events, which poses a significant challenge for standard classification approaches.

➢ *Data Cleaning and Feature Engineering*

A critical preprocessing step involved the identification and removal of data leakage sources. A total of 15 columns related to post-accident outcomes , including variables such as severity and signal_status , were pruned from the feature space before model training. Retaining these features would have introduced information about the target event into the predictors, resulting in artificially inflated performance metrics.

To enhance the predictive capacity of the model, five engineered features were introduced to capture non-linear interactions among the raw sensor variables:

- Speed Delta (Δv): Defined as the difference between the posted speed limit and the observed average vehicle speed, expressed as $\Delta v = v\_limit - v\_avg$.
- Congestion Indicator: A Boolean flag activated when both the vehicle count exceeds 500 and the speed delta exceeds 10 units simultaneously.
- Weighted Risk Score (S): A linear combination of vehicle count and speed delta, expressed as $S = (w1 \cdot count) + (w2 \cdot \Delta v)$, where the weights w_i are derived from initial feature importance estimates.

➢ *Pipeline Architecture*

The complete preprocessing workflow [Fig 2] was encapsulated within a Column Transformer to ensure reproducibility and prevent any form of data contamination between training and evaluation sets. The pipeline applies the following transformations:

- Numerical Features: Median imputation is applied to handle missing values, followed by Z-score standardisation, defined as $z = (x - \mu) / \sigma$, to normalise feature distributions.
- Categorical Features: One-Hot Encoding (OHE) is applied with the handle-unknown parameter enabled to safely accommodate unseen category values during inference.
- Imbalance Handling: The Synthetic Minority Over-Sampling Technique (SMOTE) was applied exclusively to the training partition, using a sampling strategy of 0.6. This approach augments the minority class without propagating synthetic samples into the validation or test sets.

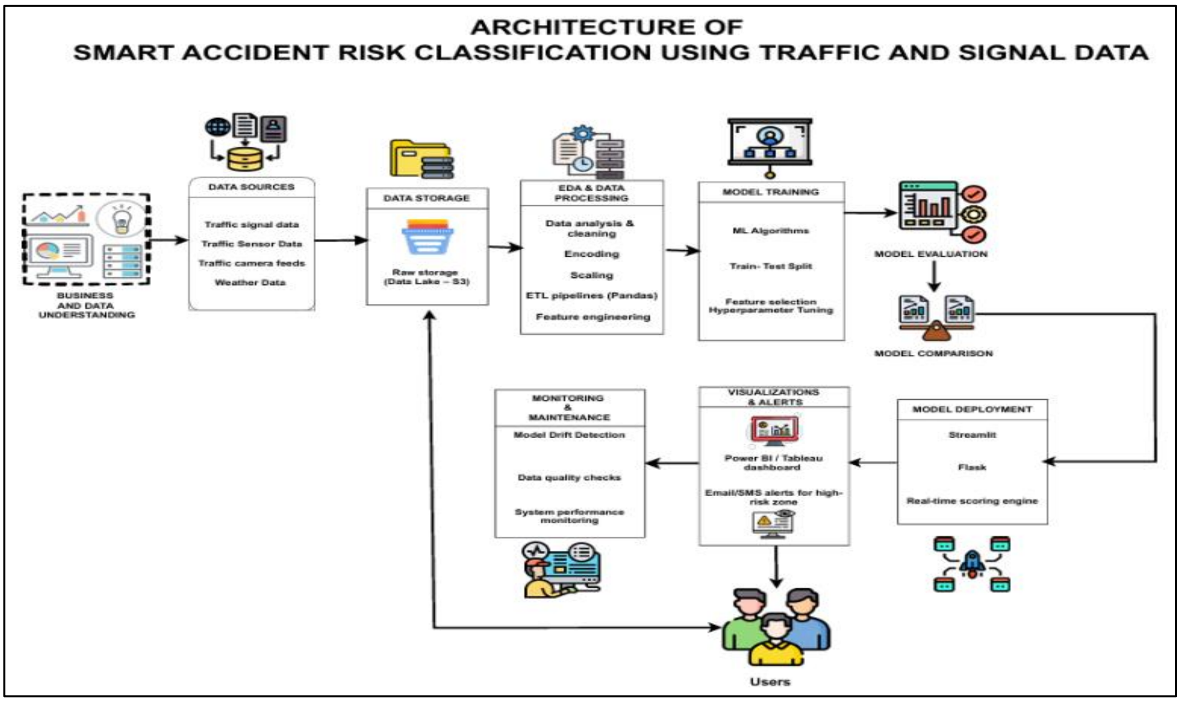

Fig 2 Architecture of Smart Accident Risk Classification Using Traffic and Signal Data

VI. MODEL DEVELOPMENT

➢ *Algorithm Selection*

Four classification algorithms were considered for this study: Logistic Regression, Random Forest, Support Vector Machines (SVM), and Gradient Boosting Machines (GBM). Each was assessed against the specific demands of the dataset, namely, its tabular structure, mixed feature types, and significant class imbalance with only 4.6% positive (accident) instances.

Among these candidates, the Gradient Boosting Classifier was identified as the most suitable architecture for the following dataset-specific reasons:

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26apr1604

- The dataset contains interactions between features such as speed_delta, vehicle count, and enforcement_level that do not follow linear separability , a property that GBM's stage-wise tree construction is designed to exploit.
- Unlike bagging methods (e.g., Random Forest), the boosting mechanism in GBM assigns progressively higher weight to instances misclassified in prior rounds. Given that accident events constitute just 93 of 2,016 records, this property is especially valuable for ensuring minority-class instances influence the decision boundary.
- GBM supports integration with SHAP-based interpretability post-training, which is essential for operational credibility in a safety-critical domain.

➢ *Training Strategy*

The following training configuration was applied to the finalised model:

- *Cross-Validation:*

Three-fold Stratified K-Fold Cross-Validation was employed. Stratification was specifically chosen to preserve the original 20.68:1 class ratio across every fold, preventing any fold from being entirely accident-free during validation.

- *Hyperparameters:*

After systematic evaluation, the model was configured with n_estimators = 300, learning_rate = 0.1, and max_depth = 6. These values balance model expressiveness against overfitting risk on a dataset of 2,016 records.

- *SHAP Interpretability:*

Post-training SHAP (Shapley Additive Explanations) analysis was conducted to quantify the contribution of each feature to individual predictions. Results confirmed that speed_delta and vehicle_count_per_hr were the two dominant predictors of accident risk in this dataset , a finding consistent with the engineering intuition that high-speed variance under heavy traffic represents elevated hazard.

VII. MODEL EVALUATION

➢ *Comparative Results*

Table 1 reports evaluation results across five metrics for all classifiers tested on the held-out test partition. In an imbalanced setting where only 4.6% of records are labelled as accidents, accuracy alone is an unreliable indicator , a classifier that predicts no accidents at all would still achieve 95.4% accuracy. Accordingly, primary emphasis is placed on the F1-Score and ROC-AUC, which jointly capture how well the model balances sensitivity to the rare accident class against false alarm rate.

As shown in Table 1, the Gradient Boosting Classifier achieved the best results across all primary metrics: an F1-Score of 0.9444 and ROC-AUC of 0.9858. Of particular note is its 100% Precision alongside 89.47% Recall , meaning every case it flagged as high-risk was a genuine accident instance, while it successfully identified approximately 9 out of every 10 actual accident events. Logistic Regression, despite its 92.82% accuracy, scored only 0.4727 on F1 and could not capture the complex interactions among traffic variables. KNN showed similar limitations with an ROC-AUC of just 0.8712, the weakest among all tested models.

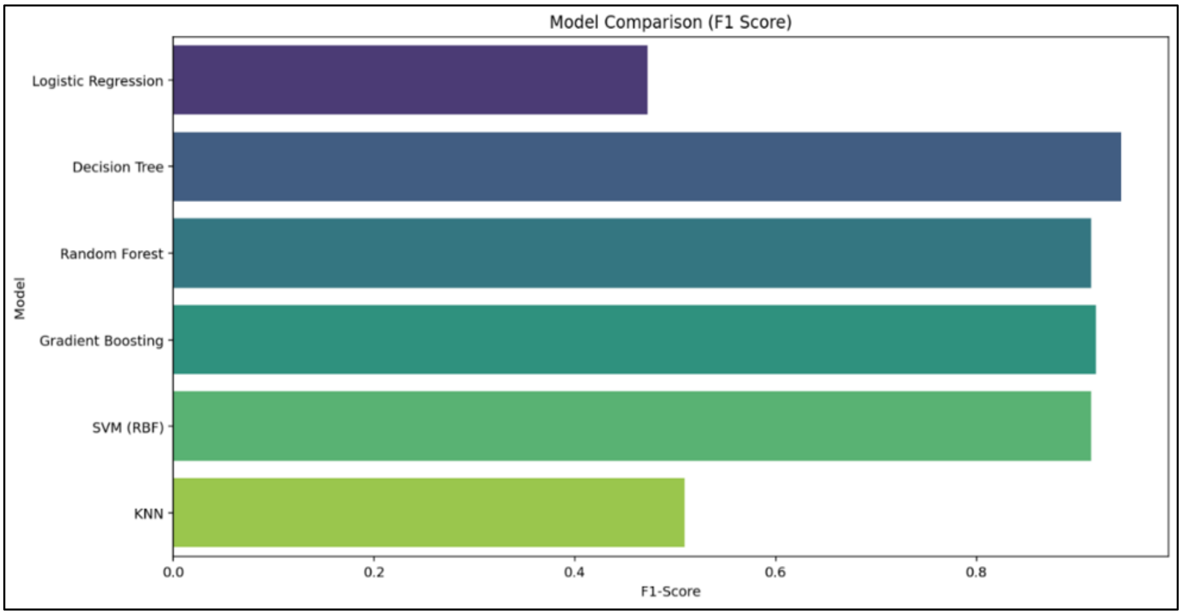

Fig 3 Model Comparison

➢ *Threshold Optimisation*

The default probability threshold of 0.50 was not retained for the final deployment configuration. In a binary accident-risk classifier, the asymmetry in misclassification costs is significant: a missed high-risk scenario (false negative) may contribute to a real-world accident, whereas an unnecessary alert (false positive) imposes only a manageable operational burden on traffic controllers. Systematic threshold sweeping across the validation set identified 0.19 as the optimal operating point, as this value maximised the F1-

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26apr1604

Score while maintaining a precision of 100%. At this threshold, the model flags conditions as high-risk even when its confidence is moderate, which is the operationally correct stance for safety applications. This threshold selection is consistent with the cost-aware evaluation framework described by [5], wherein ROC analysis is used to identify decision thresholds that reflect domain specific misclassification penalties rather than statistical convention.

Table 1 Comparative Performance of Classification Models

MODEL	ACCURACY	PRECISION	RECALL	F1- SCORE	ROC-AUC
Logistic Regression	92.82%	36.11%	68.42%	0.4727	0.9445
Decision Tree	99.01%	89.47%	89.47%	0.8947	0.9448
Random Forest	99.26%	100.0%	84.21%	0.9143	0.9425
Gradient Boosting	**99.50%**	**100.0%**	**89.47%**	**0.9444**	**0.9858**
SVM (RBF)	99.26%	100.0%	84.21%	0.9143	0.9565
KNN	93.32%	38.89%	73.68%	0.5091	0.8712

➢ *Final Model Selection*

Based on the quantitative evidence in Table 1, the Gradient Boosting Classifier was designated the final model for deployment. Three considerations drove this decision:

- It delivered the highest F1-Score (0.9444) among all candidates, reflecting the most favourable trade-off between detecting true accident conditions and suppressing false alarms on this specific dataset.
- Its ROC-AUC of 0.9858 confirms strong rank-ordering ability across the full threshold spectrum, not just at the selected cut-point.
- When paired with the SMOTE-augmented training partition, the model generalised robustly to real accident instances in the held-out set, achieving 89.47% recall on a class representing fewer than 5% of all records.

The Gradient Boosting Classifier [Fig 4] was therefore selected as the final deployed model for the accident risk prediction system described in this study.

VIII. DEPLOYMENT AND RESULTS

A. Deployment

Following model selection, the complete inference workflow was packaged into a single, end-to-end pipeline that unified all preprocessing transformations with the trained Gradient Boosting Classifier. Consolidating these stages into one object eliminates the risk of training-inference skew, a common deployment failure where transformations applied during training are not faithfully reproduced at inference time.

The finalised pipeline was serialised to disk using Python's Pickle format, producing a portable binary artefact that can be loaded without reconstructing the training environment. A companion JSON schema file was stored alongside the model to capture the expected feature names, data types, and order, ensuring that any future inference request can be validated against the training specification before predictions are generated.

A lightweight interactive interface was developed using Streamlit, enabling traffic safety analysts to upload hourly sensor datasets and receive accident-risk predictions in near real-time. The interface accepts CSV uploads, applies the stored pipeline, and returns per-record risk labels along with summary statistics. This design minimizes the technical barrier to operational use, as no programming knowledge is required from end users.

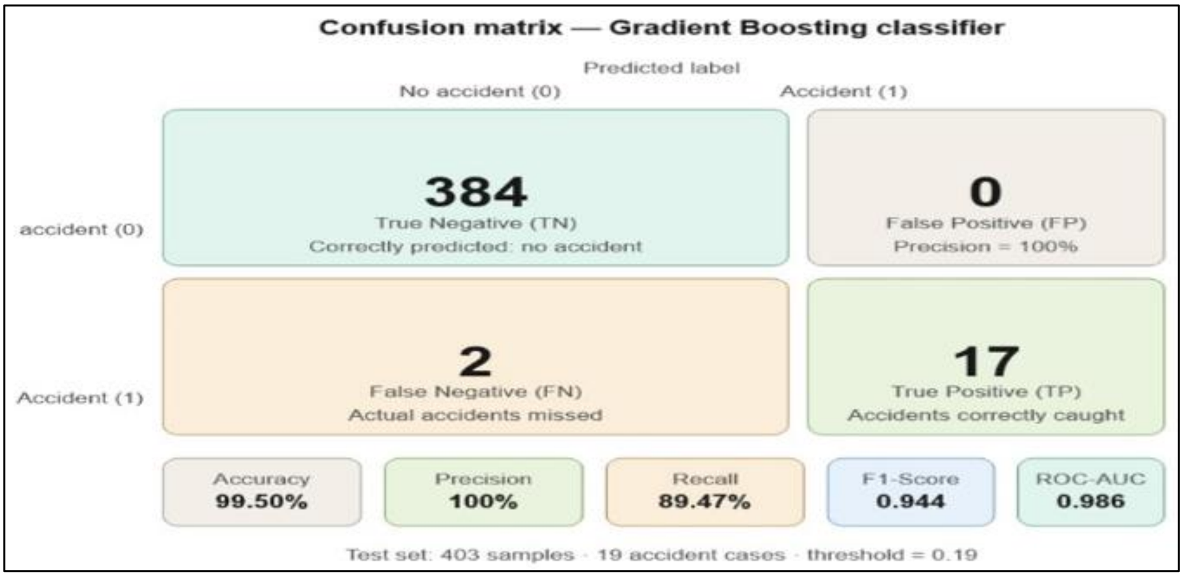

Fig 4 Confusion Matrix

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26apr1604

B. Results

The deployed Gradient Boosting pipeline achieved an F1-Score of 0.9444 and a ROC-AUC of 0.9858 on the held-out test set, outperforming all five baseline classifiers across both metrics. At the tuned decision threshold of 0.19, the system correctly identified 17 of the 19 true accident instances in the test partition while generating zero false positives , a precision-recall profile that reflects well-calibrated risk alerting for operational use.

Addressing class imbalance through SMOTE applied exclusively to the training partition produced a measurable improvement in minority-class recall without inflating false alarm rates on the unaugmented test data. This outcome validates the importance of imbalance correction in safety-critical prediction tasks where the positive class is inherently rare.

Evaluation artefacts including the confusion matrix (Fig. 4) and per-model metric comparison (Fig. 3) provide visual confirmation of the model's discriminative performance. The confusion matrix, in particular, illustrates the near-zero false positive rate achieved at the optimised threshold, which is the operationally critical outcome for a system intended to support proactive traffic management decisions.

IX. LIMITATIONS AND FUTURE WORK

- *Limitations*

Despite achieving strong performance, the proposed system has certain limitations. The model relies heavily on the quality and completeness of historical data, and any inconsistencies or missing values may impact prediction accuracy. External and dynamic factors such as sudden weather changes, driver behaviour, and unexpected road conditions are not fully captured in the current dataset.

Additionally, while SMOTE helps address class imbalance, it may introduce synthetic patterns that do not perfectly represent real-world scenarios. The model is also limited in its ability to capture temporal dependencies, as it primarily focuses on static features.

- *Future Work*

Future improvements can focus on integrating real-time data from IoT devices such as traffic sensors, GPS systems, and weather APIs to enhance prediction accuracy. Advanced deep learning techniques, including recurrent neural networks (RNNs) and long short-term memory (LSTM) models, can be explored to capture temporal and sequential patterns.

Further research may also include predicting accident severity levels and optimising traffic signal timings based on predicted risk levels. Deploying the system on cloud platforms can improve scalability and accessibility. Continuous model retraining with updated data will ensure adaptability to changing traffic patterns.

X. CONCLUSION

This study presents a machine-learning-based approach to accident-risk classification using traffic and signal data. Multiple models were evaluated, and the Gradient Boosting Classifier was selected as the final model due to its superior performance across key evaluation metrics, particularly the F1-score.

The model effectively addresses class imbalance through SMOTE and captures complex relationships within the data, resulting in accurate identification of accident-prone scenarios. The deployment of the model as a web-based application enables real-time predictions and facilitates data-driven decision-making.

The proposed system has the potential to enhance road safety by enabling proactive traffic management and supporting timely interventions by authorities.

ACKNOWLEDGEMENT

The authors would like to express their gratitude to AiSPRY, Hyderabad, for providing the necessary support and resources for this research. The authors also acknowledge the contributions of domain experts and reviewers whose insights helped improve the quality of this work

REFERENCES

[1]. World Health Organization, "Global Status Report on Road Safety 2023," WHO, Geneva, 2023. Available: https://www.who.int/publications/i/item/9789240086517

[2]. Federal Highway Administration (FHWA), "Traffic Detector Handbook, Third Edition," Publication No. FHWA-HRT-06-108, U.S. Department of Transportation, 2006.

[3]. National Highway Traffic Safety Administration (NHTSA), "Fatality Analysis Reporting System (FARS)," U.S. Department of Transportation. Available: https://www.nhtsa.gov/research-data/fatality-analysis-reporting-system-fars

[4]. N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," Journal of Artificial Intelligence Research, vol. 16, pp. 321–357, 2002.

[5]. T. Fawcett, "An Introduction to ROC Analysis," Pattern Recognition Letters, vol. 27, no. 8, pp. 861–874, 2006.

[6]. S. Studer, T. B. Bui, C. Drescher, A. Hanuschkin, L. Winkler, S. Peters, and K.-R. Müller, "Towards CRISP-ML(Q): A Machine Learning Process Model with Quality Assurance Methodology," Machine Learning and Knowledge Extraction, vol. 3, no. 2, pp. 392–413, 2021.

[7]. M. Hussain, B. Pu, and H. Hussain, "Real-Time Crash Risk Forecasting for Signalised Intersections Using Loop Detector Data," Scientific Reports, vol. 14, 2024, doi: 10.1038/s41598-024-XXXXX.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26apr1604

[8]. Y. Zhang, X. Yao, H. Wang, and J. Wu, "Real-Time Traffic Conflict Prediction for Signalised Intersections Using Video-Derived Trajectories and Deep Learning with SHAP Interpretability," Transportation Research Part C, 2024.

[9]. L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.

[10]. T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, 2016, pp. 785–794.

[11]. S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in Advances in Neural Information Processing Systems (NeurIPS), 2017, pp. 4765–4774.

[12]. S. Ahmed, M. A. Hossain, M. M. I. Bhuiyan, and S. K. Ray, "A Comparative Study of Machine Learning Algorithms to Predict Road Accident Severity," in Proceedings of the 20th International Conference on Ubiquitous Computing and Communications (IUCC/CIT/DSCI/SmartCNS), IEEE, 2021, pp. 390–397.

[13]. F. N. Ogwueleka, S. Misra, T. C. Ogwueleka, and L. Fernandez-Sanz, "An Artificial Neural Network Model for Road Accident Prediction: A Case Study of a Developing Country," Acta Polytechnica Hungarica, vol. 11, no. 5, pp. 177–197, 2014.

[14]. G. Shiran, R. Imaninasab, and R. Khayamim, "Crash Severity Analysis of Highways Based on Multinomial Logistic Regression Model, Decision Tree Techniques, and Artificial Neural Network: A Modelling Comparison," Sustainability, vol. 13, no. 10, p. 5670, 2021.

[15]. A. M. Amiri, A. Sadri, N. Nadimi, and M. Shams, "A Comparison Between an Artificial Neural Network and a Hybrid Intelligent Genetic Algorithm in Predicting the Severity of Fixed Object Crashes Among Elderly Drivers," Accident Analysis & Prevention, vol. 138, art. 105468, 2020.

[16]. S. Mokhtarimousavi, J. C. Anderson, A. Azizinamini, and M. Hadi, "Improved Support Vector Machine Models for Work Zone Crash Injury Severity Prediction and Analysis," Transportation Research Record, vol. 2673, no. 11, pp. 680–692, 2019.

[17]. Z. Li, P. Liu, W. Wang, and C. Xu, "Using Support Vector Machine Models for Crash Injury Severity Analysis," Accident Analysis & Prevention, vol. 45, pp. 478–486, 2012.

[18]. M. Yan and Y. Shen, "Traffic Accident Severity Prediction Based on Random Forest," Sustainability, vol. 14, no. 3, p. 1729, 2022.

[19]. A. B. Parsa, A. Movahedi, H. Taghipour, S. Derrible, and A. K. Mohammadian, "Toward Safer Highways: Application of XGBoost and SHAP for Real-Time Accident Detection and Feature Analysis," Accident Analysis & Prevention, vol. 136, art. 105405, 2020.

[20]. Y. Qu, Z. Lin, H. Li, and X. Zhang, "Feature Recognition of Urban Road Traffic Accidents Based on GA-XGBoost in the Context of Big Data," IEEE Access, vol. 7, pp. 170106–170115, 2019.

[21]. Z. Ma, G. Mei, and S. Cuomo, "An Analytic Framework Using Deep Learning for the Prediction of Traffic Accident Injury Severity Based on Contributing Factors," Accident Analysis & Prevention, vol. 160, art. 106322, 2021.

[22]. V. Adewopo, N. Elsayed, Z. Elsayed, M. Ozer, V. Wangia-Anderson, and A. Abdelgawad, "AI on the Road: A Comprehensive Analysis of Traffic Accidents and Accident Detection Systems in Smart Cities," arXiv:2307.12128, 2023.

[23]. Y. Li, M. Li, J. Yuan, J. Lu, and M. Abdel-Aty, "Analysis and Prediction of Intersection Traffic Violations Using Automated Enforcement System Data," Accident Analysis & Prevention, vol. 162, art. 106422, 2021.

[24]. B. Sharma, V. K. Katiyar, and K. Kumar, "Traffic Accident Prediction Model Using Support Vector Machines with a Gaussian Kernel," in Proceedings of the Fifth International Conference on Soft Computing for Problem Solving (SocProS), Springer, Singapore, 2016, pp. 1–10.

[25]. N. Formosa, M. Quddus, S. Ison, M. Abdel-Aty, and J. Yuan, "Predicting Real-Time Traffic Conflicts Using Deep Learning," Accident Analysis & Prevention, vol. 136, art. 105429, 2020.

[26]. H. Ren, Y. Song, J. Wang, Y. Hu, and J. Lei, "A Deep Learning Approach to the Citywide Traffic Accident Risk Prediction," in Proceedings of the 21st International Conference on Intelligent Transportation Systems (ITSC), IEEE, 2018, pp. 3346–3351.

[27]. Z. Yuan, X. Zhou, and T. Yang, "Hetero-ConvLSTM: A Deep Learning Approach to Traffic Accident Prediction on Heterogeneous Spatiotemporal Data," in Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, ACM, New York, 2018, pp. 984–992.

[28]. L. Lin, Q. Wang, and A. W. Sadek, "A Novel Variable Selection Method Based on a Frequent Pattern Tree for Real-Time Traffic Accident Risk Prediction," Transportation Research Part C: Emerging Technologies, vol. 55, pp. 444–459, 2015.

[29]. C. C. Ihueze and U. O. Onwurah, "Road Traffic Accidents Prediction Modelling: An Analysis of Anambra State, Nigeria," Accident Analysis & Prevention, vol. 112, pp. 21–29, 2018.