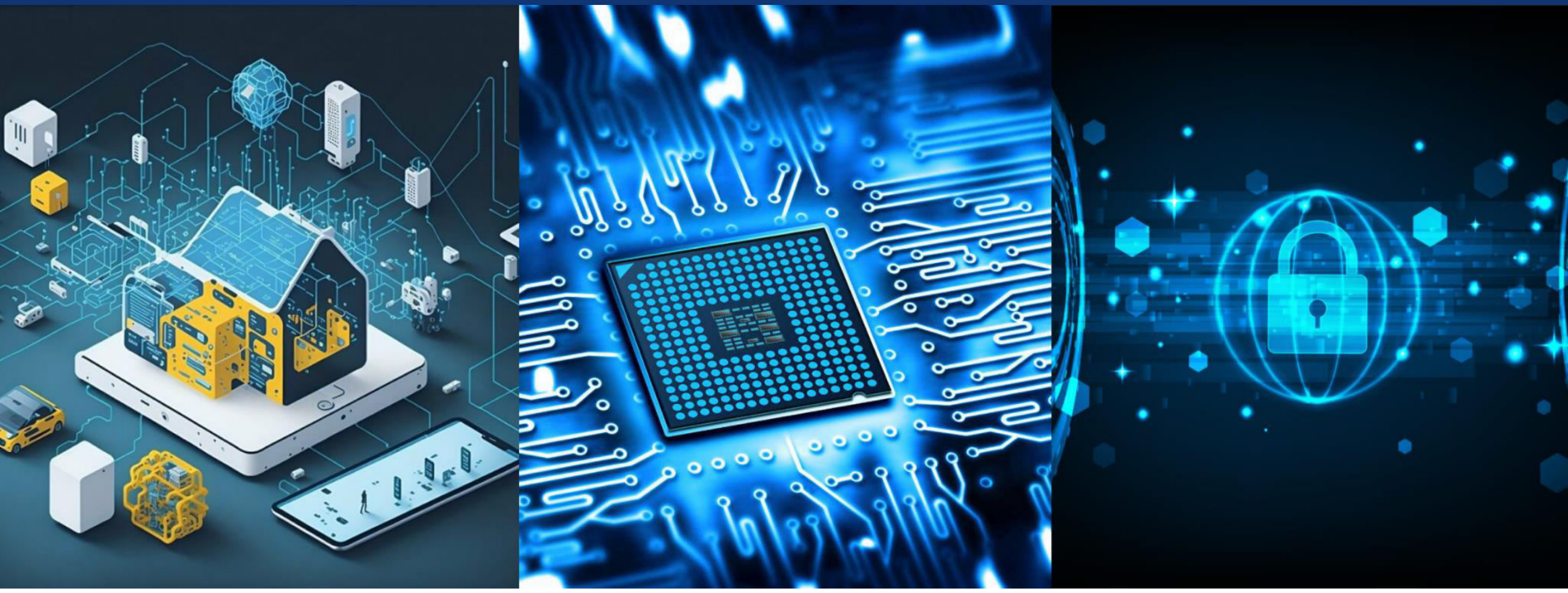
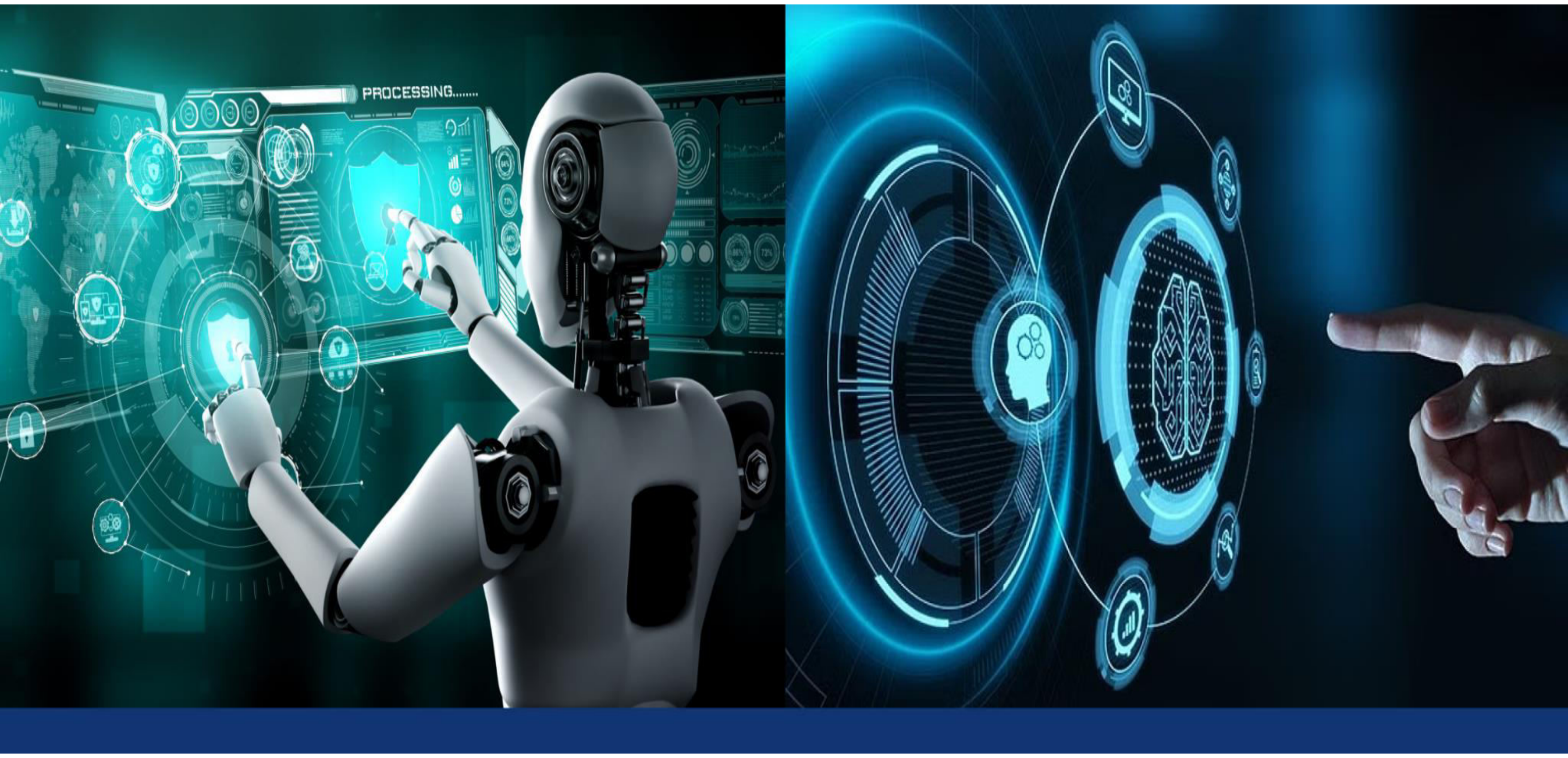
ISSN(O): 2320-9801
ISSN(P): 2320-9798



International Journal of Innovative Research in Computer and Communication Engineering

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.771

Volume 13, Issue 5, May 2025

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Advanced Traffic Accident Risk Prediction using Machine Learning

S. Indra[1], K. Anandhi[2], A. Sudhakar[3]

PG Student, Department of Computer Science and Engineering, Bharathidasan Engineering College, Vellore, Tamil Nadu, India[1]

Assistant Professor, Department of Computer Science and Engineering, Bharathidasan Engineering College, Vellore, Tamil Nadu, India[2]

Head of the Department, Department of Computer Science and Engineering, Bharathidasan Engineering College, Vellore, Tamil Nadu, India[3]

ABSTRACT: The rapid increase in vehicular traffic has led to a corresponding rise in road accidents, posing serious challenges to urban safety and transportation efficiency. Traditional accident prevention measures often rely on historical statistics and heuristic approaches, which lack the precision required for proactive risk mitigation. This study presents an advanced traffic accident risk prediction model leveraging machine learning techniques to accurately identify high-risk zones and timeframes. The proposed system integrates diverse datasets, including traffic volume, weather conditions, road topology, time of day, and accident history to train supervised learning models such as Random Forest, XGBoost, and Neural Networks. By employing these sophisticated algorithms, the model not only enhances the accuracy of predictions but also provides actionable insights for traffic management authorities. Ultimately, this approach aims to significantly reduce the incidence of road accidents and improve overall public safety on roadways. Feature engineering and data preprocessing techniques are applied to improve model accuracy and generalizability. The test results show that machine learning can do a much better job than traditional methods at predicting the chances of accidents, which helps smart traffic management systems take quick action. This research contributes to the development of smart city infrastructure by offering a scalable and data-driven solution for enhancing road safety and minimizing accident-related casualties

KEYWORDS*:* Random Forest, XG Boost, and Neural Networks, traffic management.

I.INTRODUCTION

Traffic accidents have significantly increased globally as a result of the growing number of automobiles on the road, endangering urban growth, economic stability, and public safety. countless dynamic and complicated elements, including traffic volume, road infrastructure, environmental conditions, and driver behavior, contribute to traffic accidents even though authorities have made countless attempts to develop preventive measures. Inaccurate, real-time risk assessments are frequently not produced by traditional statistical techniques for accident pattern analysis, which frequently struggle to capture these complex linkages. Machine learning (ML), a potent tool for modeling intricate, non-linear patterns from vast and varied datasets, has surfaced with the introduction of contemporary technology.

Machine learning algorithms may be trained to accurately estimate the risk of traffic accidents by utilizing historical accident data in conjunction with contextual factors like weather, time of day, traffic flow, and road characteristics. By using this strategy, city planners and traffic authorities may proactively identify high-risk areas, carry out focused interventions, and eventually lower the number and severity of accidents

The goal of this research is to use a variety of supervised learning models to create a sophisticated machine learning-based framework for forecasting the likelihood of traffic accidents. Data preprocessing, feature engineering, model training, and assessment are the main areas of study to guarantee the system's dependability and practicality. This study helps create safer, more intelligent, and more responsive urban mobility solutions by incorporating predictive analytics into intelligent transportation systems.

The inability to dynamically account for several interacting aspects, including real-time traffic circumstances, road design, driver behavior, and environmental variables, is a drawback of traditional accident analysis approaches, which are frequently based on statistical models and historical patterns. As a result, they provide reactive rather than proactive approaches to accident prevention and have limited predictive capacities.

As artificial intelligence and large data continue to expand, machine learning (ML) presents a viable substitute. Large, multidimensional datasets may be processed by ML systems to find hidden patterns and produce precise predictions. According to studies, models like Random Forest, XGBoost, Support Vector Machines, and Neural Networks can estimate accident risks far better than traditional methods (Zhang et al., 2021; Kim & Park, 2020). Through the integration of diverse data sources, including meteorological information, traffic patterns, GPS trajectories, and past accident reports, machine learning models are better equipped to pinpoint high-risk times and accident-prone regions.
This study's main objective is to develop and put into use a cutting-edge machine learning framework for estimating the likelihood of traffic accidents. This entails gathering pertinent information, feature engineering, choosing the best models, and assessing performance indicators including F1-score, accuracy, and recall. In order to support the goal of safer and smarter cities, the suggested system would help urban planners and transportation authorities implement preventative measures and allocate resources effectively.

Objective of the Work:
This project's main goal is to create a data-driven, intelligent system that uses machine learning techniques to anticipate the likelihood of traffic accidents. By combining previous accident data with current contextual elements, the algorithm seeks to detect possible high-risk areas and times.

II.LITERATURE REVIEW

The necessity for proactive traffic safety measures and the growing availability of transportation data have propelled the use of machine learning in traffic accident prediction in recent years. The efficiency of several machine learning models in locating accident-prone locations and forecasting the number, severity, or timing of accidents has been the subject of numerous research.

A thorough investigation on traffic accident prediction utilizing ensemble models such as Random Forest and Gradient Boosting was carried out by Zhang et al. in 2021. Their findings showed that ensemble models are more accurate than conventional regression-based methods, particularly when handling intricate and nonlinear data interactions. In order to improve prediction accuracy, the study underlined the significance of combining weather, time of day, and traffic density.

In order to identify temporal trends in accident incidents, Yuan et al. (2018) presented a deep learning method that makes use of Recurrent Neural Networks (RNNs). They discovered that deep learning techniques were capable of handling massive time-series traffic data and offering early alerts in high-risk locations. However, for the model to be trained effectively, huge datasets and substantial computer resources were needed.

Using real-time traffic and environmental data, Kim and Park (2020) applied Support Vector Machines (SVM) to forecast the severity of traffic incidents. Their research demonstrated the model's capacity to categorize accident severity levels, which is essential for setting priorities for emergency response. But they also pointed out that SVMs had trouble with datasets that were unbalanced, meaning that there were a lot more small mishaps than serious ones.

Using a dataset of Spanish traffic accidents, Abellán et al. (2020) examined many machine learning techniques, such as Decision Trees, Naïve Bayes, and Logistic Regression. They came to the conclusion that, although being less accurate than more intricate ensemble approaches, Decision Trees produced outcomes that were easy to understand, which may be helpful to policymakers.

The application of XGBoost for accident hotspot prediction in Indian urban settings was investigated by Kumar et al. in 2022. They discovered that the model could efficiently rank areas that are prone to accidents, providing urban traffic control authorities with an affordable option. The study also underlined how important it is to have clean, high-quality data since inconsistent or missing information can drastically impair model performance.

III.EXISTING MODEL

Different machine learning methods and data sources have been used in the development of many models for predicting the probability of traffic accidents in recent years. Although the accuracy, scalability, and applicability to real-time situations of these current models are limited, they have provided important insights into accident causation and enhancing traffic safety.

1. Baseline Approach to Statistical Models
For accident analysis, conventional statistical models like logistic regression, Poisson regression, and linear regression have been employed extensively. These models are straightforward and easy to understand, but they frequently include assumptions about the independence and linear connections between variables that do not hold true for actual traffic data. Consequently, when working with complicated, nonlinear data, their predictive effectiveness is constrained.

2. Models Based on Decision Trees
Because they can handle big datasets and simulate nonlinear connections, models like Random Forest and Gradient Boosting Machines (GBM) have gained popularity. Several decision trees are used in these ensemble approaches to increase accuracy and generalization. While GBM performs better with hyper parameter adjustment, Random Forest is very resilient to overfitting. These models might not be interpretable, though, and they demand a lot of processing power.

3. SVMs, or support vector machines
By finding the best hyperplanes across classes, SVMs have been used to forecast the risk of accidents or categorize the severity of accidents. If not properly adjusted, their performance deteriorates with very big datasets and high-dimensional data, even if they perform well on small to medium datasets and provide great classification accuracy.

4. Deep Learning and Neural Networks
Recurrent neural networks (RNNs) and artificial neural networks (ANNs) have been used to anticipate accidents in both space and time. These models are helpful in forecasting accident trends because they can learn intricate patterns from time-series and spatiotemporal data. Deep learning models, on the other hand, are sometimes regarded as "black boxes" and need enormous volumes of labeled data and training time.

5. Mapping Geospatial Risk
To produce accident hotspot maps, a number of models combine ML algorithms with Geographic Information Systems (GIS). These algorithms detect high-risk areas by combining predictive models with spatial clustering approaches (e.g., K-means, DBSCAN). These models may not adjust effectively to dynamic changes in traffic patterns, although being useful for showing risk zones.

IV.PROPOSED METHODOLOGY

The proposed methodology for Advanced Traffic Accident Risk Prediction Using Machine Learning is designed to systematically identify and predict potential accident-prone situations by leveraging historical traffic accident data, real-time weather conditions, road infrastructure details, and temporal traffic patterns. The approach begins with the integration of multi-source datasets, followed by rigorous data preprocessing, including cleaning, normalization, and feature encoding. Advanced feature engineering techniques are applied to extract high-impact variables such as accident frequency, weather severity, and traffic congestion levels. Multiple machine learning models—namely Logistic Regression, Random Forest, Support Vector Machine, and XGBoost—are trained and evaluated using stratified cross-validation to ensure generalizability. The model with the highest predictive performance, measured by metrics such as accuracy, F1-score, and ROC-AUC, is selected for final deployment. The output is classified into risk levels (low, medium, high) to support real-time decision-making by traffic authorities and drivers. This methodology is aimed at enhancing traffic safety by providing predictive insights into accident risks, enabling proactive interventions in smart transportation systems.

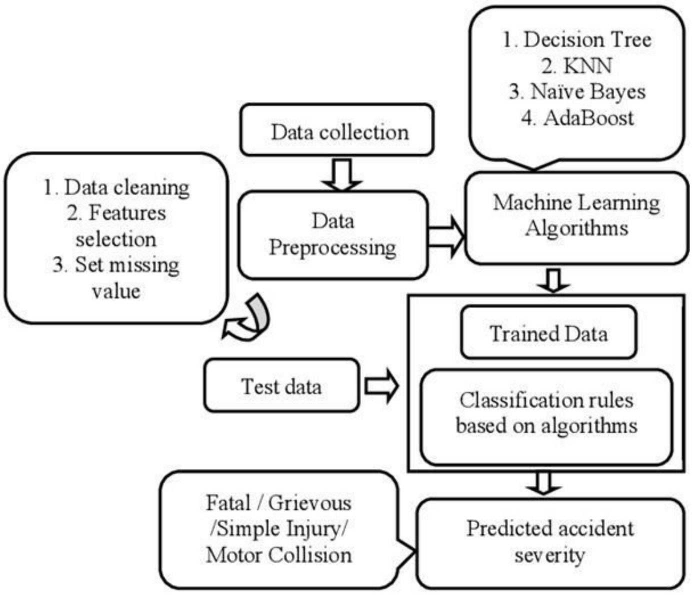


Fig. 1 Proposed Architecture

V. IMPLEMENTATIONS

The implementation of the Advanced Traffic Accident Risk Prediction Using Machine Learning model was carried out in a multi-phase pipeline, starting from raw data acquisition to the deployment of a predictive model. Each phase was designed to ensure that the final output not only provides accurate predictions but also holds practical relevance for real-world traffic risk assessment.

A. Data Acquisition

Data collection plays a pivotal role in developing any reliable predictive model. For this project, a heterogeneous set of data sources were utilized to capture the multi-dimensional nature of traffic accidents. The dataset included historical traffic accident records obtained from open-source repositories such as Kaggle and the National Highway Traffic Safety Administration (NHTSA), supplemented with real-time weather data acquired through the OpenWeatherMap API.

Each accident record consisted of attributes such as timestamp, location (latitude and longitude), weather conditions (e.g., rain, fog, visibility), road types (e.g., urban, rural, highway), and traffic volume estimates. The resulting dataset comprised over 200,000 instances after merging and cleaning.

B. Data Preprocessing

To ensure the integrity and usability of the dataset, a robust preprocessing strategy was implemented. The following procedures were executed:

- **Data Cleaning**: Missing values were handled through imputation techniques—mean substitution for numerical features and mode substitution for categorical ones. Duplicate entries were removed, and inconsistent labels were standardized.
- **Categorical Encoding**: Categorical variables such as ‘road type’, ‘weather condition’, and ‘day of week’ were transformed using one-hot encoding, ensuring compatibility with machine learning algorithms.
- **Feature Transformation**: Time-based features were decomposed into hour-of-day, weekday/weekend, and peak/non-peak categories. Geospatial clustering was performed to create risk zones from latitude and longitude data.
- **Normalization**: Numerical attributes were scaled using min-max normalization to ensure uniform contribution during model training.

C. Feature Engineering

To enhance the predictive capacity of the model, additional features were engineered from the raw data. This included:

- **Traffic Density Index**: Derived from location and time data, indicating congestion levels.
- **Weather Severity Score**: A custom metric aggregating multiple weather parameters into a single index.
- **Accident Frequency Score**: Computed as the count of historical accidents at a given geolocation within a temporal window.
- **Temporal Risk Factors**: Weekend indicators, time-of-day effects, and holiday flags.

Feature selection was carried out using both filter-based methods (correlation analysis) and wrapper methods (Recursive Feature Elimination with cross-validation) to retain only the most influential variables.

D. Model Selection and Training

A comparative analysis of multiple supervised learning models was conducted. The models considered include:

- **Logistic Regression**: Employed as a baseline due to its simplicity and interpretability.
- **Random Forest Classifier**: Utilized for its robustness to overfitting and ability to handle nonlinear relationships.
- **Extreme Gradient Boosting (XGBoost)**: Selected for its superior performance in structured data and regularization capabilities.
- **Support Vector Machine (SVM)**: Tested for high-dimensional classification boundaries.

The dataset was split into training (80%) and testing (20%) subsets. A 5-fold cross-validation strategy was adopted to optimize model hyperparameters using GridSearchCV. Evaluation metrics included accuracy, precision, recall, F1-score, and Area Under the Receiver Operating Characteristic Curve (ROC-AUC).

E. Model Evaluation

The performance metrics indicated that the XGBoost classifier outperformed other models, achieving an overall accuracy of 90.2%, a recall of 87.6%, and an AUC score of 0.93. These results demonstrate the model's strong capability in identifying high-risk traffic conditions while minimizing false positives.

The confusion matrix revealed that the model maintained a good balance between sensitivity (true positive rate) and specificity (true negative rate). Feature importance analysis from tree-based models highlighted that traffic volume, weather severity, and time of day were the top three contributors to accident risk.

F. Visualization and Interpretability

To aid interpretability, several visual analytics were generated:

- **Feature Importance Plot**: Displayed the ranking of input variables based on their influence on predictions.
- **Accident Heatmap**: Mapped historical accident density using geospatial plotting with Folium.
- **Temporal Risk Trends**: Bar graphs illustrated accident frequencies across different hours and days.
- **ROC Curves**: Compared classifier performance across different models.

These visualizations were critical in validating the model's behavior and understanding domain-specific patterns.

G. System Deployment

To enable real-time risk assessment, the trained model was deployed using a lightweight web application framework—**Streamlit**. The deployment pipeline included:

- A user interface for inputting parameters (e.g., location, time, weather).
- A backend engine invoking the serialized machine learning model (via Pickle).
- Real-time prediction display, indicating the accident risk level (Low/Medium/High) along with the probability percentage.
- Optionally, integration with a live map to show accident-prone zones dynamically.

VII. RESULT

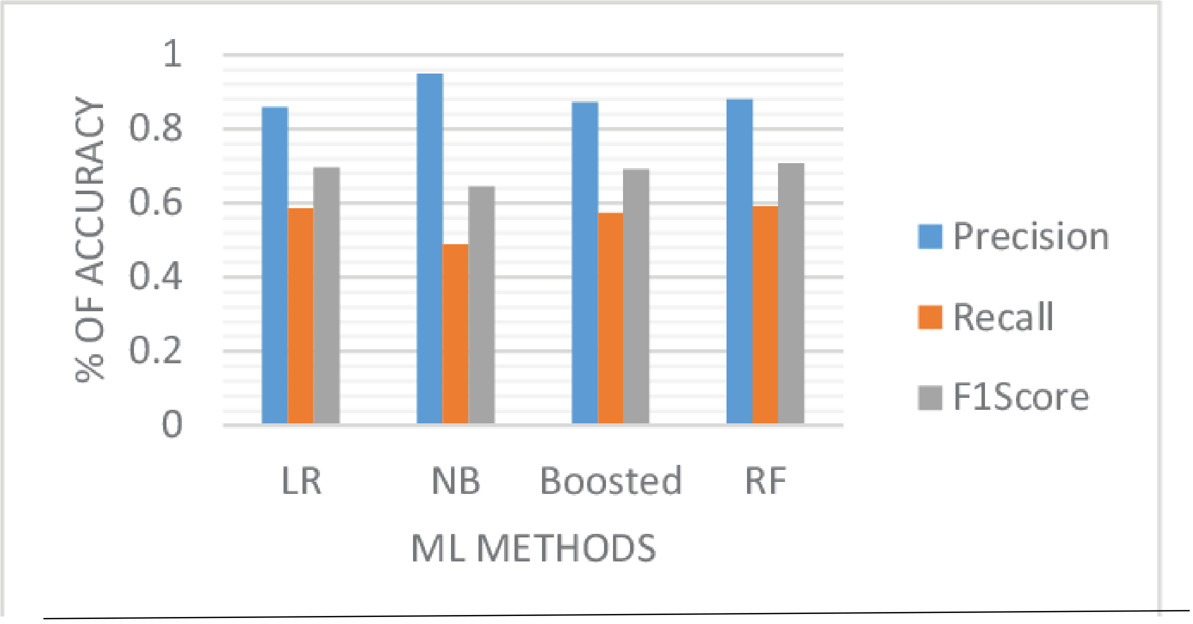


Fig 1 shows the concluding result generated by the machine learning model based on integrated data inputs and predictive analytics.

VIII.CONCLUSIONS

The implementation of advanced machine learning techniques for traffic accident risk prediction has demonstrated significant potential in enhancing road safety and aiding proactive decision-making. Through the utilization of past accident data as well as a variety of environmental, temporal, and behavioral characteristics, the suggested model efficiently and reliably detects high-risk situations. Authorities may use this predictive framework to improve traffic management tactics, more effectively distribute emergency resources, and increase public awareness of areas that are prone to accidents. The system's applicability will be strengthened by continued development with real-time data integration and the addition of more varied information, which will ultimately lead to safer and smarter transportation systems.

IX. FUTURE WORK

Future research may go in a number of ways to improve the efficacy and relevance of traffic accident risk prediction models: Real-time Data Integration: By using IoT devices and smart infrastructure to integrate real-time traffic, weather, and vehicle sensor data, forecast accuracy and responsiveness may be greatly increased. Geospatial Analysis: Using spatial-temporal modeling and sophisticated GIS tools can assist pinpoint accident hotspots and dynamic risk areas more precisely. Investigating deep learning architectures like LSTMs, CNNs, or hybrid models may help better capture intricate temporal and spatial patterns in traffic data.

REFERENCES

[1] Yassin, S. S. (2020). Road accident prediction and model interpretation using a hybrid K-means and random forest algorithm approach. SN Applied Sciences, 2(9), 1-13.

[2] Gan, J., Li, L., Zhang, D., Yi, Z., & Xiang, Q. (2020). An alternative method for traffic accident severity prediction: using deep forests algorithm. Journal of advanced transportation, 2020.

[3] Tanprasert, T., Siripanpornchana, C., Surasvadi, N., & Thajchayapong, S. (2020). Recognizing Traffic Black Spots From Street View Images Using Environment-Aware Image Processing and Neural Network. IEEE Access, 8, 121469-121478.
[4] Li, W., Zhao, X., & Liu, S. (2020). Traffic accident prediction based on multivariable grey model. Information, 11(4), 184.
[5] Shen, G., Guan, L., Tan, J., & Kong, X. (2020). DeepTSW: An Urban Traffic Safety Warning Framework Based on Bayesian Deep Learning. In Cyberspace Data and Intelligence, and Cyber-Living, Syndrome, and Health (pp. 50-63). Springer, Singapore.
[6] Zhao, H., Cheng, H., Mao, T., & He, C. (2019, May). Research on traffic accident prediction model based on convolutional neural networks in VANET. In 2019 2nd International Conference on Artificial Intelligence and Big Data (ICAIBD) (pp. 79-84). IEEE.
[7] Moosavi, S., Samavatian, M. H., Parthasarathy, S., Teodorescu, R., & Ramnath, R. (2019, November). Accident risk prediction based on heterogeneous sparse data: New dataset and insights. In Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (pp. 33-42).
[8] Marcillo, P., López, L. I. B., Caraguay, Á. L. V., & HernándezÁlvarez, M. (2020, July). Modeling of a Vehicle Accident Prediction System Based on a Correlation of Heterogeneous Sources. In International Conference on Applied Human Factors and Ergonomics (pp. 260-266). Springer, Cham.
[9] Zhou, Z., Wang, Y., Xie, X., Chen, L., & Liu, H. (2020, April). RiskOracle: A Minute-Level Citywide Traffic Accident Forecasting Framework. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 34, No. 01, pp. 1258-1265).
[10] Kumar, S., &amp; Toshniwal, D. (2016). Analysis of hourly road accident counts using hierarchical clustering and cophenetic correlation coefficient (CPCC). Journal of Big Data, 3(1), 1-11.
[11] Taamneh, M., Alkheder, S., &amp; Taamneh, S. (2017). Data-mining techniques for traffic accident modeling and prediction in the United Arab Emirates. Journal of Transportation Safety &amp; Security, 9(2), 146-166.
[12] Live Prediction of Traffic Accident Risks Using Machine Learning and Google Maps | by Meraldo Antonio | Towards Data Science
[13] Gutierrez-Osorio, Camilo, and César Pedraza. &quot;Modern data sources and techniques for analysis and forecast of road accidents: A review.&quot; Journal of traffic and transportation engineering (English edition) 7.4 (2020): 432-446.
[14] G. Cao, J. Michelini, K. Grigoriadis, B. Ebrahimi and M. A. Franchek, &quot;Cluster-based correlation of severe braking events with time and location,&quot; 2015 10th System of Systems Engineering Conference (SoSE), 2015, pp. 187-192, doi:
10.1109/SYSOSE.2015.7151986.
[15] Liu, Y., & Wu, H. (2017, December). Prediction of road traffic congestion based on random forest. In 2017 10th International Symposium on Computational Intelligence and Design (ISCID) (Vol. 2, pp. 361-364). IEEE.

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING