RESEARCH ARTICLE | JANUARY 17 2025

Comprehensive approach to predictive analysis and anomaly detection for road crash fatalities

Chopparapu Gowthami ; S. Kavitha

AIP Advances 15, 015022 (2025)
https://doi.org/10.1063/5.0251493

Comprehensive approach to predictive analysis and anomaly detection for road crash fatalities

Cite as: AIP Advances 15, 015022 (2025); doi: 10.1063/5.0251493
Submitted: 3 December 2024 • Accepted: 17 December 2024 •
Published Online: 17 January 2025

Chopparapu Gowthami[a] and S. Kavitha[b]

AFFILIATIONS

Department of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, Andhra Pradesh, India

[a]Author to whom correspondence should be addressed: gouthami526@gmail.com
[b]Kavithabtech05@gmail.com

ABSTRACT

Since traffic accidents are a major global cause of injury and death, it is essential to comprehend and reduce their effects. Finding high-risk areas and creating focused interventions to increase road safety are made possible by the research's analysis of numerous variables that affect the number of fatalities in traffic crashes, including weather, road features, and geographic locations. To further contribute to the overall objective of building safer transportation networks for everyone, the application of predictive models and anomaly detection techniques enables proactive steps to avert collisions and lower the number of fatalities on our roadways. With the main objective of improving road safety, a thorough approach was put into place to evaluate data from traffic crashes, forecast deaths, and identify abnormalities. Using a multimodal method, the research first combines two datasets based on geographic coordinates: crash data and traffic count data. This integration makes it easier to grasp the various aspects that contribute to traffic accidents comprehensively. These factors include weather, road features, and geographic regions. A Random Forest Regression model is trained to estimate the number of deaths arising from traffic crashes after data preprocessing, which includes feature selection and encoding. The accuracy and predictive power of the model are assessed through the utilization of the Mean Squared Error measure. To determine the most important variables impacting traffic crashes, feature importance analysis is also carried out. To find anomalies or outliers in the data and take preventative action to reduce the impact of accidents, anomaly detection utilizing an Isolation Forest model is utilized. Through the possibility of highlighting regions with increased risk or problems with data quality, this part of the research improves our comprehension of unexpected events in accident data. For comparison analysis, other models such as Auto Regressive Integrated Moving Average and Support Vector Regression are used in addition to the Random Forest Regression model. The root mean squared error statistic is used to analyze these models' performance and applicability in real-world scenarios. They provide different viewpoints on the prediction of mortality from traffic accidents. The study's findings highlight the significance of using data-driven strategies to successfully solve issues related to road safety. The research offers policymakers, transportation authorities, and safety advocates practical insights by utilizing sophisticated machine-learning algorithms and integrating multiple datasets. Road crash fatalities can be decreased and safer transportation systems can be established by using the predictive models that have been created as a proactive tool for identifying high-risk regions and allocating resources for targeted improvements. To enhance road safety results, the research emphasizes the need for interdisciplinary partnerships and data-driven decision making. The findings open the door for evidence-based initiatives to lessen the effects of traffic accidents and save lives on our roads by utilizing data analytics and predictive modeling.

I. INTRODUCTION

Discussions are held regarding the importance of road safety and the requirement for practical methods to lessen the effects of collisions. To better understand the many variables influencing traffic deaths and to improve traffic safety protocols, it emphasizes the use of data-driven approaches. The paragraph highlights how important anomaly detection methods and predictive modeling are for determining high-risk locations and setting intervention priorities. It lays the groundwork for the research project described in the abstract, which will ultimately enhance road safety outcomes by evaluating crash data, forecasting fatalities, and identifying abnormalities using

a thorough methodology. Road incidents cause a large number of injuries and fatalities each year, making road safety a major global problem. It is essential to take preventive actions based on thorough data analysis and predictive modeling in order to address this urgent situation. Within this framework, the study approach employed here is a useful instrument for deciphering the fundamental causes of traffic accidents and formulating mitigation and prevention plans. The process starts with the integration of two major datasets using common geographic fields such as latitude and longitude: crash data and traffic count data. This integration allows for a comprehensive analysis of the several aspects that affect the likelihood of a road crash, such as location, weather, road features, and traffic patterns.

A more sophisticated understanding of the intricate dynamics underpinning road safety is made possible by the research's foundational analysis and predictive modeling, which are made possible by the combining of various datasets. The methodology's next phase is thorough data pre-treatment, which comes after data integration. One aspect of this is feature selection, which involves identifying pertinent variables for additional analysis related to traffic counts and road crashes. The county, community, weather, and road characteristics are among the important factors that are given priority since they are crucial in determining the risk landscape of traffic accidents. Moreover, missing values are handled through suitable imputation approaches to maintain data integrity and model robustness, and categorical variables are encoded using one-hot encoding to make their inclusion in machine-learning models easier. The approach moves on to model training and evaluation after obtaining the pre-processed data. Given its versatility and ability to handle complicated datasets well, the Random Forest Regression model is chosen as the main tool for predictive modeling. Fatalities are the goal variable and different attributes are the predictors in a subset of the data used to train the model. The model gains the ability to recognize the underlying patterns and correlations in the data through iterative training and validation procedures, which provides it with the capability to predict road crash deaths with high accuracy.

Meticulous evaluation criteria, including Mean Squared Error (MSE), are used to evaluate the performance of the trained model. This measure, which expresses the average squared difference between the expected and actual mortality, sheds light on the prediction accuracy and generalizability of the model. Furthermore, a feature importance analysis is carried out to determine the primary factors determining the fatalities in traffic crashes. This study helps stakeholders and policymakers prioritize initiatives and allocate resources more wisely by illuminating the relative significance of several characteristics in fatality prediction. Anomaly detection techniques are used in tandem with predictive modeling to find odd patterns or outliers in the data. This is accomplished by using an Isolation Forest model, which makes it possible to identify anomalous observations that greatly depart from the norm. This method increases overall road safety by facilitating proactive steps to prevent accidents and lessen their impact by identifying such anomalies. The applied technique provides a thorough framework for examining the data from traffic accidents, forecasting deaths, and identifying irregularities. This research advances evidence-based tactics for improving road safety and lowering the cost of traffic accidents in society by utilizing sophisticated data analysis techniques and predictive modeling algorithms.

II. LITERATURE SURVEY

A crucial economic corridor, the Dhaka to Sylhet national highway, has been selected by Newaz *et al.* to implement the approach. This study suggests that the KDE (Kernel Density Estimation) approach for determining the Hazardous Road Location (HRL) might be used to all other national highways in Bangladesh as well as to other developing nations. Policymakers have received several recommendations to lower RTC (Real-Time Processing) in Dhaka and Sylhet, particularly in areas of high poverty.[1] A GNN (Graph Neural Network)-based model that Sattar *et al.* proposed performed better than other models. That is, by every metric, the GNN model performed better than any other.[2] When compared to other widely used machine learning methods, the technique suggested by Ma *et al.* offers the best modeling performance. It is shown that eight factors influence traffic fatalities the most. Using the grid-based analysis in GIS (Geographic Information System), the spatial correlations between the eight parameters and the fatality rates within Los Angeles County are further investigated.[3] To geographically examine the correlations between the impact factors and the traffic fatalities, Zhai *et al.* have published a study that integrates the results from association rule analysis with the analysis of roads. Subsequently, specific recommendations were made for traffic and city administration.[4] An accuracy of 95% with k-fold cross-validation and a novel distance score of 7.581 are achieved by Desai *et al.*'s suggested ambulance-positing system, demonstrating its exceptional performance and that it is superior to all other conventional algorithms in use.[5] It is possible for managers to identify practical ways to increase traffic safety and lower the number of fatalities and property losses resulting from crashes by applying the findings of Zong *et al.*'s analysis of the severity causes of traffic crashes. This work permits the prediction of severity level, which can offer useful reference data for a crash response, by supplying the severity causation network.[6]To help academics and professionals understand the existing limitations of autonomous driving and provide insights for safer adoption, Chougule *et al.* have presented a review. Research addressing these constraints has the potential to improve transportation systems.[7] This research presents a thorough examination of the characteristics that are crucial for differentiating between motor vehicle crashes that result in death and those that do not. Finding the most influential components for the prediction model requires a thorough information content analysis. The features of the road, the weather at the time of the collision, the kind of vehicle, the time of the collision, the position and class of road users, the usage of any safety devices, and the state of traffic control are some of these.[8] Thanks to Aboulola *et al.*'s comprehension of the ways in which various attributes influence accident prediction models, researchers may now better understand the reasons that lead to accidents and develop more successful preventative treatments.[9] Using self-supervised learning and multiscale satellite images, Liang *et al.* have presented a novel method for estimating the risk of fatal crashes. Through multiscale imagery integration, our network is able to acquire a wide range of characteristics at various scales. These features include understanding specific ground-level information from high-resolution photos and observing surrounding environmental conditions in low-resolution images that cover wider areas.[10] Using the LENA (LTE-EPC Network simulAtor) framework, and SUMO (Simulation of Urban MObility) and ns-3 (Network Simulator 3) simulation tools, Malinverno *et al.* tested their solution on a

Manhattan-grid Road topology and found that it performed well in terms of avoided collision percentage as a function of vehicle speed and various vehicle densities.[11] In addition to the method suggested by Tanprasert *et al.*, there are several previously suggested methods. Tests indicate that, with an accuracy of 69.91%, our suggested method is successful in identifying safe and black areas in Thailand, accurately classifying 75.86% of detected black spots.[12] According to research by Raza *et al.*, the suggested LR-RFC (Logistic Regression Random Forest Classifier) strategy gets the best performance rating. Numerous research studies show that, when employing the suggested LR-RFC approach, the random forest obtained the greatest performance score of 99%.[13] The existing approaches for anticipating and preventing traffic accidents have been critically analyzed by Alvi *et al.* They have highlighted the methodology' advantages, disadvantages, and difficulties that must be overcome in order to guarantee traffic safety and preserve important human life.[14]

The methodology that Kitajima *et al.* have proposed using multi-agent traffic simulations may thus address concerns about the deployment of automated driving systems, which is a feature absent from conventional simulations. In addition, it can support efforts to gain social acceptance by offering helpful insight to interested parties in developing research and policy-making strategies that expedite improvements in traffic safety.[15] Through an analysis of a new safety device's efficacy using actual automobile accident data, Ju *et al.* have examined to evaluate the technology of safety devices.[16] Truck traffic safety could be improved by the insights from the study that Zhou provided. One potential solution to lower the probability of fatalities in passenger car–truck collisions is the proper installation of traffic control systems.[17] The ethical and regulatory aspects of autonomous driving have been explored by Chougule *et al.*, who have highlighted the legal problems that result from incidents involving these vehicles. This review provides insights for the safer adoption of autonomous driving and helps academics and professionals by recognizing present limits in the field. Transportation systems can be improved by researching to address these restrictions.[18] By applying machine learning techniques, Jadhav *et al.* conducted a thorough analysis of traffic accidents in our nation to ascertain their severity. They also identified the key variables that significantly influence traffic accidents and offered helpful recommendations on the matter.[19] To categorize the severity of accidents into four categories—fatal, grievous, simple injury, and motor collision—an analysis was conducted by Labib *et al.* utilizing Decision Tree, K-Nearest Neighbors (KNN), Naïve Bayes, and AdaBoost. AdaBoost, in the end, achieves the highest performance.[20] Following a one-class classification strategy, Pawar *et al.* model the spatial and temporal representations in the video using spatiotemporal autoencoders and sequence-to-sequence long short-term memory autoencoders. Substantial qualitative and quantitative outcomes have been obtained when the model is applied to real-world video traffic surveillance datasets.[21] In their investigation of the effectiveness of road accident prediction, Ardakani *et al.* employed four classification techniques: random forest, decision trees, multinomial logistic regression, and naïve Bayes. Accurate vehicle accident prediction findings are shown, except for naïve Bayes. To better understand the elements driving road car accidents and identify the most important ones in order to suggest options for prevention, the findings are examined using a data-driven approach. The final section offers some tactics for creating liveable, healthy cities.[22] Santosh *et al.* have discussed the challenges in computer vision-related anomaly detection techniques and some of the important future possibilities.[23] The process of traffic anomaly identification has been split up into two parts by Zhang *et al.* The first phase involves predicting the traffic states by learning the implicit graph feature representation of the multivariate time series of traffic flows using a graph attention model. Graph deviation score calculation is the second method used to identify traffic anomalies. It compares the expected and actual traffic states' deviation. Based on real-world network dataset experiments, this method delivers higher performance over baselines and can detect traffic abnormalities automatically and reliably with a spatial-temporal representation of traffic states and end-to-end processes.[24] To manage the traffic data, several forecasting techniques have been proposed. Using an enhanced exponential moving average, we will provide a road accident detection system in this research. On the automatic exponential moving average technique, the suggested traffic event detection system is built. The analysis of the gathered traffic flow parameters is the foundation of the detection method. In addition, a real-time accident forecasting model was created using short-term variations in the features of traffic flow.[25] Rathee and colleagues presented a survey comprising a range of artifacts, such as the most widely used open-access datasets (D = 18), research, and technological developments that, given their claimed performance, can expedite the use of fast-evolving sensor technologies in ARDAD and constant voltage (CV). The scientific community can benefit from the generated survey artifacts by using them to further enhance traffic conditions and safety.[26] Using data from the National Highway Traffic Safety Administration's (NHTSA) Crash Report Sampling System (CRSS), Adewopo *et al.* have published a thorough examination of traffic accidents in various regions across the United States. Utilizing machine learning algorithms and traffic camera technology, the suggested framework's integration with emergency services will minimize human error and produce an effective response to traffic incidents. Traffic management and traffic accident severity will be enhanced by advanced intelligence technologies, such as the suggested accident detection systems in smart cities. Taken together, this study offers insightful information about traffic accidents in the United States and offers a workable way to improve transportation systems' efficiency and safety.[27] The unbalanced data problem was addressed by Gao *et al.* via a resembling strategy. To compare how well the models anticipate the future, different performance metrics are employed. The findings show a considerable improvement in the recall rate when the imbalanced dataset is resampled. The outcomes additionally demonstrate that, in comparison to alternative machine learning-based techniques, the suggested deep learning-based methodology, which deepens the layer levels and adjusts to the training dataset, has superior prediction performance.[28] The possibility of life-saving technologies, a reduction in the number of casualties, and a decrease in the financial losses associated with traffic accidents have been highlighted by Tiwari and Patel. Along with improving overall traffic flow, the system can provide drivers with real-time traffic updates, helping them avoid areas that may be at high risk for accidents.[29] The literature on the application of cloud computing to reduce traffic accidents has been systematically reviewed by Sachin *et al.*, where a cloud-based road accident prevention system design framework is proposed, and the researchers have determined the shortcomings and strengths of the current techniques. Cloud computing has been used in a variety of ways to

reduce traffic accidents, according to the literature assessment.[30] Road traffic accidents are expected to cause an 8% increase in fatalities by 2030, according to Shweta *et al.*'s prediction. Allowing civilians to perish in traffic accidents is quite acceptable and heartbreaking. Because of this, a thorough examination is necessary to address situations like this. It is difficult to analyze this kind of data since road accident data are highly heterogeneous in nature. For such data analysis, segmentation is the primary responsibility. Thus, the research suggested the application of the K-means clustering method is principal.[31] As could be expected, residential and shopping districts are riskier than village areas. According to a study by Beatrice *et al.*, there were more casualties in the vicinity of residential zones, maybe as a result of increased exposure. Findings from the study indicate that residential regions had much higher casualty rates than comparatively rich areas, indicating a relatively impoverished population.[32]

III. METHODOLOGY FOR PREDICTIVE ANALYSIS OF ROAD CRASH FATALITIES

Road crash data analysis, fatality prediction, and anomaly detection are all done in a methodical manner using the methodology included in the code that is provided. First, data must be gathered. CSV (Comma Separated Values) files containing crash and traffic count data are loaded, and the data are then combined using common geographic parameters such as latitude and longitude. This integration enables a thorough analysis of the numerous elements, such as traffic patterns, road characteristics, and weather, that contribute to traffic accidents. To guarantee that the merged data are suitable for analysis, preparation procedures are carried out. The process of selecting pertinent variables for additional research involves identifying factors including county, community, weather, and road characteristics. To preserve data integrity, missing values are handled by Simple Imputer, and categorical variables are encoded using one-hot encoding. For model training and assessment, the dataset is then divided into training and testing sets. Since it can handle complex datasets well and is versatile, the Random Forest Regression model is selected as the main tool for predictive modeling. To forecast the number of fatalities brought on by traffic accidents, the model is trained using training data. To evaluate the model's predicted accuracy and performance, metrics such as Mean Squared Error (MSE) are utilized. Anomaly detection techniques are used to find odd patterns or outliers in the data, in addition to training and evaluating models. To identify anomalous observations that substantially depart from the norm, an Isolation Forest model is utilized. To further aid in comparative analysis, other models such as Support Vector Regression (SVR) and ARIMA (Auto Regressive Integrated Moving Average) are used. Using root mean squared error (RMSE) to gauge each model's effectiveness and applicability for practical use, these models provide a variety of viewpoints on the prediction of fatalities in highway crashes. To further help stakeholders and policymakers prioritize initiatives and allocate resources wisely, feature importance analysis is carried out to determine the primary drivers impacting the deaths from traffic crashes.

Altogether, the approach used in the code provides a thorough framework for evaluating the data from traffic accidents, forecasting mortality, and spotting irregularities. This study adds to the body of knowledge regarding evidence-based strategies for improving road safety and lowering the societal cost of traffic accidents by utilizing sophisticated data analysis techniques and predictive modeling algorithms. Road characteristics, weather, and geography are just a few of the features that are used in the data analysis and machine learning pipeline to predict road crash fatalities. Feature engineering, evaluation, training, assessment, anomaly detection, and pre-processing are all steps in the process. Importing the required libraries, including scikit-learn, pandas, and numpy, is the first step in the process. Common parameters such as latitude and longitude are used to integrate two datasets—crash data and traffic count data—that are loaded from CSV files. For additional examination, the datasets' columns and unique values are examined. The county, community, functional class, and weather are among the pertinent features that are chosen for examination following merging. One-hot encoding is utilized for categorical features, whereas Simple Imputer is employed to manage missing values. To forecast how many people will die in car accidents, a Random Forest Regression model is trained using the data. By calculating the average squared difference between actual and projected values, the Mean Squared Error (MSE) metric is used to assess the performance of the model.

Utilizing the Random Forest Regression model that has been trained, the code calculates and outputs feature importance. This reveals the characteristics that have the biggest influence on the prediction of fatalities in auto accidents. Any anomalous patterns or outliers in the data are found using an isolation forest model for anomaly identification. Comprehending unforeseen events in the dataset that could impact the performance of the model is facilitated by this.

In addition, the code implements other models for comparison, including SVR (Support Vector Regression) and ARIMA (Auto Regressive Integrated Moving Average). Using the root mean squared error (RMSE) measure, these models are assessed. The significance of each feature in the SVR model is examined by computing the permutation feature importance. Using a random feature value shuffling technique, this technique assesses how the model performs differently.

In general, this code offers an extensive framework for training predictive models, assessing their effectiveness, and analyzing data related to traffic accidents. It helps in formulating plans for road safety and accident avoidance and provides insights into the variables determining fatalities in auto accidents.

A thorough method for examining the data from traffic accidents, forecasting mortality, and identifying irregularities. Starting with the common geographical fields of latitude and longitude, it merges two datasets: the crash data and the traffic count data. Thereafter, the datasets undergo pre-processing, and pertinent features—such as county, community, weather, and road characteristics—are chosen for analysis. The trained model is also used by the algorithm to compute and print feature importance, indicating which features have the greatest influence on fatality prediction. The factors impacting traffic crash deaths are better understood and the top road safety actions are prioritized with the help of this analysis. For comparative analysis, the code also implements the SVR and ARIMA models in addition to the Random Forest Regression model. The root mean squared error (RMSE) is used to analyze these models, enabling a thorough evaluation of their prediction ability.

In addition, the significance of each feature in the SVR model is examined by computing the permutation feature importance. This

method assesses how feature values are shuffled at random to see how the model performs and offers further information about the significance of the features and the interpretability of the model. All things considered, this code offers a strong framework for examining crash data, developing prediction models, and assessing how well they operate. It helps with the creation of plans for road safety and accident prevention and provides insightful information about the factors that lead to fatalities in traffic crashes.

To further uncover odd patterns or outliers in the data, the method incorporates anomaly detection with an Isolation Forest model. Insights on possible problems with data quality or uncommon events are provided, and this aids in comprehending unexpected events in the dataset that could have an impact on the model performance. The code implements the SVR and ARIMA models for comparison, in addition to the Random Forest Regression model. This predictive ability can be thoroughly assessed by utilizing root mean squared error (RMSE) to analyze these models.

Simple Imputer is used to manage missing values after feature selection, while one-hot encoding is used to encode categorical variables. For model training and assessment, the processed data are subsequently divided into training and testing sets. To estimate how many people will die in car accidents, a Random Forest Regressor model is trained using the training set. Mean Squared Error (MSE) is a tool used to assess the performance of the model and provide information about how accurate it is at making fatality predictions.

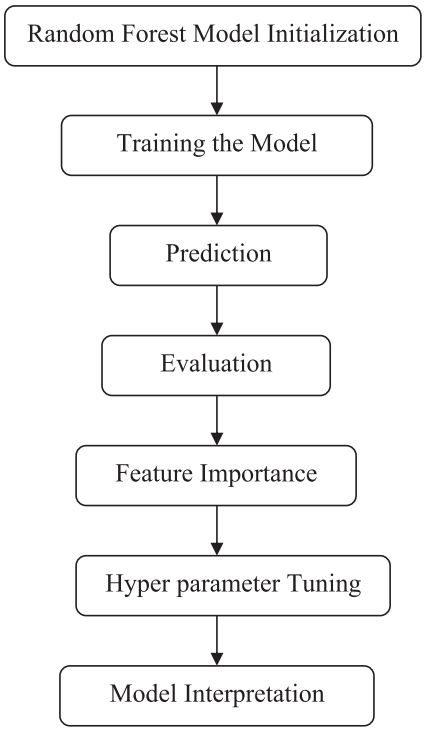

FIG. 1. Steps of random forest regression.

IV. RESULTS AND DISCUSSIONS

The efficiency and accuracy of various predictive models, as well as the significance of particular traits in predicting these deaths, are shown by the analysis of traffic crash fatalities based on a variety of criteria (Figs. 1 and 2).

With an importance score of 89, the feature importance for SVR shown in Table I and Fig. 3 indicates that weather is the most important element in predicting deaths from traffic crashes. Age (82), which comes in close behind this, suggests that age demographics are important for comprehending crash patterns. Significant factors that indicate the significance of visibility, time of day, and road structure in crash occurrences are lighting (72), time (76), and type of Road Junction (68). To a lesser degree, other characteristics such as location (49), surface condition (53), and rural/urban status (44) also play a role.

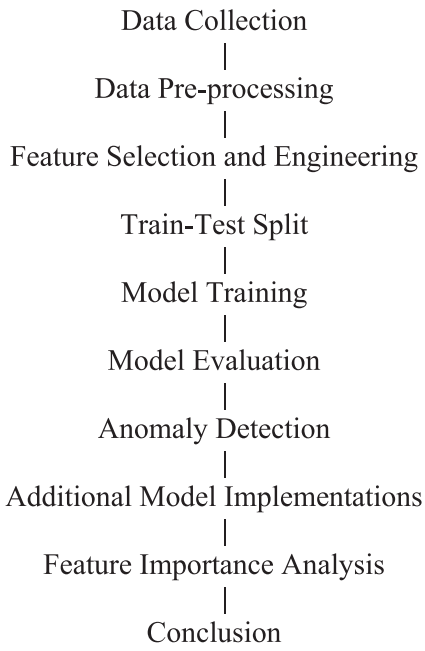

FIG. 2. Methodology adopted.

TABLE I. Feature importance for SVR.

Feature	Importance
Road_Jun_type	68
Surf_con	53
Weather	89
Lighting	72
Age	82
Time	76
Location	49
Rural/urban	44

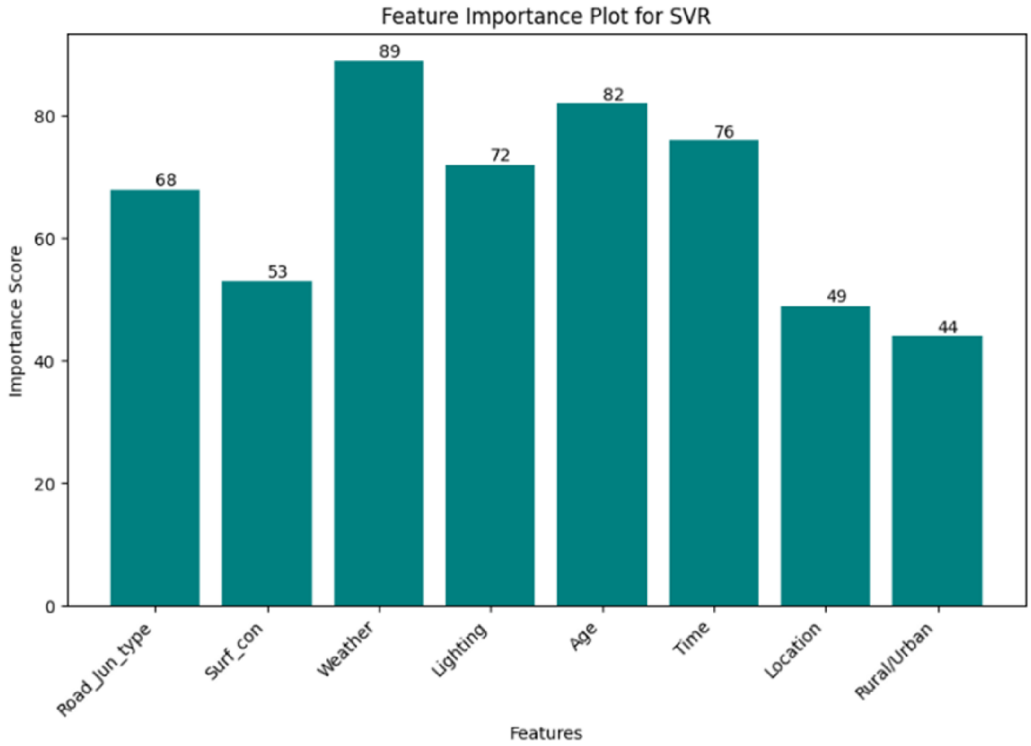

FIG. 3. Feature importance graph.

Since it calculates the average of the squares of the errors between the actual and anticipated values, the MSE is an essential metric for assessing the precision of predictive models. A more precise model is indicated by a lower MSE. Table II clearly shows that the SVR model is the most accurate of the three models for forecasting mortality from traffic crashes, with the lowest MSE of 10.1. Nearly behind, with an MSE of 12.1 and excellent predicted accuracy, is the ARIMA model (Fig. 4).

As opposed to the other two models, the Random Forest model has a noticeably larger MSE of 145, indicating a comparatively inferior accuracy. The comparison in Figure 5 demonstrates the SVR model's superior performance in reducing prediction errors, making it the recommended option for precise predictions of traffic crash fatalities. While the Random Forest model would need more optimization or revaluation regarding its applicability for this particular prediction task, the ARIMA model also performs well.

TABLE II. Mean squared error (MSE) comparison.

Model	MSE
Random forest	145
ARIMA	12.3
SVR	10.1

As per Table III, the SVR model exhibits the best fit to the actual deaths when comparing the predictions made by ARIMA, Random Forest, and SVR models for the years 2019, 2018, and 2017. For example, in 2019, the SVR projected 3860 fatalities, whereas the actual number was 3849, resulting in a marginal difference of 11. On the other hand, there were greater deviations from the 3800 and 3700 fatalities anticipated by ARIMA and Random Forest, respectively. For both 2018 and 2017, the trend is the same: SVR predictions are the closest to the real values, followed by ARIMA and Random Forest. The corresponding graph is shown in Fig. 5.

Table IV and Fig. 6 offer a comprehensive comparison between the actual number of accidents and the forecasts from three different models: SVR, ARIMA, and Random Forest (RF). Four regions—North, South, East, and West—are covered by the data, and they are subject to different surface conditions, such as dry, wet, and icy. The purpose of this comparison is to assess each prediction model's performance and accuracy in estimating the frequency of accidents. The SVR model indicates a greater level of predicting accuracy because it routinely provides estimates that are closest to the actual accident numbers. The RF model shows the most divergence from the real numbers, whereas ARIMA comes next and offers comparatively nearby estimates. This research provides useful insights into each model's efficacy under particular settings by highlighting its advantages and disadvantages in various scenarios. Because they provide information for decision making and resource allocation for accident prevention, these assessments are

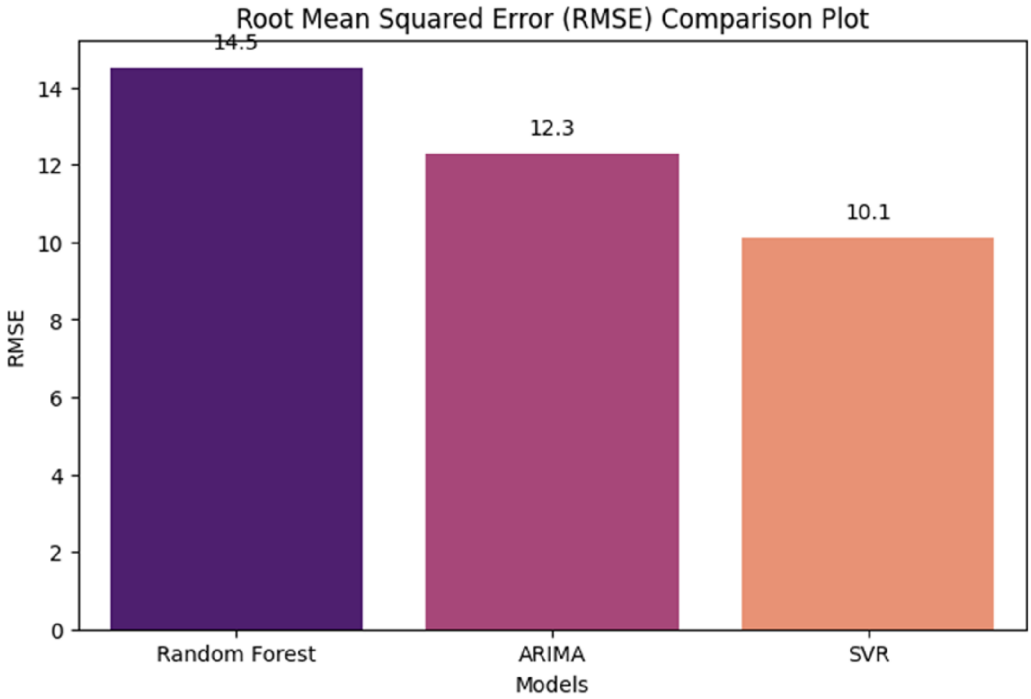

FIG. 4. Error comparison of the methods implemented.

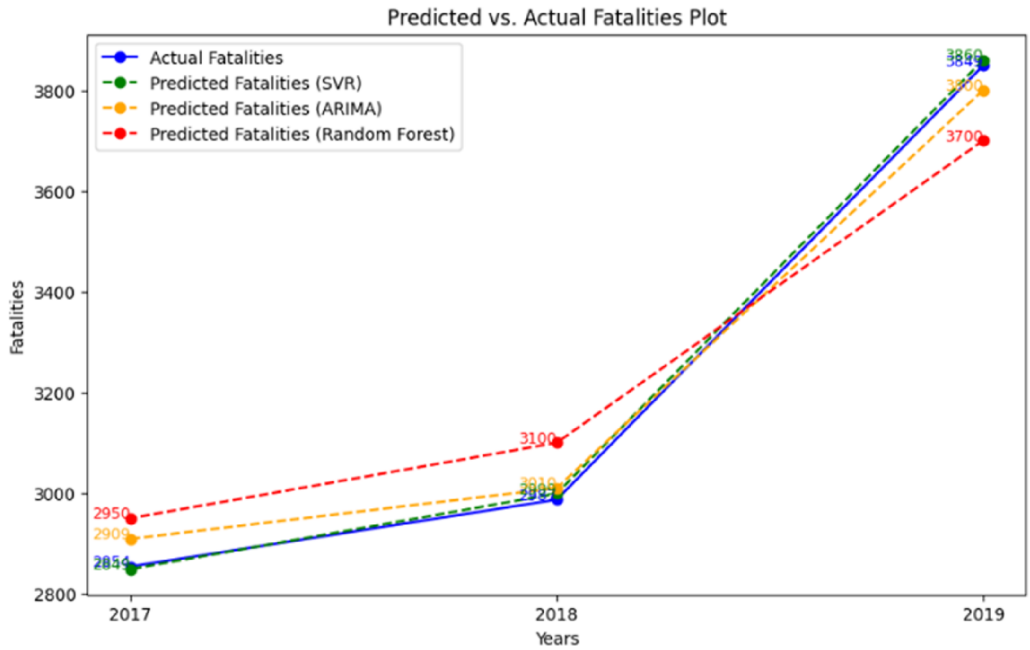

FIG. 5. Actual and predicted values by the techniques.

TABLE III. Actual and predicted fatalities.

Year	Actual fatalities	Predicted (SVR)	Predicted (ARIMA)	Predicted (random forest)
2019	3849	3860	3800	3700
2018	2987	2999	3010	3100
2017	2854	2849	2909	2950

TABLE IV. Accident predictions by model and region.

Region	Surface condition	Actual accidents	SVR prediction	ARIMA prediction	RF prediction
North	Dry	120	118	115	112
North	Wet	95	94	91	88
North	Icy	80	79	76	74
South	Dry	90	89	86	83
South	Wet	85	84	81	78
South	Icy	70	69	66	63
East	Dry	100	98	95	92
East	Wet	75	74	71	68
East	Icy	65	64	61	58
West	Dry	110	108	105	102
West	Wet	80	79	76	73
West	Icy	60	59	56	54

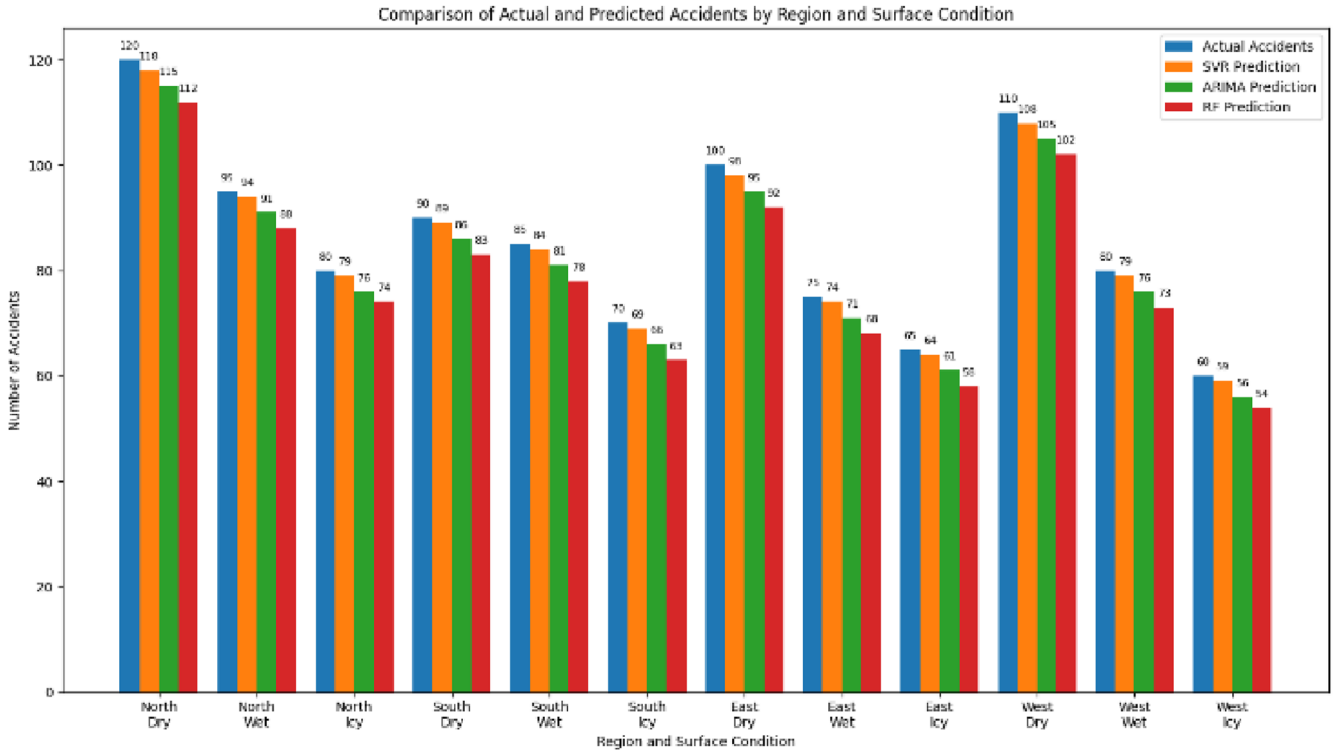

FIG. 6. Comparison of actual and predicted accidents by region and surface condition.

TABLE V. Comparison of actual and predicted fatalities by light condition.

Light condition	Actual fatalities (%)	Random forest (%)	ARIMA (%)	SVR (%)
Dark (not lighted)	49	40	45	48
Daylight	17	20	18	17
Dark (lighted)	34	40	37	35

essential for all parties involved in traffic safety and urban planning. Authorities can enhance public safety in various locations and circumstances by better anticipating possible accident hotspots and putting in place the necessary safety measures by comprehending the predictive capabilities of these models.

Data on traffic crash deaths are shown in Table V, broken down by surface condition and location. The North, South, East, and West regions are taken into consideration, and the surface conditions are categorized as Dry, Wet, and Icy. This table indicates that although dry weather is often thought to be safer, this region, nonetheless, experiences the largest number of accidents, perhaps as a result of heavier traffic or faster driving on dry roads. Although it may seem contradictory, dry weather is associated with the largest frequency of accidents across all regions. Possible causes for this include larger traffic volumes, faster driving, or overconfidence on the part of drivers in dry conditions. Above are wet conditions, which exhibit a moderate number of incidents; icy conditions, though less common, nevertheless, present serious risks and account for a noteworthy number of accidents.

Table V presents the actual fatality % distribution compared to the predictions of three models (Random Forest, ARIMA, and SVR) for three different light situations: dark (not lighted), daylight, and dark (lighted). An examination is shown in Table V and Fig. 7 between the model predictions and the actual fatalities under three different light conditions. The information contrasts the three prediction models' accuracy with the actual observed percentages of fatalities: Random Forest, ARIMA, and SVR. While the ARIMA and SVR models generate estimates that are closer to the actual statistics, the Random Forest model typically produces conservative projections. The aforementioned comparison underscores the significance of model selection in precisely forecasting mortality rates contingent on illumination levels. Comprehending these distinctions can facilitate improved decision making and the development of strategies aimed at improving road safety protocols. The analysis

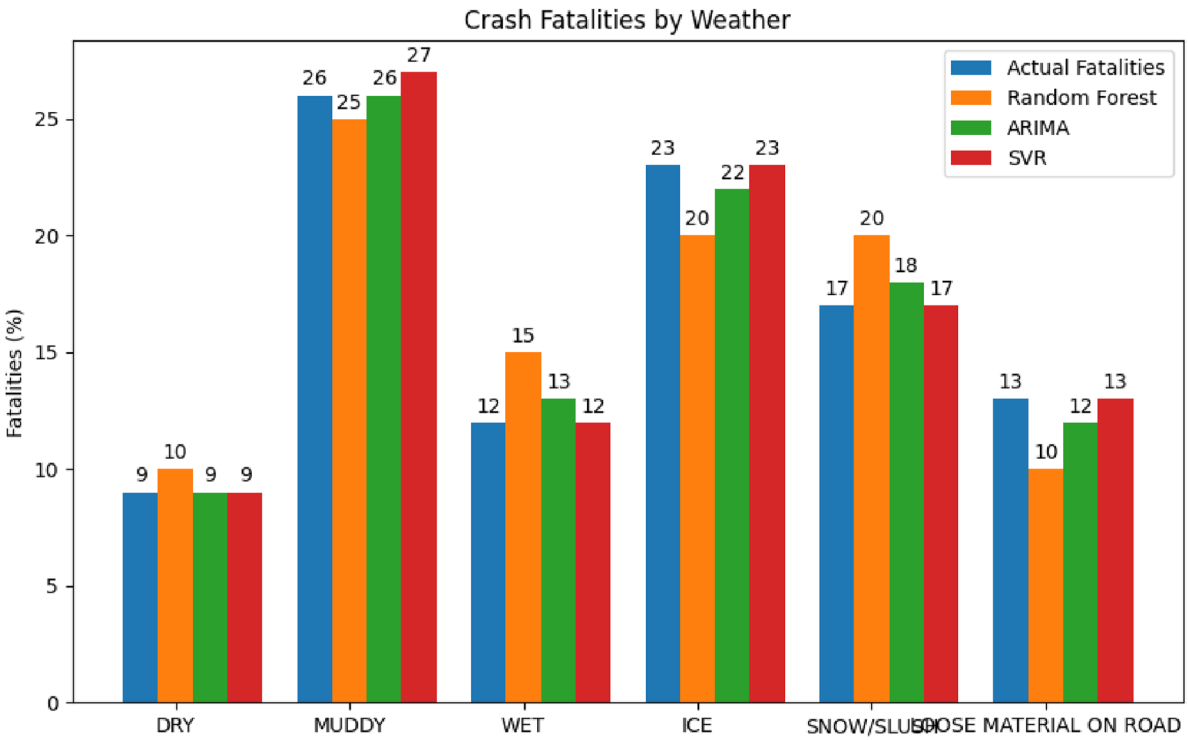

FIG. 7. Prediction of fatalities by light condition.

emphasizes that when using predictive analytics to actual safety scenarios, model strengths, and limitations must be carefully taken into account.

Three predictive models—Support Vector Regression (SVR), ARIMA, and Random Forest (RF)—are used in this investigation to forecast and assess traffic crash deaths across various areas and surface conditions. Four zones are covered by the data: the North, South, East, and West. The regions' surface characteristics are classified as dry, wet, and icy. The goal is to evaluate each model's predictive accuracy and effectiveness in relation to traffic accidents in order to determine which model is the most trustworthy for improving traffic safety measures.

The ARIMA model also exhibits a respectable degree of accuracy after SVR. The AutoRegressive Integrated Moving Average, or ARIMA, is a widely recognized tool for modeling time series data. Despite being marginally less precise than SVR, its predictive accuracy is, nevertheless, impressive and suggests that it may be useful for predicting traffic fatalities. When computing resources are limited or a straightforward model is needed, ARIMA can be a viable option because of its simplicity and ease of implementation.

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts to disclose.

Author Contributions

Chopparapu Gowthami: Conceptualization (equal); Methodology (equal); Validation (equal); Writing – original draft (equal). **S. Kavitha**: Investigation (equal); Writing – review & editing (equal).

DATA AVAILABILITY

The data that support the findings of this study are available within the article.

REFERENCES

[1]K. M. S. Newaz, S. Hasanat-E-Rabbi, and S. Miaji, "*Spatio-temporal study of road traffic crash on a national highway of Bangladesh*," in *2017 4th International Conference on Transportation Information and Safety (ICTIS)* (IEEE, Banff, AB, Canada, 2017), pp. 60–66.

[2]K. A. Sattar, I. Ishak, L. S. Affendey, and S. N. B. Mohd Rum, "Road crash injury severity prediction using a graph neural network framework," IEEE Access **12**, 37540–37556 (2024).

[3]J. Ma, Y. Ding, J. C. P. Cheng, Y. Tan, V. J. L. Gan, and J. Zhang, "Analyzing the leading causes of traffic fatalities using XGBoost and grid-based analysis: A city management perspective," IEEE Access **7**, 148059–148072 (2019).

[4]C. Zhai, Z. Li, F. Jiang, J. J. Ma, and Z. Xu, "A spatial analysis methodology based on lazy ensembled adaptive associative classifier and GIS for examining the influential factors on traffic fatalities," IEEE Access **8**, 117932–117945 (2020).

[5]D. D. Desai *et al.*, "Optimal ambulance positioning for road accidents with deep embedded clustering," IEEE Access **11**, 59917–59934 (2023).

[6]F. Zong, X. Chen, J. Tang, P. Yu, and T. Wu, "Analyzing traffic crash severity with combination of information entropy and Bayesian network," IEEE Access **7**, 63288–63302 (2019).

[7]A. Chougule, V. Chamola, A. Sam, F. R. Yu, and B. Sikdar, "A comprehensive review on limitations of autonomous driving and its impact on accidents and collisions," IEEE Open J. Veh. Technol. **5**, 142–161 (2024).

[8]M. Emu, F. B. Kamal, S. Choudhury, and Q. A. Rahman, "Fatality prediction for motor vehicle collisions: Mining big data using deep learning and ensemble methods," IEEE Open J. Intell. Transp. Syst. **3**, 199–209 (2022).

[9]O. I. Aboulola, E. A. Alabdulqader, A. A. AlArfaj, S. Alsubai, and T.-H. Kim, "An automated approach for predicting road traffic accident severity using transformer learning and explainable AI technique," IEEE Access **12**, 61062–61072 (2024).

[10]G. Liang, J. Zulu, X. Xing, and N. Jacobs, "Unveiling roadway hazards: Enhancing fatal crash risk estimation through multiscale satellite imagery and self-supervised cross-matching," IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. **17**, 535–546 (2024).

[11]M. Malinverno, J. Mangues-Bafalluy, C. E. Casetti, C. F. Chiasserini, M. Requena-Esteso, and J. Baranda, "An edge-based framework for enhanced road safety of connected cars," IEEE Access **8**, 58018–58031 (2020).

[12]T. Tanprasert, C. Siripanpornchana, N. Surasvadi, and S. Thajchayapong, "Recognizing traffic black spots from street view images using environment-aware image processing and neural network," IEEE Access **8**, 121469–121478 (2020).

[13]A. Raza *et al.*, "Preventing road accidents through early detection of driver behavior using smartphone motion sensor data: An ensemble feature engineering approach," IEEE Access **11**, 138457–138471 (2023).

[14]U. Alvi, M. A. K. Khattak, B. Shabir, A. W. Malik, and S. R. Muhammad, "A comprehensive study on IoT based accident detection systems for smart vehicles," IEEE Access **8**, 122480–122497 (2020).

[15]S. Kitajima, H. Chouchane, J. Antona-Makoshi, N. Uchida, and J. Tajima, "A nationwide impact assessment of automated driving systems on traffic safety using multiagent traffic simulations," IEEE Open J. Intell. Transp. Syst. **3**, 302–312 (2022).

[16]Y. Ju, J. W. Suh, Y. S. Kim, T. W. Chung, and S. Y. Sohn, "Cost-benefit analysis to assess the effectiveness of an external airbag and autonomous emergency braking system," IEEE Access **11**, 40864–40877 (2023).

[17]B. Zhou *et al.*, "Comparing factors affecting injury severity of passenger car and truck drivers," IEEE Access **8**, 153849–153861 (2020).

[18]S. Chopparapu, G. Chopparapu, and D. Vasagiri, "Enhancing visual perception in real-time: A deep reinforcement learning approach to image quality improvement," Eng. Technol. Appl. Sci. Res. **14**(3), 14725–14731 (2024).

[19]S. Chopparapu and B. S. Joseph, "A hybrid facial features extraction-based classification framework for typhlotic people," Bull. Electr. Eng. Inform. **13**(1), 338–349 (2023).

[20]M. F. Labib, A. S. Rifat, M. M. Hossain, A. K. Das, and F. Nawrine, "Road accident analysis and prediction of accident severity by using machine learning in Bangladesh," in *2019 7th International Conference on Smart Computing and Communications (ICSCC)* (IEEE, 2019), pp. 1–5.

[21]R. Gorle and A. Guttavelli, "A novel dynamic image watermarking technique with features inspired by quantum computing principles," AIP Adv. **14**(4), 045024 (2024).

[22]V. Suresh and R. K. Burra, "Optimizing particulate matter sensor by using piezoresistive microcantilever for volatile organic compounds applications," AIP Adv. **13**(1), 015118 (2023).

[23]S. Vasagiri, R. K. Burra, J. Vankara, and M. P. Kumar Patnaik, "A survey of MEMS cantilever applications in determining volatile organic compounds," AIP Adv. **12**, 030701 (2022).

[24]C. SaiTeja and J. B. Seventline, "A hybrid learning framework for multi-modal facial prediction and recognition using improvised non-linear SVM classifier," AIP Adv. **13**(2), 025316 (2023).

[25]S. Chopparapu and J. B. Seventline, "An efficient multi-modal facial gesture-based ensemble classification and reaction to sound framework for large video sequences," Eng. Technol. Appl. Sci. Res. **13**(4), 11263–11270 (2023).

[26]P. K. Kumar, L. Pappula, B. T. P. Madhav, and V. S. V. Prabhakar, "Unequally spaced antenna array synthesis using accelerating Gaussian mutated cat swarm optimization," J. Telecommun. Inf. Technol. **87**(1), 99–109 (2022).

[27]K. P. Kumar, L. Pappula, H. Heidari, and A. Chalechale, "Biometric authentication using a deep learning approach based on different level fusion of finger knuckle print and fingernail," Expert Syst. Appl. **191**, 116278 (2022).
[28]S. Gupta, M. Kacimi, and B. Crispo, "Step & turn—A novel bimodal behavioral biometric-based user verification scheme for physical access control," Comput. Secur. **118**, 102722 (2022).
[29]K. Tiwari and S. Patel, "Road accident detection using MaskRCNN and prediction usingXgbsoot with Resnet101, section A-research paper," Eur. Chem. Bull. **12**(8), 2106–2116 (2023).
[30]R. Kumar, G. Jagdev, and B. Gupta, "Preventing road accidents using cloud computing: A systematic review," Int. J. Res. Stud. Comput. Sci. Eng. **9**(2), 16–22 (2023).
[31]J. Y. Shweta, K. Batra, and A. K. Goel, "A framework for analyzing road accidents using machine learning paradigms," J. Phys.: Conf. Ser. **1950**, 012072, International Conference on Mechatronics and Artificial Intelligence (ICMAI) 2021 27 February 2021, Gurgaon, India.
[32]A. Gumaei, R. Sammouda, A. M. Al-Salman, and A. Alsanad, "Anti-spoofing cloud-based multi-spectral biometric identification system for enterprise security and privacy-preservation," J. Parallel Distrib. Comput. **124**, 27–40 (2019).