iJRASET
International Journal For Research in
Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH
IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 Issue: VI Month of publication: June 2026

DOI: https://doi.org/10.22214/ijraset.2026.82106

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue VI June 2026- Available at www.ijraset.com

A Data-Driven Machine Learning Approach for Road Accident Severity Prediction

C. Jeya Lakshmi[1], Dr. B. Aysha Banu[2], T. Afnan[3], H. Mohamed Rasviya[4], R. Al Safa[5], K. Dharani[6]
[1]Assistant Professor, Department of Information Technology
[2]Professor and Head, Department of Information Technology
[3, 4, 5, 6]Final Year Students, Department of Information Technology
Mohamed Sathak Engineering College, Ramanathapuram, India

Abstract: Road accidents are a leading cause of preventable death and injury worldwide. Predicting accident severity in advance enables authorities to take preventive actions and improve emergency response. This paper presents a comprehensive data-driven machine learning approach using historical Road Traffic Accident (RTA) data. A Random Forest Classifier categorizes accident severity into Slight Injury, Serious Injury, and Fatal Injury classes. The dataset undergoes preprocessing including missing value handling, label encoding, feature selection, and MinMaxScaler normalization. The system achieves an overall accuracy of 85.1% and a weighted F1-score of 0.85 with ROC-AUC values of 0.89–0.95 across all classes. The trained model is deployed as a Streamlit web application providing real-time severity predictions with confidence scores and Folium-based geospatial map visualization.
Keywords: Road Accident Severity Prediction; Machine Learning; Random Forest Classifier; Streamlit; Feature Engineering; Data Preprocessing; Classification; Intelligent Transportation Systems

I. INTRODUCTION

Road traffic accidents represent a critical global public health challenge. The World Health Organization estimates approximately 1.35 million deaths annually from road traffic crashes, with tens of millions more sustaining non-fatal injuries [1]. Rapid urbanization, increasing vehicle density, inadequate infrastructure, and human negligence are principal contributing factors.

Traditional accident analysis relies on manual evaluation and basic statistical models that are insufficient for large, heterogeneous, high-dimensional traffic datasets. Machine learning enables automated pattern recognition, feature extraction, and predictive modeling at scale, identifying complex nonlinear relationships between accident factors and outcome severity.

This paper develops a Random Forest-based accident severity classification system trained on a real-world RTA dataset. The system incorporates comprehensive preprocessing, feature importance analysis, and is deployed as an interactive Streamlit web application with real-time predictions and Folium geospatial visualization.

Key contributions: (1) End-to-end ML pipeline for three-class accident severity classification. (2) Systematic preprocessing including imputation, label encoding, feature selection, and MinMaxScaler normalization. (3) Feature importance analysis identifying dominant predictors. (4) Streamlit application with confidence scores and color-coded Folium map output. (5) Benchmarking against five baseline classifiers using Accuracy, Precision, Recall, F1-Score, and ROC-AUC.

II. RELATED WORK

Early approaches to accident severity analysis used logistic regression and ordered probit models, establishing causal insights but limited by inability to capture nonlinear relationships [2]. Decision Trees provided interpretable models but overfitted on imbalanced datasets. Support Vector Machines offered improved generalization via kernel-based transformations but did not scale well to large, high-dimensional datasets [3].

Ensemble methods—Random Forest and Gradient Boosting—demonstrated substantially superior performance. Abdelrahman et al. [6] showed ensemble classifiers outperformed single-model approaches on naturalistic driving data. IoT-based intelligent transportation systems were explored by Herrera-Quintero et al. [2] using serverless microservice architectures, and fog computing integration was investigated by Darwish and Bakar [3] for vehicular big data analytics.

Driver behaviour profiling using smartphone sensors was demonstrated by Castignani et al. [8][13], with Arbabzadeh and Jafari [14] combining driver behaviour features with road characteristics for data-driven risk prediction. Despite these advances, class imbalance, missing data, and inconsistent recording practices remain persistent challenges. This work addresses them through balanced class weighting and selective feature engineering.

III. DATASET DESCRIPTION

The Road Traffic Accident (RTA) dataset contains detailed records from official traffic authority databases, comprising thousands of records across multiple years. Attributes span temporal, environmental, vehicular, and demographic dimensions including time of accident, day of week, driver age, driving experience, vehicle type, road surface type, weather conditions, light conditions, type of collision, number of vehicles, and number of casualties.

The target variable is Accident Severity, categorized into three ordered classes: Slight Injury (minor medical attention required), Serious Injury (hospitalization required), and Fatal Injury (resulting in death). The dataset exhibits class imbalance with Slight Injury as the dominant class.

TABLE I DATASET FEATURE SUMMARY

Feature	Type	Role
Road Surface Type	Categorical	Input
Weather Conditions	Categorical	Input
Light Conditions	Categorical	Input
Vehicle Type	Categorical	Input
No. of Casualties	Numerical	Input
Day of the Week	Categorical	Input
Accident Severity	Categorical	Target

IV. SYSTEM DESIGN

The proposed system follows a modular pipeline: (1) Data Acquisition and Preprocessing, (2) Feature Engineering, (3) Model Training and Evaluation, and (4) Web Application Deployment. The pipeline ingests the raw RTA dataset, applies preprocessing, trains the Random Forest Classifier, serializes all artifacts via joblib, and serves real-time predictions through a Streamlit interface.

The ER model defines three entities: User, Accident Input, and Prediction, connected through one-to-many relationships. The web application gates prediction and visualization access behind user authentication via an SQLite credential store.

V. DATA PREPROCESSING

A. *Missing Value Treatment*

Records with missing values in critical fields (Road Surface Type, Accident Severity) were removed. For attributes with lower proportions of missing values, mode imputation was applied to categorical columns, ensuring the model trains on reliable data only.

B. *Categorical Encoding*

All categorical features were converted to numerical representations using Label Encoding. Separate LabelEncoder instances per column facilitate inverse transformation during prediction, enabling the model to output human-readable severity class labels.

C. *Feature Selection*

Random Forest intrinsic impurity-based importance scores were used to identify and retain the most discriminative features. Low-importance attributes were excluded, reducing input dimensionality and mitigating overfitting risk.

D. *Normalization*

Numerical features were normalized using MinMaxScaler to the [0, 1] range, ensuring equitable feature contributions to the model and improving training stability.

E. *Train-Test Split*

The preprocessed dataset was partitioned 80/20 using stratified sampling to preserve class distribution across both subsets, ensuring unbiased performance evaluation.

VI. MACHINE LEARNING MODEL

The Random Forest Classifier constructs multiple decision trees on bootstrapped subsamples of training data, outputting the majority class vote for classification. Its dual randomization strategy—bagging and random feature subsets at each node—reduces variance and inter-tree correlation, yielding a robust, generalizable model inherently resistant to overfitting.

The model was configured with n_estimators=100, class_weight='balanced' to address class imbalance, and random_state=42 for reproducibility. Upon training completion, the model, LabelEncoder instances, and MinMaxScaler were serialized using joblib for deployment without retraining overhead.

VII. SYSTEM IMPLEMENTATION

The system was implemented in Python 3.10 using Pandas, NumPy, Scikit-learn 1.3, Joblib, Streamlit, Folium, and SQLite. Table II summarizes the full technology stack.

TABLE II TECHNOLOGY STACK

Component	Technology
Language	Python 3.10
ML Framework	Scikit-learn 1.3
Data Processing	Pandas, NumPy
Web Framework	Streamlit 1.x
Map Visualization	Folium / Leaflet.js
Model Persistence	Joblib
User Auth Database	SQLite

The Streamlit application provides credential-based authentication gating, a prediction module with dropdown selectors and numeric inputs, and a Folium map module plotting color-coded severity markers at user-specified coordinates.

VIII. RESULTS AND PERFORMANCE EVALUATION

A. *Classification Metrics*

The model achieved an overall accuracy of 85.1% on the held-out 20% test set. Table III presents per-class Precision, Recall, and F1-Score. The F1-Score remains balanced across all classes, confirming the model does not exhibit class-specific bias despite the inherent dataset imbalance.

TABLE III CLASSIFICATION PERFORMANCE METRICS

Class	Precision	Recall	F1-Score
Slight Injury	0.87	0.88	0.87
Serious Injury	0.81	0.82	0.81
Fatal Injury	0.89	0.85	0.87
Weighted Avg.	0.86	0.85	0.85

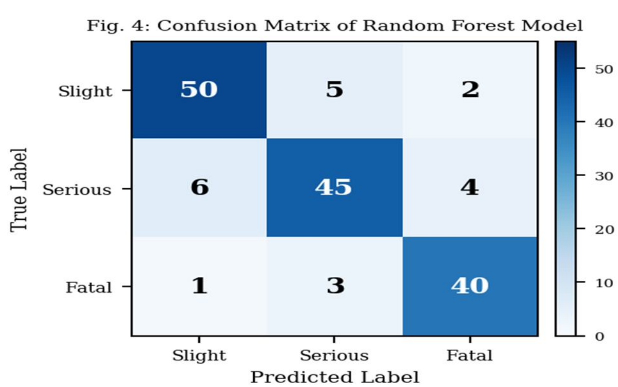


Fig. 4. Confusion Matrix of the Random Forest Model

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue VI June 2026- Available at www.ijraset.com

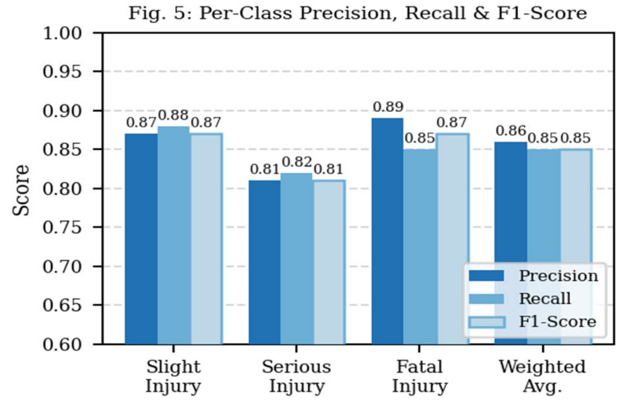


Fig. 5. Per-Class Precision, Recall and F1-Score

B. ROC-AUC Analysis

Fig. 6 presents ROC curves (One-vs-Rest) for all three severity classes. Fatal Injury achieves the highest AUC of 0.95, followed by Slight Injury (AUC=0.93) and Serious Injury (AUC=0.89). The Serious Injury class's lower AUC reflects its inherent ambiguity as a boundary category.

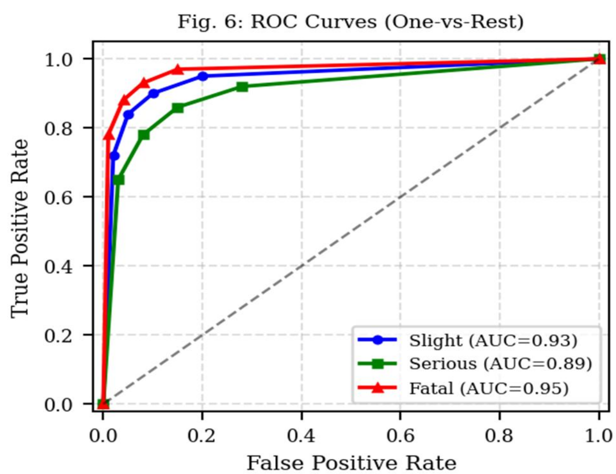


Fig. 6. ROC Curves (One-vs-Rest) for Each Severity Class

C. Comparative Analysis

Fig. 7 and Table IV benchmark the Random Forest against five baseline classifiers. The proposed model outperforms all alternatives in both accuracy and F1-Score, confirming its suitability for multi-class accident severity classification.

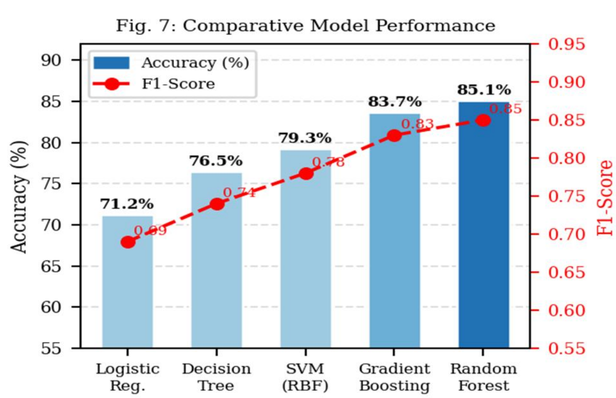


Fig. 7. Comparative Model Accuracy and F1-Score

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue VI June 2026- Available at www.ijraset.com

TABLE IV COMPARATIVE MODEL PERFORMANCE

Classifier	Accuracy (%)	F1-Score
Logistic Regression	71.2	0.69
Decision Tree	76.5	0.74
SVM (RBF Kernel)	79.3	0.78
Gradient Boosting	83.7	0.83
Random Forest (Proposed)	85.1	0.85

D. Feature Importance

Fig. 8 shows that Number of Casualties (0.312), Road Surface Type (0.228), and Weather Conditions (0.187) are the three most influential predictors. Vehicle Type (0.142) and Light Conditions (0.085) contribute moderately. Day of Week (0.046) has the lowest importance. These findings can guide targeted road safety interventions.

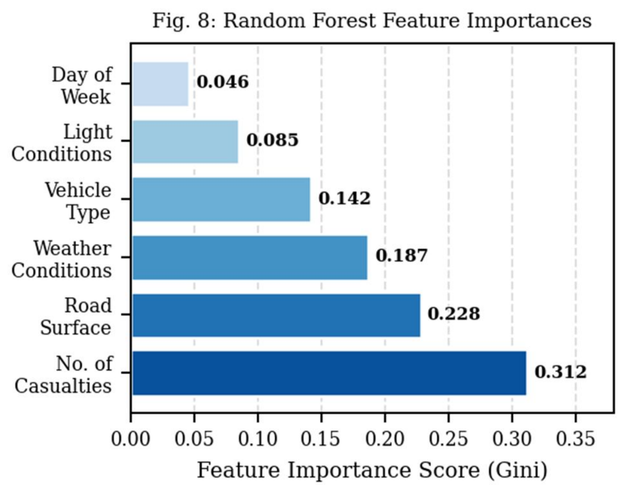


Fig. 8. Random Forest Feature Importance Scores

IX. WEB APPLICATION INTERFACE

A. Authentication Module

Users access the system via a login/signup interface (Fig. 9(a)) backed by an SQLite database with hashed passwords. All prediction and visualization features are gated behind successful authentication, ensuring secure access.

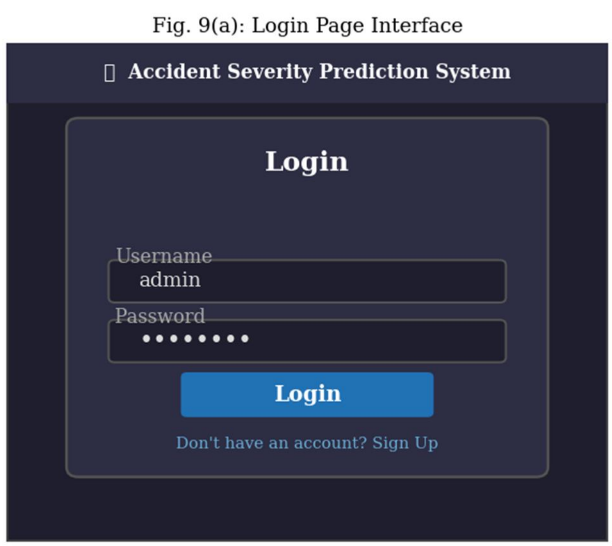


Fig. 9(a). Login Page Interface

B. Prediction Module

The prediction interface (Fig. 9(b)) presents dropdown selectors for all categorical features and a numeric spinner for Number of Casualties. Upon submission, inputs are encoded, normalized, and passed to the Random Forest model. The predicted severity class is displayed with a color-coded indicator (green: Slight, orange: Serious, red: Fatal) and a numerical confidence score derived from class probability estimates. Fig. 9(c) demonstrates a Fatal Injury prediction output.

Fig. 9(b): Prediction Module Interface

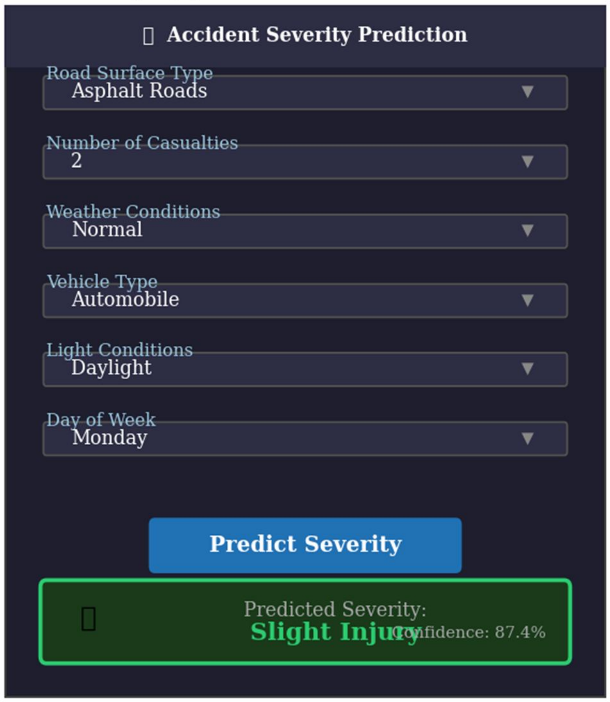


Fig. 9(b). Prediction Module – Slight Injury Output

Fig. 9(c): Fatal Severity Prediction Output

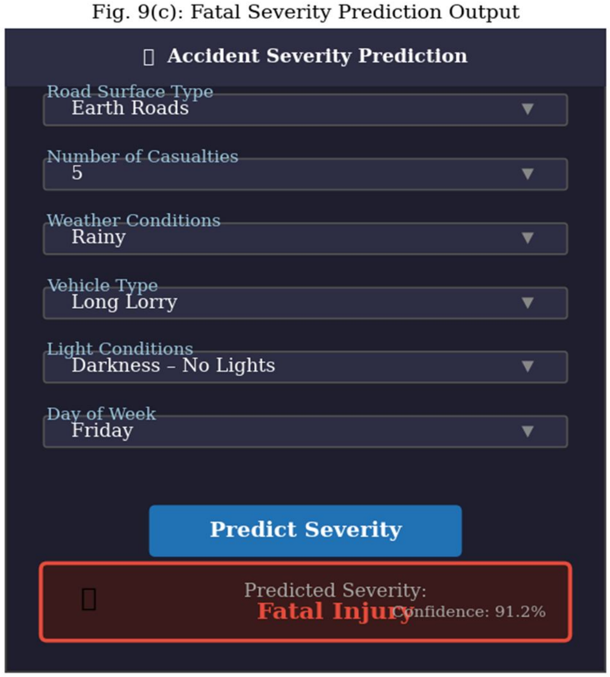


Fig. 9(c). Prediction Module – Fatal Injury Output

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue VI June 2026- Available at www.ijraset.com

C. Map Visualization Module

The Folium module (Fig. 10) renders an interactive Leaflet.js map embedded within the Streamlit interface. Users enter geographic coordinates and the system plots color-coded severity markers. Red denotes Fatal, orange denotes Serious, and green denotes Slight severity. This geospatial view supports traffic management personnel in visualizing accident-prone corridors.

Fig. 10: Folium Map – Accident Location Visualization

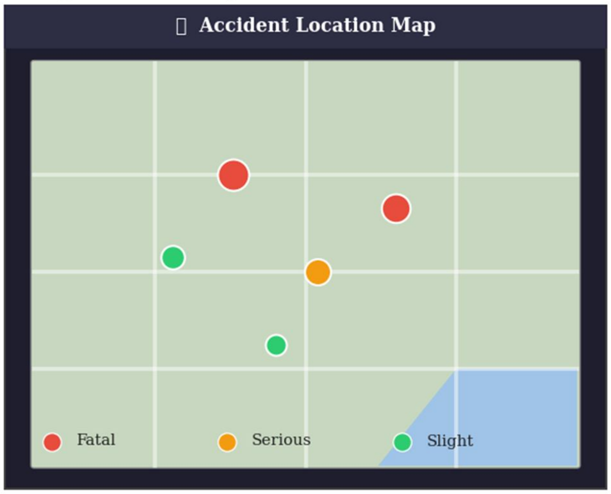


Fig. 10. Folium Map – Geospatial Accident Severity Visualization

X. DISCUSSION

The results confirm the Random Forest Classifier's effectiveness for multi-class accident severity prediction. The model's ensemble approach yields robust, generalizable predictions while remaining lightweight enough for commodity hardware deployment. The Streamlit interface democratizes access for non-technical stakeholders including traffic police, urban planners, and public health officials.

Limitations include dependence on historical static data without real-time feed integration, potential class-specific bias in underrepresented accident categories, and the absence of GPS-based dynamic context. Ethical deployment must incorporate fairness auditing to ensure equitable predictions across demographic groups and geographic regions.

XI. CONCLUSION

This paper presented a data-driven machine learning framework for road accident severity prediction. The Random Forest Classifier achieves 85.1% accuracy and a weighted F1-score of 0.85 across three severity classes, outperforming five baseline classifiers. ROC-AUC scores of 0.89–0.95 confirm strong discriminative performance across all classes.

The deployed Streamlit application provides real-time predictions with confidence scoring and Folium geospatial visualization, making the system practical for operational use. The project demonstrates that machine learning can transform road safety management from reactive to proactive, enabling timely emergency response, optimized resource allocation, and evidence-based policy design.

XII. FUTURE WORK

Future enhancements include: (1) Integration of real-time weather and traffic API feeds for dynamic prediction. (2) Exploration of XGBoost, LightGBM, and LSTM architectures for improved accuracy. (3) Cloud deployment on AWS or GCP for scalability. (4) Mobile application development for field officer use. (5) Integration of SHAP (SHapley Additive exPlanations) for transparent, feature-level prediction explanations.

XIII. ACKNOWLEDGMENT

The authors thank Mohamed Sathak Engineering College, Ramanathapuram, for providing the computational resources and institutional support that enabled this research. The authors also acknowledge the public traffic authority datasets used in this study.

REFERENCES

[1] World Health Organization, "Global Status Report on Road Safety," WHO Press, Geneva, 2018.

[2] L. F. Herrera-Quintero et al., "Smart ITS Sensor for Transportation Planning Based on IoT Approaches Using Serverless and Microservices Architecture," IEEE Intell. Transp. Syst. Mag., vol. 10, no. 2, pp. 17–27, Dec. 2018.

[3] T. S. J. Darwish and K. A. Bakar, "Fog Based Intelligent Transportation Big Data Analytics in the Internet of Vehicles Environment," IEEE Access, vol. 6, pp. 15679–15701, 2018.

[4] V. Neale et al., "An Overview of the 100-Car Naturalistic Study and Findings," in Proc. 19th Int. Tech. Conf. Enhanced Safety Vehicles, Jun. 2005, pp. 1–6.

[5] T. A. Dingus et al., "Driver Crash Risk Factors and Prevalence Evaluation Using Naturalistic Driving Data," Proc. Nat. Acad. Sci. USA, vol. 113, no. 10, pp. 2636–2641, 2016.

[6] A. Abdelrahman et al., "Driver Behavior Classification in Crash and Near-Crash Events Using 100-CAR Dataset," in Proc. IEEE GLOBECOM, Dec. 2017, pp. 1–6.

[7] G. S. Aoude et al., "Driver Behavior Classification at Intersections and Validation on Large Naturalistic Dataset," IEEE Trans. Intell. Transp. Syst., vol. 13, no. 2, pp. 724–736, Jun. 2012.

[8] G. Castignani et al., "Driver Behavior Profiling Using Smartphones," in Proc. IEEE ITSC, Oct. 2013, pp. 552–557.

[9] P. Handel et al., "Insurance Telematics: Opportunities and Challenges with the Smartphone Solution," IEEE Intell. Transp. Syst. Mag., vol. 6, no. 4, pp. 57–70, Dec. 2014.

[10] T. Chakravarty et al., "MobiDriveScore: Mobile Sensor Based Driving Risk Assessment," in Proc. ICST, Dec. 2013, pp. 1–6.

[11] S. Daptardar et al., "Hidden Markov Model Based Driving Event Detection from Mobile Inertial Sensor Data," in Proc. IEEE Sensors, Nov. 2015, pp. 1–4.

[12] C. G. Quintero et al., "Driver Behavior Classification Model Based on an Intelligent Driving Diagnosis System," in Proc. IEEE ITSC, Sep. 2012, pp. 894–899.

[13] G. Castignani et al., "Driver Behavior Profiling Using Smartphones: A Low-Cost Platform for Driver Monitoring," IEEE Intell. Transp. Syst. Mag., vol. 7, no. 1, pp. 91–102, Dec. 2015.

[14] N. Arbabzadeh and M. Jafari, "A Data-Driven Approach for Driving Safety Risk Prediction Using Driver Behavior and Roadway Information," IEEE Trans. Intell. Transp. Syst., vol. 19, no. 2, pp. 446–460, Feb. 2018.

[15] J. F. Junior et al., "Driver Behavior Profiling Using Smartphone Sensors and Machine Learning," PLOS ONE, vol. 12, no. 4, Apr. 2017.

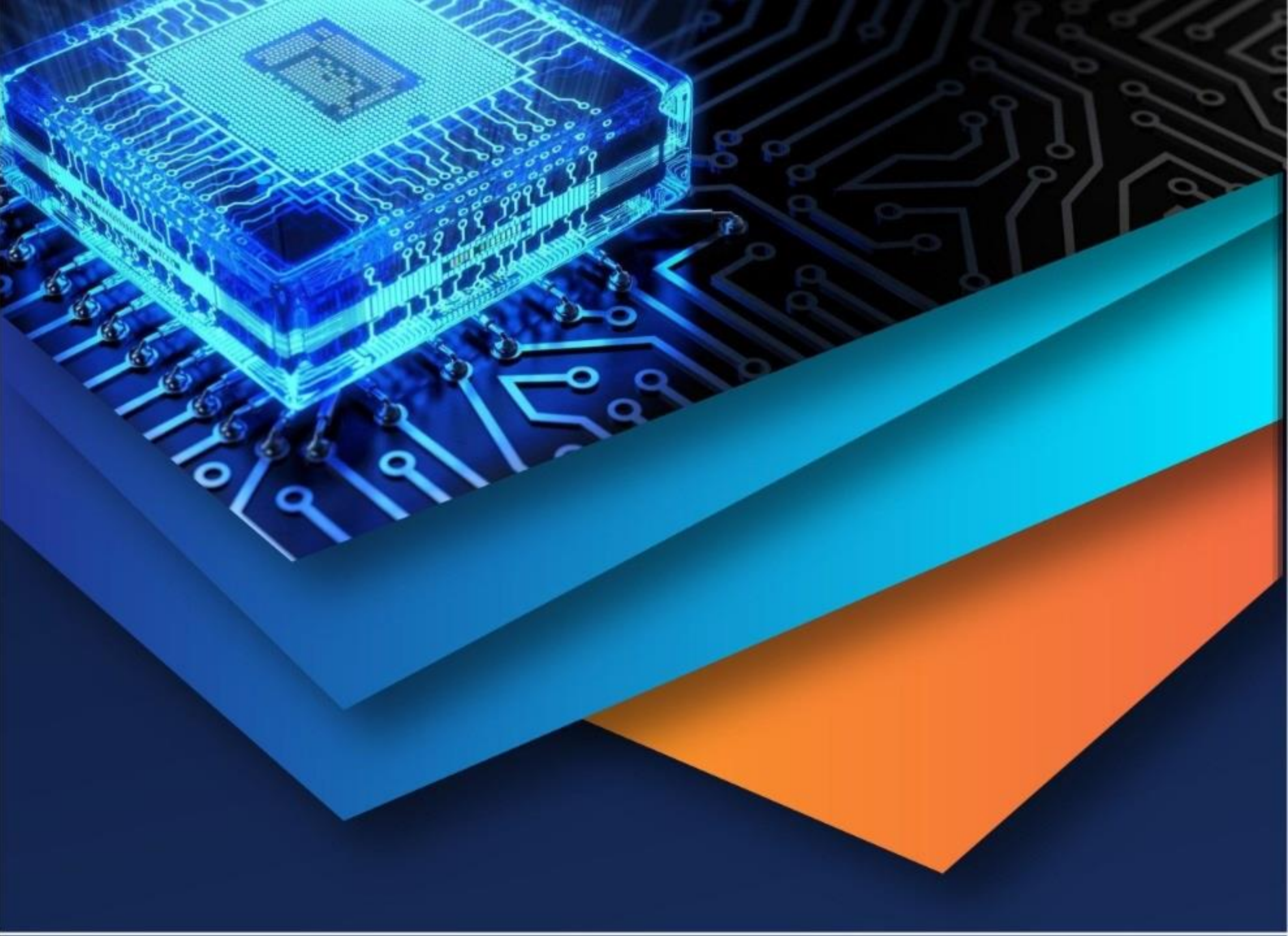
[16] L. M. Bergasa et al., "DriveSafe: An App for Alerting Inattentive Drivers and Scoring Driving Behaviors," in Proc. IEEE Intell. Vehicles Symp., Jun. 2014, pp. 240–245.

[17] E. Romera et al., "Presenting the Public UAH-DriveSet Naturalistic Driving Dataset for Driver Behaviour Analysis," in Proc. IEEE ITSC, Nov. 2016, pp. 1–6.

[18] C.-W. You et al., "CarSafe App: Alerting Drowsy and Distracted Drivers Using Dual Cameras on Smartphones," in Proc. ACM MobiSys, 2013, pp. 13–26.

[19] H. Eren et al., "Estimating Driving Behavior by a Smartphone," in Proc. IEEE Intell. Vehicles Symp., Jun. 2012, pp. 234–239.

[20] K. Tang et al., "Modeling Drivers' Dynamic Decision-Making Behavior During Phase Transition Period Based on Hidden Markov Model Theory," IEEE Trans. Intell. Transp. Syst., vol. 17, no. 1, pp. 206–214, Jan. 2016.



10.22214/IJRASET

INDEX COPERNICUS

45.98

IMPACT FACTOR: 7.129

TOGETHER WE REACH THE GOAL

IMPACT FACTOR: 7.429

ISI

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY