

Weather-Based Road Accident Risk Prediction Using Machine Learning and Explainable Artificial Intelligence

Vaishnavi Raju Chavan¹ and Ms. Yogita kadbane²

¹Student, Master of Computer Applications,

²Assistant Professor, Master of Computer Applications,

Sadhu Vaswani Institute of Management Studies for Girls, Pune¹⁻²,

Savitribai Phule Pune University, Pune, Maharashtra, India¹⁻².

vaishnavichavan2211@gmail.com¹, yogita.raskar@gmail.com²

Abstract: Road accidents are a common safety problem, and weather conditions can affect road and driving conditions. In this study, machine learning is used to predict road accident risk based on weather information. The study uses 12,000 records from the US Accidents dataset collected between April and September 2022. Seven weather factors are considered such as temperature, humidity, pressure, visibility, wind speed, precipitation, and weather conditions. Accident severity is grouped into three risk levels: Low, Medium, and High. Three machine learning models, Logistic Regression, Random Forest, and XGBoost, are developed and compared using Accuracy, Precision, Recall, Weighted F1-score, and Macro F1-score. Among these models, Random Forest gives the highest performance, with 69.38% Accuracy, 65.33% Precision, 69.38% Recall, 65.80% Weighted F1-score, and 62.91% Macro F1-score. Therefore, Random Forest is selected for the final prediction. SHAP is used to understand how different weather factors affect the model's prediction. The results show that pressure, humidity, weather conditions, temperature, and wind speed have higher importance in the model. An interactive prototype is also developed where users can enter weather conditions and get a predicted risk level along with a simple explanation.

Keywords: Road Accident Prediction, Machine Learning, Weather Conditions, Random Forest, XGBoost, Logistic Regression, Explainable AI, SHAP, Accident Risk.

I. INTRODUCTION

Road accidents are a serious transportation safety concern worldwide, and their occurrence can be influenced by several factors, including traffic conditions, road characteristics, vehicle conditions, driver behaviors, geographical location, and environmental factors. Among these, weather conditions can have a considerable effect on driving safety by influencing visibility, road surface conditions, vehicle control, and driver response. Weather conditions such as rainfall, fog, strong winds, reduced visibility, and high humidity can create challenging driving environments and may increase the possibility of accidents. Therefore, analyzing weather conditions for accident risk prediction can provide useful insights for improving road safety and supporting intelligent transportation applications. Traditional statistical methods can be used to examine relationships between variables, but machine learning techniques are better suited to identifying complex and nonlinear patterns among multiple weather-related factors. However, machine learning models, particularly ensemble-based approaches, may be difficult to interpret because their decision-making processes are not always transparent. Explainable Artificial Intelligence (XAI) can address this limitation by providing meaningful explanations for model predictions. In this research, SHAP (SHapley Additive exPlanations) is employed to determine how individual weather factors influence the predicted accident risk. The proposed system focuses specifically on



weather-based accident risk prediction, where weather parameters are provided as input, and the model classifies the accident risk into three categories: Low, Medium, or High. Along with the predicted risk level, SHAP explanations indicate which weather factors contribute toward or away from the prediction, making the model's results easier to understand and interpret.

II. OBJECTIVE OF STUDY

The objective of this study is to analyze how weather conditions affect road accident risk using factors such as temperature, humidity, pressure, visibility, wind speed, precipitation, and weather conditions. It aims to prepare the accident data for machine learning and classify accident risk into Low, Medium, and High categories. The study also aims to develop and compare Logistic Regression, Random Forest, and XGBoost models and evaluate their performance using different classification metrics. In addition, SHAP is used to understand the importance of weather factors and explain individual risk predictions. Finally, an interactive prototype is developed to provide the predicted risk level and its explanation based on the entered weather conditions.

III. LITERATURE REVIEW

Song et al. (2023) [1] investigated accident risk during dense fog conditions by combining weather observations with traffic and road information. Their work indicates the importance of environmental conditions at accident risk although it is limited to fog-related situations.

Sukonna and Swapnil (2023) [2] examined accident duration using weather and road-related variables with Random Forest and LightGBM. They also applied SHAP to interpret the models, demonstrating how explainability can help identify influential factors. However, their objective was accident duration rather than risk-level classification.

Rifat et al. (2024) [3] applied several machine learning techniques with SHAP to analyse traffic accident fatality. Their study showed that explainable models can provide better understanding of the factors behind predictions, but the analysis relied mainly on accident, vehicle, road, and location characteristics.

Aboulola (2024) [4] combined deep learning with SHAP for accident severity prediction. The study demonstrated the usefulness of XAI for interpreting complex models although its focus was on deep-learning-based severity prediction rather than weather-specific risk assessment.

Patel et al. (2026) [5] compared multiple machine learning and deep learning approaches for accident prediction and considered environmental information among other factors. Their work highlights the potential of AI for road safety but does not specifically develop a multi-weather-factor model with SHAP-based risk explanations.

Research Gap: From the reviewed studies the following research gaps are identified:

[1] Weather information is frequently analyzed together with road, traffic, vehicle, driver, and geographical variables. This makes it difficult to examine the contribution of weather conditions independently.[2] Previous studies commonly focus on accident severity, fatality, or duration, whereas comparatively less attention has been given to classifying accident outcomes into Low, Medium, and High-risk categories using weather information.[3] Machine learning predictions may be difficult to interpret, particularly when complex models are used. Therefore, prediction performance alone may not provide sufficient understanding of the factors influencing a model's output. [4] Individual-level explanations are not consistently incorporated into weather-focused accident classification. Users may benefit from knowing which weather conditions influence a particular prediction and the direction of their contribution.[5] Weather-only multi-class classification presents a modelling challenge because weather variables can interact with one another and may not have a simple or uniform relationship with accident outcomes.

Based on these gaps, the present study focuses on weather-based accident-risk classification and integrates SHAP-based explainability to make model predictions more interpretable.



IV. RESEARCH METHODOLOGY

Research Design

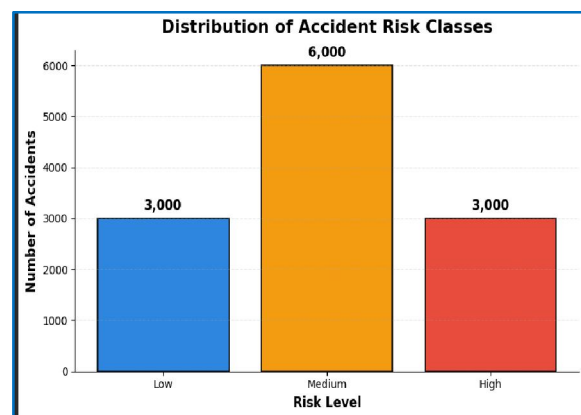
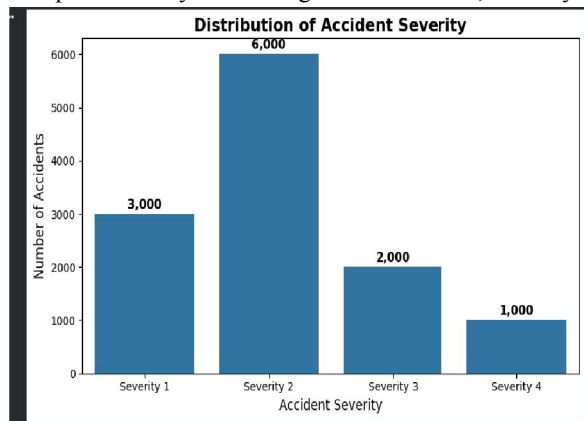
The proposed research methodology is designed to predict road accident risk based on weather conditions using Machine Learning (ML) and Explainable Artificial Intelligence (XAI). The methodology consists of data collection, data selection, preprocessing, risk-level generation, feature transformation, model development, model evaluation, selection of the best model, and SHAP-based explanation.

Data Collection and Feature Selection

The US Accidents dataset was used as the primary data source. A total of 12,000 records from April to September 2022 were selected. Seven weather-related features were used: Temperature, Humidity, Pressure, Visibility, Wind Speed, Precipitation, and Weather Condition. Road, traffic, vehicle, and location-related attributes were excluded to maintain the weather-focused scope, Start_Time was used only for selecting the required time period.

Risk-Level Classification

The original accident severity value was converted into three risk categories to make the model output easier to interpret. Severity 1 was assigned to Low Risk, Severity 2 to Medium Risk, and Severities 3 and 4 to High Risk.



Data Preprocessing

The selected dataset was processed before model training. Missing numerical values were replaced using median imputation, while missing categorical values in Weather_Condition were handled using the most frequent category. Numerical features were standardized using StandardScaler, and the categorical weather condition was converted into numerical form using one-hot encoding. These preprocessing steps were applied consistently during model training and prediction.

Model Training and Machine Learning Models

The prepared dataset was divided into 80% training data and 20% testing data using a stratified split. This helped maintain a similar distribution of the Low, Medium, and High-risk categories in both datasets. A fixed random state was used to support reproducibility. The training data was used to develop the models, while the testing data was used for performance evaluation.

Three supervised machine learning algorithms were developed and compared:

Logistic Regression: Used as a baseline model to estimate the probability of different accident-risk categories from weather features.

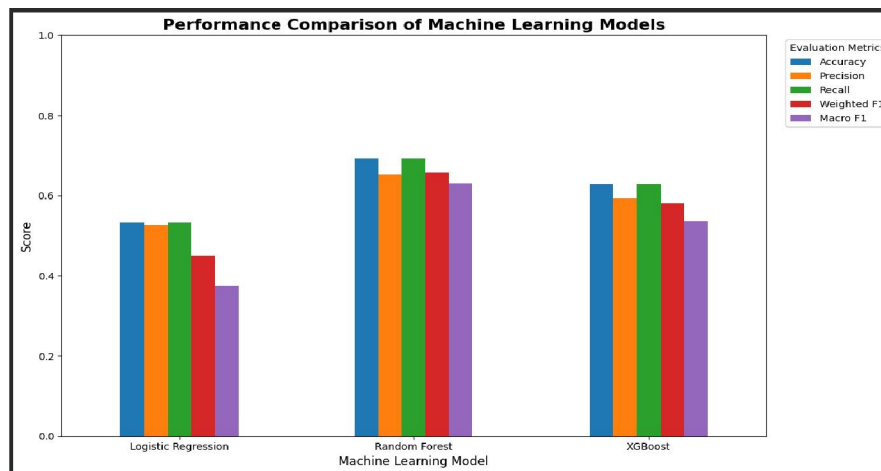
Random Forest: An ensemble algorithm that combines multiple decision trees. It was selected for its ability to capture nonlinear relationships between weather conditions and accident risk. It achieved the best overall performance in this study.



XGBoost: A gradient-boosting algorithm that builds decision trees sequentially to reduce prediction errors. It was included to compare its performance with Logistic Regression and Random Forest.

Model Evaluation

The models were compared using Accuracy, Weighted Precision, Weighted Recall, Weighted F1-score, and Macro F1-score. These metrics were used to evaluate the prediction performance of the models across the three risk categories. Accuracy represents the percentage of correct predictions. Precision measures the reliability of the predicted classes, while Recall shows how effectively the models identified the actual classes. Weighted F1-score accounts for class frequency, whereas Macro F1-score treats Low, Medium, and High risk equally. Confusion matrices were also used to study correct and incorrect classifications for each model.



Experimental Results: The three models produced the following results on the 12,000-record dataset:

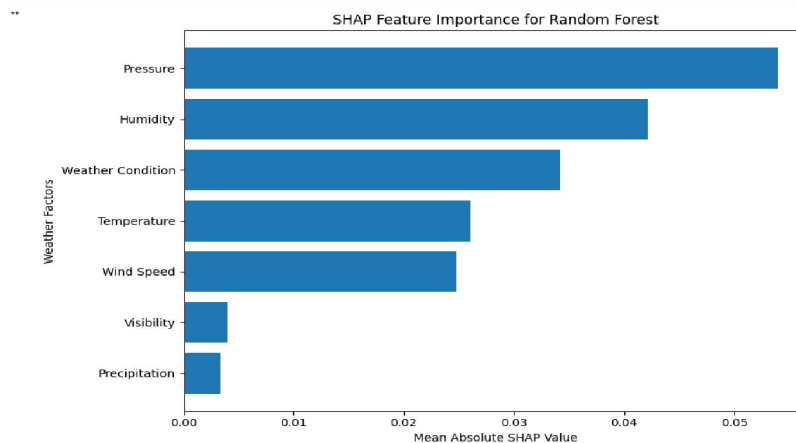
	Model	Accuracy	Precision	Recall	Weighted F1	Macro F1
0	Logistic Regression	53.25	52.61	53.25	44.96	37.59
1	Random Forest	69.38	65.33	69.38	65.80	62.91
2	XGBoost	62.79	59.32	62.79	58.18	53.64

Based on these evaluation results, Random Forest was selected as the best-performing model, as it achieved the highest Accuracy, Precision, Recall, Weighted F1-score, and Macro F1-score among the three models.

SHAP-Based Model Interpretation

After identifying Random Forest as the selected model, SHAP (SHapley Additive exPlanations) was applied to interpret the contribution of the weather-related features to the model's predictions. SHAP values were calculated for each of the three risk classes to investigate how individual input variables influenced the predicted outcomes. Feature importance was summarized using the mean of the absolute SHAP values. Because the Weather Condition variable was transformed into several one-hot encoded features during preprocessing, the SHAP contributions of these encoded categories were aggregated to obtain an overall measure of the influence of the original Weather Condition feature.





The analysis indicates that Pressure, Humidity, Weather Condition, Temperature, and Wind Speed have relatively greater influence on the model predictions, while Visibility and Precipitation show lower overall SHAP importance in the obtained results.

Individual Risk Prediction and Explanation

The Random Forest classifier was used to predict accident risks for different weather conditions. It uses seven weather inputs: Temperature, Humidity, Pressure, Visibility, Wind Speed, Precipitation, and Weather Condition. The entered values were prepared using the same preprocessing steps applied during model training. The system provided the following results: Predicted risk level: Low, Medium, or High, Probability score for each risk class, SHAP-based explanation of the selected prediction. The SHAP results indicated how each weather variable affected the model's output, either increasing or decreasing the predicted risk. This approach enabled the system to provide an individual risk estimate together with an interpretable explanation.

V. LIMITATIONS

This research has certain limitations. [1] The study is based on 12,000 accident records from the US Accidents dataset. Hence, the findings may not fully represent accident conditions in different regions or situations. [2] The model uses only weather information, while other important factors such as road quality, traffic density, vehicle type, driver behaviour, and accident location are not considered. [3] The original four severity levels were grouped into three risk classes Low, Medium, and High. This grouping makes the prediction process simpler but may remove some details from the original severity information. [4] The reported performance depends on the selected data and train-test division. Different results may be obtained with another dataset, time period, or data split. [5] SHAP identifies the influence of input features on the model's output; however, it cannot confirm that a particular weather condition directly causes an accident. [6] The developed model is intended for research and experimental use. It should not be considered a fully reliable real-time safety warning or emergency response system.

VI. CONCLUSION

In this study, weather information was used to understand and predict different levels of road accident risk. The results show that machine learning can be used to classify accidents into High, Medium, and Low risk categories using selected weather factors. Among the three models tested, Random Forest gave the best results, with an accuracy of 69.38%. The SHAP analysis also helped in understanding the model by showing that Pressure, Humidity, Weather Condition, Temperature, and Wind Speed had more influence on the predictions. The interactive prototype makes the system easier to use by allowing weather values to be entered and providing both the predicted risk level and its



explanation. Overall, this study shows that combining machine learning with SHAP can make weather-based accident-risk prediction more understandable and useful for further road-safety research.

REFERENCES

- [1] Song et al., "Road accident risk prediction under dense fog conditions using weather, traffic, and road-related information," 2023.
- [2] R. T. Sukonna and S. I. Swapnil, "A Bi-level Framework for Traffic Accident Duration Prediction: Leveraging Weather and Road Condition Data within a Practical Optimum Pipeline," 2023. The study used machine learning models including Random Forest and LightGBM and applied SHAP for interpretation.
- [3] M. A. K. Rifat, A. Kabir, and A. S. Huq, "An Explainable Machine Learning Approach to Traffic Accident Fatality Prediction," *Procedia Computer Science*, vol. 246, pp. 1905–1914, 2024.
- [4] O. I. Aboulola, "Improving Traffic Accident Severity Prediction Using MobileNet Transfer Learning Model and SHAP XAI Technique," *PLOS ONE*, vol. 19, no. 4, article e0300640, 2024.
- [5] T. Patel, Y. Shah, N. Doshi, R. Gupta, and A. Sharma, "AI Enabled Road Accident Prediction Analysis," *Procedia Computer Science*, vol. 282, pp. 2745–2753, 2026

