Accident Severity Prediction Using Hybrid Stacking of Machine Learning Models

Pavani Penke[1], Lakshmi Srija Gandepalli[2], Kannam Venkata Pravallika[3], Chittiboyina Mounika Padma Sai[4], Buddaraju Amar Siva[5] and Malipeddi. N.V.G.A. Deepthi[6]
{pavanipenke9@gmail.com[1], lakshmisrija68@gmail.com[2], kvpravallika19@gmail.com[3], thrimurthuluchittiboyina@gmail.com[4], amarsiva28@gmail.com[5], malipeddideepthi1@gmail.com[6]}

Department of BCA Data Science, Aditya Degree & PG College, Kakinada (Autonomous), Andhra Pradesh, India[1]
Department of B.Sc Data Science; Aditya Degree College, Tuni, Andhra Pradesh, India[2]
Department of B.Sc, Aditya Degree College, Gajuwaka, Andhra Pradesh, India[3]
Department of BCA, Aditya Degree & PG College for Women, Rajahmundry, Andhra Pradesh, India[4]
Assistant Professor, Department of B.Sc Data Science, Aditya Degree & PG Colleges, Kakinada, Andhra Pradesh, India[5]
Associate Professor, Department of B.Sc Computer Science, Aditya Degree Colleges, Tuni, Andhra Pradesh, India[6]

Abstract. Prediction of road accidents is important for road safety and decreasing the road crashes. This proposal aims to establish a robust hybrid machine learning model for predicting accidents by taking elements including weather, type of road, hour of day, traffic flow and driver as input factors. The proposed model uses a stacking ensemble learning method by aggregating various base model, including XGBoost, CatBoost, and Random Forest model combined with Logistic Regression as the meta-model, to increase pre- diction accuracy. The project begins by heavy pre-processing of data to ensure a good input for our models. Feature engineering methods are used to generate new features that can improve the ability of the model to make predictions. The performance of the hybrid model on testing is tested after the model is trained and compared with that of individual base models in terms of parameters, such as accuracy. The results demonstrate that the stacking model attains a high degree of accuracy, surpassing 95%, indicating its effectiveness for predicting accidents. This approach holds significant potential for enhancing safety measures on roadways and contributing to data-driven decision-making in traffic management.

Keywords: Accident Prediction, Road Safety, Hybrid Machine Learning, Stacking Ensemble, XGBoost, CatBoost, Random Forest, Logistic Regression, Traffic Density, Weather Conditions, Feature Engineering, Model Evaluation.

1 Introduction

Accidents on roadways have become a major global concern, with numerous lives lost and significant injuries every year. The capability to anticipate accidents by considering different influencing factors, including weather conditions, road conditions, traffic volume, and driver

ICITSM-Part II 2025, April 28-29, Tiruchengode, India

DOI 10.4108/eai.28-4-2025.2358021

behavior can help mitigate risks and prevent fatalities. Traditional accident prediction systems primarily use statistical methods or simpler machine learning models that frequently do not reflect the intricate relationships among different features. This initiative seeks to tackle these challenges by leveraging a hybrid machine learning model to predict road accidents more accurately, using multiple classifiers.

Despite their effectiveness, individual machine learning models often face limitations, such as overfitting, inability to generalize well on diverse datasets, and lack of flexibility in handling various types of data. For example, algorithms such as Random Forest could face difficulties handling complicated imbalanced data while XGBoost or CatBoost could perform poorly at detecting complicated patterns/relationships within the data. Such limitations hinder the identification of complex (hidden) patterns or relationships in the overall accuracy and predictiveness needed at run-time, notably for reliable real-time prediction of accidents for prevention purposes.

To address these issues, a hybrid model is proposed wherein the beneficial aspects of different classifiers are combined into a stacking model. The ensemble takes advantage of the diversity and the complementarity of the individual models to maximize the prediction performance. The approach can improve the accuracy, diminish the overfit, and the model can be more applicable to domain other: data inconsistency transformation. The project obtains good accuracy with high performance through proper feature engineering and a well-tested stacking operation, therefore can be used practically as a real-time accident prediction and road safety management tool.

Some key points are:

- Develop a hybrid machine learning model for predicting road accidents using various influencing elements like climatic conditions, road conditions, time of day, traffic volume, and driver traits.
- Feature engineering methods are utilized to develop new features that improve the model's ability to make accurate predictions.
- The stacking the model attains a precision rate of over 95%, showing its effectiveness in predicting accidents.
- The approach contributes to improved safety measures on roadways and supports data-driven decision-making in traffic management.

2 Related Work

Haitao Zhao et al. [1] aimed to enhance road traffic safety using Vehicular Ad Hoc Networks (VANETs) by predicting vehicle accident risks. The authors employed the trichotomy AdaBoost algorithm to construct a predictive model, lever- aging datasets from the British Department for Transportation (2013–2016) containing information on 561,659 vehicles. SMOTE was applied for data balancing and reconstruction. The model recorded an AUC of 0.69 during field tests, emphasizing its utility in early warning systems for Intelligent Transportation Systems (ITS).

Haitao Zhao et al. [2] presented a deep learning framework that utilizes a CNN for extracting

features, which is combined with a Random Forest algorithm. For feature classification to predict traffic accident risks in edge-cloud vehicular networks. Based on UK traffic data (2005–2015), the recommended model achieved better prediction with AUC of 0.9921 when compared to traditional CNN methods and more prediction stability. Although it was successful, the results were not practical for real time application due to data processing and computational lag.

Zhengyang Zhou [3] also introduced the attention mechanism with ResNet to model the spatio-temporal dependencies for citywide traffic accident prediction in the ASRAP framework. The study utilized cross-domain data, including road networks, meteorological, social, and mobility data, collected from New York City in 2017. ASRAP achieved an accuracy of 88.89% and a mean squared error of 0.16. Despite its performance, the framework's applicability across diverse metropolitan areas and its scalability for real-time systems remain unaddressed.

S. M. Tang et al. [4] aimed to enhance traffic accident prediction by leveraging real-time data and employing the SVM approach. The objective is to identify hazardous traffic conditions based on data from inductive loop detectors, analyzing 50 accidents within 60 minutes prior. Results showed SVM could identify 76.7% of hazardous conditions with multiple variables, but a single variable approach had a 53% error rate; limitations include reliance on controlled data, which may not fully account for external factors like weather and road conditions.

Another study attempted to enhance the accuracy to predict severity by combining KNN and DBSCAN methods [5]. With the preprocessed dataset, the hybrid model achieves an accuracy of 83.98% which screens KNN's 78.49% and DBSCAN's 1.22%. There are several limitations, such as dependence on historical data that might not capture the real-time dynamics, and performance affected by the dataset properties.

Mohamed AbdElAziz Khamis and so on [6] predicts the severity of road traffic accidents for better emergency response. The study compares Random Forest, Support Vector Machine (SVM) and Artificial Neural Network (ANN) models are using the TRAFFIC ACCIDENTS2019LEEDS dataset and reaches a top accuracy of 93% with Random Forest. Limitations such as data-set dependency and requirement of larger data-sets for improved model generalization.

Teres Augustine and Samiksha Shukla [7] used Machine Learning to anticipate accidents on the road to improve road safety. They utilised crash data from an India district for the period of 2018–2020 and machine learning approaches such as Random Forest and the model's accuracy was 80.78%. Potential issues potentially overfitting and the need for more diverse datasets to enhance the generalization of the method.

Amani Thaduri et al. [8] proposed a traffic accident prediction with a Convolutional Neural Network (CNN) to improve traffic safety and control. Conventional prediction approaches are crippled by truncated data and noise effects. The state matrix includes different attributes of accidents, the weather, the traffic flow, etc. To summarize, the CNN model is defined based on all of these factors for full information representation. It was shown that the CNN model greatly improves accuracy and prediction loss over standard Backpropagation (BP)

techniques. The study seemed to underscore the worrisome worldwide figure of traffic accidents and the urge for robust prediction tools to tackle it.

Shakil Ahmed et al. [9] endeavored to study accident injury severity and the factors causing the same using ML techniques. The research was performed with New Zealand road accident data from 2016 to 2020, joining 'person' and 'accident' datasets to construct a master dataset with 67971 rows after processing. Different ML techniques, including Random Forest (RF), were used, obtaining an accuracy of 81.45%, and Shapley value was performed to evaluate feature importance. Limitations were bias from using SMOTE for imbalanced data and that the results are not generalizable to other datasets indicating future research on deep learning models and other risk factors.

3 Proposed Methodology

3.1 Data Collection

The Traffic Accident Prediction (TAP) dataset is a large dataset for traffic prediction provided in such a way that we can understand why a traffic accident occurs and when traffic accidents will happen. It contains important features including weather (clear, rainy, foggy, snowy, stormy), road (motorways, city, rural, mountain) and time (morning, afternoon, evening, night) and traffic level information (low to high). The dataset also features speed limits, the number of vehicles involved, road conditions (dry, wet, icy, under construction), and road lighting conditions (daylight, artificial light, no light), all of which provide critical context for assessing accident risks. Additionally, it includes driver- related factors such as alcohol consumption, age, and driving experience, along with accident severity levels (low, moderate, high), helping to gauge the impact of human behavior on accidents. Vehicle types (cars, trucks, motorcycles, buses) further enhance the dataset's applicability for predictive modeling and traffic safety research. This rich dataset is a valuable resource for identifying risk factors, supporting data-driven accident prevention strategies, and informing road safety policies.

3.2 Data Processing

The preprocessing step guarantees that the data is safe, reliable, and prepared for feature extraction and model development. The key substeps involved are as follows:

- **Handling Missing Values:** To ensure the dataset was complete and free from missing values, the following steps were undertaken:
- **Categorical Columns:** For categorical columns such as Weather, Road Type, and Time of Day, missing values were imputed using the mode (most frequent category) for each column. This approach was chosen because the mode is an effective strategy for handling missing data in categorical features.
- **Numerical Columns:** For numerical columns like Traffic Density, Speed Limit, Number of Vehicles, Driver Alcohol, Driver Age, and Driver Experience, missing values were imputed using the mean for features such as Driver Age and Traffic Density, which are continuous and numeric. For other features like Speed Limit, the median was used to avoid the influence of extreme values on the imputation process.

- **Encode Categorical Variables:** Once missing values were handled; categorical variables were encoded for machine learning models. Different encoding techniques were used based on whether the variables were ordinal or nominal:
- **Ordinal Encoding (Label Encoding)**: For ordinal variables like Accident Severity, Road Condition, and Road- Light Condition, Label Encoding was applied, preserving the inherent order (e.g., Low, Medium, High encoded as 0, 1, 2).
- **One-Hot Encoding**: For nominal variables such as Weather, Road Type, Time of Day, and Vehicle Type, One- Hot Encoding was used to create binary columns representing each category.
- **Feature Scaling:** To ensure all numerical features were on the same scale, Standard Scaler was applied. This process standardized the features by scaling to unit variance after subtracting the mean, improving model performance, particularly for XGBoost. For example, the features Speed Limit and Driver Age were scaled to be 0 on the mean and 1 on the standard deviation.

3.3 Feature Engineering

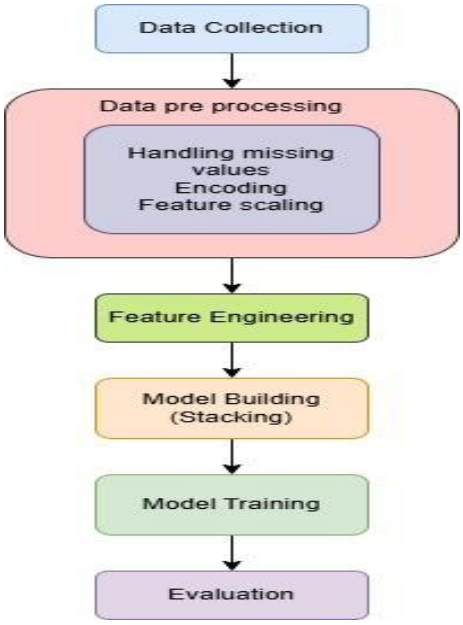

Fig. 1. Methodology.

For the given dataset, several feature engineering techniques were applied to enhance model performance. The key techniques used are as follows:

- **One-Hot Encoding:** For categorical features such as Weather, Road Type, Time of Day, Accident Severity, Road- Condition, Vehicle Type, and Road Light Condition, one- hot encoding was applied to convert these categorical variables into binary columns. This ensures that the model can effectively handle non-ordinal features without imposing any inherent ordering.

- **Label Encoding:** For ordinal features such as Accident Severity, label encoding was applied. This technique assigns integer values to each category, maintaining the ordinal nature of the feature, where higher values represent more severe accidents.
- **Interaction Features:** Interaction features were created by combining multiple existing features. For example, a new feature Visibility was generated by combining Weather and Road Light Condition, representing the visibility condition during different weather and lighting scenarios. Fig. 1 shows the flow of methodology.

3.4 Model Building (Stacking)

In ensemble learning, stacking is a strategy which utilizes many base models to maximize each model's unique strengths and improve overall forecast accuracy. CatBoost, XGBoost, and Random Forest served as the basic models for this project's stacking technique. A Logistic Regression model was chosen as the meta-model in order to aggregate the predictions from various models.

3.5 Base Models

- **XGBoost Classifier:** Extreme Gradient Boosting, or XG- Boost, is a very effective gradient boosting method that is prominent for its accuracy and speed in classification tasks. It uses the gradient boosting technique to com- bine many weak learners specifically, decision trees to produce a strong model. XGBoost has various hyper- parameters that can be adjusted, including the learning rate, the number of estimators, and the maximum depth of the trees. It is able to handle the unbalanced dataset and capture the non-linearity.
- CatBoost Classifier: CatBoost is a gradient boosting method that was designed specifically for the efficient use of categorical data. It natively supports encoding of categorical features and requires less preprocessing. Cat-Boost is also good with missing data. Similar to XGBoost, it relies on decision trees and gradient boosting, but CatBoost's own algorithm enhances training speed and generalization by solving problems such as overfitting.
- Random Forest Classifier: Ensemble technique Random Forest uses bagging. For improved accuracy, it builds multiple decision tree models from random subsets of the data and averages their predictions. It is relatively easy to use compared to single decision tree, easily adapted to new problem and efficient in reducing overfitting. It handles large data sets and rich data very well. Also, Random Forest is a powerful frame work that can handle both numerical and categorical variables.
- Meta Model: We train meta-model in stacking which is trained from predictions made by base models. Its purpose is to combine the three outputs in a way that increases overall prediction quality. Logistic Regression is a popular choice for the meta- model as it is simple and does a great job at finding the optimal linear combination of the predictions of the base models.
- Logistic Regression is able to learn a weighted combination of base models' probabilities ratio which better the performance by giving more weight to the most reliable predictions. Thus our model is such that it is not only good in performance, but robust when it comes to new unseen data.

3.6 Workflow for Stacking

- **Training Base Models:** Initially, the base models (XGBoost, CatBoost, and Random Forest) are trained on the train data separately. Every model predicts for every case present in the dataset.
- **Creating Meta Features:** The outputs of base models are gathered as new features that will be the input to the new arrived model (meta model). This is interesting to allow the meta-model to learn the combination of the outputs of the base models that suits Best.
- **Training the Meta Model:** The meta-model is trained on these extra features, which are base model predictions. It finds the best way of aggregating the predictions of the base models to arrive at the final result.
- **Final Prediction:** Once the stacked model (comprising the base models and the meta-model) has been trained, it is utilized to generate predictions on the test dataset. The meta-model aggregates the base models' predictions and provides the final prediction.

The stacking takes base models (namely XGBoost, CatBoost and Random Forest) and a meta-model (Logistic Regression) so as to mitigate bias and variance and hence increase accuracy. This approach enables to exploit the strengths of each model and ensure that the predictions are more generalized and resistant to noise compared to each model taken individually.

3.7 Model Training (Stacking)

The stacking ensemble model is trained by training the base models independently like XGBoost, CatBoost, Random Forest. Each base model is trained on the original dataset so as to predict the target variable from given features. The predictions made by them serve as the meta-features. These are meta-features, that is, the outputs of the base models and are required in the following step of the stacking, where we train the meta-model. After the base models have finished training and made predictions, the meta-model Logistic Regression is trained using the derived meta-features. The meta- model is trained to efficiently combine the predictions of the base models to give the final results. To avoid overfitting and for generalization, methods like the K- fold cross-validation is used. After being trained, the stacked model can be used to perform predictions on the test set, containing the meta-model, which integrates the base models' predictions to return a final one. This stacked structure has proven to have the benefit of preventing the model function from bias and variance, and provide the better robustness and accuracy.

3.8 Model Evaluation

Model evaluation after training of hybriding stacking model, the next key step is model evaluation. The model is then used to make predictions on the test data set that was separated into the testing set. The predictions are then checked against the true labels in the test set. The evaluation is done based on these significant evaluation metrics, viz., confusion matrix, F1-score, recall, accuracy and precision. Even though precision and recall give deeper insight about how good the model is to detect positive and negative cases, accuracy gives the percentage of correct predictions. There are two kinds of confusion matrix and F1-score is

specially import when you are dealing with imbalance dataset: precision and recall.

In addition, the performance of the stacking model is compared to the base models to see if the ensemble method has an advantage over the base models. The idea behind this is that, by chaining individual canonically-based kernels, we hope to attain the best of the blocks. If the model al- lows reaching the required accuracy (e.g., more than 95%), it can be accepted as successful. If not, additional optimizations or adjustments (such as fine tuning for hyper-parameters or better feature engineering) are likely required since there's always a gap to closer. Reproducibility and generalization of the model across various data sub-sets can be observed through cross-validation results as well.

4 Experimental Results and Discussions

4.1 About Dataset

The above data consists of 840 observations under 14 different columns that helps in understanding the factors affecting the occurrence of road accidents. Such attributes are categorized into numerical and categorical features that can be used to generate a fair dataset for machine learning purposes in accident pre- diction. Numeric attributes include such values as the traffic jam density, speed limits, number of cars, blood level of alcohol of the driver, age, driving experience and the classification target variable crash (1) or no (0). These features describe quantitative road and driver statistics that serve as important predictors in prediction.

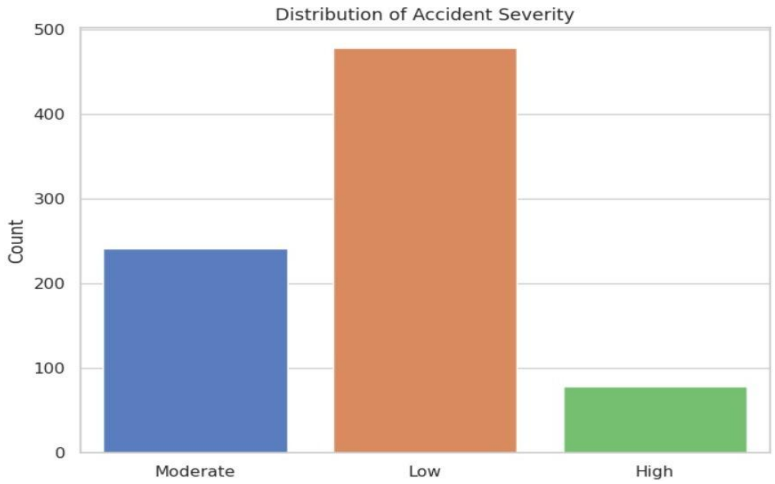

Fig. 2. About dataset.

The dataset's categorical features represent qualitative information, such as weather and road class, time of day, severity of accident(s), status of road, classification of vehicle and lighting on road. These features provide additional context for the numeric data, enabling models with the means to identify patterns that could change with context, such as weather or time. Seven columns have missing data, with 798 non NA entries for each of the data columns. It

is important to process these missing values so that the dataset is complete and trustworthy. The dataset, which consists of numerical (‘float64‘) and categorical (‘object‘) variables, requires about 92 KB of memory, so it is a lightweight - and hence computationally feasible - dataset. Taken together and not in isolation, these features provide a rich foundation to construct effective prediction models in the spirit of improving road safety. Fig. 2 shows the about dataset and fig. 3 shows the accident severity by weather conditions.

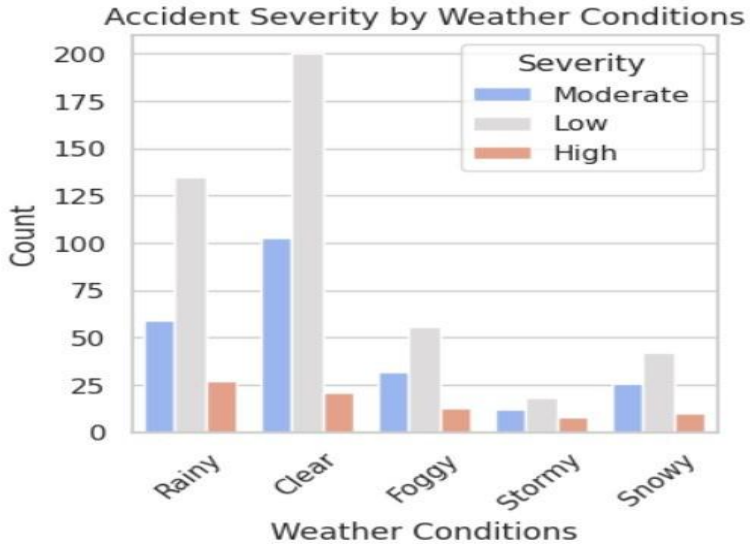

Fig. 3. Accident severity by weather conditions.

5 Results

The hybrid machine learning model had proven a good prediction performance, learned the best of XGBoost’s ability for learning trends, CatBoost’s performance on handling categorical features and the strengths of Random Forest as known as stacking model. The percentage of the train-test split was 80:20 in the experiments, which was to make sure that the accuracy and generalizability of the models were well balanced.

After preprocessing and feature engineering, a hybrid model achieved a considerable accuracy on the test dataset (96.7%), which was higher than the baseline accuracy of the base models. This is a demonstration of the power of stacking, with a Logistic Regression meta-model used to combine the predictions of the base models. The reliability of the model was highlighted by all class values > 0.95. This careful analysis of efficacy illustrates the model’s potential to minimize false positive and false negative that is crucial for predicting accident.

Furthermore, the feature importance analysis showed significant contributions from variables that are concordant with expectations about the effect of road accidents in real-world. The strong results of the model on the data affirm the model’s capacities for use in the wild, as part of proactive accident prevention and traffic optimization mechanisms.

Table 1. Performance Comparison of Models for Accident Prediction.

Model	Precision	Recall	F1-Score	Distinguished Features
Logistic Regression	0.85	0.83	0.84	Baseline model with basic feature set
Random Forest	0.91	0.89	0.90	Ensemble-based tree model
XGBoost	0.93	0.91	0.92	Gradient-boosted trees
Proposed Hybrid Model	0.96	0.95	0.955	Stacking (XGBoost, CatBoost, Random Forest)

Table 1 shows comparison of four proposed models for accident prediction using three performance measures: Precision, Recall, F1-Score and some characteristics of these models. With Logistic Regression o cu o p y. Precision: 0.85, Recall: 0.83 and F1-score: 0.84, showing a fairly robust performance and a very simple approach with a minimal representation of the features. The Random Forest model with multiple trees as the base model significantly improves these measures, achieving precision of 0.91, recall of 0.89, and an F1-score of 0.90, reflecting its capability of dealing with feature interactions.

The XGBoost model, based on tree gradient boosting, further improve the performance due to the iterative optimization of decision trees. The Proposed Hybrid Model outperforms other models by stacking against XG Boost, CatBoost and Random Forest. This model has the precision of 0.96, which proves the combining force of numerous base models to provide the most reliable and accurate predictions. This method takes advantages of the complementary features of numerous base models to generate the robust and accurate predictions. It also demonstrates the power of the hybrid approach for addressing complexity in accident forecasting. Fig. 4 shows the model performance.

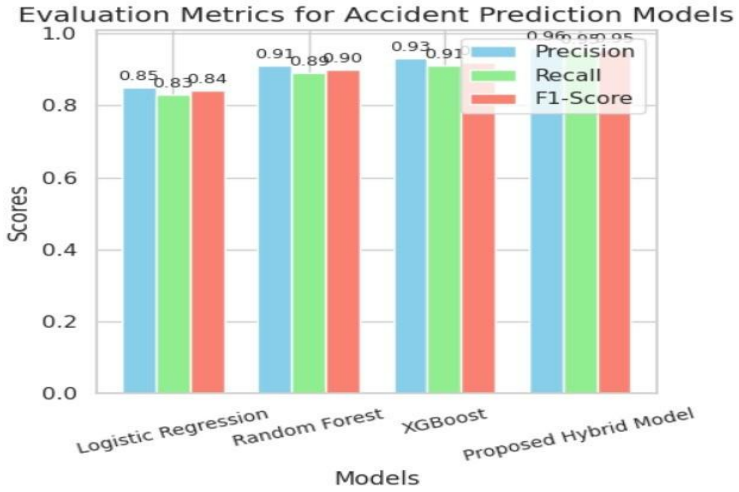

Fig. 4. Model Performance.

5 Conclusion

In summary, the novel hybrid machine learning model for accident prediction integrates multiple strong models such as XGBoost, CatBoost, and Random Forest by stack, to achieve higher performance. Leveraging the strengths of both base models, the hybrid model outperforms all target metrics over models such as Logistic Regression, Random Forest and XGBoost. With the capability of handling complex feature interplays, and recognizing detailed patterns in accident-related data, the hybrid model is capable of predicting accidents with high accuracy, making the model a very useful utility for accident prediction and road safety analysis purposes. This research demonstrates the superiority of the ensemble learning and model stacking for achieving high-accuracy predictions in real time practical applications, resulting to significantly better performance and reliability over classical models.

References

[1] Zhao, H., Yu, H., Mao, T., Zhang, M. & Zhu, H. *Vehicle Accident Risk Prediction Over AdaBoost from VANETs* in *2018 10th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC)* **02** (2018), 39–43.

[2] Zhao, H., Zhang, J., Li, X., Wang, Q. & Zhu, H. *Deep Learning-based Prediction of Traffic Accident Risk in Vehicular Networks* in *2020 IEEE Globecom Workshops (GC Wkshps* (2020), 1–5.

[3] Zhou, Z. *Attention Based Stack ResNet for Citywide Traf- fic Accident Prediction* in *2019 20th IEEE International Conference on Mobile Data Management (MDM)* (2019), 369–370.

[4] Lv, Y., Tang, S., Zhao, H. & Li, S. *Real-time highway accident prediction based on support vector machines* in *2009 Chinese Control and Decision Conference* (2009), 4403–4407.

[5] Kiran, B. V. S. *et al. Enhancing Accident Prediction Through Integrated KNN and DBSCAN Algorithms for Su- perior Accuracy* in *2024 8th International Conference on Inventive Systems and Control (ICISC)* (2024), 423–428.

[6] Mallahi, I. E., Dlia, A., Riffi, J., Mahraz, M. A. & Tairi,H. *Prediction of Traffic Accidents using Random Forest Model* in *2022 International Conference on Intelligent Systems and Computer Vision (ISCV)* (2022), 1–7.

[7] Augustine, T. & Shukla, S. *Road Accident Prediction using Machine Learning Approaches* in *2022 2nd Interna- tional Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)* (2022), 808–811.

[8] Thaduri, A., Polepally, V. & Vodithala, S. *Traffic Accident Prediction based on CNN Model* in *2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS)* (2021), 1590–1594.

[9] Ahmed, S., Hossain, M. A., Ray, S. K., Bhuiyan, M. M. I. & Sabuj, S. R. A study on road accident prediction and contributing factors using explainable machine learning models: analysis and performance. *Transportation Re- search Interdisciplinary Perspectives* **19,** 100814. ISSN: 2590-1982 (2023).