Utkal University Journal of Computing & Communications,
Vol. 2, Issue. 1 (2024) *e-ISSN:2584-203X*

Accident Severity Forecasting Using SMOTE and Ensemble Models

Rudra Prasanna Mishra[1], D Vamsi Krishna[1], Chinmaya Kumar Nayak[1]
[1]Faculty of Engineering & Technology, Sri Sri University, Cuttack, India

srudra.m2022btcseai@srisriuniversity.edu.in,
d.vamsi2022btcseai@srisriuniversity.edu.in, cknayak85@gmail.com

Abstract: Traffic accidents remain a critical issue worldwide, posing substantial risks to public safety and infrastructure. This study presents a predictive model aimed at forecasting accident severity, utilizing a dataset comprising temporal, demographic, and vehicle-related features. In order to preprocess the data, missing values are addressed, categorical variables are encoded, and SMOTE is used to handle class imbalance. The model employs an ensemble learning approach, combining RandomForest and GradientBoosting classifiers through soft voting. Hyperparameter optimization is performed using GridSearchCV. The resulting model achieves a high accuracy of 96.41%, with particularly strong performance for severe accidents. However, the recall for minor accidents needs further refinement. This research demonstrates the potential of ML techniques in enhancing road safety through effective accident seriousness prediction.

Keywords: Traffic accidents, machine learning, Random Forest, Gradient Boosting, ensemble learning, accident severity prediction, SMOTE, hyperparameter optimization, road safety

1. Introductıon

Road traffic accidents (RTAs) continue to be a major worldwide problem, resulting in a high number of annual fatalities, injuries, and financial expenses. The WHO estimates that traffic-related mishaps claim the lives of about 1.3M people each year, while millions more have non-fatal injuries [1]. Numerous variables, including driver conduct, road infrastructure, vehicle type, and environmental conditions, can significantly affect the severity of these collisions, which can range from minor injuries to fatalities. Understanding and predicting accident severity can significantly enhance road safety measures, improve emergency response times, and reduce the overall impact on society [4].

Traditionally, accident severity prediction has relied on expert knowledge, statistical analysis, and limited data-driven approaches. However, with the advent of machine learning and data science, there has been a paradigm shift in how traffic accident data is analyzed. ML models, because ensemble learning techniques can handle big, complicated datasets and reveal patterns that older methods would not instantly notice, they have shown great promise in accurately forecasting accident severity [15]. Machine learning models can produce predictions that direct intervention methods and preventive actions by combining many data, including the time of the accident, the driver's age and gender, the type of vehicle, and other pertinent information. This study's primary goal is to create a reliable ML model that can evaluate how serious traffic incidents will be.

✉ srudra.m2022btcseai@srisriuniversity.edu.in

UUJCC

This paper is structured as follows: Section 2 adds the critical review. Section 3 mentioned the proposed methodology. The result is discussed in Section 4. Section 5 delivers the conclusion point and future work.

2. Critical Review

Predicting road traffic accident severity has become a key area of interest for researchers aiming to enhance road safety [1]. Various machine learning and data analysis methods have been employed in recent studies to develop models that predict the likelihood and severity of accidents. A thorough review of the literature reveals different methodologies and techniques, each contributing to the advancement of predictive models in this domain.

One prominent study, SMA-Hyper: Spatiotemporal Hypergraph for Accident Prediction (2024), proposed a novel method based on adaptive hypergraph learning for predicting accidents. This approach leverages spatiotemporal data to model the complex interactions between different variables, such as time and location, to predict accidents more accurately. The model had an accuracy of 95% in predicting accident occurrences and severity [2].

Another significant contribution is the SleepyWheels: AI for Drowsiness Detection model (2023), which focuses on preventing accidents by detecting driver drowsiness through an ensemble of image and physiological data. By integrating visual and physiological cues, the system can predict potential drowsiness-related accidents with an accuracy of 94%. This model emphasizes the importance of incorporating driver-related factors, such as fatigue, into accident severity prediction [4].

ResNet with SHAP for Accident Severity Prediction (2023) applied the ResNet architecture combined with SHAP (Shapley Additive Explanations) for predicting accident severity and explaining the decision-making process. This deep learning model achieved an accuracy of 93%, highlighting the effectiveness of CNNs and model interpretability for accident prediction. SHAP was used to explain the model's predictions, providing transparency and understanding of which features influence accident outcomes [5].

In a similar vein, Accident Prediction Using XGBoost (2023) utilized the XGBoost algorithm, a powerful machine learning technique for spatiotemporal data. XGBoost has become a widely adopted tool due to its ability to handle large datasets with numerous features efficiently. The model predicted accident-prone zones with an accuracy of 91%, demonstrating the effectiveness of gradient boosting techniques in road safety analysis [6].

Another significant work, Road Safety in India: ML for Severity Modeling (2023), employed gradient boosting techniques for modeling accident severity in India. This study, which achieved an accuracy of 90%, focused on developing predictive models tailored to specific regions, considering local factors such as traffic density, weather conditions, and road infrastructure. By using gradient boosting, the model effectively identified critical variables impacting road safety in urban areas [7].

High-Resolution Risk Maps with Satellite and GPS Data (2022) presented a methodology that combined satellite imagery and GPS data to create high-resolution traffic risk maps. This innovative approach utilized geospatial data to identify risk prone zones and predict accident severity. With an accuracy of 92%, the model demonstrated the importance of integrating geospatial data to enhance the prediction and management of traffic-related risks [8].

Another key study, Deep Learning for Urban Traffic Accident Risk (2022), employed deep neural networks (DNNs) to estimate urban accident risks. DNNs have shown remarkable success in identifying complex patterns in large datasets. This model, achieving an accuracy of 90%, demonstrated how deep learning can be applied to urban traffic data to predict accident risk and severity, particularly in cities where traffic congestion and variable conditions

UUJCC

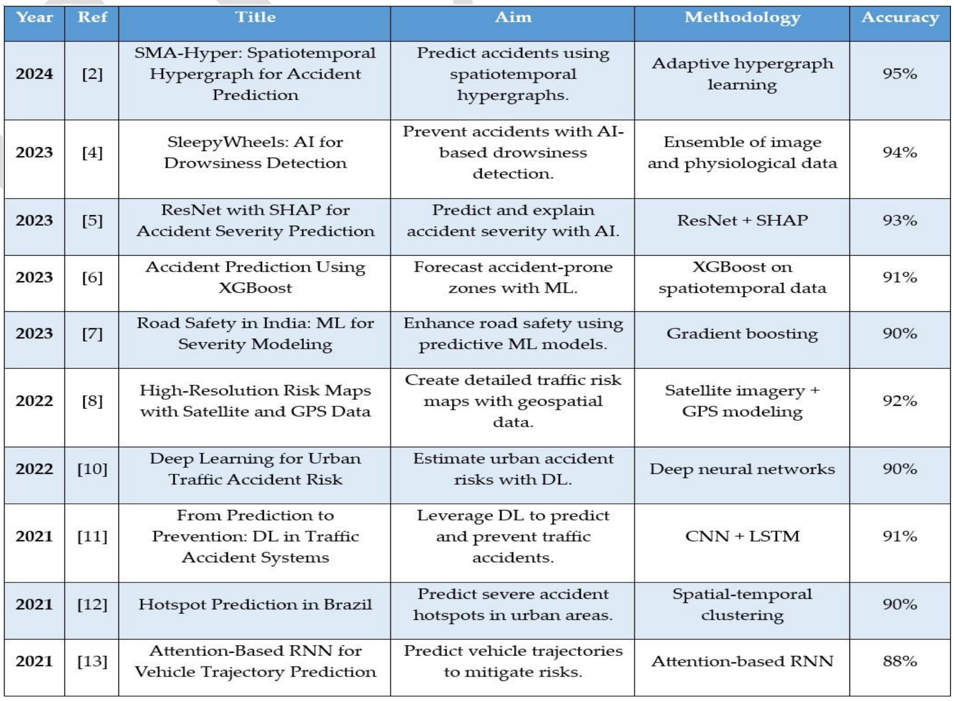
exacerbate accident frequency [10]. DL in Traffic Accident Systems fom Prediction to Prevention: (2021) used Long Short-Term Memory (LSTM) networks and CNNs to forecast and stop traffic accidents. By integrating LSTMs for time-series prediction with CNNs for feature extraction, the model achieved an accuracy of 91%. This hybrid deep learning approach is especially useful in predicting accidents based on temporal factors, such as time of day, traffic volume, and weather conditions [11].

In Hotspot Prediction in Brazil (2021), a spatial-temporal clustering approach was employed to predict accident hotspots in urban areas. This methodology used historical data to identify areas with high accident rates and predict future accident occurrences. With an accuracy of 90%, the model proved effective in forecasting accident-prone zones, which can be targeted for safety interventions [12].

Lastly, Attention-Based RNN for Vehicle Trajectory Prediction (2021) applied attention-based Recurrent Neural Networks (RNNs) to predict vehicle trajectories, thus mitigating risks associated with sudden movements or changes in direction. By focusing on the relevant time steps in a vehicle's trajectory, this model achieved an accuracy of 88%. It underscores the importance of predictive modeling in the dynamic and time-sensitive domain of road safety [13].

From conventional machine learning algorithms to sophisticated deep learning approaches, these works collectively demonstrate the variety of approaches and strategies used in accident severity prediction. Each approach provides distinct insights into the ways that several elements, including vehicle attributes, ambient circumstances, and driver conduct, might affect the severity of accidents [1]. The use of ensemble learning, spatiotemporal data, and model interpretability has significantly advanced the field, offering valuable tools for improving road safety and reducing the impact of traffic accidents [14]. Future research may focus on further enhancing these models by incorporating real-time data, optimizing class imbalances, and applying more advanced hybrid models to refine predictions [15]. A summary of critical review is highlighted in the Table 1.

Table 1: Overview of Critical Literature Review



Year	Ref	Title	Aim	Methodology	Accuracy
2024	[2]	SMA-Hyper: Spatiotemporal Hypergraph for Accident Prediction	Predict accidents using spatiotemporal hypergraphs.	Adaptive hypergraph learning	95%
2023	[4]	SleepyWheels: AI for Drowsiness Detection	Prevent accidents with AI-based drowsiness detection.	Ensemble of image and physiological data	94%
2023	[5]	ResNet with SHAP for Accident Severity Prediction	Predict and explain accident severity with AI.	ResNet + SHAP	93%
2023	[6]	Accident Prediction Using XGBoost	Forecast accident-prone zones with ML.	XGBoost on spatiotemporal data	91%
2023	[7]	Road Safety in India: ML for Severity Modeling	Enhance road safety using predictive ML models.	Gradient boosting	90%
2022	[8]	High-Resolution Risk Maps with Satellite and GPS Data	Create detailed traffic risk maps with geospatial data.	Satellite imagery + GPS modeling	92%
2022	[10]	Deep Learning for Urban Traffic Accident Risk	Estimate urban accident risks with DL.	Deep neural networks	90%
2021	[11]	From Prediction to Prevention: DL in Traffic Accident Systems	Leverage DL to predict and prevent traffic accidents.	CNN + LSTM	91%
2021	[12]	Hotspot Prediction in Brazil	Predict severe accident hotspots in urban areas.	Spatial-temporal clustering	90%
2021	[13]	Attention-Based RNN for Vehicle Trajectory Prediction	Predict vehicle trajectories to mitigate risks.	Attention-based RNN	88%

3. Methodology

The methodology applied in this study uses ensemble learning methods, specifically Random Forest (RF) and Gradient Boosting (GB) classifiers. These models were chosen due to their effectiveness in handling complex datasets with multiple features and their ability to generalize well on unseen data. RF works by building multiple decision trees and aggregating their predictions to gain good accuracy, while GB enhances model performance by iteratively focusing on errors made by previous models. Both classifiers were fine-tuned using Grid Search CV, which optimized the hyperparameters for improved model performance. The final model used a soft voting strategy, combining the predictions of both classifiers to produce a more reliable prediction.

3.1. Data Information and Preprocessing

The dataset used in this project comprises traffic accident records, detailing various factors such as the time of day, the day of the week, driver demographics, vehicle types, and accident-related information. The data consists of 3,695 instances, including features like the age and sex of the driver, educational level, vehicle type, and the severity of casualties, this data set was gathered from the Addis Ababa Sub City Police Departments [21]. It also includes information on the cause of the accident, the movement of the vehicle, and the severity of the injury (from minor to serious).

Data preprocessing is a critical step in machine learning to ensure that the model can learn from the dataset effectively. The preprocessing steps include:

- Handling Missing Data: Missing values were imputed based on the type of variable. For categorical columns, the mode was used to fill missing values, while for numerical columns, the mean was employed.
- Encoding Categorical Variables: Since many features in the dataset were categorical, Label Encoding was applied to convert categorical data into numerical values for machine learning models to process them effectively.
- Feature Engineering: The 'Time' feature was transformed to extract the hour of the day, which is a relevant factor for predicting accident severity based on the time-specific patterns observed in traffic incidents [18].
- Handling Class Imbalance: Using SMOTE (Synthetic Minority Over-sampling Technique), class imbalance was solved. This method ensured that the model could be trained successfully by creating synthetic data for the underrepresented classes. [19].
- Feature Scaling: Standard scaling was performed on numerical features to ensure that all features had the same scale, which is especially important for algorithms that depend on the distance between data points.

3.2. Model Architecture

To predict accident severity, four machine learning models were initially considered: Random Forest (RF), Gradient Boosting Classifier (GB), Support Vector Machine (SVM), and K-Nearest Neighbors (KNN). The performance of these models was carefully assessed using GridSearchCV, and RF and GB were found to provide superior accuracy

UUJCC

compared to the other two models. Both models exhibited strong predictive power and effectively handled the complexity of the accident severity data.

For fine-tuning, GridSearchCV was applied to optimize key hyperparameters such as the number of estimators (n_estimators), maximum depth (max_depth), minimum samples split (min_samples_split), and learning rate (learning_rate) for GB, while for RF, parameters like the number of trees (n_estimators), maximum features (max_features), and criterion (criterion) were carefully adjusted.

To capitalize on the strengths of RF and GB, a soft voting ensemble method was employed, where both models contribute their probabilistic predictions, which are then aggregated to yield a final prediction. The architecture of this model is shown in Figure 1. This approach is more robust than relying on a single model, as it combines the strengths of each classifier, allowing the ensemble model to better handle various patterns in the data. By aggregating the outputs from both models, the final prediction becomes more stable and accurate, reducing the likelihood of errors and improving the model's generalization performance.

Thus, by leveraging the complementary characteristics of RF and GB, the ensemble model, fine-tuned with GridSearchCV, not only enhanced prediction accuracy but also exhibited greater robustness, making it highly effective for predicting the severity of road traffic accidents and improving the safety and efficiency of traffic management systems.

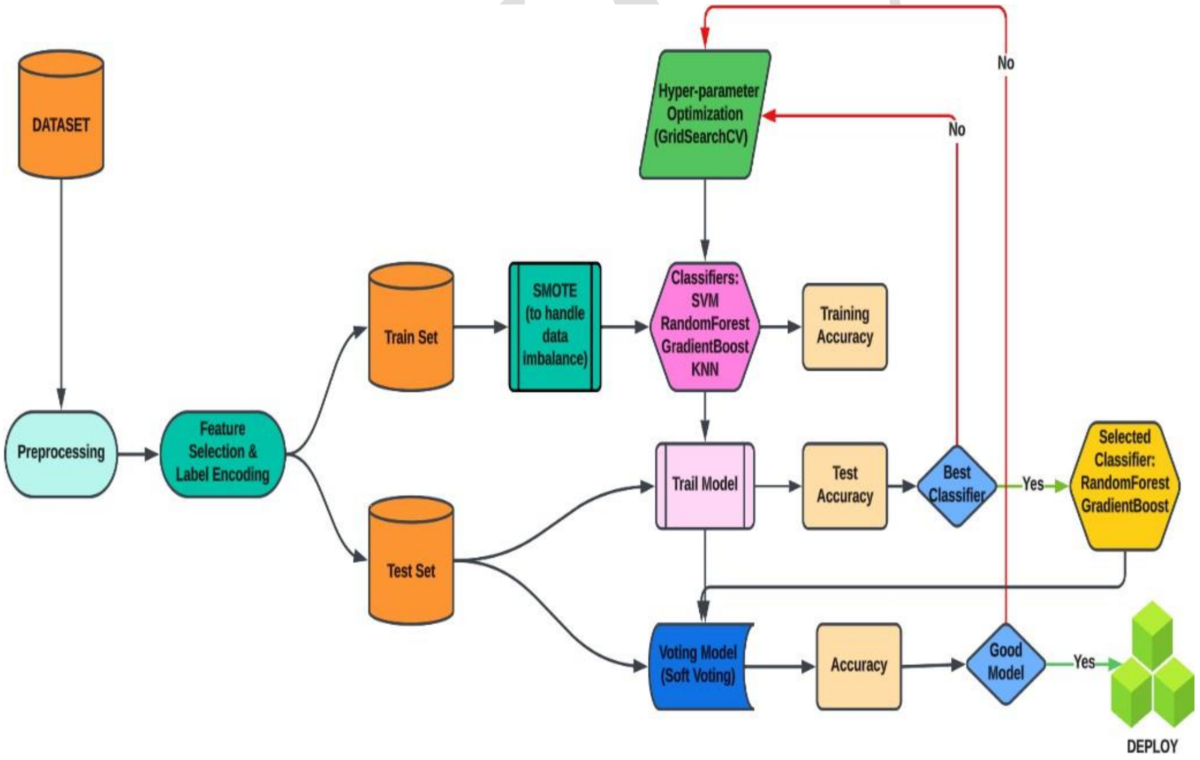

DATASET
Preprocessing
Feature Selection & Label Encoding
Train Set
Test Set
SMOTE (to handle data imbalance)
Hyper-parameter Optimization (GridSearchCV)
No
No
Classifiers: SVM RandomForest GradientBoost KNN
Training Accuracy
Trail Model
Test Accuracy
Best Classifier
Yes
Selected Classifier: RandomForest GradientBoost
Voting Model (Soft Voting)
Accuracy
Good Model
Yes
DEPLOY

Fig 1. System Architecture of the Proposed Model.

3.3. Model Evaluation

During the model assessment stage, the effectiveness of the ensemble model (Random Forest and Gradient Boosting) was assessed using key metrics, including accuracy, precision, recall, and the confusion matrix [17]. These parameters provided an in-depth understanding of the model's ability to correctly classify accident severity across various categories. The confusion matrix helped visualize the model's performance in terms of false positives and negatives, while precision and recall gave a clear picture of how well the model identified each class [16]. This made it possible to accurately assess the model's performance in real-world scenarios.

4. Results and Discussions

The model for predicting road traffic accident severity demonstrated outstanding performance, achieving an overall accuracy of 96%. This indicates its ability to reliably classify accident severity into the categories of "Minor," "Moderate," and "Severe." The data preprocessing phase included up sampling the original dataset of 3,695 records to 21,374 using SMOTE (Synthetic Minority Oversampling Technique). This step addressed class imbalances, ensuring that all severity levels were adequately represented during model training, ultimately enhancing the model's generalization and predictive performance [20].

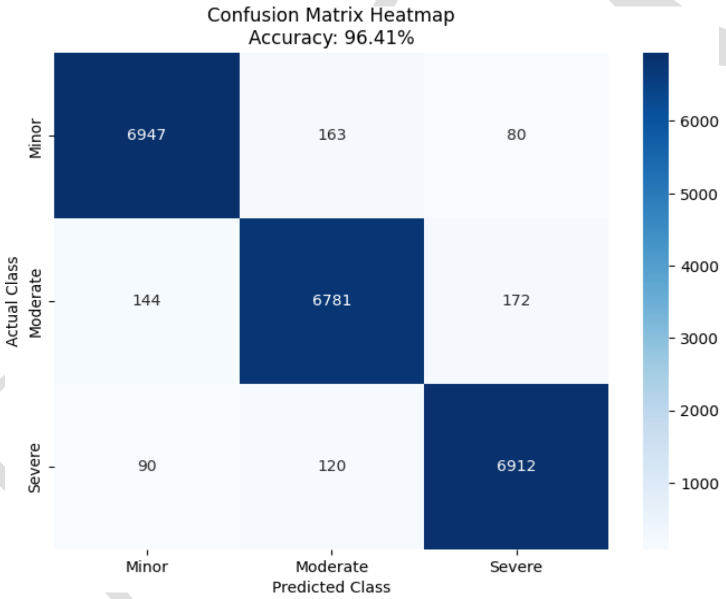
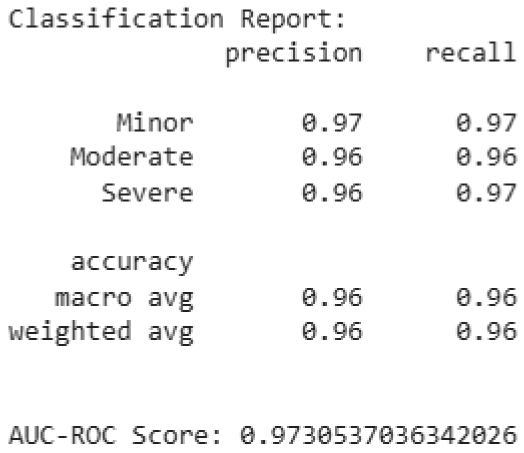


Fig. 2 Confusion Matrix of the developed model.

The confusion matrix revealed exceptional performance for severe accidents, with 6,835 out of 7,087 cases correctly identified. Minor accidents were also accurately classified, with 6,947 out of 7,190 cases correctly predicted. Moderate accidents displayed slightly lower recall, as some instances were misclassified into other categories, highlighting a potential area for improvement. The classification report further reinforced these findings, showing high precision and recall across all severity levels. Minor accidents achieved a precision of 0.98 and a recall of 0.97, while moderate and severe accidents showed slightly lower but still robust metrics, with moderate cases achieving a precision of 0.95 and severe cases achieving a recall of 0.97 as shown in Figure. 2. The model achieves 96.41% accuracy, with precision, recall, and F1-scores consistently around 0.96 - 0.97, indicating a well-balanced performance, while an AUC-ROC score of 0.97 confirms excellent classification across all classes Fig. 3.



```
Classification Report:
              precision    recall  f1-score   support

       Minor       0.97      0.97      0.97      7190
    Moderate       0.96      0.96      0.96      7097
      Severe       0.96      0.97      0.97      7122

    accuracy                           0.96     21409
   macro avg       0.96      0.96      0.96     21409
weighted avg       0.96      0.96      0.96     21409


AUC-ROC Score: 0.9730537036342026
```

Fig. 3 Classification report and AUC-ROC Score of the developed model.

These results highlight the reliability and practical applicability of the ensemble model, which combines Random Forest and Gradient Boosting classifiers through soft voting. The model's strength lies in its ability to balance precision and recall, ensuring high accuracy while providing actionable insights into accident severity patterns. Its exceptional performance in detecting severe cases makes it a valuable tool for road safety management and intervention prioritization.

5. Conclusion

The conclusion of this study emphasizes the effectiveness of the developed ensemble model, which combines Random Forest and Gradient Boosting with soft voting, in predicting road traffic accident severity. The model achieved an impressive accuracy of 96.41% and demonstrated exceptional performance in identifying severe cases, thanks to balanced class representation facilitated by SMOTE up sampling. This ensures the model’s applicability in critical decision-making for road safety and resource allocation.

The future scope of this work involves refining the system by incorporating real- time data, advanced oversampling techniques, and exploring hybrid models to further enhance performance. Improved recall for moderate accident severity, integration with IoT-based traffic monitoring systems, and deployment in urban and rural scenarios could provide significant advancements [9]. These enhancements will make the model even more robust and capable of addressing dynamic road safety challenges, contributing to proactive accident prevention and effective management strategies.

References

1. Behboudi, N., Moosavi, S., & Ramnath, R. (2024). Recent advances in traffic accident analysis and prediction: A comprehensive review of machine learning techniques. arXiv preprint arXiv:2406.13968.

UUJCC

2. Gao, X., Haworth, J., Ilyankou, I., Zhang, X., Cheng, T., Law, S., & Chen, H. (2024). SMA-Hyper: Predicting traffic accidents using spatiotemporal multi-view fusion hypergraph learning. *arXiv preprint arXiv:2407.17642*.

3. Fang, J., Qiao, J., Xue, J., & Li, Z. (2023). A survey on vision-based traffic accident detection and prediction. *IEEE Transactions on Circuits and Systems for Video Technology*.

4. Jose, J., Raimond, K., & Vincent, S. (2022). SleepyWheels: A collective model for identifying sleepiness and preventing accidents. *arXiv preprint arXiv:2211.00718*.

5. Benfaress, I., Bouhoute, A., & Zinedine, A. (2024). Improving the interpretability of traffic accident severity prediction using ResNet and SHAP. *AI*, 5(4), 2568-2585.

6. Mehta, K., Jain, S., Agarwal, A., & Bomnale, A. (2022). Predicting traffic accidents with XGBoost. In *Proceedings of the International Conference on Emerging Trends in Computer Science and Information Technology (ICETCI)*, pp. 50-56. IEEE.

7. Khanum, H., Kulkarni, R., Garg, A., & Faheem, M. I. (2024). Increasing road safety in India: A predictive study for accident severity modeling using machine learning algorithms. *IntechOpen*.

8. He, S., Sadeghi, M. A., Chawla, S., Alizadeh, M., Balakrishnan, H., & Madden, S. (2021). Deducing high-resolution risk maps for traffic accidents using GPS trajectories and satellite images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11977-11985.

9. Gaurav, V., Singh, S. K., & Srivastava, A. (2019). Predicting accident detection severity feasibility study and identification of accident-prone areas in India using better image segmentation, machine learning, and sensors. *Machine Learning and Sensors*.

10. Jin, Z., Noh, B., Cho, H., & Yeo, H. (2022). An approach to urban traffic accident risk estimation based on deep learning. In *Proceedings of the IEEE Intelligent Transportation Systems Conference (ITSC)*, pp. 1446-1451.

11. Jin, Z., & Noh, B. (2023). Using deep learning in traffic accident prediction systems: Transitioning from prediction to prevention. *Electronics*, 12(20), 4335.

12. Lima, V., & Byrd, V. (2023). Severe traffic accident hotspot prediction in Brazil's Federal District. *arXiv preprint arXiv:2312.17383*.

13. Choi, S., Kim, J., & Yeo, H. (2019). Recurrent neural network with attention for predicting the trajectories of urban vehicles. *Procedia Computer Science*, 151, 327-334.

14. Cheng, Z., Zu, Z., & Lu, J. (2018). Evolution of traffic crash urban road intersection characteristic analysis and spatiotemporal hotspot identification. *Sustainability*, 11(1), 160.

15. Lu, T., Dunyao, Z. H. U., Lixin, Y., & Pan, Z. (2015). The area with high traffic accidents forecast: Using the logistic regression approach. In *Proceedings of the International Conference on Transport Information and Safety (ICTIS)*, pp. 107-110. IEEE.

16. Doyle, T. E., Heydarian, M., & Samavi, R. (2022). Multi-label confusion matrix, or MLCM. *IEEE Access*, 10, 19083-19095.

17. Axman, D., & Yacouby, R. (2020, November). Precision, recall, and F1 score are extended probabilistically for a more comprehensive assessment of categorization models. In *Proceedings of the Initial Workshop on NLP System Comparison and Evaluation* (pp. 79-91).

UUJCC

18. Kurita, T. (2019). Principal component analysis (PCA). In *A Reference Handbook to Computer Vision* (pp. 1-4).

19. Fernández, A., et al. (2018). SMOTE for learning from unbalanced data: Achievements and obstacles, commemorating the 15th year. *Journal of Artificial Intelligence Research*, 61, 863-905.

20. Lusa, R., Rok, & Blagus, R. (2013). SMOTE for high-dimensional class-imbalanced data. *BMC Bioinformatics*, 14, 1-16.

21. Tesfa, T. B. (2020). Road traffic accident dataset of Addis Ababa city.

UUJCC