Hybrid RFCNN Model for Accurate Prediction of Road Accident Severity

[1] Mandala Sowmya, [2] Koppala K.V.P Sekhar , [3] Tata Narasimha Murthy
Student in Dept. Of Master of Computer Applications, Miracle Educational Society Group of Institutions
Assistant Professor, Miracle Educational Society Group of Institutions
Associate Professor, Miracle Educational Society Group of Institutions

ABSTRACT
The socio-economic impact of road accidents is huge in developing countries. Employing appropriate measures of defining determinants of accident severity enhances emergency response and informed policy decisions. This study proposes a hybrid framework RFCNN that integrates deep learning and machine learning conducted by combining Random Forest (RF) and Convolutional Neural Network (CNN). Through utilizing the US Road Accident dataset, crucial parameters such as temperature, visibility and other weather were retrieved. The model was implemented with all available features as well as selected features, producing superior results echoing discoveries made with RF, AdaBoost and the CNN systems. This hybrid model helps automate accident severity prediction and offers an intelligent tool to support transportation safety management and emergency response planning.

Keywords: Road Accident Prediction, Accident Severity, Random Forest, Convolutional Neural Network, Hybrid Model

INTRODUCTION:

Road traffic accident's continues to be a global issue of huge concern due to the number of injuries, deaths, and damages to prperty. The impact of public health and infrastructure systems are critical, especially for developing countries. In modern commecial driving, safety measures are incorporated, although, due to numerous factors, the magnitude of impact in accidents will still remain a challenge. The traditional statistical approaches do not address the specific needs of high-dimensional data, thus requiring the advanced alternative of machine learning. This paper investigates the integration of Random Forest (RF) and Convolutional Neural Networks (CNN), which results in the RFCNN model. This integration utilizes fire SWI's feature selection capability while leveraging the pattern recognition prowess of CNN. This combination aids in accurately predicting the degree of collision, thus supporting rapid and reliable emergency actions. In so doing, the objective of reducing road safety accidents that result in severe injuries is achieved by real-time predictions through automated systems that use historical and environmental information.

GAP IDENTIFIED BASED ON LITERATURE SURVEY:

It is evident from literature that there is ample focus on predicting the severity of traffic related accidents. This is in spite of the gaps that arise from using both traditional and machine learning based models along with the isolating and integrating methodologies.Traditional statistical frameworks, including the Logit and Probit models, have limitations in managing multidimensional datasets, as well as accommodating changes in real-time. Existing machine learning approaches, such

as Random Forests and Gradient Boosting, display superior results but do not possess the requisite representational depth for transformative feature representation essential for accurate predictive modeling.

One primary concern is the lack of a comprehensive approach that combines feature-driven and deep pattern-driven methodologies. Most prior works either focused on machine learning or deep learning methods in isolation, neglecting the hybrid approach's potential. Very few have attempted to develop unified models with multiple learning approaches at the decision level to optimize predictive trustworthiness. Furthermore, the absence of feature selection processes leaves low-dimensional data void of relevant information, which diminishes the model's overall performance.

Key Gaps Identified:

- Lack of hybrid models integrating ML and DL for accident severity.
- Inadequate use of deep feature extraction methods.
- Poor handling of high-dimensional data and irrelevant features.
- Underutilization of ensemble models that can learn from both shallow and deep patterns in data.

PROBLEM STATEMENT:

Accurate prediction of road accident severity is essential for timely emergency response and public safety. Current systems struggle with effectively analyzing complex, high-dimensional accident datasets using a single approach.

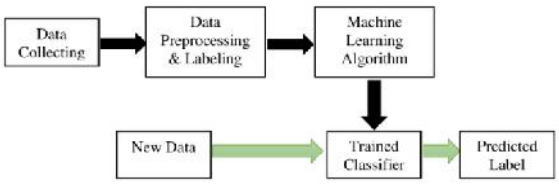
Key Challenges:

- Managing large datasets with many noisy or irrelevant features.
- Balancing model complexity with computational efficiency.
- Integrating different model outputs without losing accuracy.
- Handling real-time predictions where speed and precision are crucial.
- Ensuring generalizability across regions and different traffic scenarios.

PROPOSED METHOD:

The goal for this work is to create a novel model referred to as RFCNN, which uses a Random Forest (RF) for feature selection followed by a Convolutional Neural Network (CNN) for pattern-based deep learning. This begins with maintaining the US Road Accident dataset and selecting the appropriate features with RF. Subsequently, CNN is trained using the selected features to learn intricate representational patterns beyond the capabilities of traditional approaches. Finally, the outputs from RF and CNN are fused at the decision level to enhance accuracy and reliability. This dual-layer approach ensures both interpretability and robustness. Comparative analysis is performed against standalone classifiers like AdaBoost, Extra Trees, and Logistic Regression to validate RFCNN's superior performance across precision, recall, F1-score, and overall accuracy.

ARCHITECTURE:



DATASET:

The study uses the US Road Accident dataset, which comprises over 1.5 million records of accidents across various states. The dataset includes 47 features such as distance, weather condition, temperature, humidity, wind speed, pressure, and visibility. Boolean and categorical data, including traffic signs, junctions, and road types, are also present. Data preprocessing

involves handling missing values, encoding categorical variables, and normalization. Random Forest helps in extracting key features that significantly influence accident severity. The selected subset is used to train models more efficiently. The diversity and size of this dataset make it ideal for training and validating a robust prediction model.

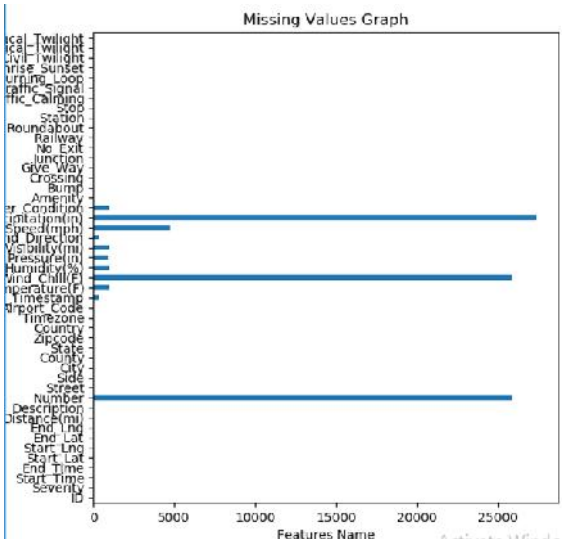
METHODOLOGY:

1. **Dataset Collection and Inspection**
    - The US Road Accident dataset was sourced, containing approximately 1.5 million accident records.
    - Each record includes 47 features such as distance, temperature, humidity, weather conditions, wind speed, and traffic-related indicators.
    - Initial inspection focused on identifying missing values, irrelevant columns, and inconsistencies.
2. **Data Preprocessing**
    - Missing values were imputed using the mean or mode, depending on feature type.
    - Categorical variables like wind direction and weather condition were encoded using label encoding.
    - Boolean features (e.g., "Traffic_Signal", "Junction") were converted to binary (0/1) format.
    - StandardScaler was applied to normalize continuous data for optimal CNN and ML model performance.
3. **Feature Selection Using Random Forest**
    - Random Forest was trained on the full dataset to calculate the importance score for each feature.
    - The top 23 most significant features were selected, such as temperature, visibility, wind chill, humidity, and traffic controls.
    - This reduced noise and improved model efficiency without compromising prediction power.
4. **Train-Test Split**
    - The dataset was split into 80% training and 20% testing sets to validate model performance.
    - Both the full-feature and selected-feature datasets followed this split structure to ensure consistency in comparison.
5. **Model Training with Traditional ML Algorithms**
    - Multiple models were trained using both full and selected features, including:
        - Random Forest (RF)
        - AdaBoost Classifier (AC)
        - Extra Trees Classifier (ETC)
        - Gradient Boosting Machine (GBM)
        - Voting Classifier (VC) using Logistic Regression and SGD
    - Evaluation metrics such as Accuracy, Precision, Recall, and F1-score were used for benchmarking.
6. **Deep Learning: Convolutional Neural Network (CNN)**
    - Input data was reshaped to match CNN's input format: [samples, features, 1, 1].
    - A CNN model with two convolutional layers, pooling

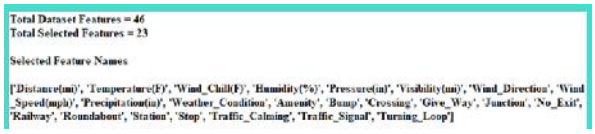
layers, and dense layers was constructed.
- Trained on both full and selected feature sets for 25 epochs using categorical cross-entropy loss.
- Achieved robust classification results, especially in learning complex patterns.

7. **RFCNN Construction: Model Fusion**
   - CNN's second last layer (feature extraction layer) was used to generate deep features.
   - These features were passed to Random Forest to classify severity levels.
   - The final RFCNN model integrates both models' strengths: RF for decision boundaries and CNN for abstract pattern recognition.
   - This fusion occurs at the decision level, leading to improved overall accuracy and robustness.
8. **Evaluation and Visualization**
   - All models were evaluated and compared using metrics and visualizations (bar graphs, tables).
   - RFCNN demonstrated the highest accuracy and F1-score on both full and selected features.
   - A GUI-based tool was developed to input test data and predict accident severity in real-time.
9. **Prediction and Deployment**
   - The trained RFCNN model predicts severity levels: Minor, Major, Severe, or Property Damage.
   - It enables authorities and emergency services to make timely decisions, helping reduce fatalities and delays.

RESULTS:

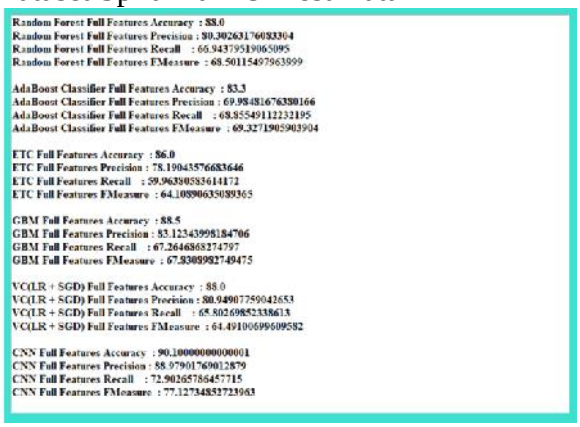


X-axis represents count of missing values and y-axis represents name of features in dataset

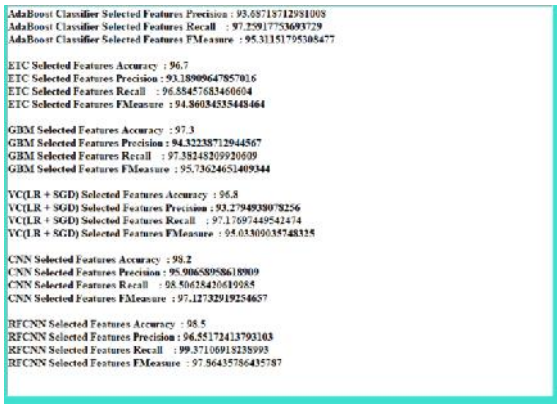


Extract Full & Selected Features

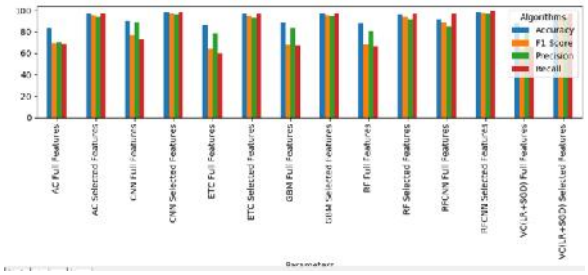
Dataset Split Train & Test Data



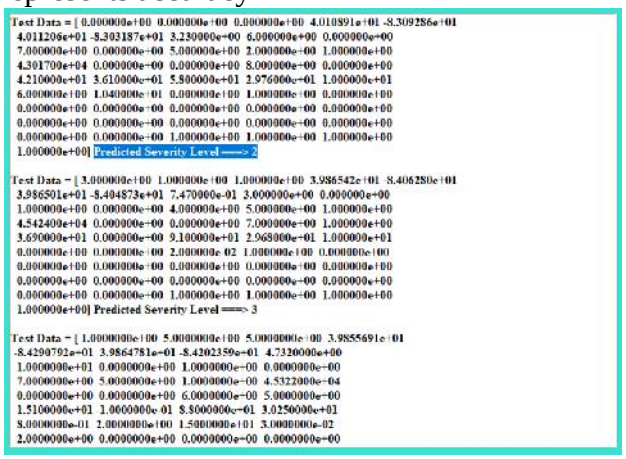
Classifiers on Full Dataset



Classifiers on Selected Features



Graph x-axis represents algorithm names on full and selected features and y-axis represents accuracy



Predicted accident severity level as 2 or 3

CONCLUSION

The RFCNN model presents a novel and effective approach to predicting road accident severity using decision-level fusion of machine learning and deep learning techniques. By combining the feature selection power of Random Forest with the pattern recognition capabilities of CNN, RFCNN offers improved prediction accuracy and reliability. The model outperforms traditional classifiers, especially in handling high-dimensional data and extracting meaningful insights. This approach holds strong potential for real-world deployment in intelligent transportation systems, enabling timely responses and proactive safety measures. The study lays the groundwork for future work on real-time implementation and cross-regional generalization of severity prediction systems.

REFERENCES:

[1] World Health Organization, Global Status Report on Road Safety 2015, World Health Org., Geneva, Switzerland, 2015.

[2] W. Gissane, ‘‘Accidents—A modern epidemic,’’ J. Inst. Health Educ., vol. 3, no. 1, pp. 16–18, 1965.

[3] H. Manner and L. Wünsch-Ziegler, ‘‘Analyzing the severity of accidents on the German autobahn,’’ Accident Anal. Prevention, vol. 57, pp. 40–48, Aug. 2013.

[4] J. C. Milton, V. N. Shankar, and F. L. Mannering, ‘‘Highway accident severities and the mixed logit model: An exploratory empirical analysis,’’ Accident Anal. Prevention, vol. 40, no. 1, pp. 260–266, 2008.

[5] N. V. Malyshkina and F. L. Mannering, ‘‘Markov switching multinomial logit model: An application to accident-injury severities,’’ Accident Anal. Prevention, vol. 41, no. 4, pp. 829–838, Jul. 2009.

[6] K. Haleem, P. Alluri, and A. Gan, ‘‘Analyzing pedestrian crash injury severity at signalized and non-signalized locations,’’

Accident Anal. Prevention, vol. 81, pp. 14–23, Aug. 2015.

[7] M. Bedard, G. H. Guyatt, M. J. Stones, and J. P. Hirdes, ‘‘The independent contribution of driver, crash, and vehicle characteristics to driver fatalities,’’ Accid Anal. Prevention, vol. 34, no. 6, pp. 717–727, 2002.

[8] S. S. Zajac and J. N. Ivan, ‘‘Factors influencing injury severity of motor vehicle–crossing pedestrian crashes in rural connecticut,’’ Accident Anal. Prevention, vol. 35, no. 3, pp. 369–379, May 2003.

[9] W.-H. Chen and P. P. Jovanis, ‘‘Method for identifying factors contributing to driver-injury severity in traffic crashes,’’ Transp. Res. Rec., J. Transp. Res. Board, vol. 1717, no. 1, pp. 1–9, Jan. 2000.

[10] S. Sarkar, S. Vinay, R. Raj, J. Maiti, and P. Mitra, ‘‘Application of optimized machine learning techniques for prediction of occupational accidents,’’ Comput. Oper. Res., vol. 106, pp. 210–224, Jun. 2019.

[11] K. Rumar, ‘‘Transport safety visions, targets and strategies: Beyond 2000,’’ First Eur. Transp. Saf. Lecture, Brussels, Eur. Transp. Saf. Council, pp. 6–8, 1999.

[12] T. Hummel, ‘‘Land use planning in safer transportation network planning,’’ Leidschendam, Inst. Road Saf. Res., Citeseer, 2001.

[13] A. S. Edlin, Per-Mile Premiums for Auto Insurance. Cambridge, MA, USA: National Bureau of Economic Research, Dec. 1998. [Online]. Available: https://ssrn.com/abstract=142667, doi: 10.2139/ssrn.142667.

[14] R. Wittink, ‘‘Promotion of mobility and safety of vulnerable road users: Final report of the European research project promising (promotion of measures for vulnerable road users),’’ Inst. Road Saf. Res., Tech. Rep., 2001.

[15] D. Mohan and G. Tiwari, ‘‘Traffic safety in low-income countries: Issues and concerns regarding technology transfer from high-income countries,’’ in Reflections of the Transfer of Traffic Safety Knowledge to Motorising Nations, vol. 56. Melbourne, Global Traffic Safety Trust, 1998.

[16] S. Ameratunga, M. Hijar, and R. Norton, ‘‘Road-traffic injuries: Confronting disparities to address a global-health problem,’’ Lancet, vol. 367, no. 9521, pp. 1533–1540, May 2006.

[17] M. N. Ghani, A. Zainuddin, R. R. Umar, and H. Hussain, ‘‘Use of exposure control methods to tackle motorcycle accidents in Malaysia research report 3/98,’’ Road Saf. Res. Centre, Serdang, Malaysia, Tech. Rep., 1998.

[18] T. Litman, ‘‘Integrating public health objectives in transportation decisionmaking,’’ Amer. J. Health Promotion, vol. 18, no. 1, pp. 103–108, Sep. 2003.

[19] P. F. Waller, ‘‘The genesis of GDL,’’ J. Saf. Res., vol. 34, no. 1, pp. 17–23, Jan. 2003.

[20] Y. Zou, B. Lin, X. Yang, L. Wu, M. M. Abid, and J. Tang, ‘‘Application of the Bayesian model averaging in analyzing freeway traffic incident clearance time for emergency management,’’ J. Adv. Transp., vol. 2021, pp. 1–9, Mar. 2021.